The many faces of AI in the Phishing-website landscape

Giovanni Apruzzese

University of St. Gallen – November 28th, 2024

UNIVERSITÄT LIECHTENSTEIN

whoami: Dr. **Giovanni Apruzzese**

- **Background:**
  - Did my academic studies (BSc, MSc, PhD) @ University of Modena, Italy.
  - In 2019, spent 6 months @ Dartmouth College, USA.
  - Joined the University of Liechtenstein in July 2020 as a PostDoc Researcher.
  - Was "promoted" to Assistant Professor in September 2022.

- **Interests:**
  - [Areas] Cybersecurity, machine learning, with a strong focus on practice
  - [Applications] Phishing, human factors, 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 an 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, deceiving end-users)

BONUS

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

UNIVERSITÄT LIECHTENSTEIN

(more recently)

Human factors in ML & Cybersecurity

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

Outline of Today

- Using Machine Learning (ML) for Phishing Website Detection
- “Trivially” evading ML-based Phishing Website Detectors
- Using ML to evade ML-based Phishing Website Detectors
- The viewpoint of human users in the above

UNIVERSITÄT LIECHTENSTEIN

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

Outline of Today

- Using Machine Learning (ML) for Phishing Website Detection
- "Trivially" evading ML-based Phishing Website Detectors
- Using ML to evade ML-based Phishing Website Detectors
- The viewpoint of human users in the above

Talk based on the following peer-reviewed papers:

- Apruzzese, Giovanni, Mauro Conti, and Ying Yuan. "Spacephish: The evasion-space of adversarial attacks against phishing website detectors using machine learning." Proceedings of the 38th Annual Computer Security Applications Conference. 2022. (ACSAC)
- Apruzzese, G., Anderson, H. S., Dambra, S., Freeman, D., Pierazzi, F., & Roundy, K. "Real attackers don't compute gradients": bridging the gap between adversarial ml research and practice. In 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)
- Draganovic, A., Dambra, S., Iuit, J. A., Roundy, K., & Apruzzese, G. (2023, November). "Do Users Fall for Real Adversarial Phishing?" Investigating the Human Response to Evasive Webpages. In 2023 APWG Symposium on Electronic Crime Research (eCrime)
- Yuan, Y., Hao, Q., Apruzzese, G., Conti, M., & Wang, G. (2024, May). " Are Adversarial Phishing Webpages a Threat in Reality?" Understanding the Users' Perception of Adversarial Webpages. In Proceedings of the ACM on Web Conference 2024 (TheWebConf)
- Lee, J., Xin, Z., See, M. N. P., Sabharwal, K., Apruzzese, G., & Divakaran, D. M. (2023, September). Attacking logo-based phishing website detectors with adversarial perturbations. In European Symposium on Research in Computer Security (ESORICS)
- Hao, Q., Diwan, N., Yuan, Y., Apruzzese, G., Conti, M., & Wang, G. (2024). It Doesn't Look Like Anything to Me: Using Diffusion Model to Subvert Visual Phishing Detectors. In 33rd USENIX Security Symposium (USENIX Security 24)

All papers are publicly accessible on my website (www.giovanniapruzzese.com)

UNIVERSITÄT
LIECHTENSTEIN

Outline of Today

- Using Machine Learning (ML) for Phishing Website Detection
- "Trivially" evading ML-based Phishing Website Detectors
- Using ML to evade ML-based Phishing Website Detectors
- The viewpoint of human users in the above

Talk based on the following peer-reviewed papers:

- Apruzzese, Giovanni, Mauro Conti, and Ying Yuan. "Spacephish: The evasion-space of adversarial attacks against phishing website detectors using machine learning." Proceedings of the 38th Annual Computer Security Applications Conference. 2022. (ACSAC)
- Apruzzese, G., Anderson, H. S., Dambra, S., Freeman, D., Pierazzi, F., & Roundy, K. "Real attackers don't compute gradients": bridging the gap between adversarial ml research and practice. In 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)
- Draganovic, A., Dambra, S., Iuit, J. A., Roundy, K., & Apruzzese, G. (2023, November). "Do Users Fall for Real Adversarial Phishing?" Investigating the Human Response to Evasive Webpages. In 2023 APWG Symposium on Electronic Crime Research (eCrime)
- Yuan, Y., Hao, Q., Apruzzese, G., Conti, M., & Wang, G. (2024, May). " Are Adversarial Phishing Webpages a Threat in Reality?" Understanding the Users' Perception of Adversarial Webpages. In Proceedings of the ACM on Web Conference 2024 (TheWebConf)
- Lee, J., Xin, Z., See, M. N. P., Sabharwal, K., Apruzzese, G., & Divakaran, D. M. (2023, September). Attacking logo-based phishing website detectors with adversarial perturbations. In European Symposium on Research in Computer Security (ESORICS)
- Hao, Q., Diwan, N., Yuan, Y., Apruzzese, G., Conti, M., & Wang, G. (2024). It Doesn't Look Like Anything to Me: Using Diffusion Model to Subvert Visual Phishing Detectors. In 33rd USENIX Security Symposium (USENIX Security 24)

All papers are publicly accessible on my website (www.giovanniapruzzese.com)

UNIVERSITÄT LIECHTENSTEIN

Two goals:

- Inspire you (to do/consider doing research in computer security)
- Entertain you (research should be fun)

Phishing Website Detection (via ML)

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!

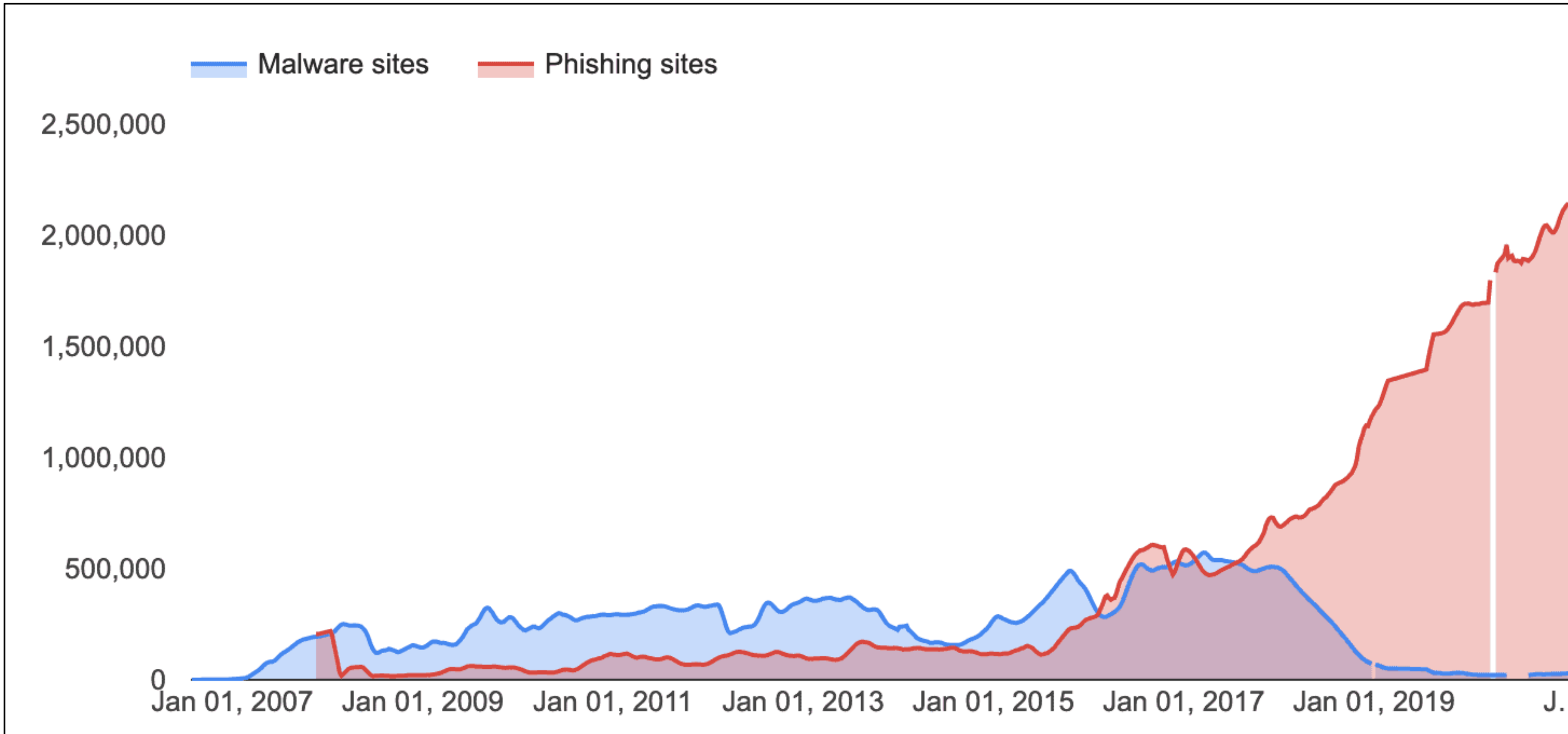


Image source: https://www.tessian.com/blog/phishing-statistics-2020/

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!

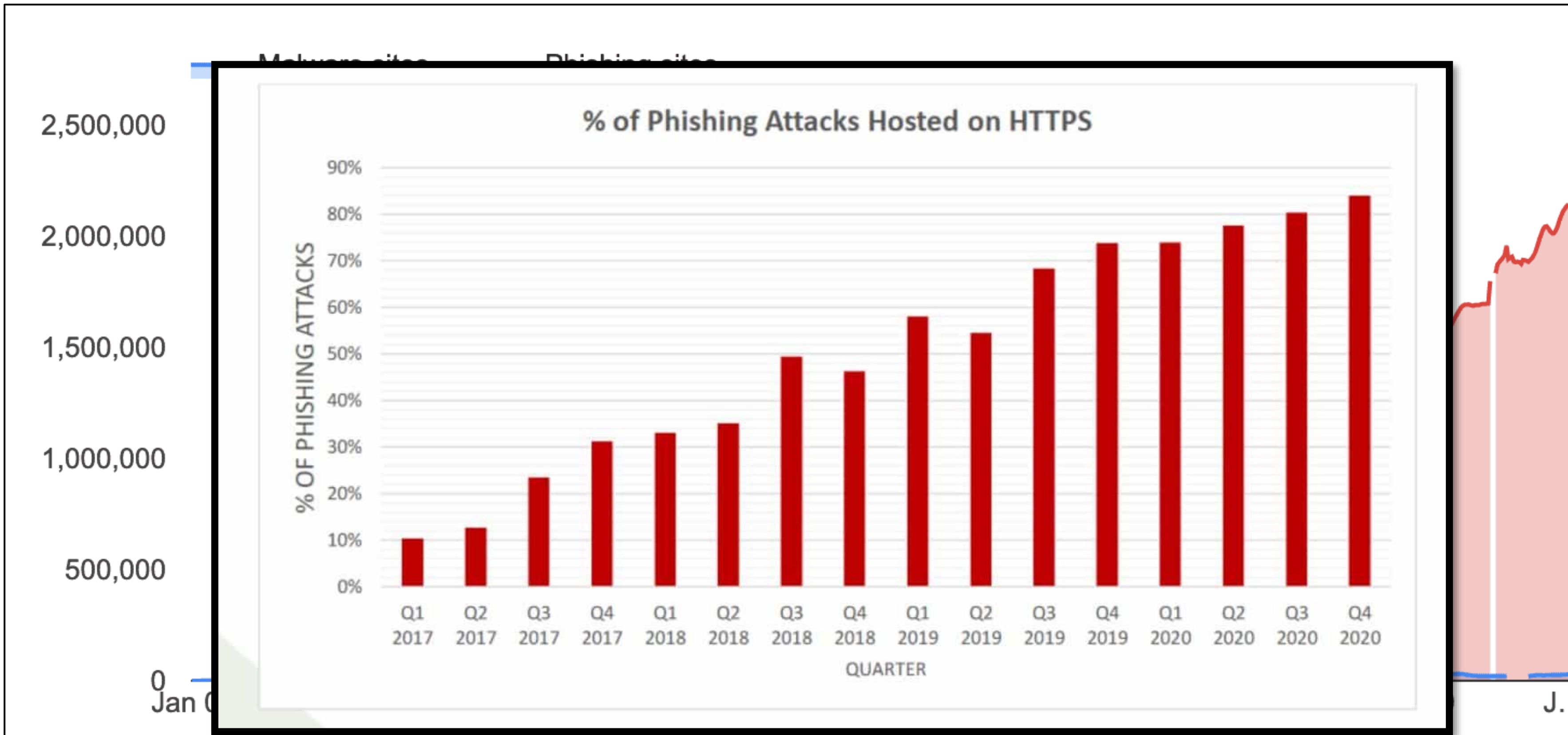


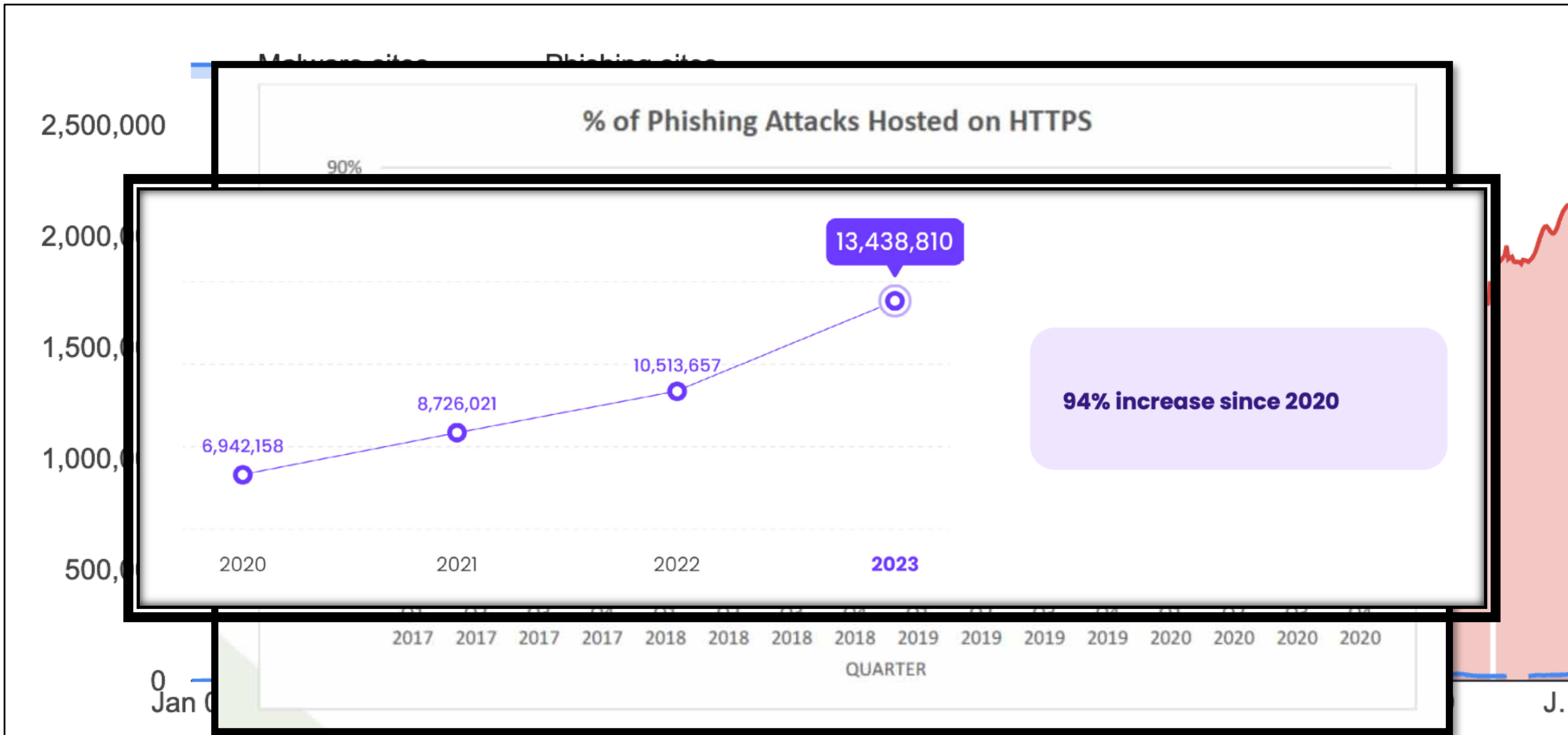
Image source: https://www.tessian.com/blog/phishing-statistics-2020/

Image source: https://cdn.comparitech.com/wp-content/uploads/2018/08/AWPG-q4-2020-phishing-over-https.jpg

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

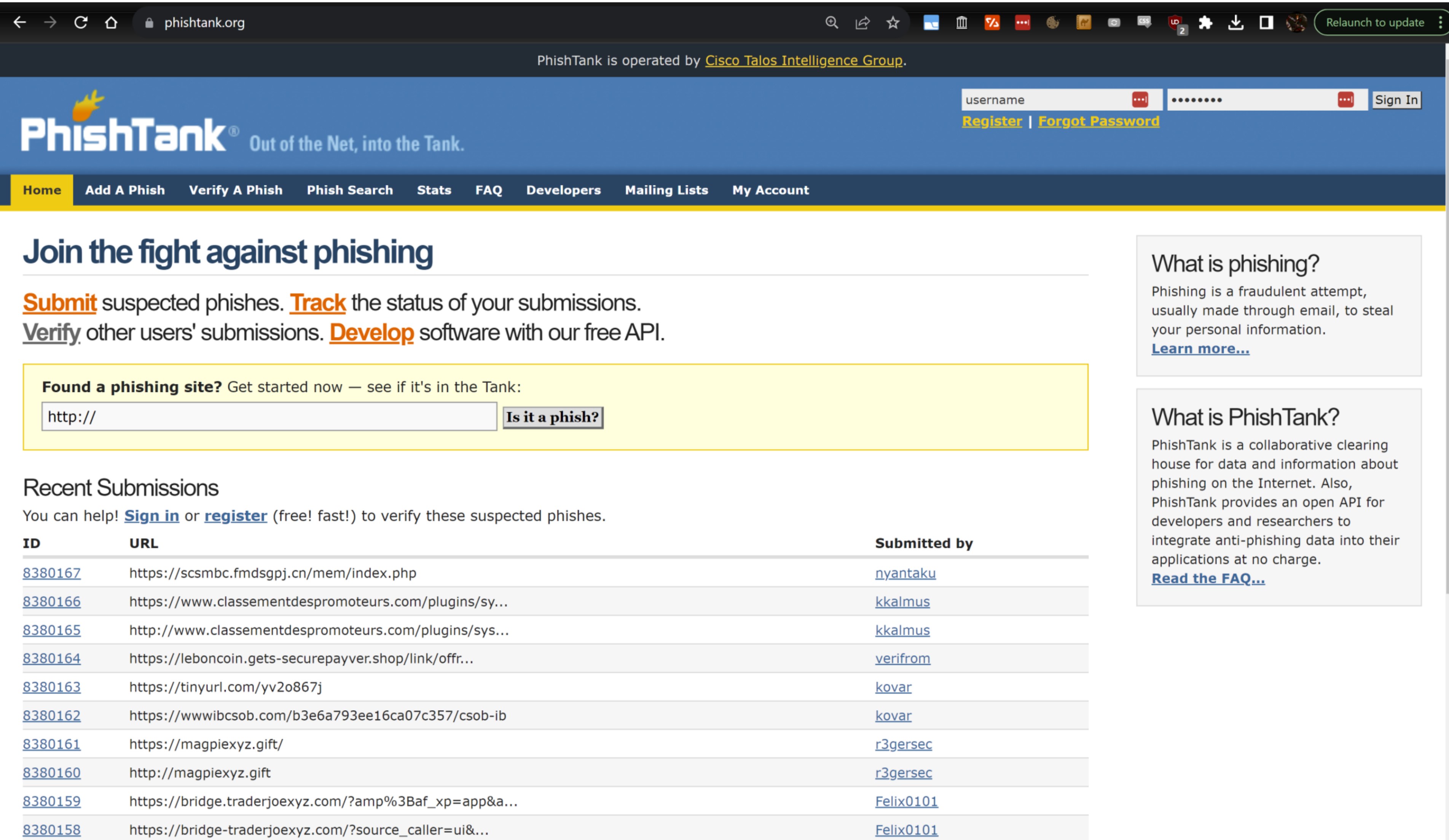
Image source: https://www.tessian.com/blog/phishing-statistics-2020/

Image source: https://cdn.comparitech.com/wp-content/uploads/2018/08/AWPG-q4-2020-phishing-over-https.jpg

Image source: https://bolster.ai/wp-content/uploads/2024/03/increase-in-phishing-and-scam-activity.png

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

Up-to-date list of phishing URLs: PhishTank (www.phishtank.org)

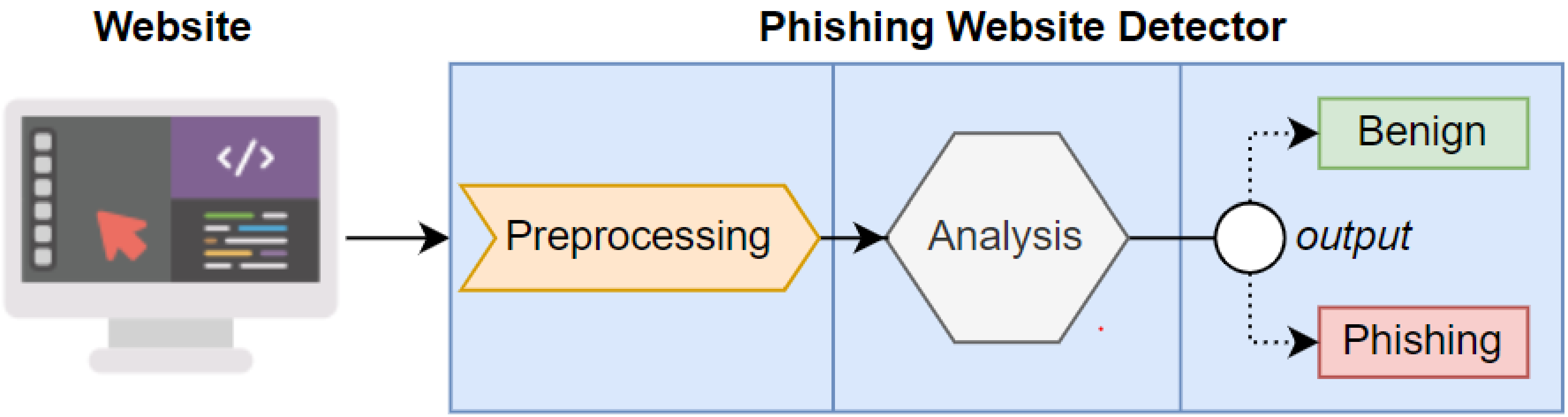


UNIVERSITÄT
LIECHTENSTEIN

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

Current Landscape of Phishing – Countermeasures

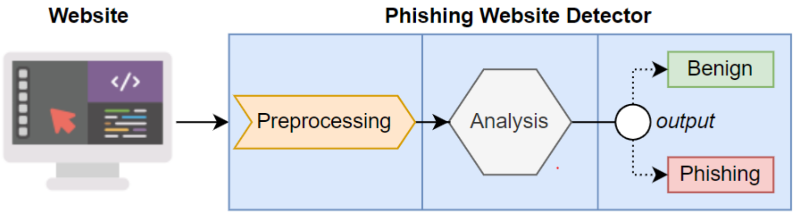
- Countering such simple (but effective) strategies can be done via *data-driven* methods



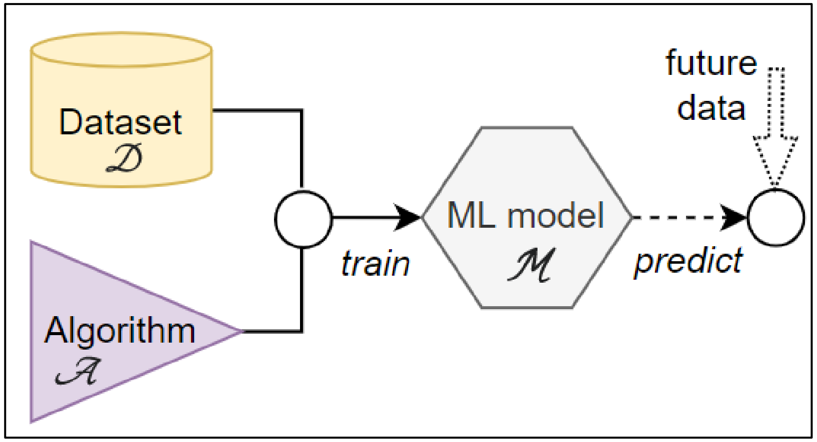
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

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, 3]

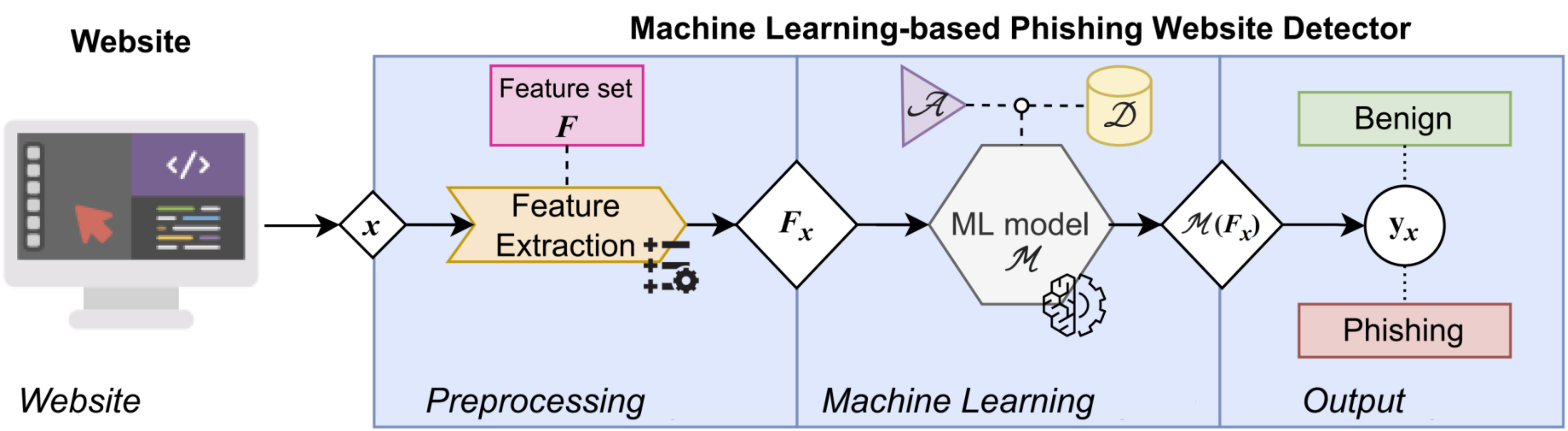
UNIVERSITÄT LIECHTENSTEIN

[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." International Conference on Networking, Information Systems & Security. 2020.
[3]: Miao, C., Feng, J., You, W., Shi, W., Huang, J., & Liang, B. (2023, November). A Good Fishman Knows All the Angles: A Critical Evaluation of Google's Phishing Page Classifier. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*

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

Phishing Website Detection (via ML)

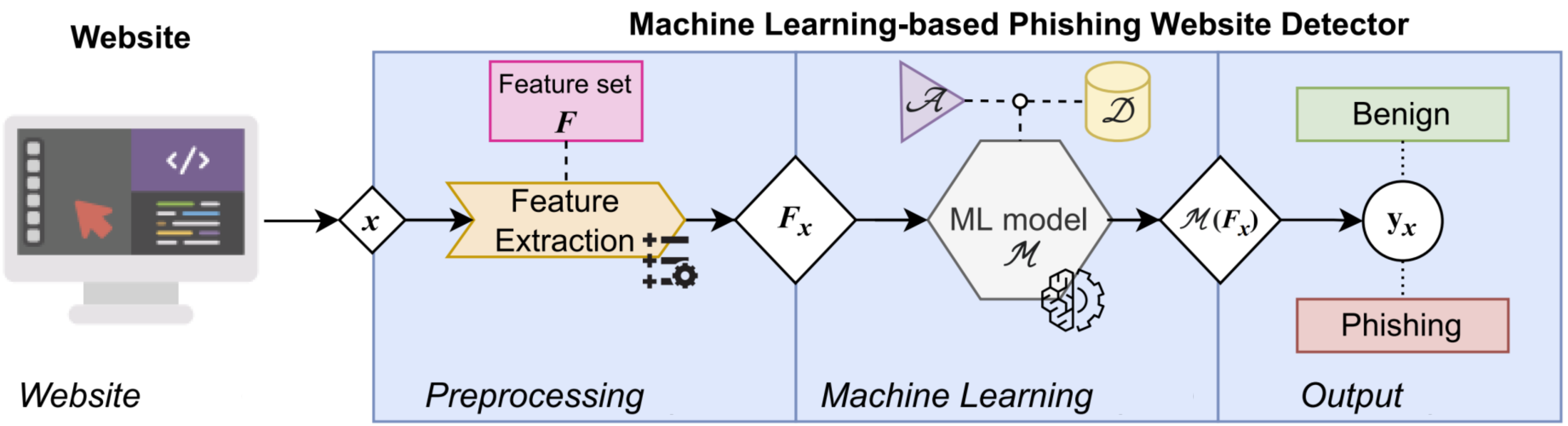
- The *detection* of a phishing webpage can entail the analysis of various elements, such as:
  - The URL of the webpage (e.g., long URLs are more likely suspicious)
  - The HTML (e.g., phishing webpages have many elements hosted under a different domain)
  - The 'reputation' of a webpage (e.g., a webpage whose domain has been active for a long time, or that is indexed in Google, is likely benign)
  - The visual representation (through *reference-based* detectors)
- These analyses can be done via *Machine Learning.*



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

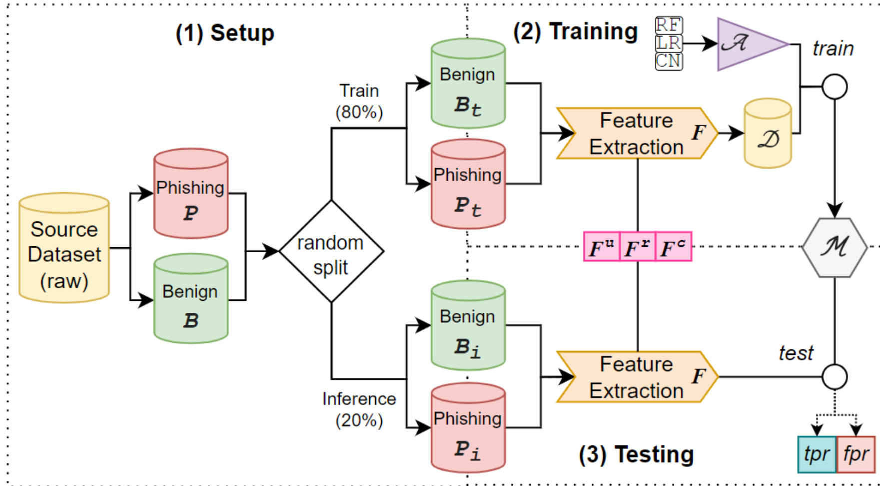
Phishing Website Detection (via ML) [cont'd]

- The most straightforward way to use ML for phishing website detection is to develop a binary classifier:
  - By training an ML model over some training data (containing both benign and phishing webpages) by means of an ML algorithm, it is possible to develop a detector that can discriminate between benign and phishing webpages.
  - Using (including training) the ML model in this way typically requires to *preprocess* any given webpage so as to extract its *feature representation*.
  - The ML model will then analyse the feature representation of any given webpage, and make its decisions depending on how similar such feature representation is w.r.t. the benign/malicious webpages seen during the training stage.



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

Empirical evidence (from my ACSAC'22 paper)



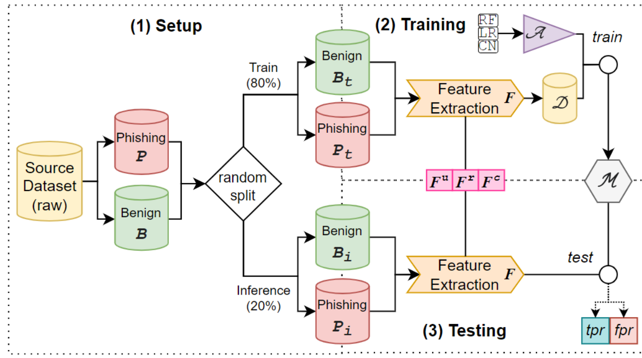
It is indeed possible to develop ML-based detectors that are highly effective (at least in a "research environment") by analysing various types of "features" (using either the URL, the HTML, or both) and by using diverse types of ML algorithms, such as random forests (RF), logistic regression (LR), or convolutional neural networks (CN)

$\mathcal{A}$	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	$\mathbf{0.98_{\pm 0.006}}$	$\mathbf{0.007_{\pm 0.0046}}$	$\mathbf{0.97_{\pm 0.007}}$	$\mathbf{0.001_{\pm 0.0011}}$
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

Empirical evidence (from my ACSAC'22 paper)



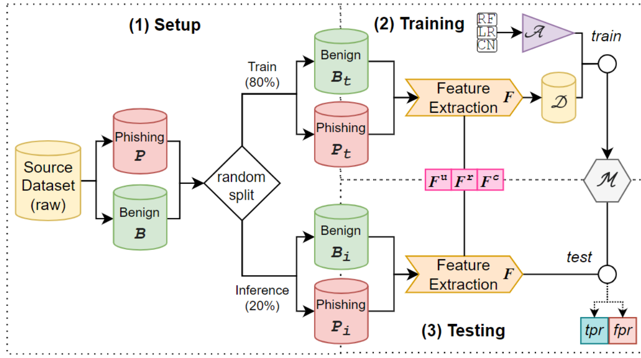
It is indeed possible to develop ML-based detectors that are highly effective (at least in a "research environment") by analysing various types of "features" (using either the URL, the HTML, or both) and by using diverse types of ML algorithms, such as random forests (RF), logistic regression (LR), or convolutional neural networks (CN)

$\mathcal{A}$	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	$0.96_{\pm 0.007}$	$0.025_{\pm 0.0077}$	$0.81_{\pm 0.020}$	$0.013_{\pm 0.0037}$

Limitation:

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

Empirical evidence (from my ACSAC’22 paper)



It is indeed possible to develop ML-based detectors that are highly effective (at least in a “research environment”) by analysing various types of “features” (using either the URL, the HTML, or both) and by using diverse types of ML algorithms, such as random forests (RF), logistic regression (LR), or convolutional neural networks (CN)

$\mathcal{A}$	F	Zenodo		δphish	
		tpr	fpr	tpr	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	$\mathbf{0.98_{\pm 0.006}}$	$\mathbf{0.007_{\pm 0.0046}}$	$\mathbf{0.97_{\pm 0.007}}$	$\mathbf{0.001_{\pm 0.0011}}$
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	$0.96_{\pm 0.007}$	$0.025_{\pm 0.0077}$	$0.81_{\pm 0.020}$	$0.013_{\pm 0.0037}$

Limitation: high number of false positives, and computationally expensive

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

Phishing Website Detection: Reference Based (visual similarity)

- Some detectors leverage the intuition that most phishing webpages try to mimic well-known brands, but they are hosted under a different domain.
- These *reference based* detectors can provide some protection against phishing websites that target a restricted set of brands (e.g., PayPal, Amazon, Google).

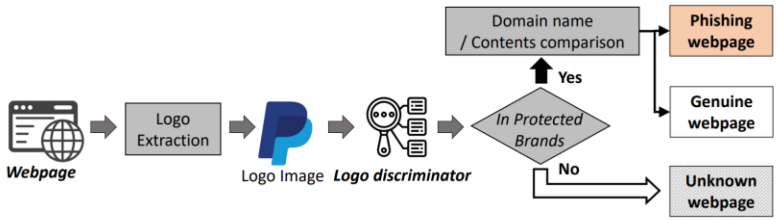


Fig. 1: Detection process of logo-based phishing detection systems

UNIVERSITÄT LIECHTENSTEIN

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

Phishing Website Detection: Reference Based (visual similarity)

- Some detectors leverage the intuition that most phishing webpages try to mimic well-known brands, but they are hosted under a different domain.
- These *reference based* detectors can provide some protection against phishing websites that target a restricted set of brands (e.g., PayPal, Amazon, Google).

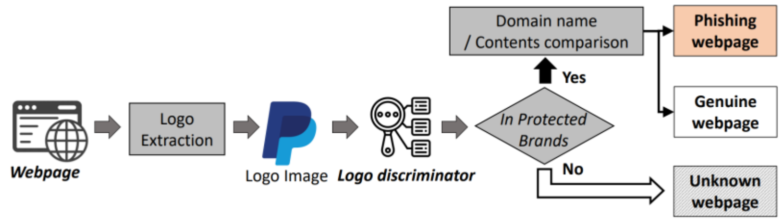


Fig. 1: Detection process of logo-based phishing detection systems

- First, they see if a webpage is visually similar to a webpage of well-known brands.
  - E.g., is this webpage similar to any webpage of PayPal, Amazon, or Google?
    - (If a match is NOT found, then the webpage is treated as benign (to avoid triggering false positives)
- Then, if a match is found, then the detector checks if the given webpage is hosted under the same domain of the well-known brand
  - E.g., is this webpage which is similar to PayPal also hosted under the same domain as Paypal?
- If yes, then the webpage is benign (i.e, it is Paypal). If not, then the webpage is phishing (i.e., it is a phishing webpage that is trying to mimic PayPal).

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

Phishing Website Detection: Reference Based (visual similarity)

- Some detectors leverage the intuition that most phishing webpages try to mimic well-known brands, but they are hosted under a different domain.
- These *reference based* detectors can provide some protection against phishing websites that target a restricted set of brands (e.g., PayPal, Amazon, Google).

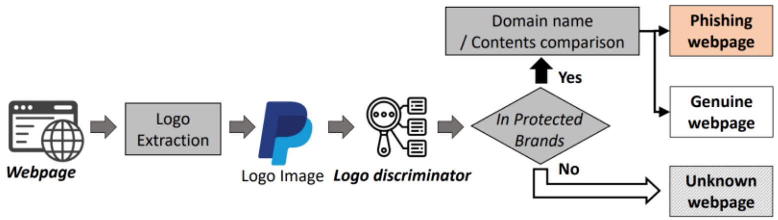


Fig. 1: Detection process of logo-based phishing detection systems

- First, they see if a webpage is visually similar to a webpage of well-known brands.
  - E.g., is this webpage similar to any webpage of PayPal, Amazon, or Google?
    - (If a match is NOT found, then the webpage is treated as benign (to avoid triggering false positives)
- Then, if a match is found, then the detector checks if the given webpage is hosted under the same domain of the well-known brand
  - E.g., is this webpage which is similar to PayPal also hosted under the same domain as Paypal?
- If yes, then the webpage is benign (i.e, it is Paypal). If not, then the webpage is phishing (i.e., it is a phishing webpage that is trying to mimic PayPal).

Limitation:

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

Phishing Website Detection: Reference Based (visual similarity)

- Some detectors leverage the intuition that most phishing webpages try to mimic well-known brands, but they are hosted under a different domain.
- These *reference based* detectors can provide some protection against phishing websites that target a restricted set of brands (e.g., PayPal, Amazon, Google).

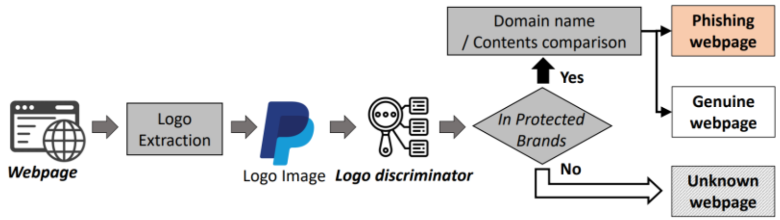


Fig. 1: Detection process of logo-based phishing detection systems

- First, they see if a webpage is visually similar to a webpage of well-known brands.
  - E.g., is this webpage similar to any webpage of PayPal, Amazon, or Google?
    - (If a match is NOT found, then the webpage is treated as benign (to avoid triggering false positives)
- Then, if a match is found, then the detector checks if the given webpage is hosted under the same domain of the well-known brand
  - E.g., is this webpage which is similar to PayPal also hosted under the same domain as Paypal?
- If yes, then the webpage is benign (i.e, it is Paypal). If not, then the webpage is phishing (i.e., it is a phishing webpage that is trying to mimic PayPal).

Limitation: these systems only work on websites in the “reference” list

Evading Phishing Website Detectors

Trivially

UNIVERSITÄT
LIECHTENSTEIN

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 [7]

UNIVERSITÄT LIECHTENSTEIN

[7] 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

- 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 [7]

- 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 [8])

UNIVERSITÄT LIECHTENSTEIN

[7] 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)
[8] Kelce S Wilson and Müge Ayse Kiy. 2014. Some fundamental Cybersecurity concepts. IEEE Access (2014).

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

Evasion Attacks against ML-based phishing website detectors

- ML-based phishing website detectors (ML-PWD) are good but…
- …the detection of ML methods *can* be bypassed via (adversarial) *evasion* attacks!

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

Evasion Attacks against ML-based phishing website detectors

- ML-based phishing website detectors (ML-PWD) are good but…
- …the detection of ML methods *can* be bypassed via (adversarial) *evasion* attacks!
- Such "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

Evasion Attacks against ML-based phishing website detectors

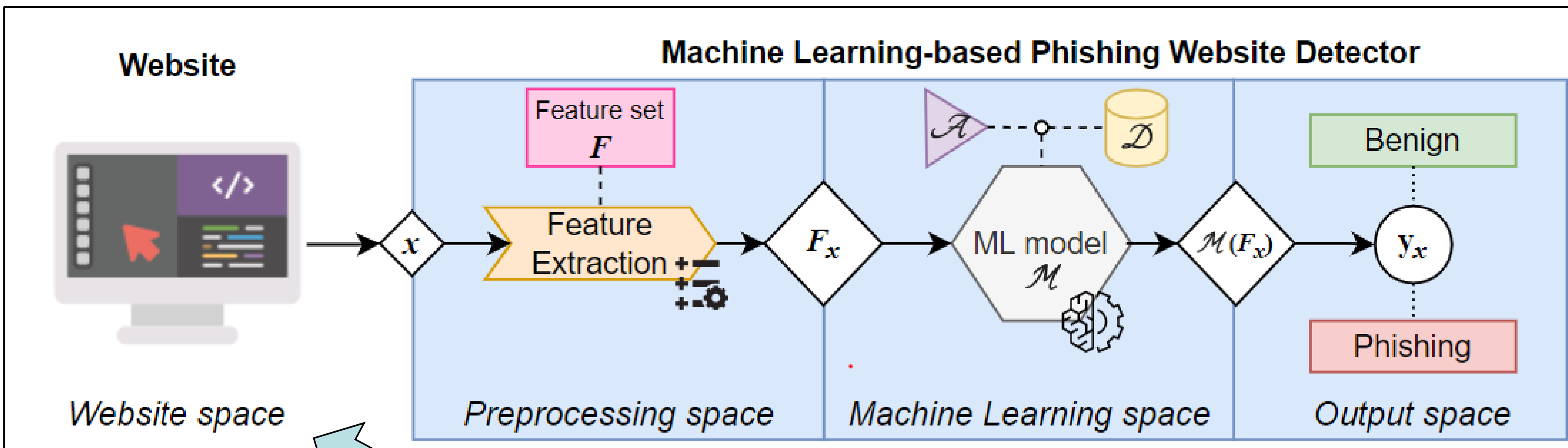
- ML-based phishing website detectors (ML-PWD) are good but…
- …the detection of ML methods *can* be bypassed via (adversarial) *evasion* attacks!
- Such "adversarial" attacks exploit a **perturbation** that induces an ML model to misclassify a given input (i.e., a phishing website) by producing an incorrect output (i.e., classified as a benign website)

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

Evasion Attacks against ML-based phishing website detectors

- ML-based phishing website detectors (ML-PWD) are good but…
- …the detection of ML methods *can* be bypassed via (adversarial) *evasion* attacks!
- Such "adversarial" attacks exploit a **perturbation** that induces an ML model to misclassify a given input (i.e., a phishing website) by producing an incorrect output (i.e., classified as a benign website)

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

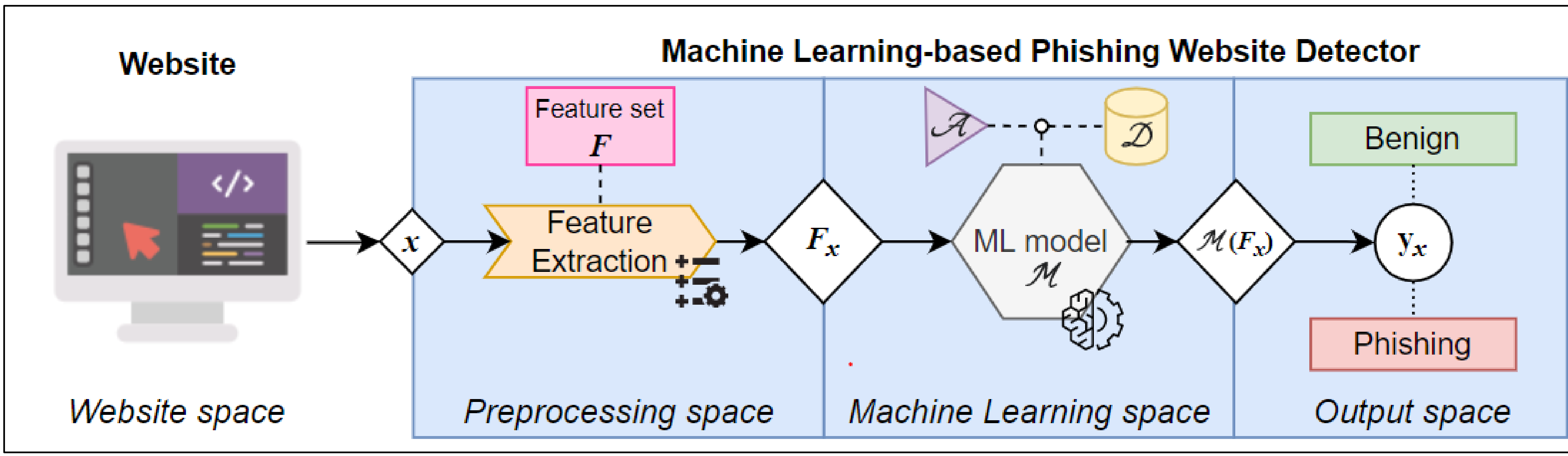


Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Evasion Attacks against ML-based phishing website detectors

- ML-based phishing website detectors (ML-PWD) are good but…
- …the detection of ML methods *can* be bypassed via (adversarial) *evasion* attacks!
- Such "adversarial" attacks exploit a **perturbation** that induces an ML model to misclassify a given input (i.e., a phishing website) by producing an incorrect output (i.e., classified as a benign website)

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



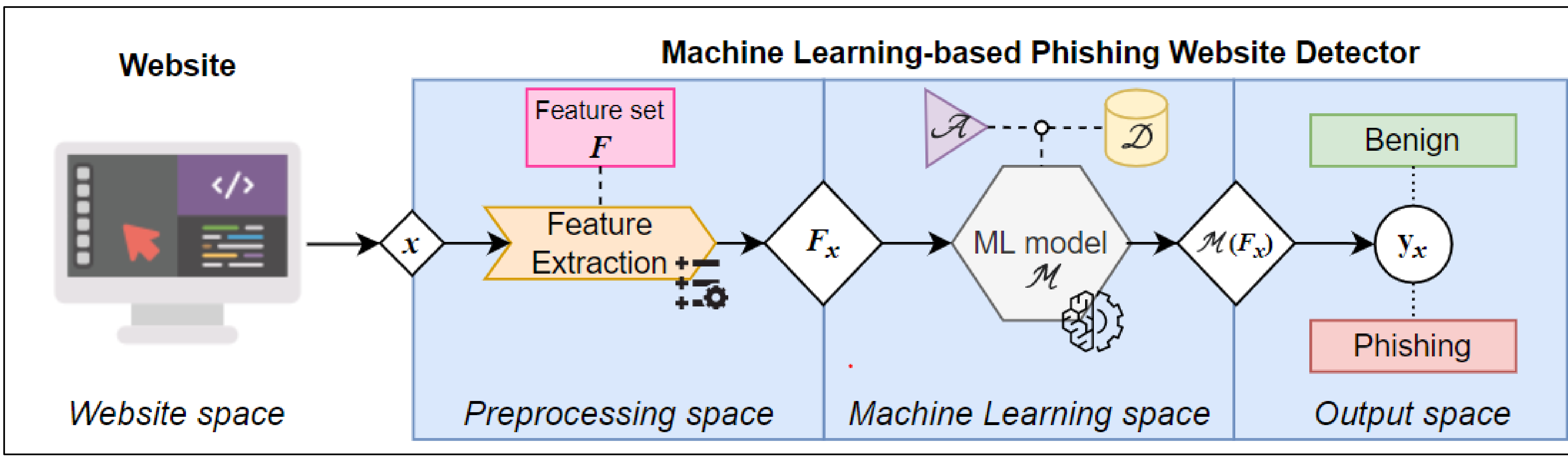
UNIVERSITÄT LIECHTENSTEIN

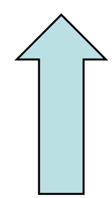
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Evasion Attacks against ML-based phishing website detectors

- ML-based phishing website detectors (ML-PWD) are good but…
- …the detection of ML methods *can* be bypassed via (adversarial) *evasion* attacks!
- Such "adversarial" attacks exploit a **perturbation** that induces an ML model to misclassify a given input (i.e., a phishing website) by producing an incorrect output (i.e., classified as a benign website)

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



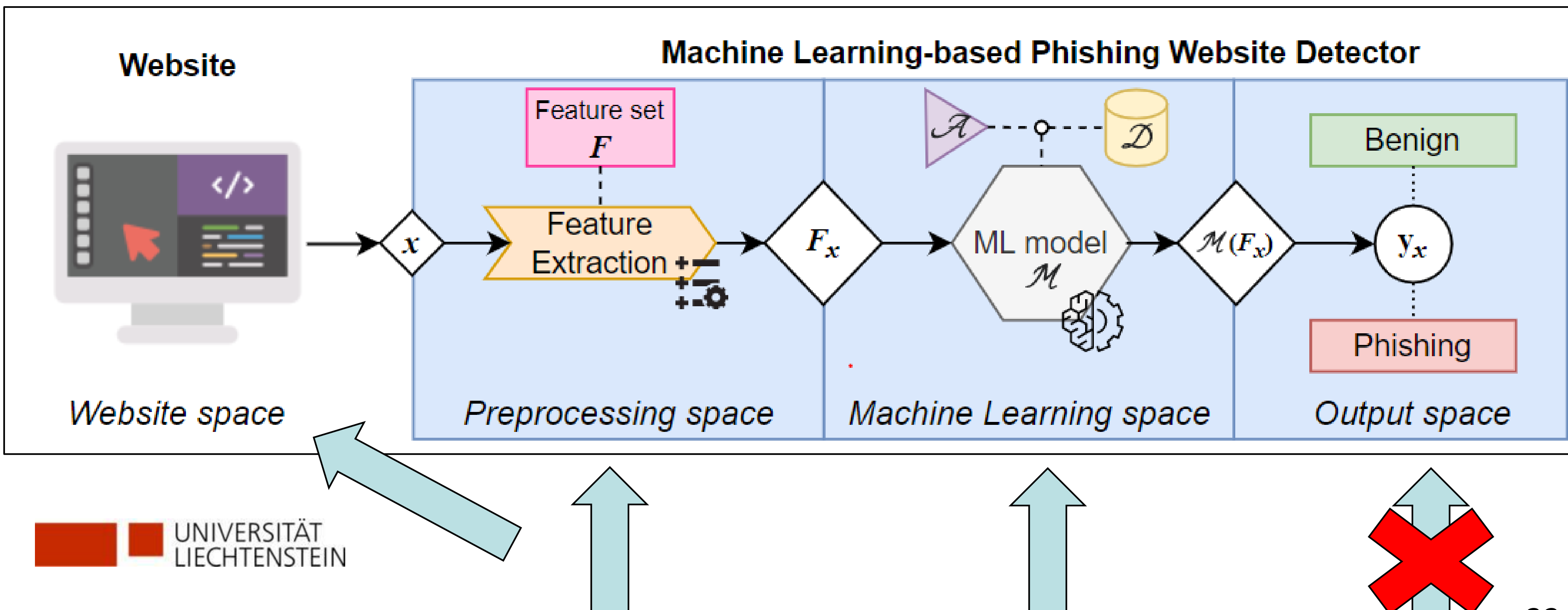


Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Evasion Attacks against ML-based phishing website detectors

- ML-based phishing website detectors (ML-PWD) are good but…
- …the detection of ML methods *can* be bypassed via (adversarial) *evasion* attacks!
- Such "adversarial" attacks exploit a **perturbation** that induces an ML model to misclassify a given input (i.e., a phishing website) by producing an incorrect output (i.e., classified as a benign website)

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



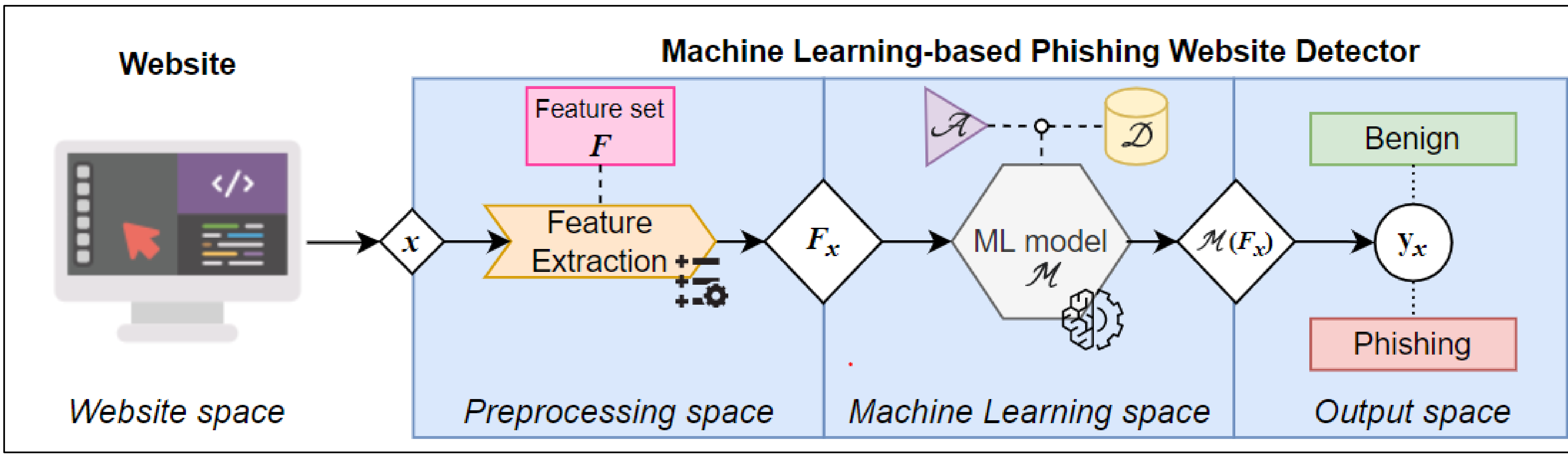
UNIVERSITÄT LIECHTENSTEIN

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

Evasion Attacks against ML-based phishing website detectors

- ML-based phishing website detectors (ML-PWD) are good but…
- …the detection of ML methods *can* be bypassed via (adversarial) *evasion* attacks!
- Such "adversarial" attacks exploit a **perturbation** that induces an ML model to misclassify a given input (i.e., a phishing website) by producing an incorrect output (i.e., classified as a benign website)

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

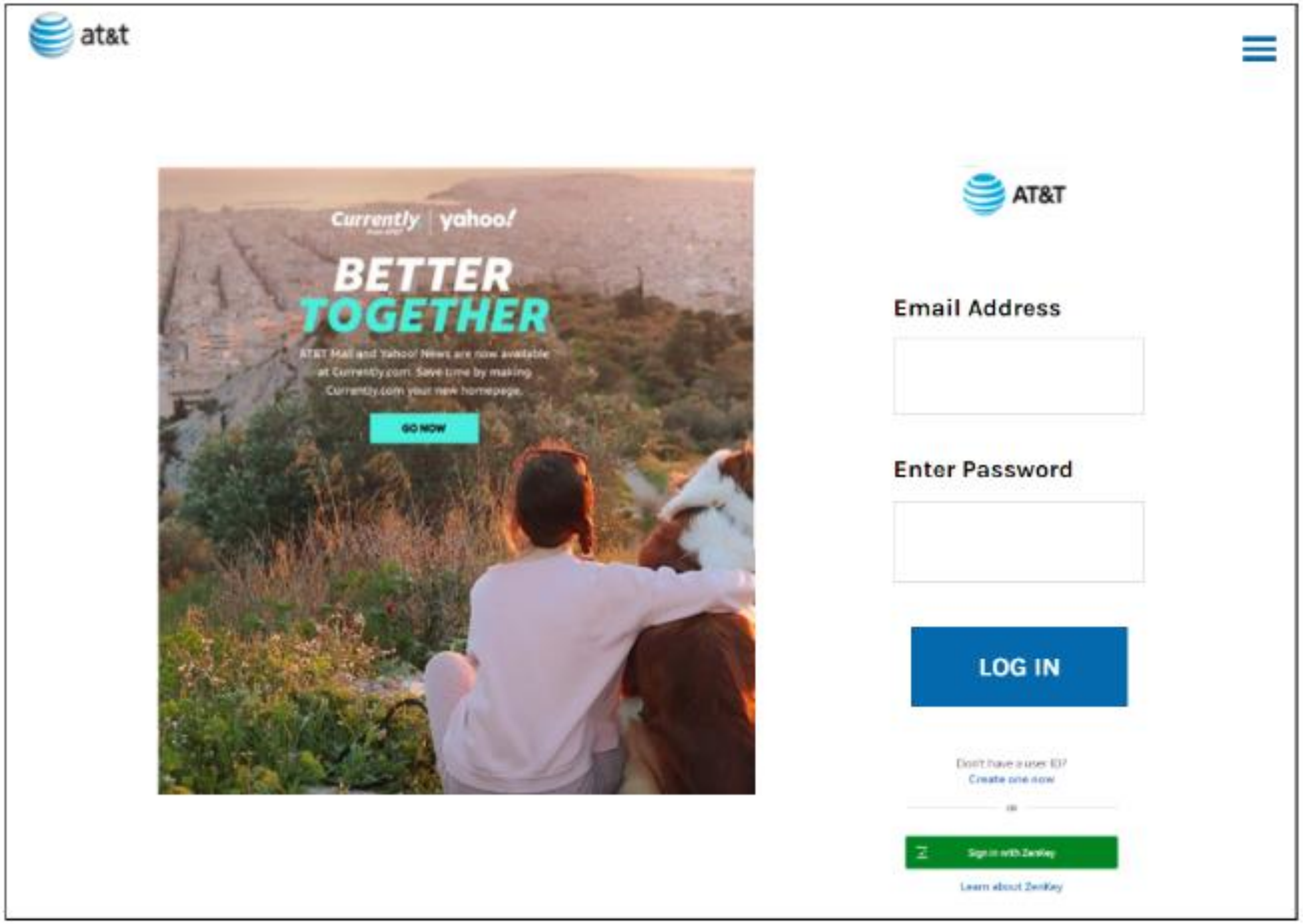


Question: Which 'space' do you think an *attacker* is **most likely** to use?

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/.



UNIVERSITÄT LIECHTENSTEIN

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

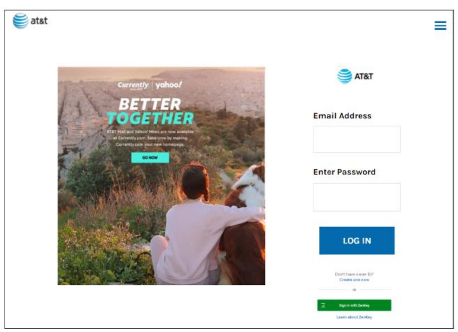
Website-space Perturbations (WsP) in practice – changing the URL

https://www.63y3hfh-fj39f30-f30if0f-f392.weebly.com/ → https://www.legitimate123.weebly.com/

UNIVERSITÄT LIECHTENSTEIN

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

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



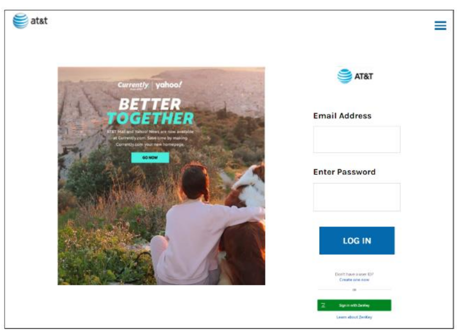

```
<div>
    <form enctype="multipart/form-data" action="//www.weebly.com/weebly/apps/formSubmit.php" method=
    "POST" id="form-723155629711391878">
        <div id="723155629711391878-form-parent" class="wsite-form-container"
            style="margin-top:10px;">
            <ul class="formlist" 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://www.legitimate123.weebly.com/



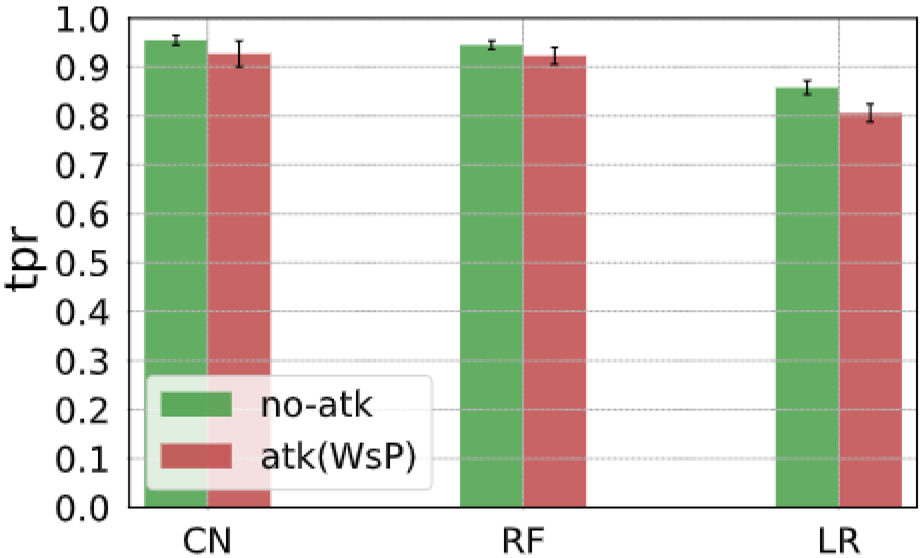

```
<div>
    <form enctype="multipart/form-data" action="//www.weebly.com/weebly/apps/formSubmit.php" method=
    "POST" id="form-723155629711391878">
        <div id="723155629711391878-form-parent" class="wsite-form-container"
                 style="margin-top:10px;">
            <ul class="formlist"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

Ɛ (WsP)

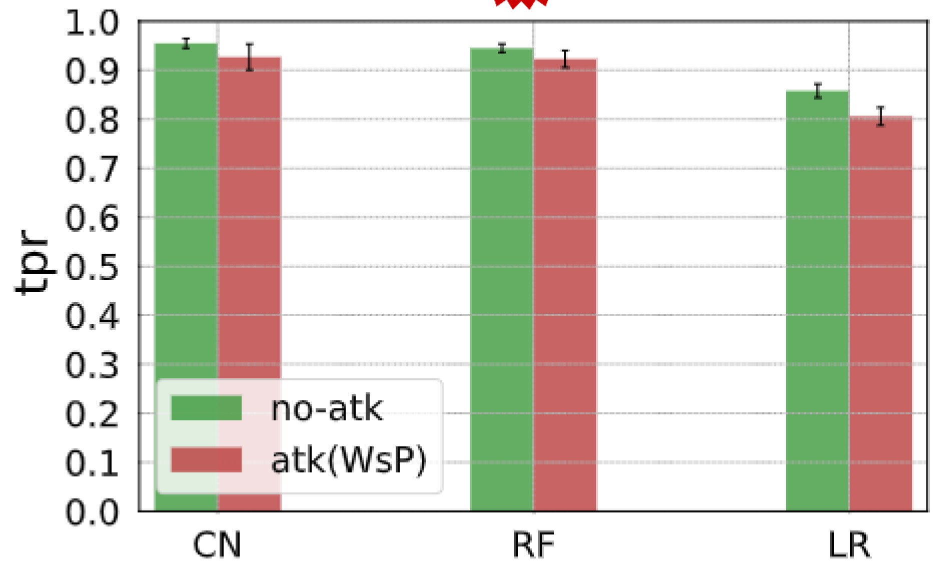
UNIVERSITÄT LIECHTENSTEIN

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

Evaluation – Are WsP effective?



(a) Zenodo. The plot shows the *tpr* before and after our WsP attack. The WsP entail invisible manipulations of the HTML. We repeat the experiments 50 times.



(b) δPhish. The plot shows the *tpr* before and after our WsP attack. The WsP entail invisible manipulations of the HTML. We repeat the experiments 50 times.

- 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 (not shown here), YES
  - This is very significant, because WsP are cheap and are likely to be exploited by attackers

UNIVERSITÄT
LIECHTENSTEIN

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

Demonstration: competition-grade ML-PWD

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

UNIVERSITÄT LIECHTENSTEIN

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

Demonstration: competition-grade ML-PWD

- https://spacephish.github.io (https://tinyurl.com/spacephish-demo)
- https://nbviewer.org/github/hihey54/acsac22_spacephish/blob/main/mlsec_folder/mlsec_artifact-manipulate.ipynb

```
def websiteAttacks_html(in_html,string,num):
    ind=in_html.find('</body>')
    content=""
    for i in range(0, num):
        content=content+string
    out_html=in_html[:ind]+content+in_html[ind:]
    return out_html
```

```
In [6]: # TEST ORIGINAL

        with open(original_fil
            original_data = f.
        original_response = re
        print(original_respons
```

```
{
  "n_models": 8,
  "p_mod_00": 0.891,
  "p_mod_01": 0.811,
  "p_mod_02": 0.891,
  "p_mod_03": 0.811,
  "p_mod_04": 0.806,
  "p_mod_05": 0.741,
  "p_mod_06": 0.806,
  "p_mod_07": 0.741
}
```

```
In [8]: # TEST ADVERSARIAL

        with open(output_file,
            adversarial_data =
        adversarial_response =
        print(adversarial_respo
```

```
{
  "n_models": 8,
  "p_mod_00": 0.426,
  "p_mod_01": 0.794,
  "p_mod_02": 0.426,
  "p_mod_03": 0.794,
  "p_mod_04": 0.864,
  "p_mod_05": 0.774,
  "p_mod_06": 0.794,
  "p_mod_07": 0.741
}
```

UNIVERSITÄT LIECHTENSTEIN

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

Demonstration: competition-grade ML-PWD

- https://spacephish.github.io (https://tinyurl.com/spacephish-demo)
- https://nbviewer.org/github/hihey54/acsac22_spacephish/blob/main/mlsec_folder/mlsec_artifact-manipulate.ipynb

```python
def websiteAttacks_html(in_html,string,num):
    ind=in_html.find('</body>')
    content=""
    for i in range(0, num):
        content=content+string
    out_html=in_html[:ind]+content+in_html[ind:]
    return out_html
```

```python
In [6]: # TEST ORIGINAL

        with open(original_fil
            original_data = f.
        original_response = re
        print(original_respons
```

```
{
  "n_models": 8,
  "p_mod_00": 0.891,
  "p_mod_01": 0.811,
  "p_mod_02": 0.891,
  "p_mod_03": 0.811,
  "p_mod_04": 0.806,
  "p_mod_05": 0.741,
  "p_mod_06": 0.806,
  "p_mod_07": 0.741
}
```

```python
In [8]: # TEST ADVERSARIAL

        with open(output_file,
            adversarial_data =
        adversarial_response =
        print(adversarial_respo
```

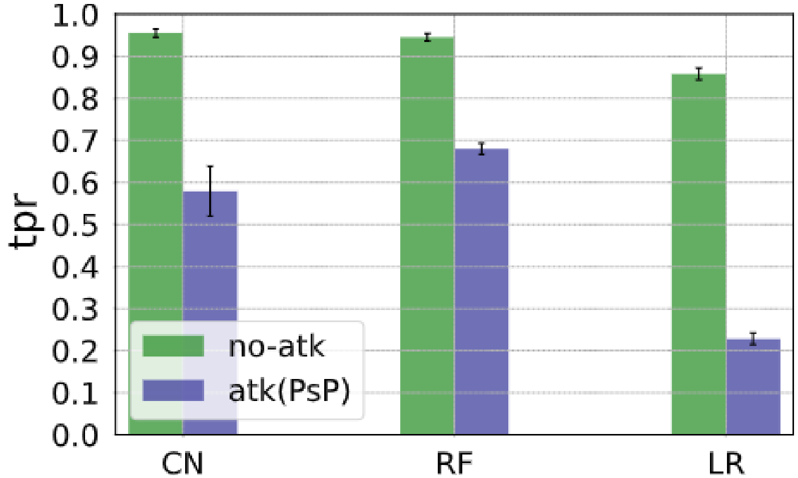
```
{
  "n_models": 8,
  "p_mod_00": 0.426,
  "p_mod_01": 0.794,
  "p_mod_02": 0.426,
  "p_mod_03": 0.794,
  "p_mod_04": 0.864,
  "p_mod_05": 0.774,
  "p_mod_06": 0.794,
  "p_mod_07": 0.741
}
```

UNIVERSITÄT LIECHTENSTEIN

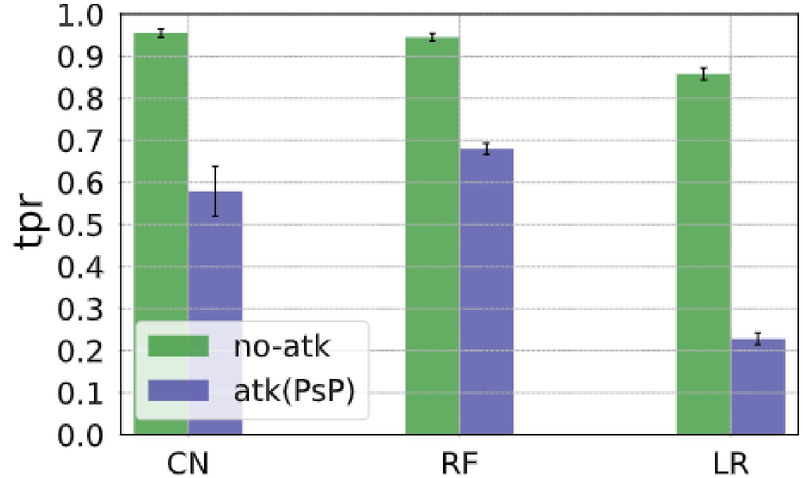
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Evaluation – What about perturbation in the other spaces?

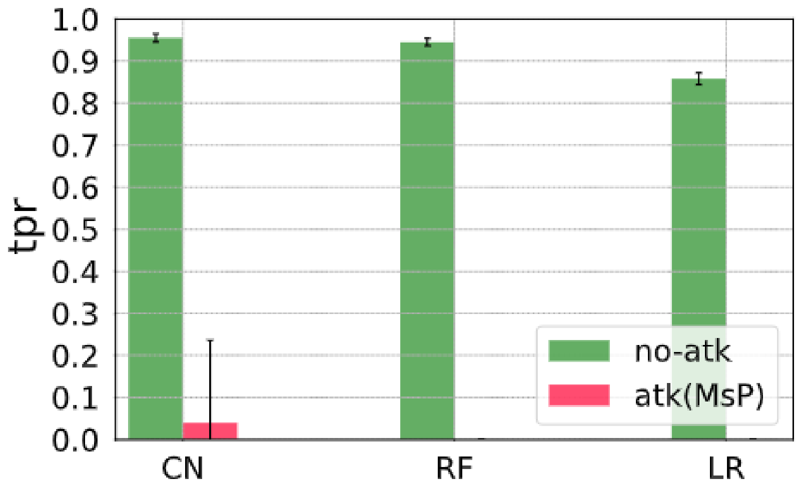
In general, attacks in the other spaces (via PsP and MsP) are more disruptive…

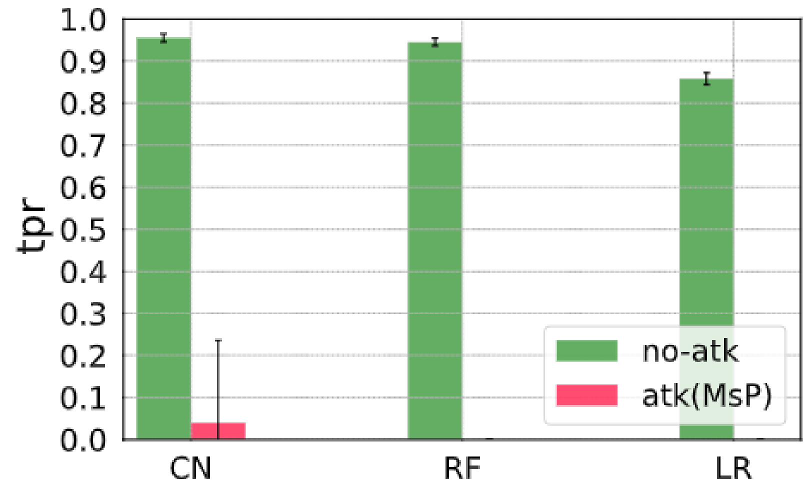


(a) Zenodo. The plot shows the *tpr* before and after our (blind) PsP attack.



(b) δPhish. The plot shows the *tpr* before and after our (blind) PsP attack.





(a) Zenodo. The plot sl … MsP attack.

LIECHTENSTE

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

What about the real world? (from [SaTML'23])

UNIVERSITÄT
LIECHTENSTEIN

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

What about the real world? (from [SaTML'23])

- We asked a well-known **cybersecurity company** to provide us with data from their (operational!) phishing website detector, empowered by *deep learning*
  - This system uses a reference-based mechanism, similar to the one in PhishIntention [6]

UNIVERSITÄT LIECHTENSTEIN

[6]: Liu, R., Lin, Y., Yang, X., Ng, S. H., Divakaran, D. M., & Dong, J. S. (2022). Inferring phishing intention via webpage appearance and dynamics: A deep vision based approach. In *31st USENIX Security Symposium (USENIX Security 22)* (pp. 1633-1650).

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

What about the real world? (from [SaTML'23])

- We asked a well-known **cybersecurity company** to provide us with data from their (operational!) phishing website detector, empowered by *deep learning*
  - This system uses a reference-based mechanism, similar to the one in PhishIntention [6]

- Just in July 2022, there were **9K samples** for which the ML detector was "uncertain"
  - In practice, these samples have been deemed as "benign" to avoid triggering false positives
  - However, they were "close to the decision boundary", and required manual triage by experts

- We **manually analyzed** these (phishing) samples, trying to understand cases of failure of these state-of-the-art phishing detection systems

What did we find?

[6]: Liu, R., Lin, Y., Yang, X., Ng, S. H., Divakaran, D. M., & Dong, J. S. (2022). Inferring phishing intention via webpage appearance and dynamics: A deep vision based approach. In *31st USENIX Security Symposium (USENIX Security 22)* (pp. 1633-1650).

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

What about the real world? (from [SaTML'23]) [cont'd]

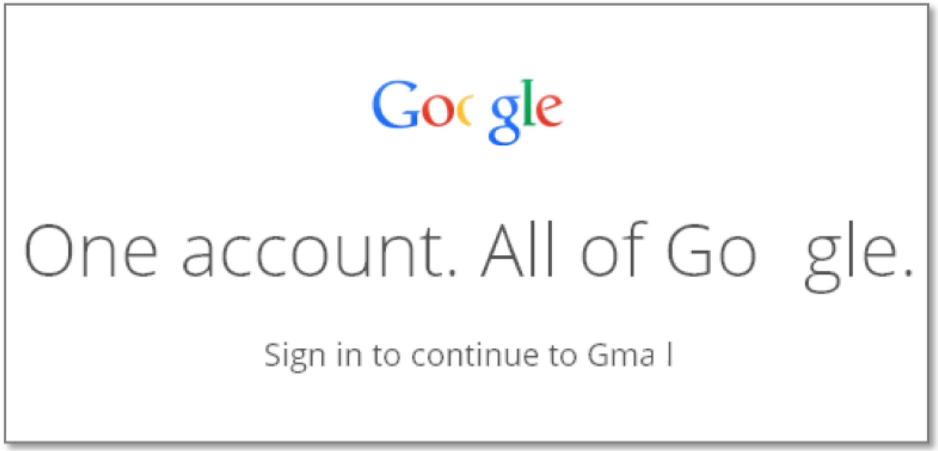
- The **vast majority** of these webpages were "out of distribution"
  - They were different from any sample in the training set
- We then looked at a small subset of the remaining ones…

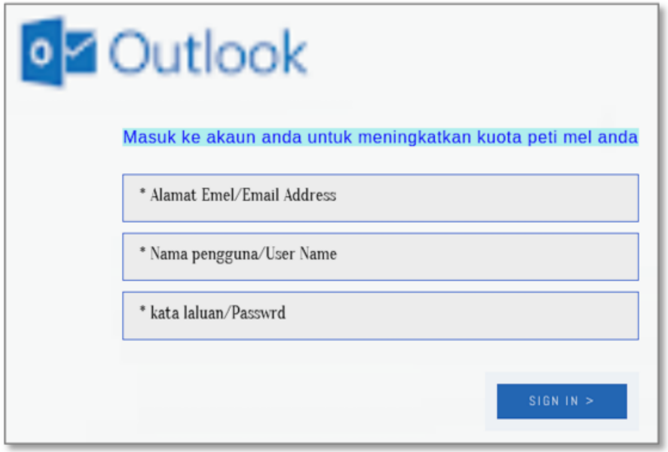
UNIVERSITÄT LIECHTENSTEIN

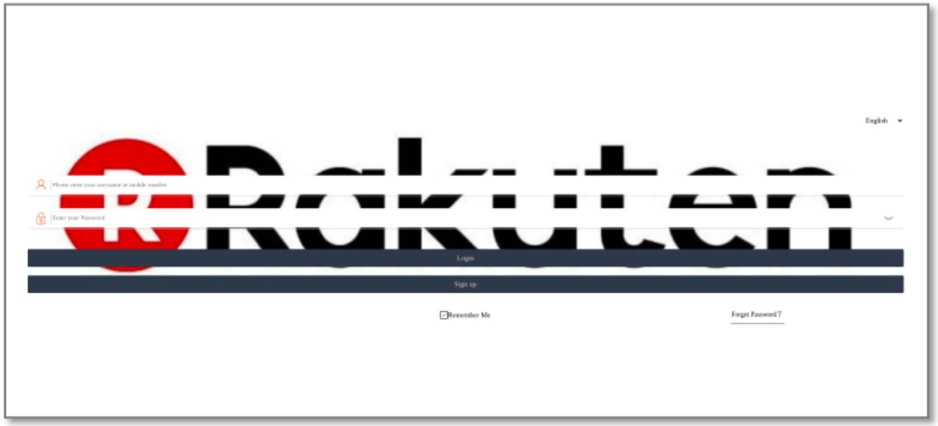
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

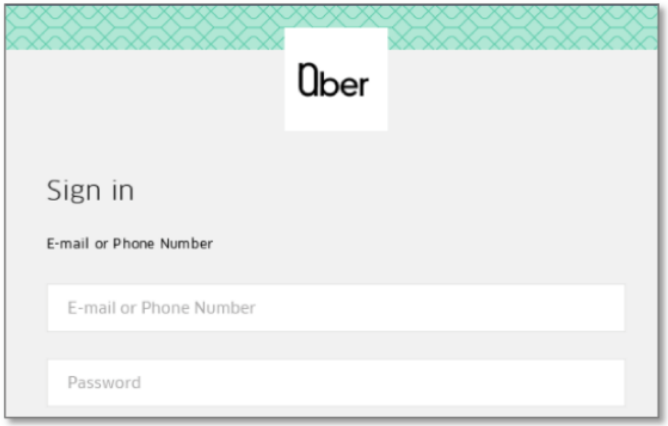
What about the real world? (from [SaTML'23]) [cont'd]

- The **vast majority** of these webpages were "out of distribution"
  - They were different from any sample in the training set
- We then looked at a small subset of the remaining ones…





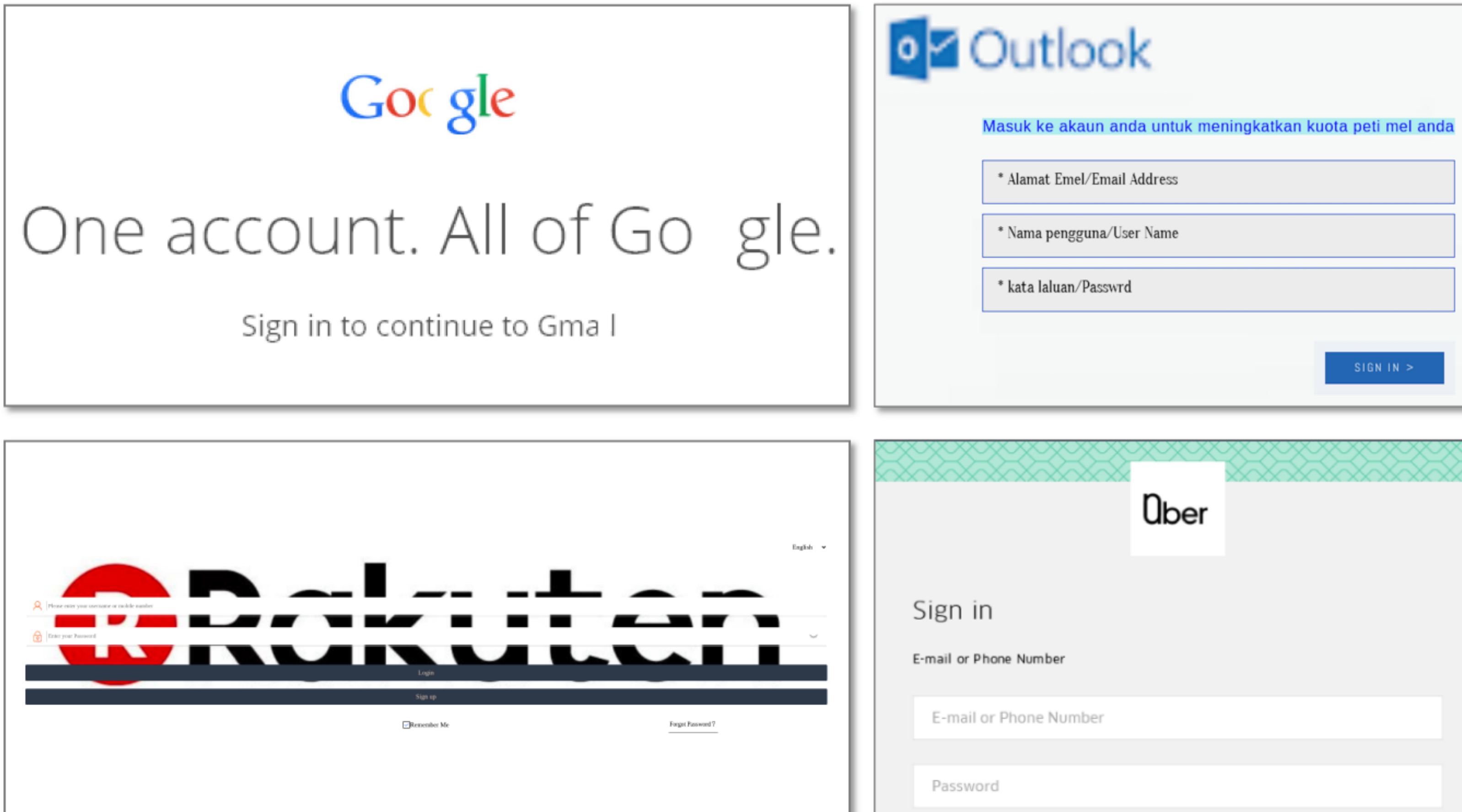




Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What about the real world? (from [SaTML'23]) [cont'd]

- The **vast majority** of these webpages were "out of distribution"
  - They were different from any sample in the training set
- We then looked at a small subset of the remaining ones…



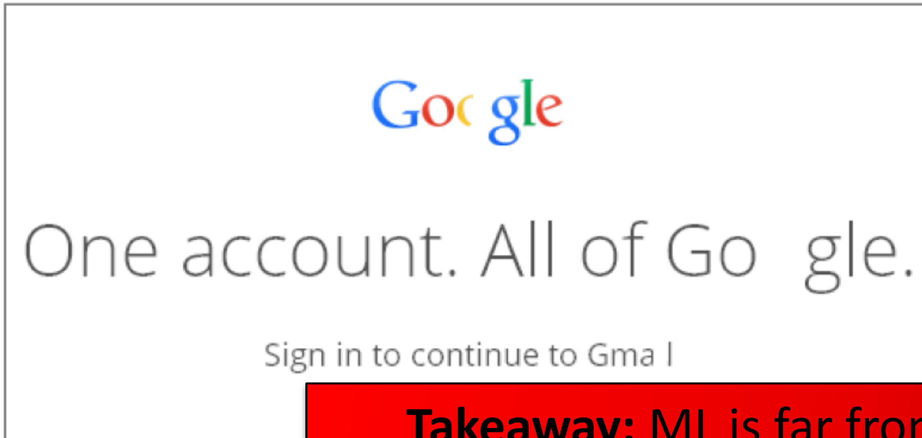
These techniques have been known for decades…
but can still evade modern (and real) *ML systems*.

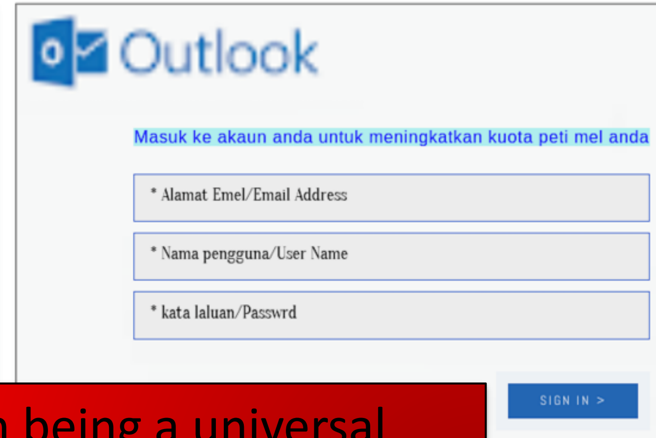
And they're cheap!

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

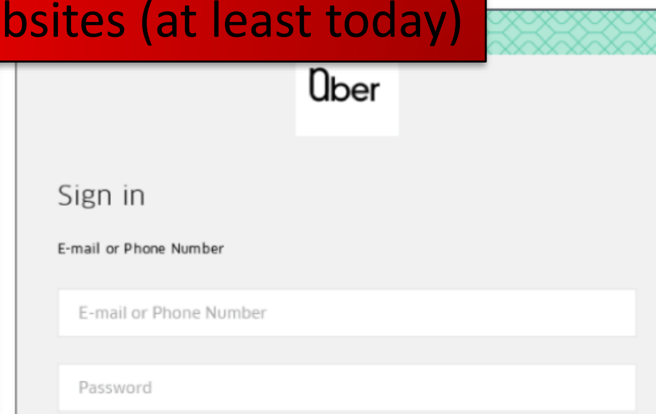
What about the real world? (from [SaTML'23]) [cont'd]

- The **vast majority** of these webpages were "out of distribution"
  - They were different from any sample in the training set
- We then looked at a small subset of the remaining ones…





Takeaway: ML is far from being a universal solution against phishing websites (at least today)



These techniques have been known for decades… but can still evade modern (and real) *ML systems*.

And they're cheap!

Evading Phishing Website Detectors

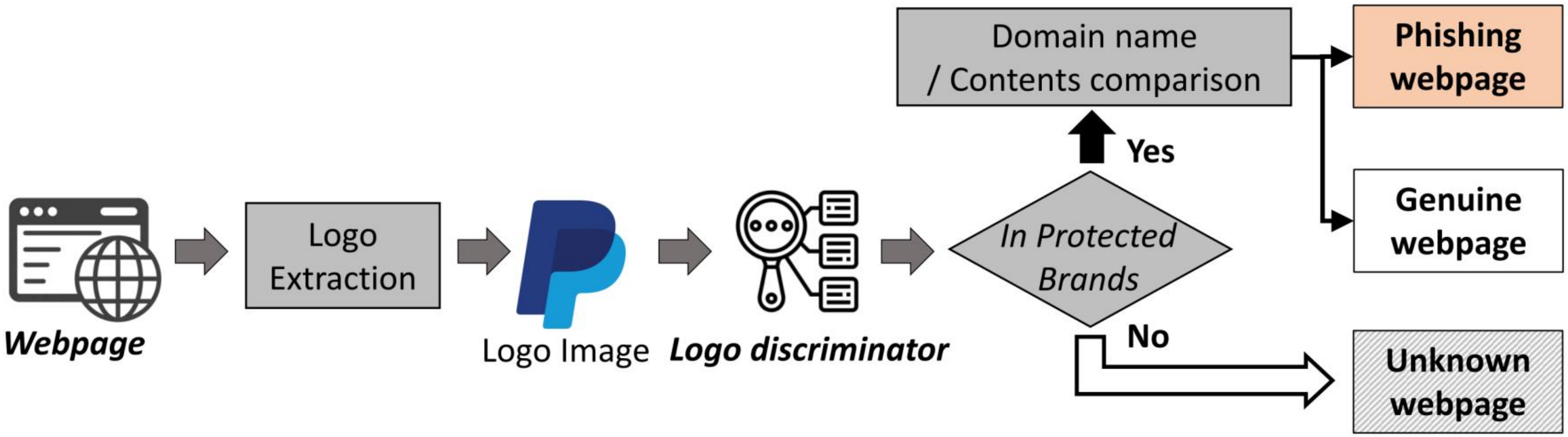
Algorithmically
(via ML)

UNIVERSITÄT
LIECHTENSTEIN

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

Logo-based Phishing Website Detection

in a nutshell



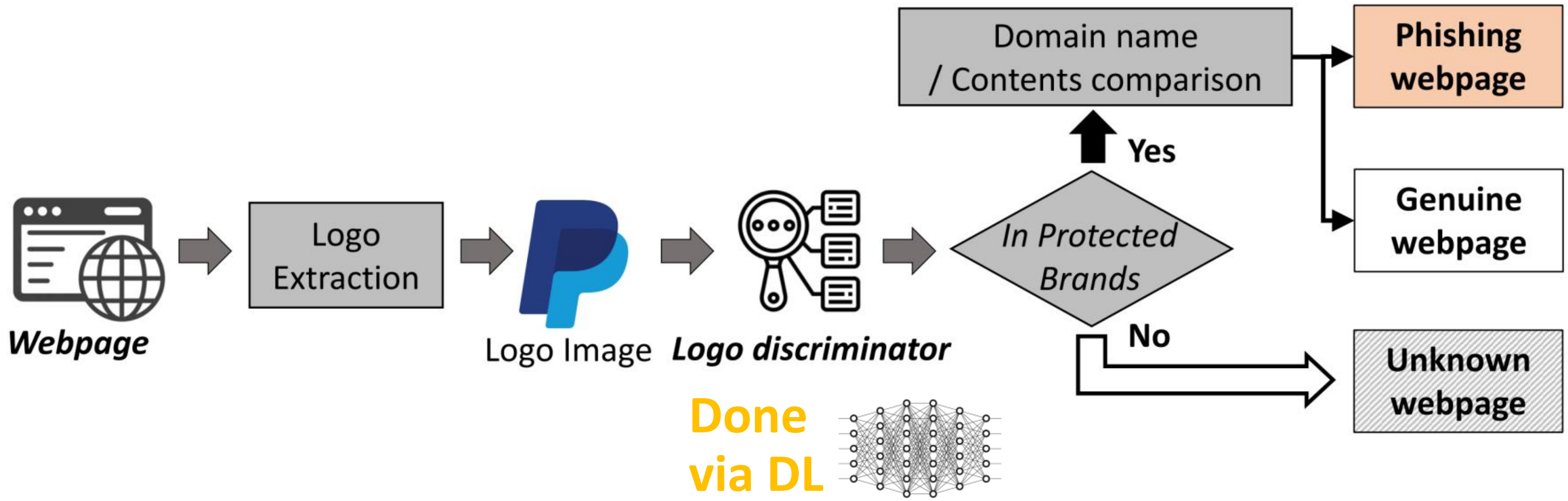
- Note: this architecture resembles that of PhishIntention [6]

[6]: Liu, R., Lin, Y., Yang, X., Ng, S. H., Divakaran, D. M., & Dong, J. S. (2022). Inferring phishing intention via webpage appearance and dynamics: A deep vision based approach. In *31st USENIX Security Symposium (USENIX Security 22)* (pp. 1633-1650).

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

Logo-based Phishing Website Detection

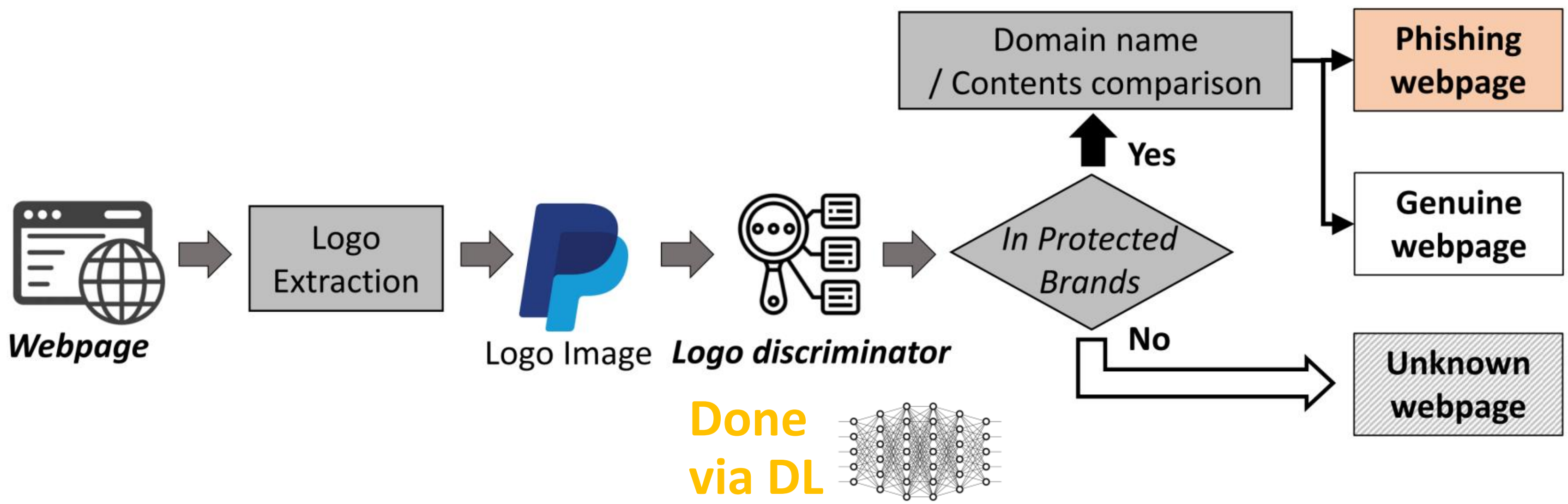
in a nutshell



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

Logo-based Phishing Website Detection

in a nutshell

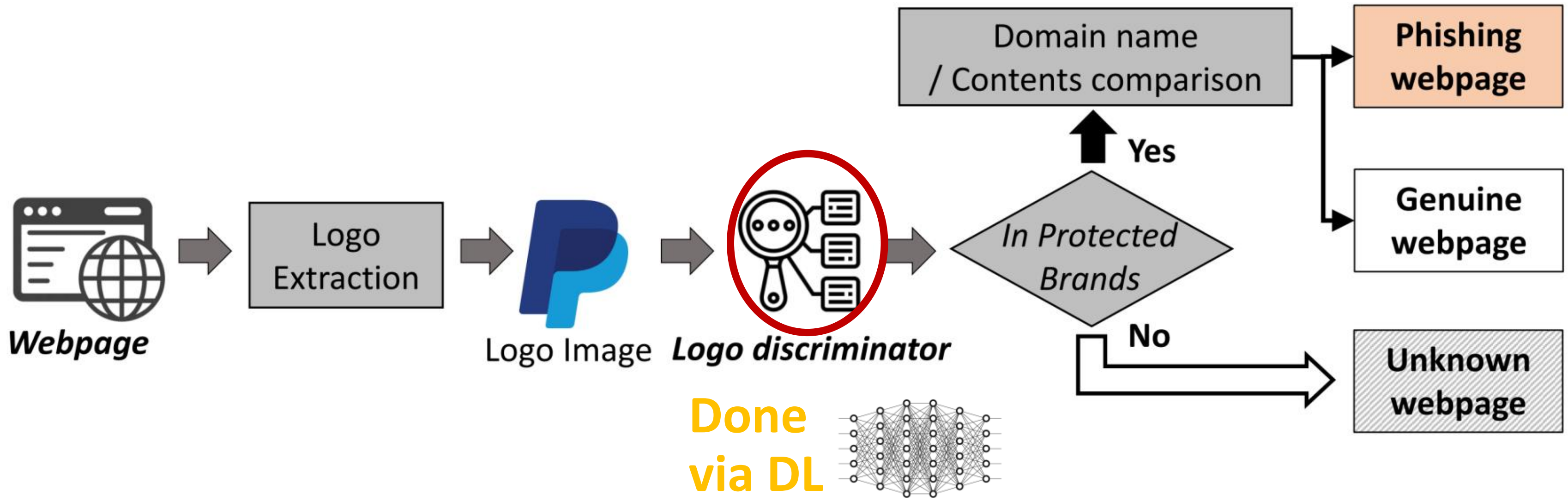


Problem: these systems are tweaked to minimize false positives.

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

Logo-based Phishing Website Detection

in a nutshell



Problem: these systems are tweaked to minimize false positives.

We focus on the Logo-discriminator.

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

Our attack: adversarial logos

Intuition: create an adversarial logo that is (i) minimally altered w.r.t. its original variant; and that (ii) misleads the logo discriminator.

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

Our attack: adversarial logos

Intuition: create an adversarial logo that is (i) minimally altered w.r.t. its original variant; and that (ii) misleads the logo discriminator.

1. **Knowledge:**

2. **Capabilities:**

3. **Strategy:**

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

Our attack: adversarial logos

Intuition: create an adversarial logo that is (i) minimally altered w.r.t. its original variant; and that (ii) misleads the logo discriminator.

No knowledge of the DL model is required!

1. **Knowledge:**
   - the attacker expects the detector to have the "phished" brand(s) in the protected set (and that its logos are inspected)
2. **Capabilities:**
3. **Strategy:**

UNIVERSITÄT LIECHTENSTEIN

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

Our attack: adversarial logos

Intuition: create an adversarial logo that is (i) minimally altered w.r.t. its original variant; and that (ii) misleads the logo discriminator.

No knowledge of the DL model is required!

1. **Knowledge:**
   - the attacker expects the detector to have the "phished" brand(s) in the protected set (and that its logos are inspected)
2. **Capabilities:**
   - the attacker can observe the decision of the detector
   - the attacker can manipulate their phishing webpages

The attacker can do nothing to the training data.

3. **Strategy:**

UNIVERSITÄT LIECHTENSTEIN

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

Our attack: adversarial logos

Intuition: create an adversarial logo that is (i) minimally altered w.r.t. its original variant; and that (ii) misleads the logo discriminator.

No knowledge of the DL model is required!

1. **Knowledge:**
   - the attacker expects the detector to have the "phished" brand(s) in the protected set (and that its logos are inspected)
2. **Capabilities:**
   - the attacker can observe the decision of the detector
   - the attacker can manipulate their phishing webpages

The attacker can do nothing to the training data.

3. **Strategy:** Manipulate the logo so that the discriminator has a lower confidence → the detector will default to a "unknown webpage"

UNIVERSITÄT LIECHTENSTEIN

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

Evaluation: Baseline

- We propose two novel methods for logo-identification: ViT and Swin
  - Both ViT and Swin leverage transformers [23, 36].

We are the first to use transformers for logo-identification (ttbook)

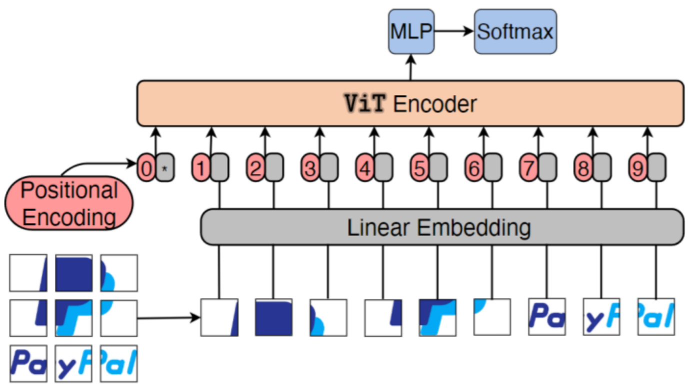


Fig. 2: ViT-based Model Architecture

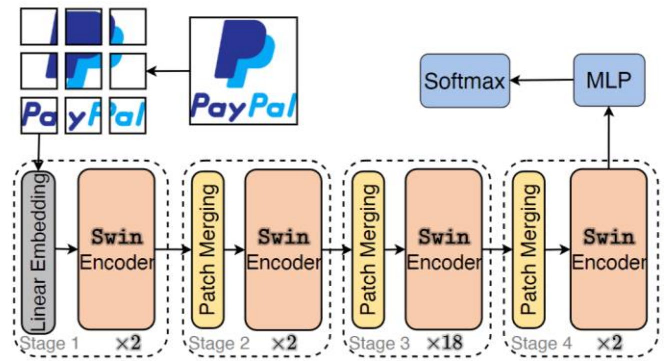


Fig. 3: Swin-based Model Architecture

[23] Dosovitskiy, A., et al.: *An image isworth 16x16 words: Transformers for image recognition at scale.* arXiv:2010.11929 (2020)
[36] Liu, Z., et al. : Swin transformer: Hierarchical vision transformer using shifted windows. IEEE/CVF ICCV (2021)

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

Evaluation: Baseline

- We propose two novel methods for logo-identification: ViT and Swin
  - Both ViT and Swin leverage transformers [23, 36].

We are the first to use transformers for logo-identification (ttbook)

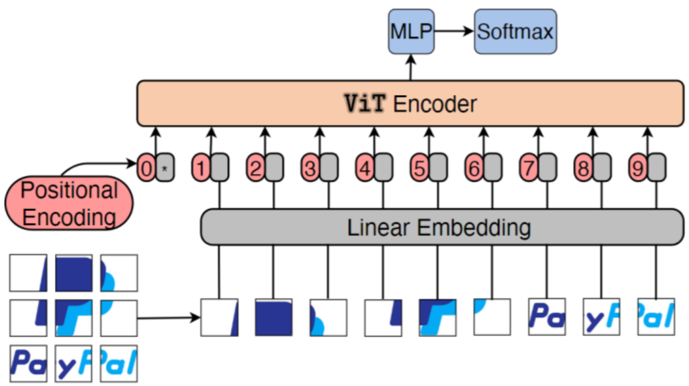


Fig. 2: ViT-based Model Architecture

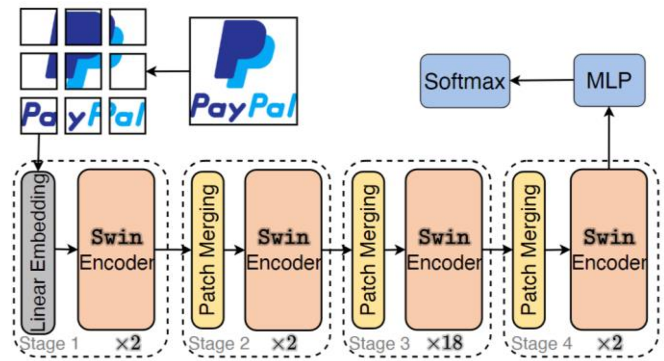


Fig. 3: Swin-based Model Architecture

- We will show that these methods reach state-of-the-art performance (currently obtained by Siamese networks [34])

UNIVERSITÄT LIECHTENSTEIN

[23] Dosovitskiy, A., et al.: *An image isworth 16x16 words: Transformers for image recognition at scale.* arXiv:2010.11929 (2020)
[36] Liu, Z., et al. : Swin transformer: Hierarchical vision transformer using shifted windows. IEEE/CVF ICCV (2021)
[34]: Lin, Y., et al.: *Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages.* USENIX Security (2021)

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

Evaluation: Attack

We are inspired by "GAN"

- Our attack applies a "Generative Adversarial Perturbations" (GAP)

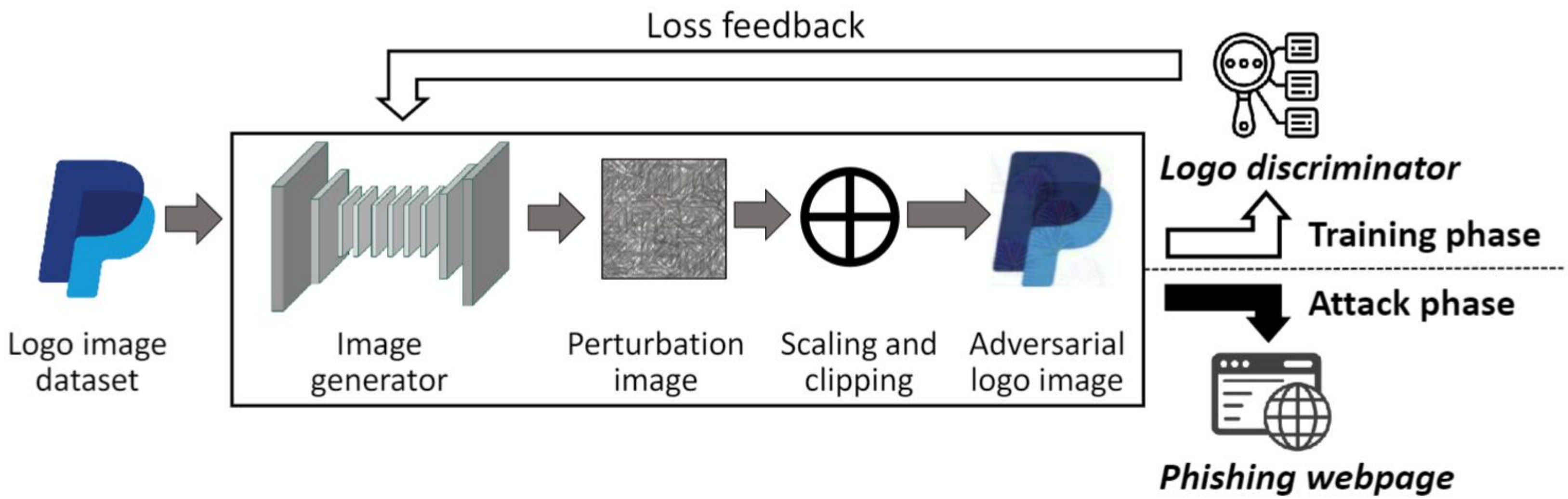


Fig. 4: Generative adversarial perturbation workflow

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

Evaluation: Attack

We are inspired by "GAN"

- Our attack applies a "Generative Adversarial Perturbations" (GAP)

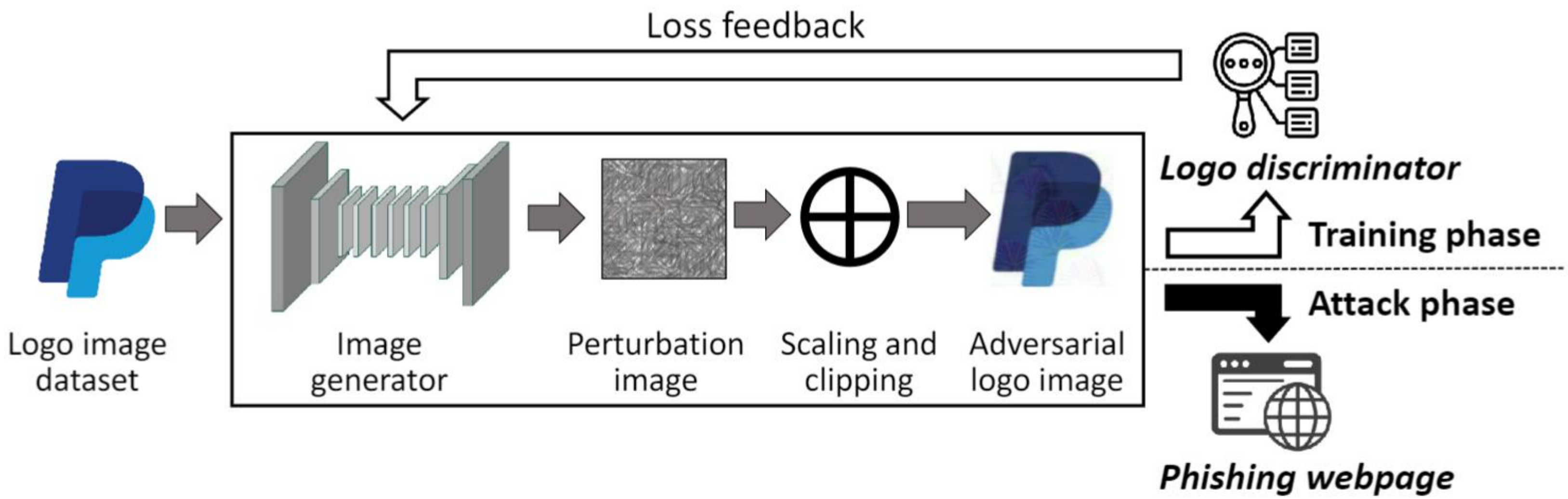


Fig. 4: Generative adversarial perturbation workflow

- The GAP automatically "learns" to craft adversarial logos that mislead the logo discriminator – while being minimally altered.

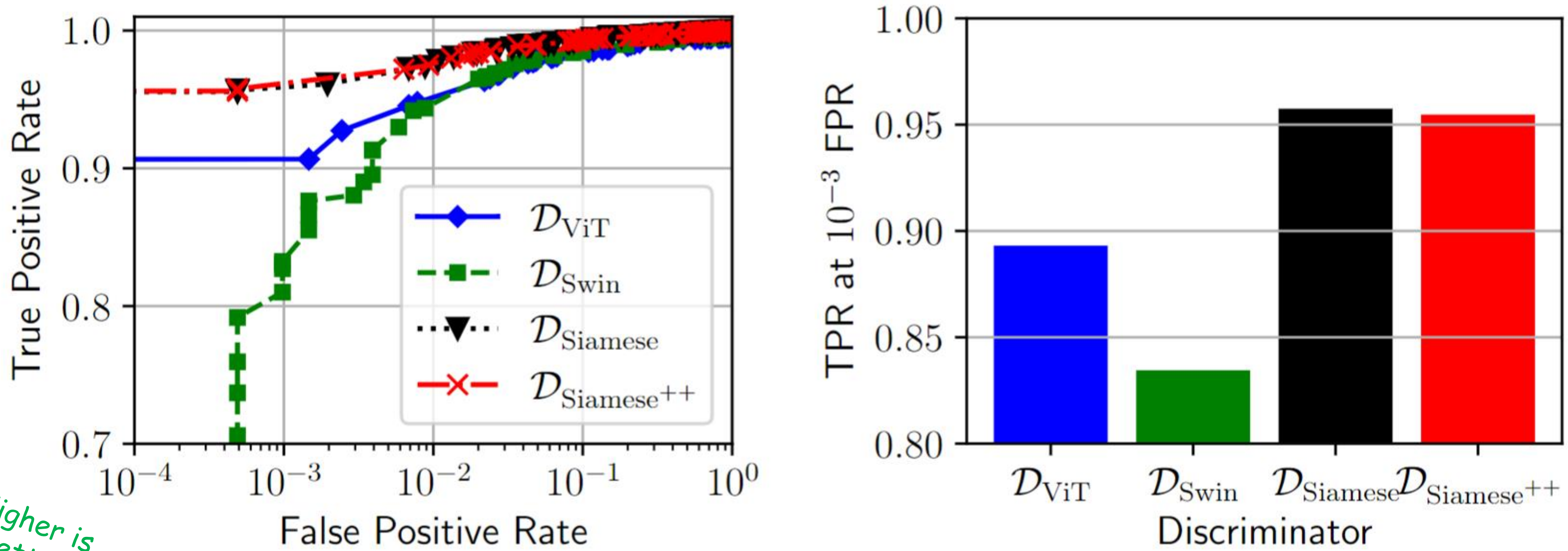
We will assess the cross-model transferability of our adversarial logos!

UNIVERSITÄT LIECHTENSTEIN

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

Results: Baseline

$\mathcal{D}_{Siamese++}$ is a "robust" version of Siamese networks



Higher is better

(a) ROC curves

(b) TPR at 10^{-3} FPR

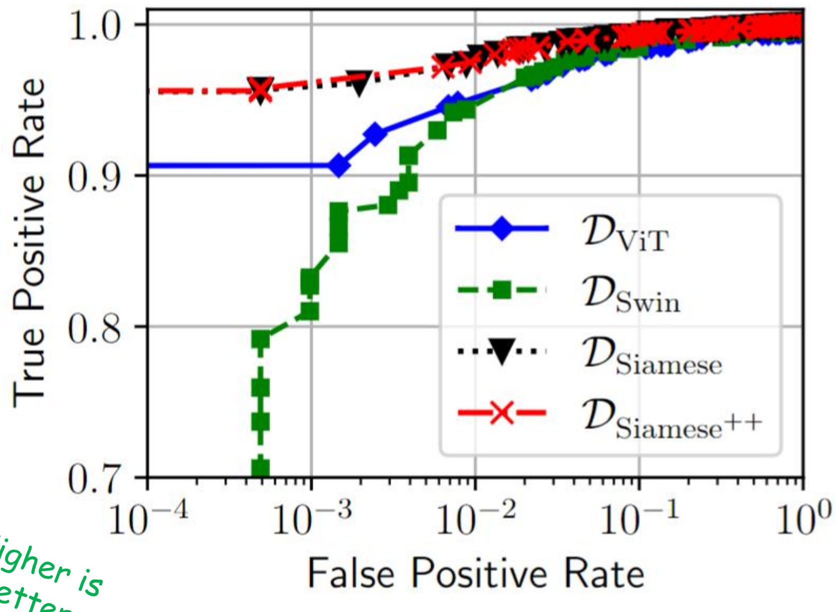
Our baselines are trained to identify 181 brands (~28k logos)

UNIVERSITÄT LIECHTENSTEIN

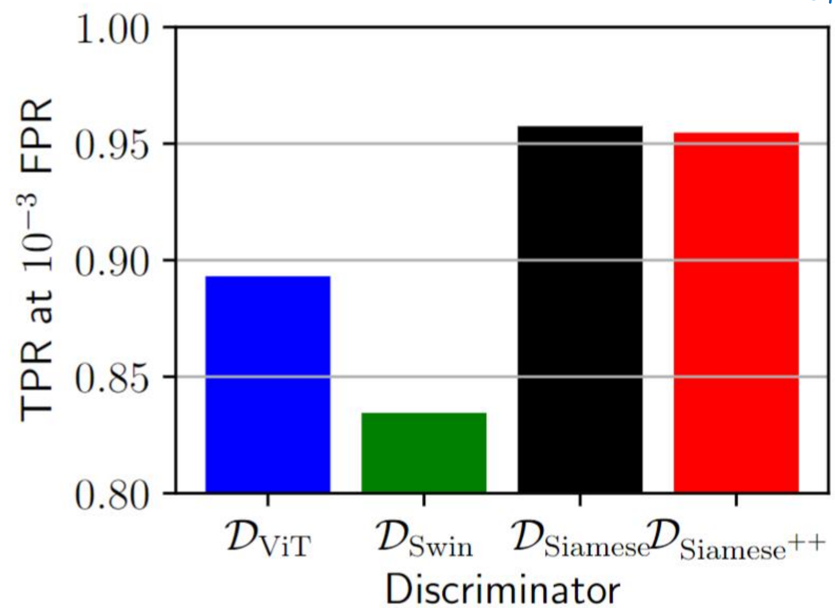
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Baseline

$\mathcal{D}_{Siamese++}$ is a "robust" version of Siamese networks



(a) ROC curves



(b) TPR at 10^{-3} FPR

Higher is better

Takeaways:

1. Our baselines "work well" (in the absence of attacks!)
2. ViT and Swin are slightly worse than Siamese…

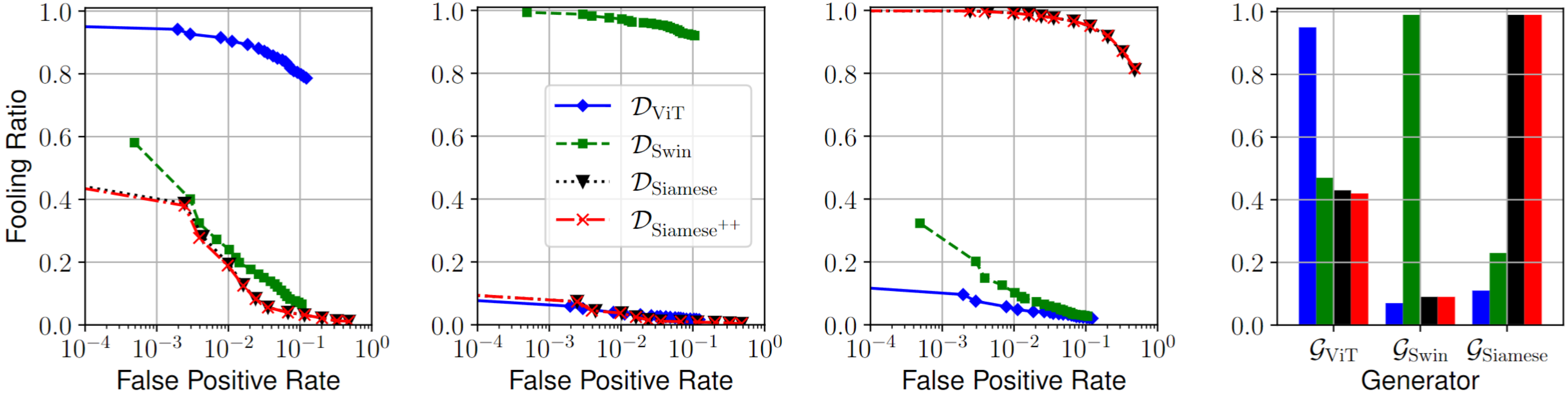
Our baselines are trained to identify 181 brands (~28k logos)

UNIVERSITÄT LIECHTENSTEIN

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

Results: Attack

Higher = stronger attack



(a) $\mathcal{G}_{\text{ViT}}$ **(b)** $\mathcal{G}_{\text{Swin}}$ **(c)** $\mathcal{G}_{\text{Siamese}}$ **(d)** at 10^{-3} FPR

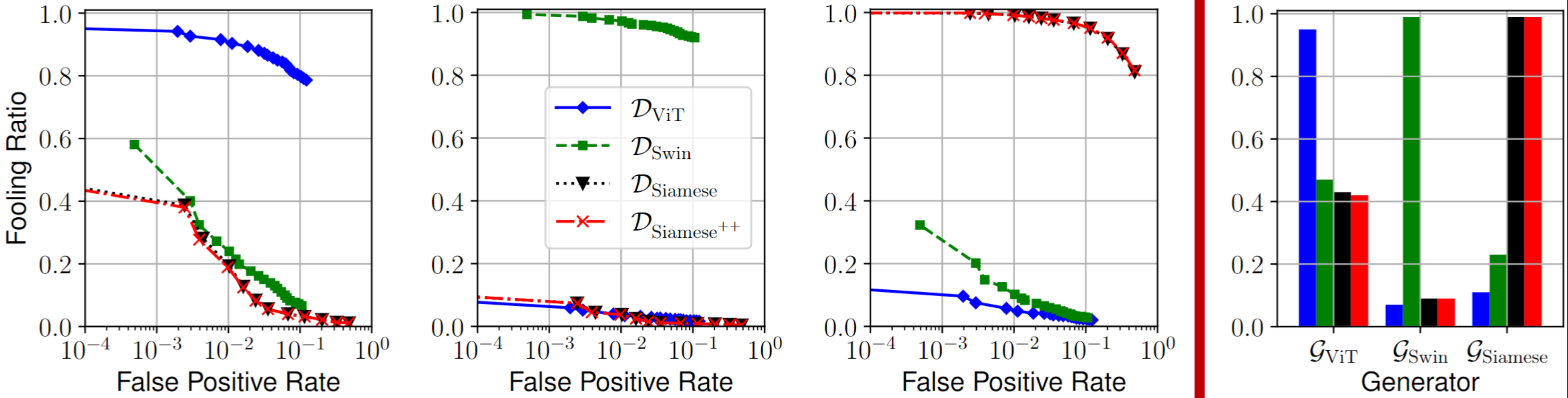
E.g.: G_{ViT} denotes the GAN trained to evade D_{ViT}

UNIVERSITÄT LIECHTENSTEIN

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

Results: Attack

Higher = stronger attack



(a) $\mathcal{G}_{\text{ViT}}$ **(b)** $\mathcal{G}_{\text{Swin}}$ **(c)** $\mathcal{G}_{\text{Siamese}}$ **(d)** at 10^{-3} FPR

E.g.: G_{ViT} denotes the GAN trained to evade D_{ViT}

Takeaways:

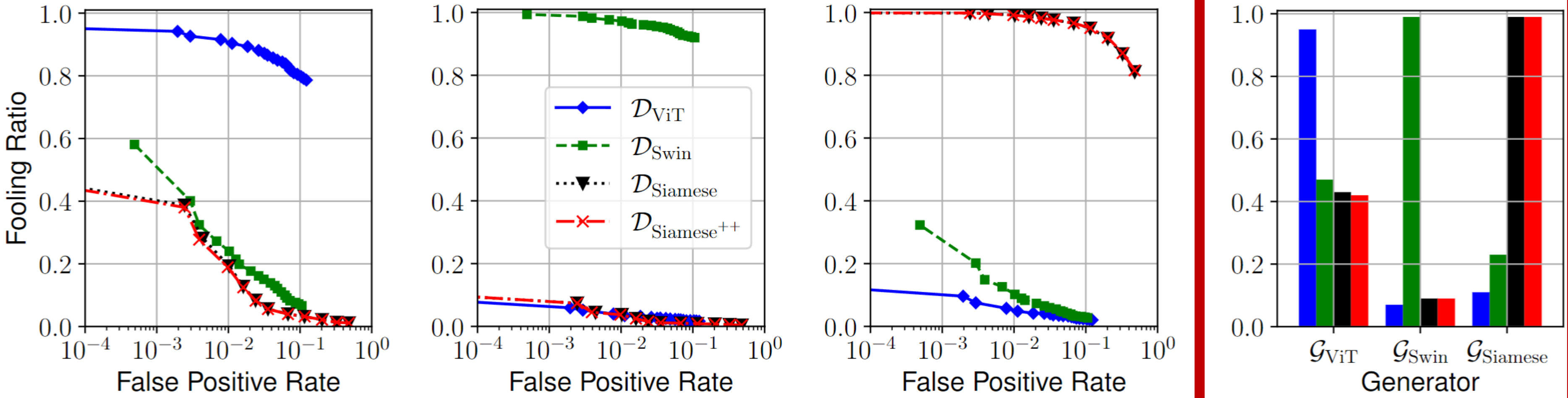
1. When the attacker and defender use the same model, the attack is ~100% effective
2. ViT is the "more robust" detector! (if the attacker is blind)

UNIVERSITÄT LIECHTENSTEIN

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

Results: Attack

Higher = stronger attack



(a) $\mathcal{G}_{\mathrm{ViT}}$ **(b)** $\mathcal{G}_{\mathrm{Swin}}$ **(c)** $\mathcal{G}_{\mathrm{Siamese}}$ **(d)** at 10^{-3} FPR

E.g.: G_{ViT} denotes the GAN trained to evade D_{ViT}

Takeaways:

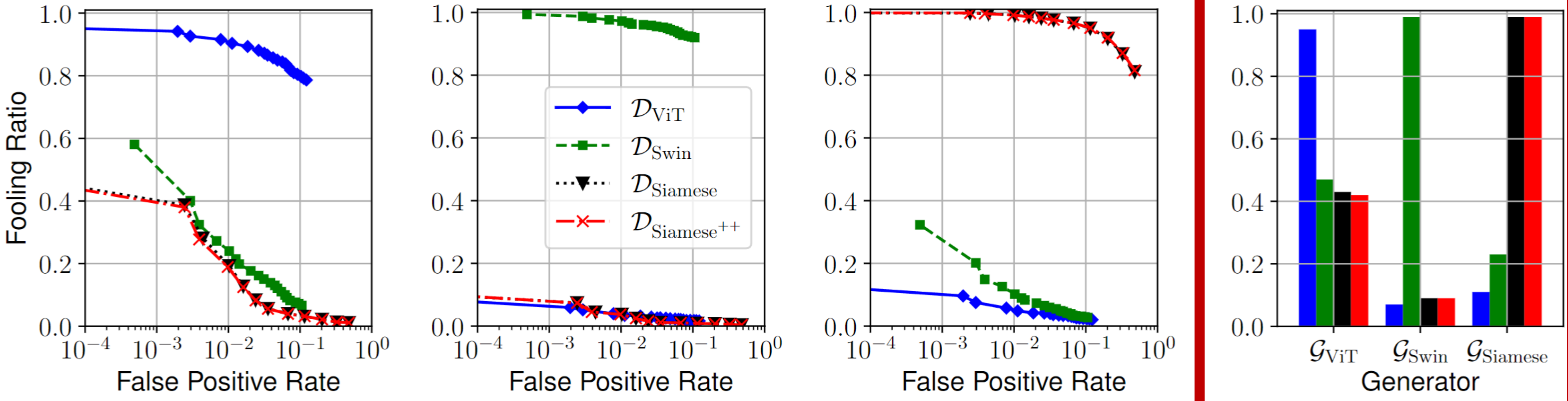
1. When the attacker and defender use the same model, the attack is ~100% effective
2. ViT is the “more robust” detector! (if the attacker is blind)

UNIVERSITÄT LIECHTENSTEIN

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

Results: Attack

Higher = stronger attack



(a) $\mathcal{G}_{\mathrm{ViT}}$ **(b)** $\mathcal{G}_{\mathrm{Swin}}$ **(c)** $\mathcal{G}_{\mathrm{Siamese}}$ **(d)** at 10^{-3} FPR

E.g.: G_{ViT} denotes the GAN trained to evade D_{ViT}

Takeaways:

1. When the attacker and defender use the same model, the attack is ~100% effective
2. ViT is the "more robust" detector! (if the attacker is blind)

UNIVERSITÄT LIECHTENSTEIN

However, these attacks only focused on the logo-discriminator: what about the overarching phishing detection system?

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

Another attack (against the end-to-end phishing detection system)

- In our USENIX Sec’24 paper, we devise a stronger attack, “LogoMorph”, which we test against various phishing website detectors reliant on visual similarity.

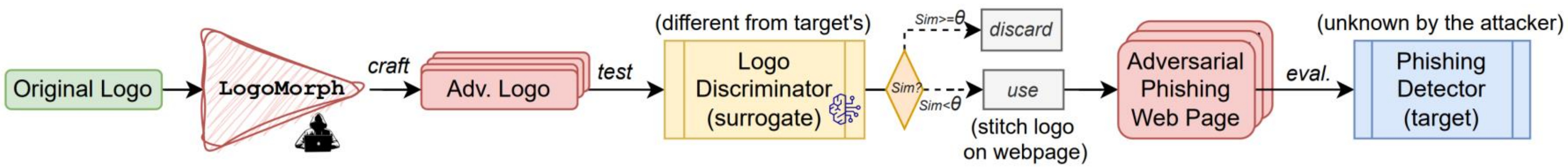


Figure 6: **Our Blackbox Experiment Setup.**—We use a surrogate logo discriminator (which is different from the one used by the target model) to generate and select adversarial logos via `LogoMorph`. Logos that bypass the surrogate discriminator (by achieving a low similarity) will be used to attack the targeted phishing detector at the webpage level.

UNIVERSITÄT LIECHTENSTEIN

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

Another attack (against the end-to-end phishing detection system)

- In our USENIX Sec'24 paper, we devise a stronger attack, "LogoMorph", which we test against various phishing website detectors reliant on visual similarity.

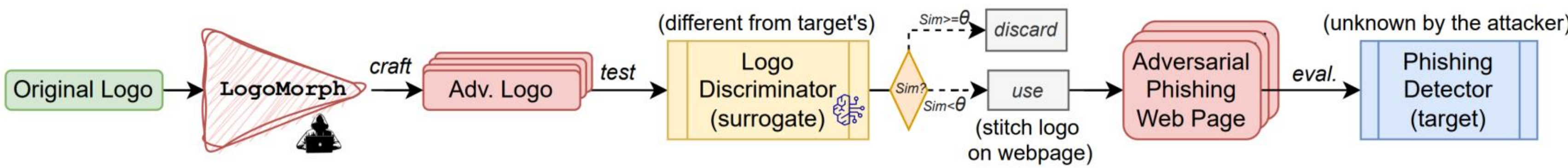


Figure 6: **Our Blackbox Experiment Setup.**—We use a surrogate logo discriminator (which is different from the one used by the target model) to generate and select adversarial logos via `LogoMorph`. Logos that bypass the surrogate discriminator (by achieving a low similarity) will be used to attack the targeted phishing detector at the webpage level.

- The attack leverages *diffusion models* to create an adversarial logo that is minimally altered, preserving its semantics, and which can fool the system end-to-end

- We also consider changing the *font* of a logo (if it has textual elements)

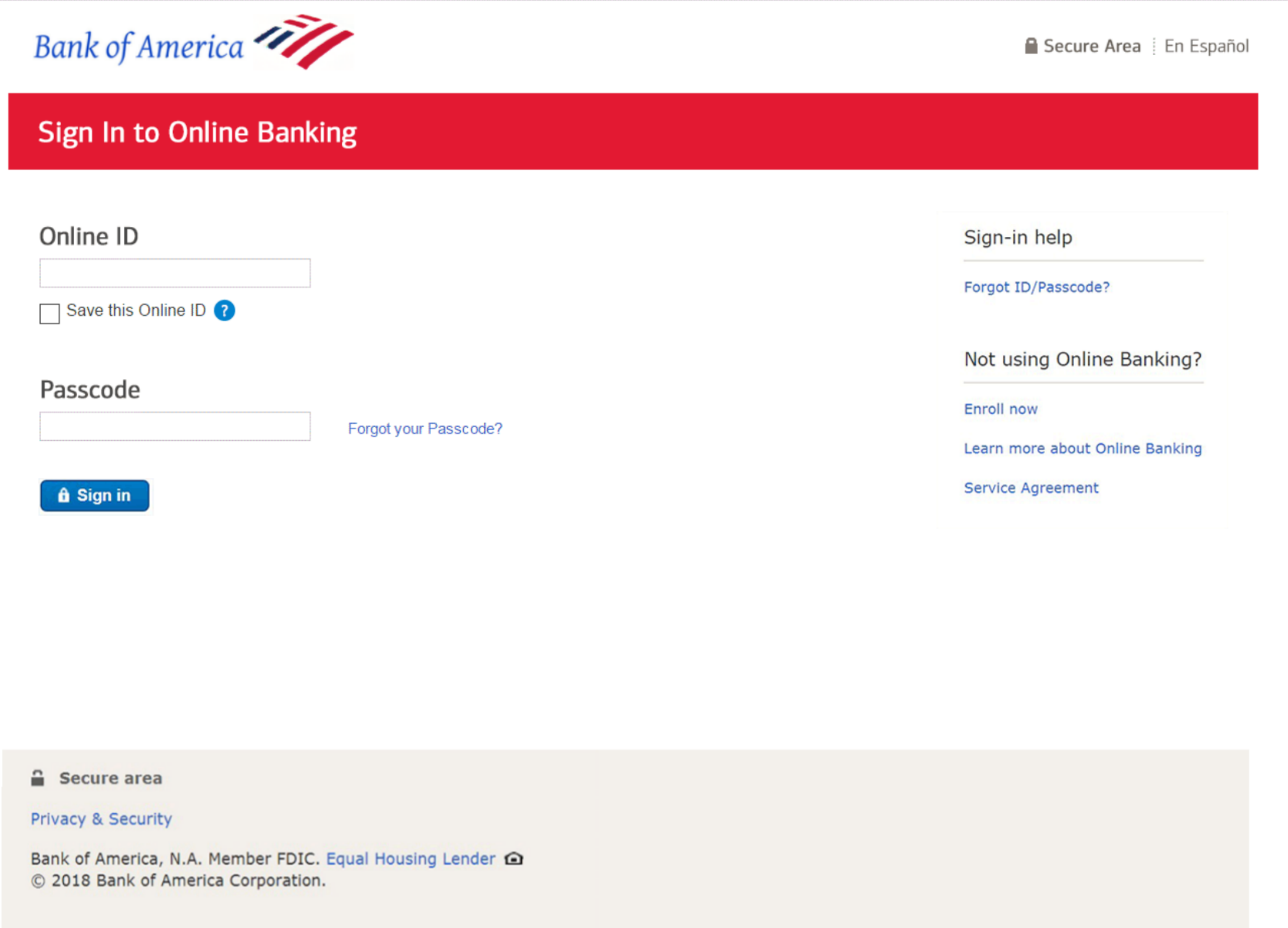


Figure 2: **Adversarial Phishing Webpage**—By using an adversarial logo crafted with `LogoMorph`, this phishing webpage bypasses detectors such as PhishIntention [32] and Phishpedia [30].

[30] Lin, Y., Liu, R., Divakaran, D. M., Ng, J. Y., Chan, Q. Z., Lu, Y., ... & Dong, J. S. (2021). Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages. In *30th USENIX Security Symposium (USENIX Security 21)* (pp. 3793-3810).

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

Another attack – results

- Of course, the attack "works". For most of the brands we considered, we were able to craft "adversarial logos" that, when put onto a webpage, would induce the entire system to believe the page to be benign.

Brand	**# of Success Logos (Rate)** Sim <0.87	0.6<Sim<0.87
Amazon	500 (1.00)	433 (0.87)
PayPal	311 (0.62)	308 (0.62)
LinkedIn	357 (0.71)	244 (0.49)
DHL	236 (0.47)	216 (0.43)
Dropbox	212 (0.42)	196 (0.39)
Chase	195 (0.39)	184 (0.37)
BOA	220 (0.44)	183 (0.37)
CIBC	188 (0.38)	152 (0.30)
AT&T	104 (0.21)	102 (0.20)
Outlook	105 (0.21)	99 (0.20)
Spotify	76 (0.15)	73 (0.15)

Table 4: **Logo-level Results (Image Logo)**—Number of generated logos images that bypass $\theta = 0.87$ threshold among 500 testing logos. We also report the number and % of logos with a similarity above 0.6 to indicate good image quality.

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

Another attack – results

- Of course, the attack "works". For most of the brands we considered, we were able to craft "adversarial logos" that, when put onto a webpage, would induce the entire system to believe the page to be benign.

Brand	# of Success Logos (Rate) Sim <0.87	0.6<Sim<0.87
Amazon	500 (1.00)	433 (0.87)
PayPal	311 (0.62)	308 (0.62)
LinkedIn	357 (0.71)	244 (0.49)
DHL	236 (0.47)	216 (0.43)
Dropbox	212 (0.42)	196 (0.39)
Chase	195 (0.39)	184 (0.37)
BOA	220 (0.44)	183 (0.37)
CIBC	188 (0.38)	152 (0.30)
AT&T	104 (0.21)	102 (0.20)
Outlook	105 (0.21)	99 (0.20)
Spotify	76 (0.15)	73 (0.15)

Table 4: **Logo-level Results (Image Logo)**—Number of generated logos images that bypass $\theta = 0.87$ threshold among 500 testing logos. We also report the number and % of logos with a similarity above 0.6 to indicate good image quality.

Brand	# Success Logos (# Tested)	Rate	Avg. Sim
Amazon	362 (362)	1.00	0.67
PayPal	308 (308)	1.00	0.67
DHL	194 (216)	0.90	0.71
Dropbox	174 (196)	0.89	0.70
BOA	154 (183)	0.84	0.73
Chase	146 (184)	0.80	0.80
CIBC	121 (152)	0.80	0.72
AT&T	81 (102)	0.79	0.76
LinkedIn	175 (244)	0.72	0.65
Spotify	50 (73)	0.68	0.83
Outlook	44 (99)	0.44	0.75

Table 5: **Webpage-Level Results (Image Logo)**— Number of logos that bypass the end-to-end detection of PhishIntention after being placed on actual webpages. We only test logos from Table 4.

Takeaway. Our method is always able to generate adversarial logo-images that bypass the logo-detector (76 in the worst case) and the end-to-end system (44 in the worst case).

UNIVERSITÄT LIECHTENSTEIN

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

Another attack – results (transferability)

- The attack also works when used against a phishing detection system that uses a different logic: PhishPedia [30]

Brand	**# Bypass Phishpedia (# Tested)**	**Rate**
DocuSign	178 (178)	1.00
Comcast	145 (145)	1.00
Yahoo	39 (39)	1.00
LinkedIn	6,172 (6,249)	0.99
Amazon	37,177 (37,970)	0.98
Google	116 (121)	0.96
Netflix	77 (80)	0.96
Instagram	192 (199)	0.96
eBay	170 (183)	0.93
Chase	17,361 (18,601)	0.93
Spotify	3,291 (3,596)	0.92
Outlook	10,361 (11,387)	0.91
AT&T	70 (81)	0.86
PayPal	5,497 (6,383)	0.86
CIBC	108 (121)	0.89
DHL	156 (194)	0.80
Dropbox	23,746 (29,773)	0.80
BOA	7,652 (13,479)	0.57

Table 7: **Transferability to Phishpedia (All Logos)**—Number of adversarial phishing webpages (bypassing PhishIntention [32]) that successfully bypass another phishing detector (Phishpedia [30]).

UNIVERSITÄT LIECHTENSTEIN

[30] Lin, Y., Liu, R., Divakaran, D. M., Ng, J. Y., Chan, Q. Z., Lu, Y., ... & Dong, J. S. (2021). Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages. In *30th USENIX Security Symposium (USENIX Security 21)* (pp. 3793-3810).

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

Another attack – results (transferability)

- The attack also works when used against a phishing detection system that uses a different logic: PhishPedia [30]

Brand	**# Bypass Phishpedia (# Tested)**	**Rate**
DocuSign	178 (178)	1.00
Comcast	145 (145)	1.00
Yahoo	39 (39)	1.00
LinkedIn	6,172 (6,249)	0.99
Amazon	37,177 (37,970)	0.98
Google	116 (121)	0.96
Netflix	77 (80)	0.96
Instagram	192 (199)	0.96
eBay	170 (183)	0.93
Chase	17,361 (18,601)	0.93
Spotify	3,291 (3,596)	0.92
Outlook	10,361 (11,387)	0.91
AT&T	70 (81)	0.86
PayPal	5,497 (6,383)	0.86
CIBC	108 (121)	0.89
DHL	156 (194)	0.80

Takeaway: these systems can be evaded

Table 7: **Transferability to Phishpedia (All Logos)**—Number of adversarial phishing webpages (bypassing PhishIntention [32]) that successfully bypass another phishing detector (Phishpedia [30]).

[30] Lin, Y., Liu, R., Divakaran, D. M., Ng, J. Y., Chan, Q. Z., Lu, Y., ... & Dong, J. S. (2021). Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages. In *30th USENIX Security Symposium (USENIX Security 21)* (pp. 3793-3810).

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

Another attack – results (transferability)

- The attack also works when used against a phishing detection system that uses a different logic: PhishPedia [30]

Brand	# Bypass Phishpedia (# Tested)	Rate
DocuSign	178 (178)	1.00
Comcast	145 (145)	1.00
Yahoo	39 (39)	1.00
LinkedIn	6,172 (6,249)	0.99
Amazon	37,177 (37,970)	0.98
Google	116 (121)	0.96
Netflix	77 (80)	0.96
Instagram	192 (199)	0.96
eBay	170 (183)	0.93
Chase	17,361 (18,601)	0.93
Spotify	3,291 (3,596)	0.92
Outlook	10,361 (11,387)	0.91
AT&T	70 (81)	0.86
PayPal	5,497 (6,383)	0.86
CIBC	108 (121)	0.89
DHL	156 (194)	0.80

Takeaway: these systems can be evaded

Table 7: **Transferability to Phishpedia (All Logos)**—Number of adversarial phishing webpages (bypassing PhishIntention [32]) that successfully bypass another phishing detector (Phishpedia [30]).

[30] Lin, Y., Liu, R., Divakaran, D. M., Ng, J. Y., Chan, Q. Z., Lu, Y., ... & Dong, J. S. (2021). Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages. In *30th USENIX Security Symposium (USENIX Security 21)* (pp. 3793-3810).

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

Another attack – results (extended evaluation)

- Overall, we generated adversarial logos pertaining to 110 different brands
  - Although in the main paper we deeply analyse only a subset of 17 popular brands

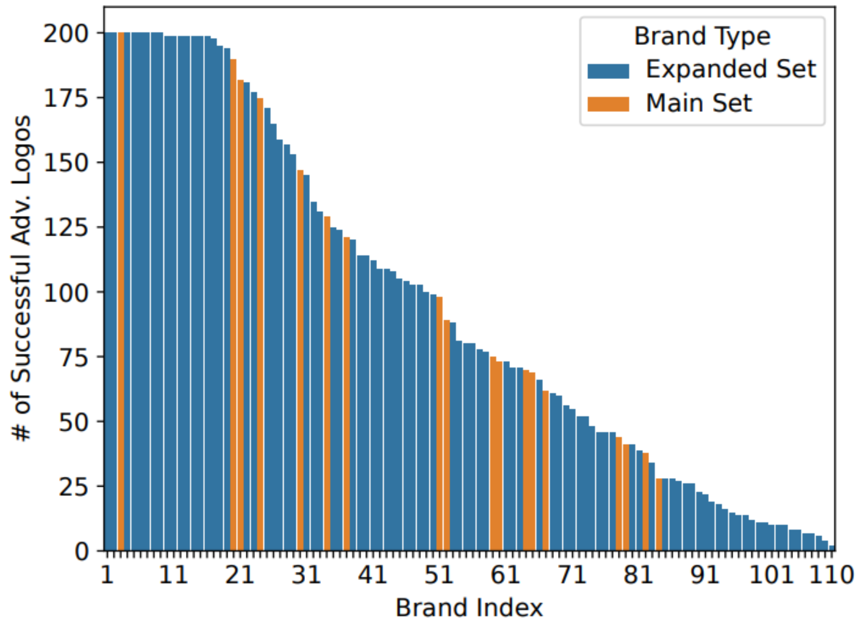


Figure 9: **Successful Adv. Logos Per Brand (110 Brands)** —We sorted the 110 brands on the x-axis based on the number of successful adversarial logos identified by `LogoMorph` (out of 200 candidate logos tested against PhishIntention).

UNIVERSITÄT LIECHTENSTEIN

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

Another attack – what about the previous attack? [ESORICS'23]

Takeaway. Of the 2,057 adversarial logos generated by PhishGAP [29], only 5.5% evade PhishIntention [32] end-to-end (despite bypassing its logo-discriminator).

UNIVERSITÄT LIECHTENSTEIN

[29] Lee, J., Xin, Z., See, M. N. P., Sabharwal, K., Apruzzese, G., & Divakaran, D. M. (2023, September). Attacking logo-based phishing website detectors with adversarial perturbations. In *European Symposium on Research in Computer Security* (pp. 162-182). Cham: Springer Nature Switzerland.

…what about humans?

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

(Phishing 101)

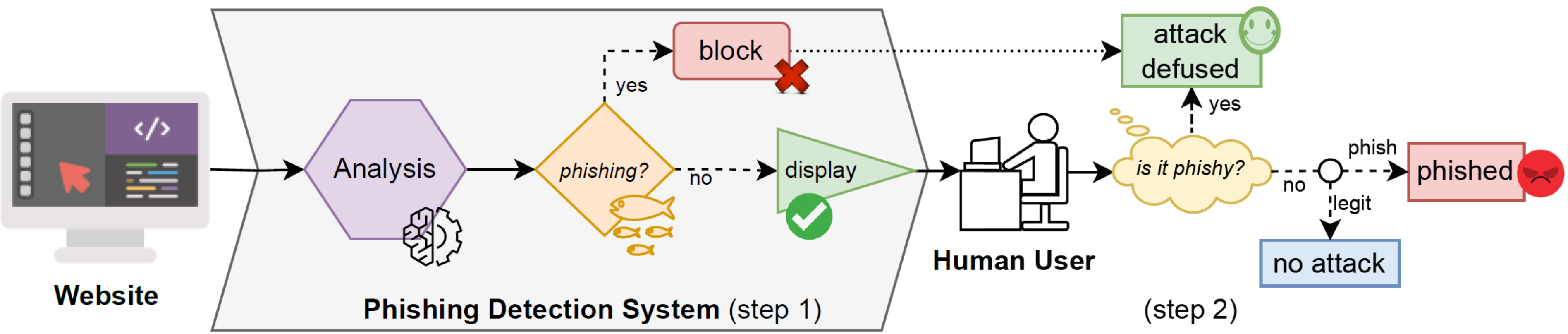


Fig. 1: Scenario: phishing detection is a two-step decision process.

UNIVERSITÄT LIECHTENSTEIN

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

Gap: Technical papers…

Typical workflow of an "adversarial machine learning" paper:

1. Propose an attack
2. Develop an ML model (trained on a benchmark dataset)

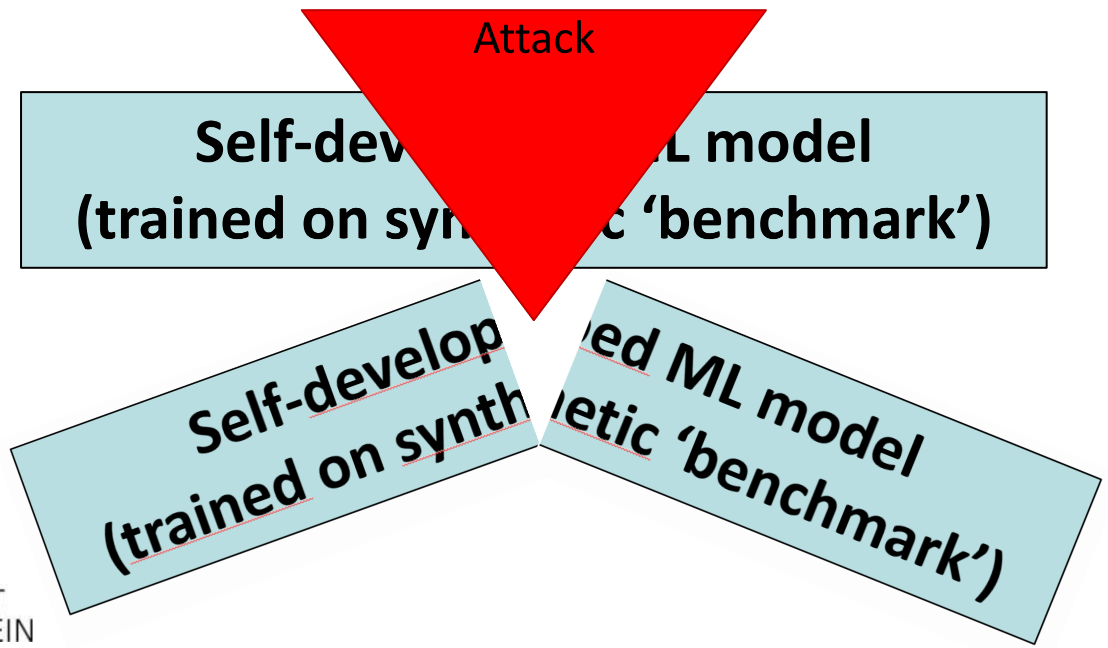
UNIVERSITÄT LIECHTENSTEIN

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

Gap: Technical papers…

Typical workflow of an “adversarial machine learning” paper:

1. Propose an attack
2. Develop an ML model (trained on a benchmark dataset)
3. Show that the attack “breaks” the ML model



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

Gap: Technical papers…

Typical workflow of an "adversarial machine learning" paper:

1. Propose an attack
2. Develop an ML model (trained on a benchmark dataset)
3. Show that the attack "breaks" the ML model

What about real ML systems?

- Evading *real* ML systems is not (always) simple [10]

UNIVERSITÄT LIECHTENSTEIN

[10] G. Apruzzese, H. S. Anderson, S. Dambra, D. Freeman, F. Pierazzi, and K. Roundy, ""Real attackers don't compute gradients": Bridging the gap between adversarial ML research and practice," in SaTML, 2023.

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

Gap: Technical papers…

Typical workflow of an "adversarial machine learning" paper:

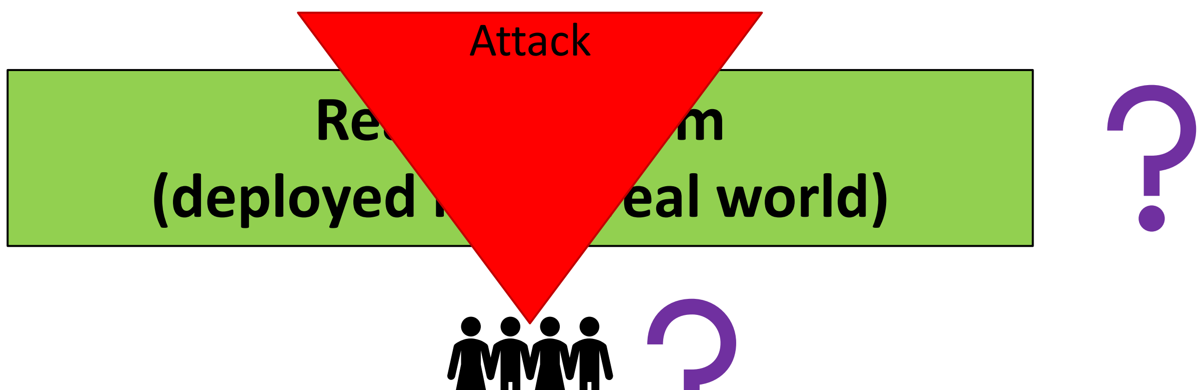
1. Propose an attack
2. Develop an ML model (trained on a benchmark dataset)
3. Show that the attack "breaks" the ML model

What about real ML systems?

- Evading *real* ML systems is not (always) simple [10]

…and are humans tricked as well?

- In some settings (e.g., phishing), humans *see* the "adversarial example"



UNIVERSITÄT LIECHTENSTEIN

[10] G. Apruzzese, H. S. Anderson, S. Dambra, D. Freeman, F. Pierazzi, and K. Roundy, "'Real attackers don't compute gradients": Bridging the gap between adversarial ML research and practice," in SaTML, 2023.

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

Gap: …and user studies

Typical workflow of a user study on “phishing assessment”:

1. Craft/collect phishing samples
2. Create a questionnaire and ask users to identify phishing samples
3. Draw conclusions

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

Gap: …and user studies

Typical workflow of a user study on “phishing assessment”:

1. Craft/collect phishing samples
2. Create a questionnaire and ask users to identify phishing samples
3. Draw conclusions

What about real (ML-based) phishing detectors?

- Maybe the samples would be trivially blocked by the detector

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

Gap: …and user studies

Typical workflow of a user study on "phishing assessment":

1. Craft/collect phishing samples
2. Create a questionnaire and ask users to identify phishing samples
3. Draw conclusions

What about real (ML-based) phishing detectors?

- Maybe the samples would be trivially blocked by the detector

…and what about priming?

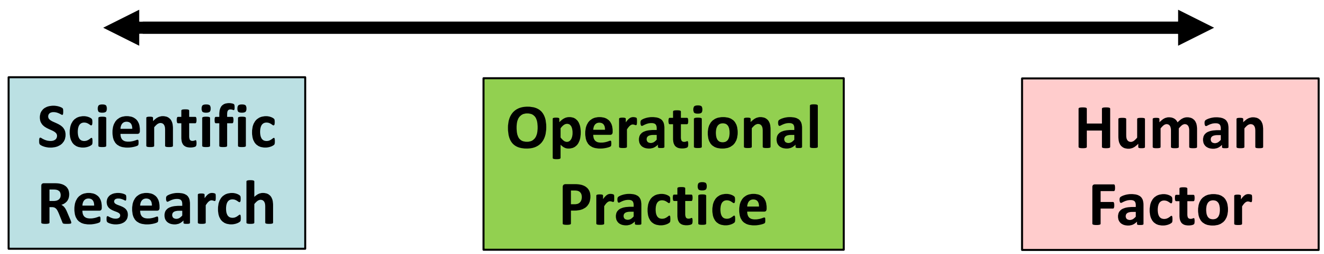
- Users are more suspicious when they are aware of being "tested" for phishing

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

What should be done

To provide more compelling studies, we should try to align

- **Research** in ML security, with
- **Operational** ML security and with
- The **human factor** in ML security



UNIVERSITÄT LIECHTENSTEIN

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

What should be done

To provide more compelling studies, we should try to align

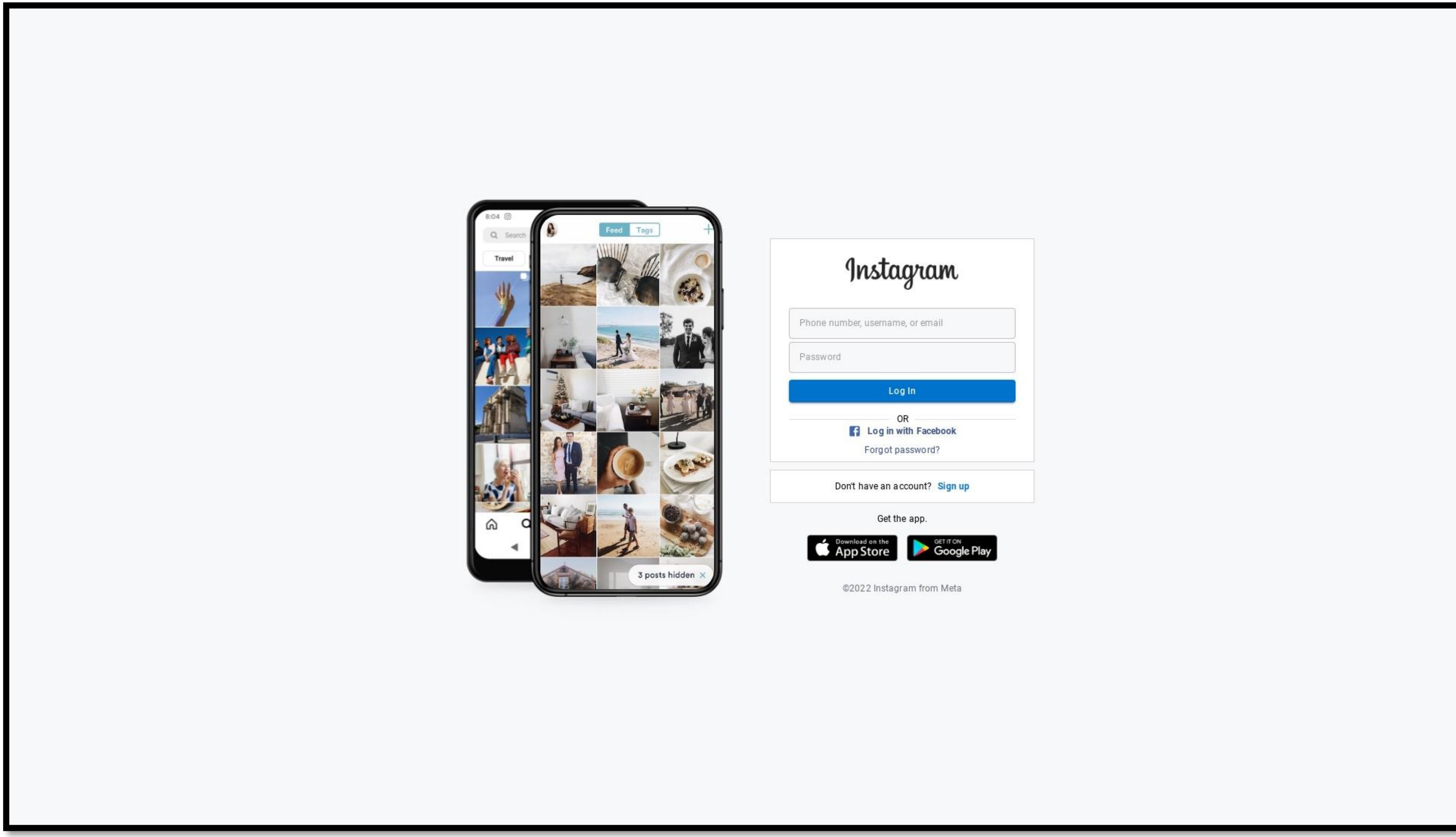
- **Research** in ML security, with
- **Operational** ML security and with
- The **human factor** in ML security

Scientific Research ⟷ **Operational Practice** ⟷ **Human Factor**

In what follows, I will show how we did the above:

- When considering the system used in [ACSAC'22]
- When considering the detector of [ESORICS'23]
- When considering the system of [USENIX'24]
- When considering the system of [SaTML'23]

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



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

What do we do? [SaTML'23]

RQ: 'Are human users deceived by phishing webpages that evade a real-world phishing website detection system?'

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

How did we do it? [SaTML'23]

1. We reach out to a well-known security company ("Sigma")
2. We ask Sigma to provide us with phishing webpages that evaded their operational Phishing Detection System (reliant on deep learning)

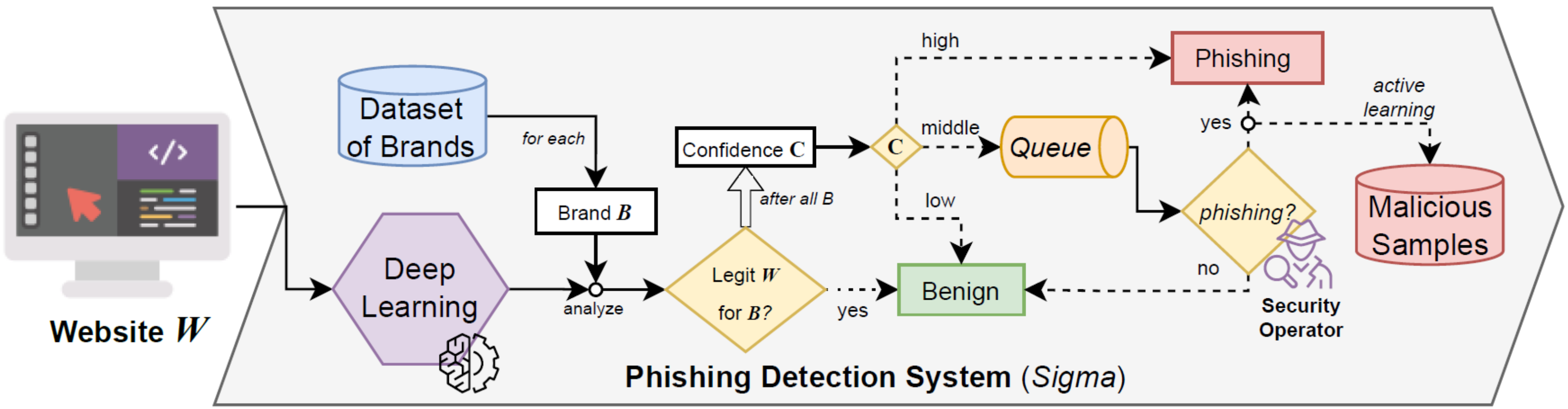


Fig. 2: The architecture of the PDS deployed by *Sigma*, used as basis for the phishing examples to include in our user-study.

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

How did we do it? [SaTML'23]

3. We select a set of 18 "adversarial" phishing webpages (mimicking brands popular in the EU)
4. We add 2 "legitimate" webpages (as a form of control)
5. We use the screenshots of these 20 webpages to carry out a user study

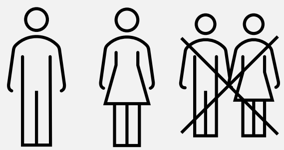
TABLE III: Sequence of screenshots in our questionnaire, and their difficulty level. The number points to the image (hosted in our repo).

#	Brand	Difficulty	Comment
1	Instagram	*Hard*	Resembles the legitimate login page, with the sole distinction being the footer's style.
2	Facebook	*Moderate*	Appears similar to the authentic version; however, suspicion may arise due to the multiple profiles that have recently logged in from the same device (specifically, six different profiles).
3	Facebook	*Hard*	Closely resembles the original, with the sole exception of a missing footer.
4	Instagram	*Hard*	Extremely challenging to distinguish, as it perfectly mirrors the original.
5	PayPal	*Hard*	Resembles the authentic site very closely.
6	Google	*Hard*	Resembles the authentic site very closely.
7	Amazon	*Hard*	Resembles the authentic site very closely.
8	Airbnb	—	It is the legitimate website.
9	Zalando	—	It is the legitimate website.
10	Netflix	*Moderate*	The website's header and logo may induce suspicion due to their uncharacteristic design.
11	Yahoo	*Hard*	Resembles the authentic site very closely.
12	Yahoo	*Hard*	Resembles the authentic site very closely.
13	Netflix	*Easy*	The font style noticeably deviates from the one typically used.
14	Uber	*Easy*	The appearance of Uber's sign-in page notably diverges from the expected layout.
15	PayPal	*Moderate*	The background color of the input fields clashes with the overall design aesthetic of the website.
16	Uber	*Easy*	The appearance suggests it might be an outdated version of Uber.
17	LinkedIn	*Easy*	The font style significantly deviates from what one would expect on a professional website, disrupting its overall look and feel.
18	Netflix	*Very easy*	No resemblance to the original sign-up page, with a starkly contrasting and distinctive styling.
19	Twitter	*Moderate*	It gives the impression of being an older version of Twitter, which could still potentially elicit trust from unfamiliar users.
20	Amazon	*Moderate*	While it bears a striking resemblance, participants might grow suspicious due to the button on the page appearing incongruous with the overall design.

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

How did we do it? [SaTML'23]

6. We advertise the questionnaire on popular social media for 3 weeks
7. **We do not prime the users (!)**
8. We received 126 responses



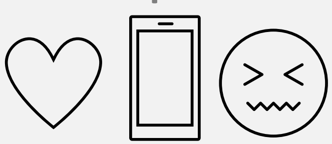
70 ***55*** ***1***

Country

78 ***19*** ***12***

Slovenia, Italy, Bosnia, Liechtenstein, Finland …

IT expertise



75 ***48*** ***3***

Education

Basic → 11

High School → 45

Bachelor's → 41

Master's → 27

PhD → 2

Age

<16 → 3

16-24 years → 44

25-34 years → 57

35-44 years → 12

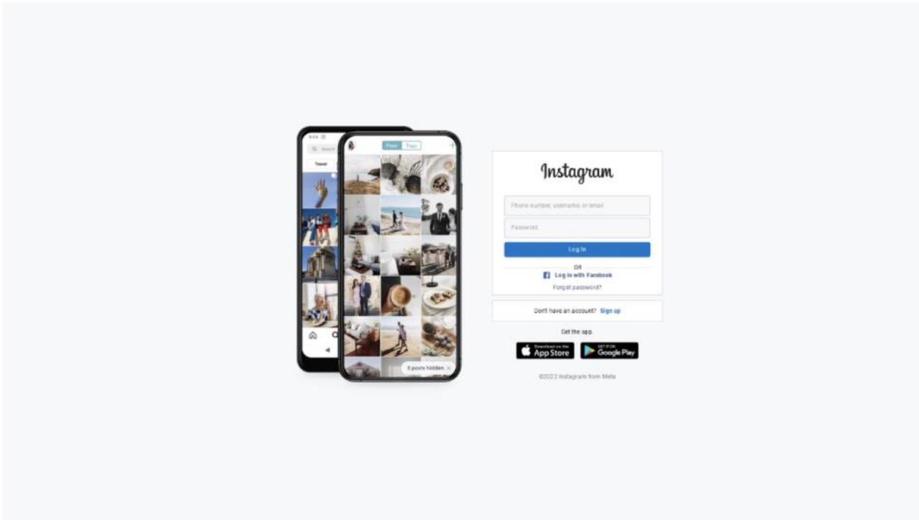
45-54 years → 4

55-64 years → 6

1. Screenshot - Please rate how much you agree with the following statement: *

"On the screenshot you see the login page of a social media platform where users can share photos, videos and messages with their followers."

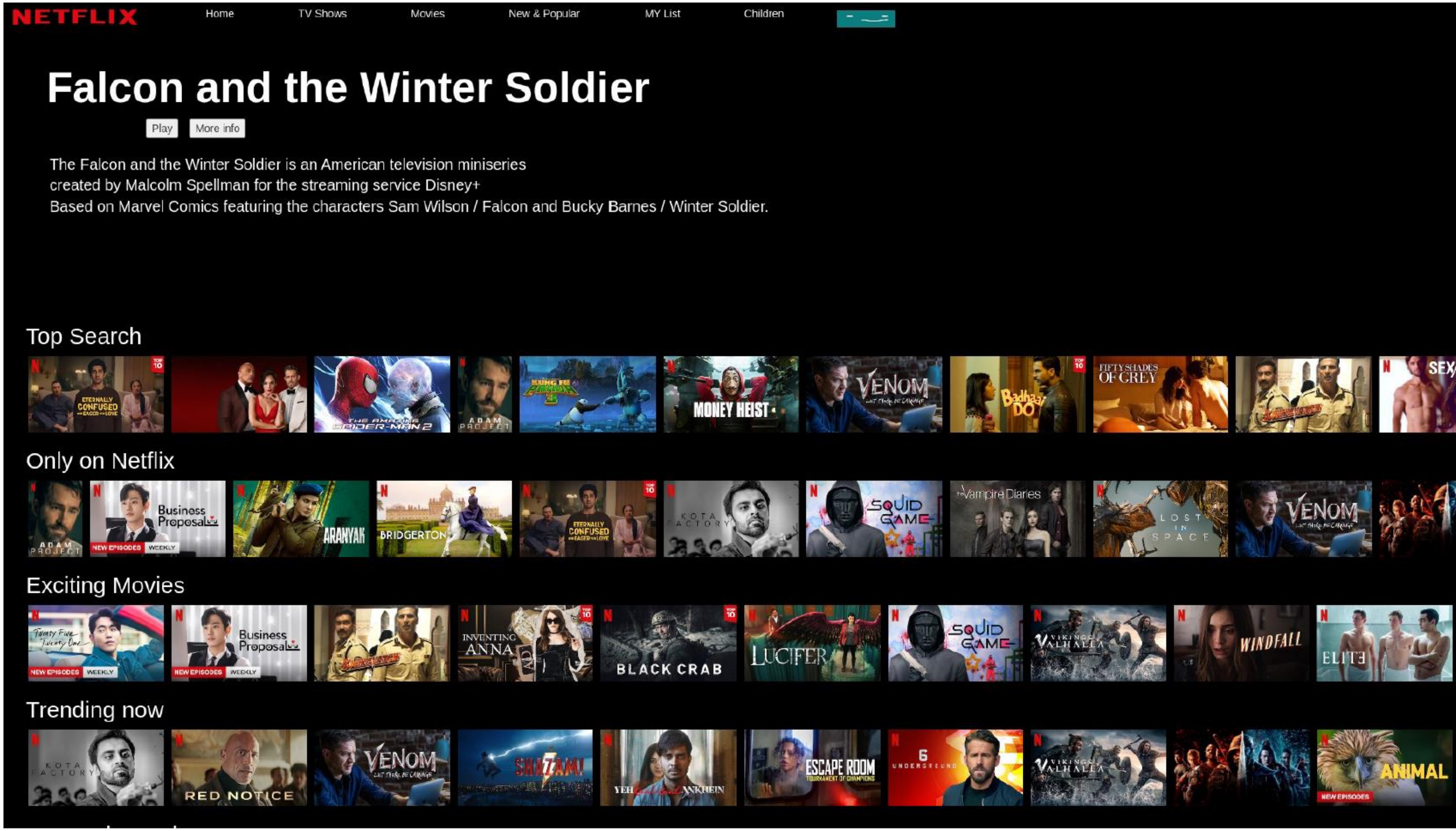
(larger image: here)



	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

Fig. 3: Exemplary question (i.e., the first) in part II of our questionnaire. The screenshot refers to an adversarial webpage.

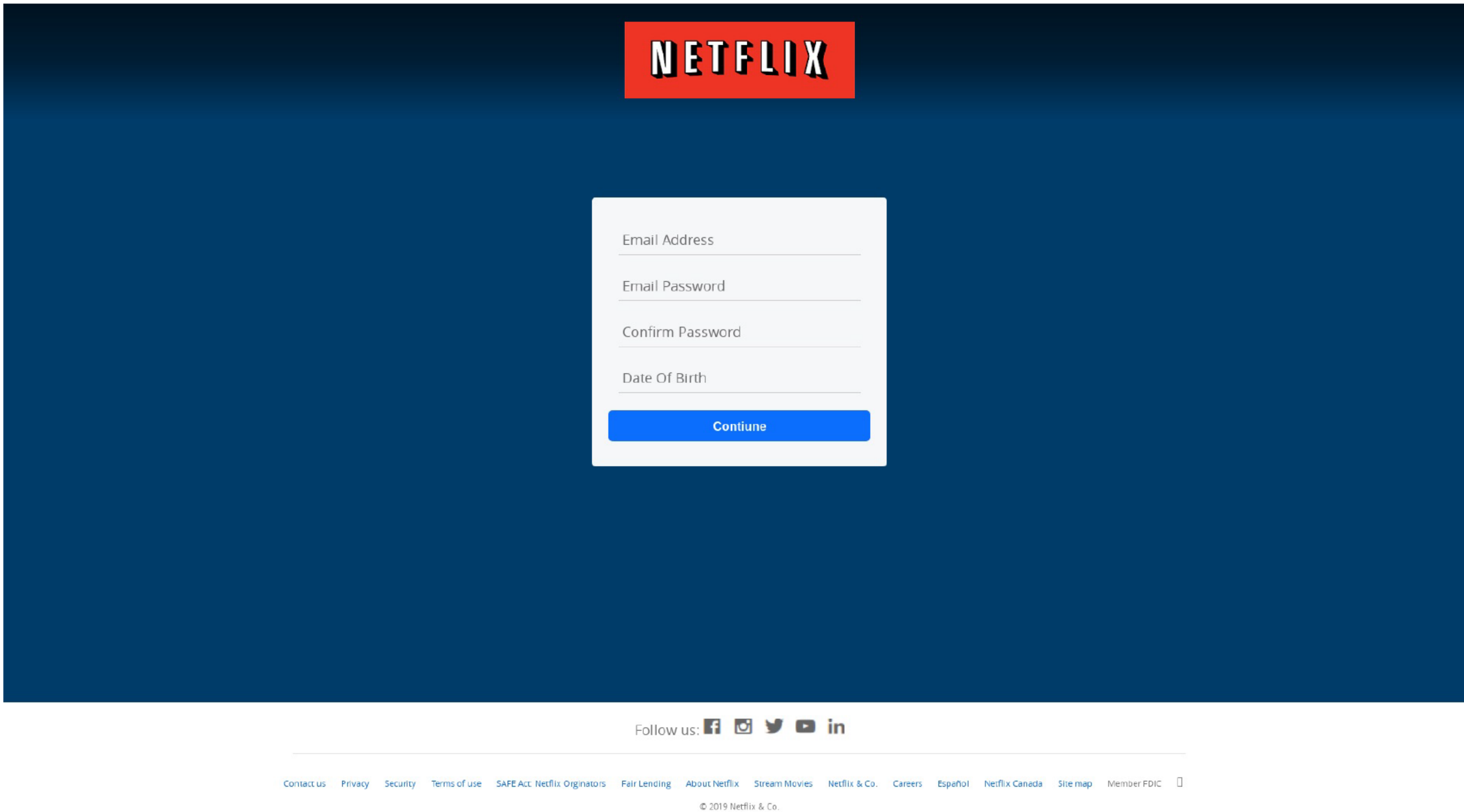
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li



(a) Screenshot 10 ("moderate difficulty" to identify as phishing—by humans).

UNIVERSITAT
LIECHTENSTEIN

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

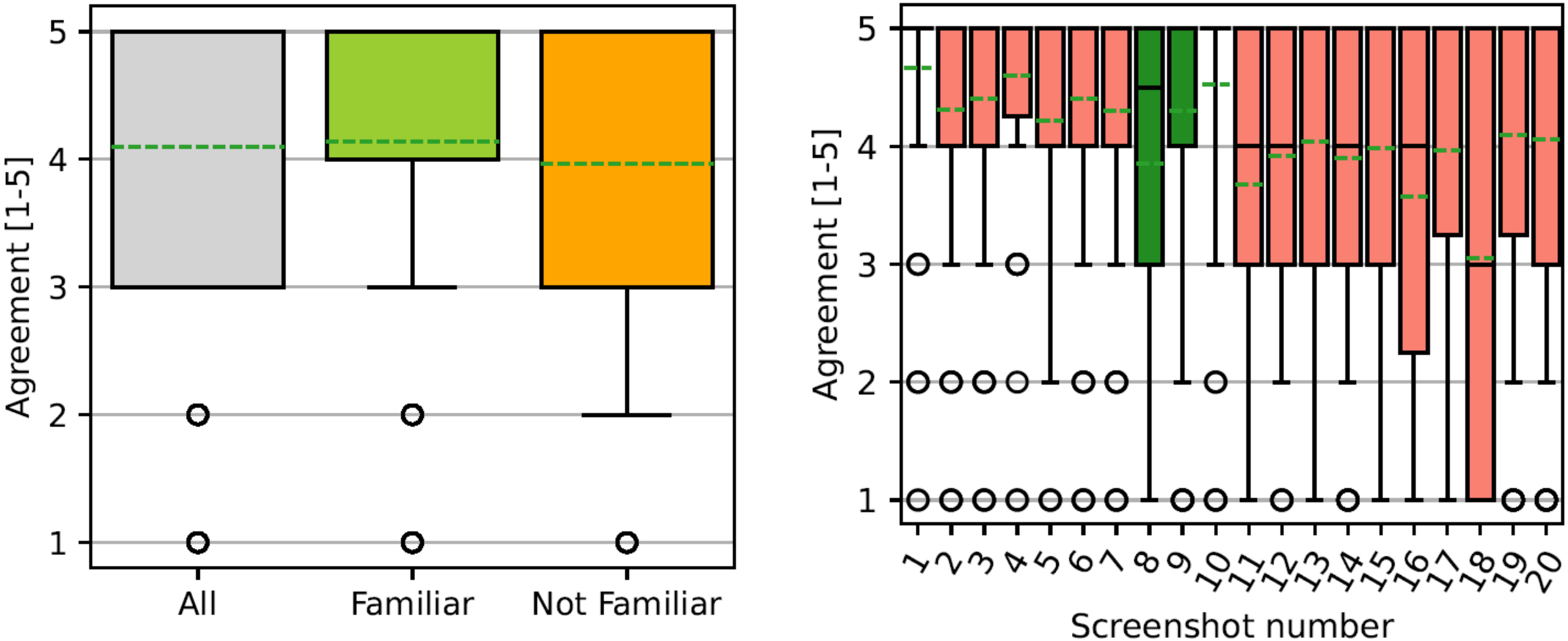


(b) Screenshot 18 (“very easy difficulty” to identify as phishing—by humans).

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

What did we find? (1) [SaTML'23]

Higher agreement = higher likelihood of being deceived



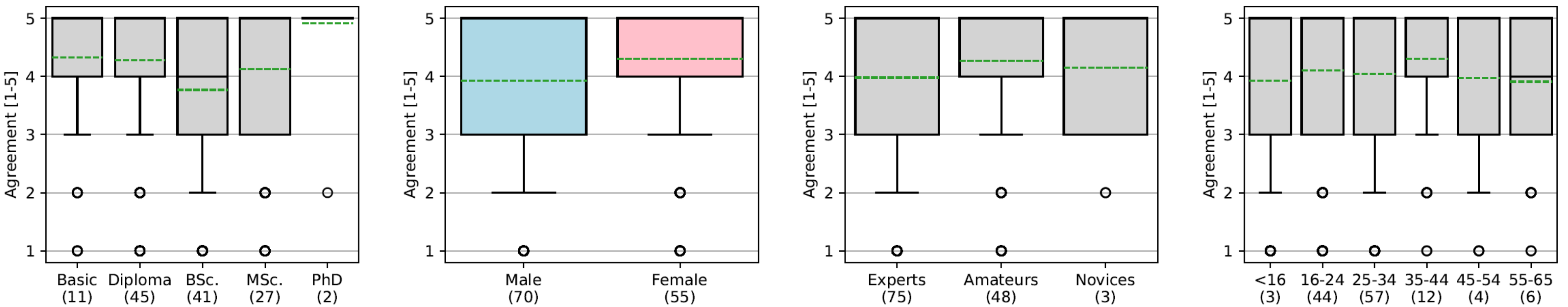
TAKEAWAY. Most of our sample cannot recognize AW, and familiarity with a brand hinders the detection skills of users.

These claims are statistically significant ($p<0.05$)

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

What did we find? (2) [SaTML'23]

Higher agreement = higher likelihood of being deceived



(a) Education. (b) Gender. (c) Expertise with IT. (d) Age.

Fig. 5: Subgroup results. The figures report the aggregated ratings (for the 18 AW) of each subgroup (the x-axis shows the size of each subgroup).

- University graduates are more suspicious
- Female appear to be less suspicious than males
- IT experts are more skeptical than amateurs
- Age is not correlated with suspiciousness

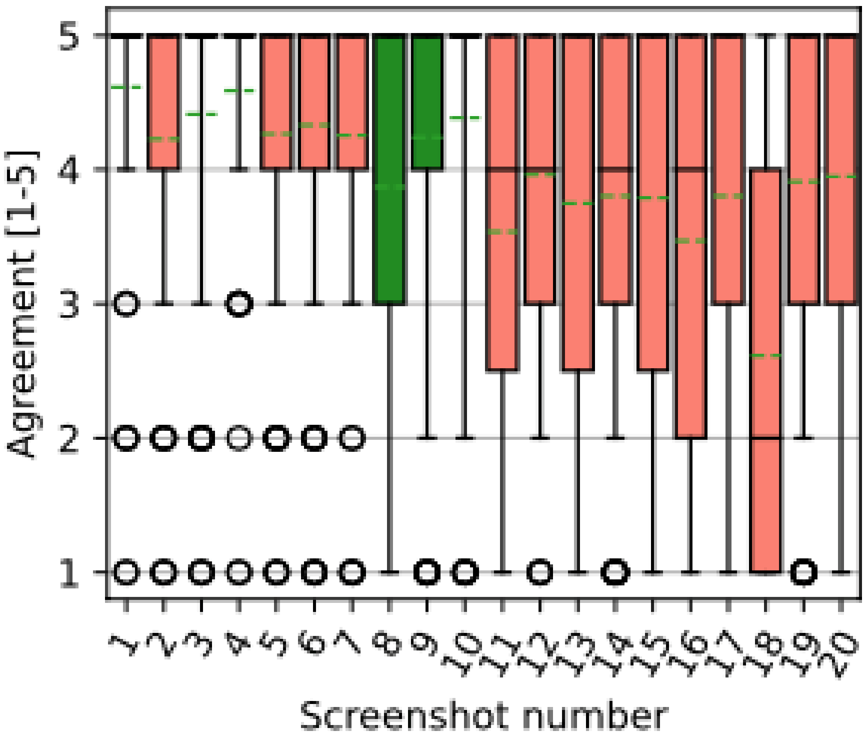
UNIVERSITÄT LIECHTENSTEIN

These claims are statistically significant ($p<0.05$)

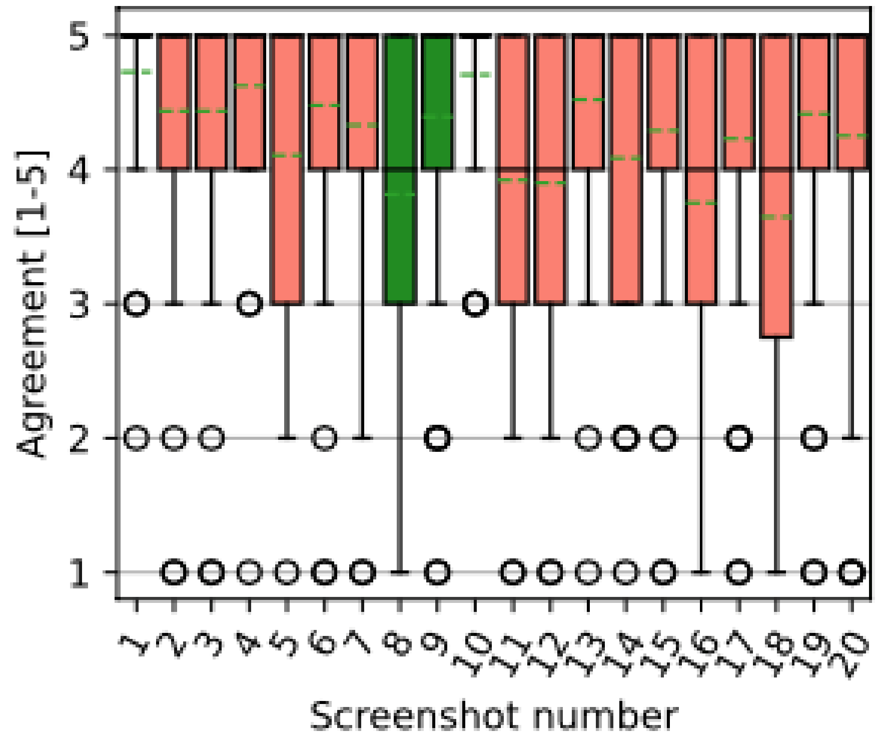
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What did we find? (3) [SaTML'23]

IT expertise influences the skepticism of participants



IT experts



IT amateurs

UNIVERSITÄT LIECHTENSTEIN

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

What do we do? [ACSAC'22]

RQ: 'Is it convenient for an attacker to create an «adversarial webpage»?'
(what if such a webpage, despite fooling the detector, can be easily recognized by humans?)

UNIVERSITÄT LIECHTENSTEIN

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

How did we do it? [ACSAC'22]

1. We take the detector we developed for [ACSAC'22]
2. We deliberately introduce "perturbations" in the webpages
3. We check if these webpages evade the detector
4. We ask users if they see anything suspicious (we prime users!)
   a. In the "non perturbed" webpages (baseline study)
   b. In the "perturbed" webpages (adversarial study)

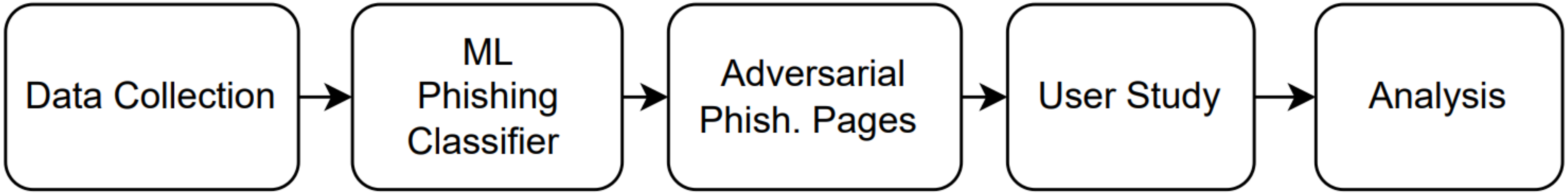


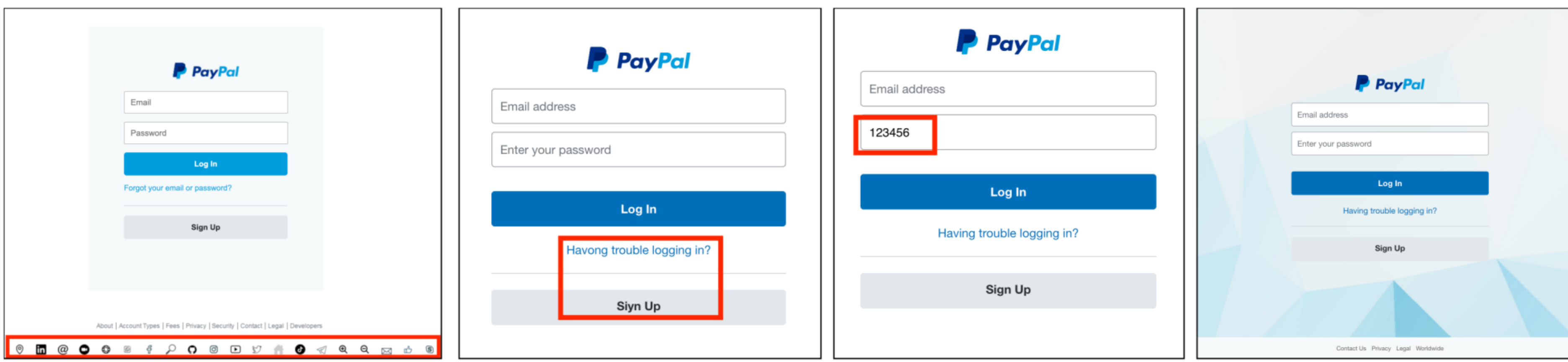
Fig. 1: Workflow of our study.

UNIVERSITÄT LIECHTENSTEIN

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

How did we do it? [ACSAC'22]

1. We take the detector we developed for [ACSAC'22]
2. We deliberately introduce "perturbations" in the webpages
3. We check if these webpages evade the detector
4. We ask users if they see anything suspicious (we prime users!)
   a. In the "non perturbed" webpages (baseline study)
   b. In the "perturbed" webpages (adversarial study)



(a) APW-Lab_img (b) APW-Lab_typo (c) APW-Lab_pswd (d) APW-Lab_bg

Fig. 4: Example screenshot of lab-generated adversarial phishing pages targeting Paypal. We include two types of perturbations: (a) adding small images to the footer, (b) introducing typos, (c) making the password visible, and (d) adding a background image.

UNIVERSITÄT LIECHTENSTEIN

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

How did we do it? [ACSAC'22]

1. We take the detector we developed for [ACSAC'22]
2. We deliberately introduce "perturbations" in the webpages
3. We check if these webpages evade the detector
4. We ask users if they see anything suspicious (we prime users!)
   a. In the "non perturbed" webpages (baseline study)
   b. In the "perturbed" webpages (adversarial study)

Study	Pages Seen by Each Participant	Participants
Baseline	7 Legitimate + 8 Unperturbed Phishing	235
Adversarial	7 Legitimate + 4 *APW-Lab* + 4 *APW-Wild*	235

Table 1: Summary of our user studies. We report the classes of webpages that *each participant views* and the number of participants.

UNIVERSITÄT LIECHTENSTEIN

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

What did we find? (1) [ACSAC'22]

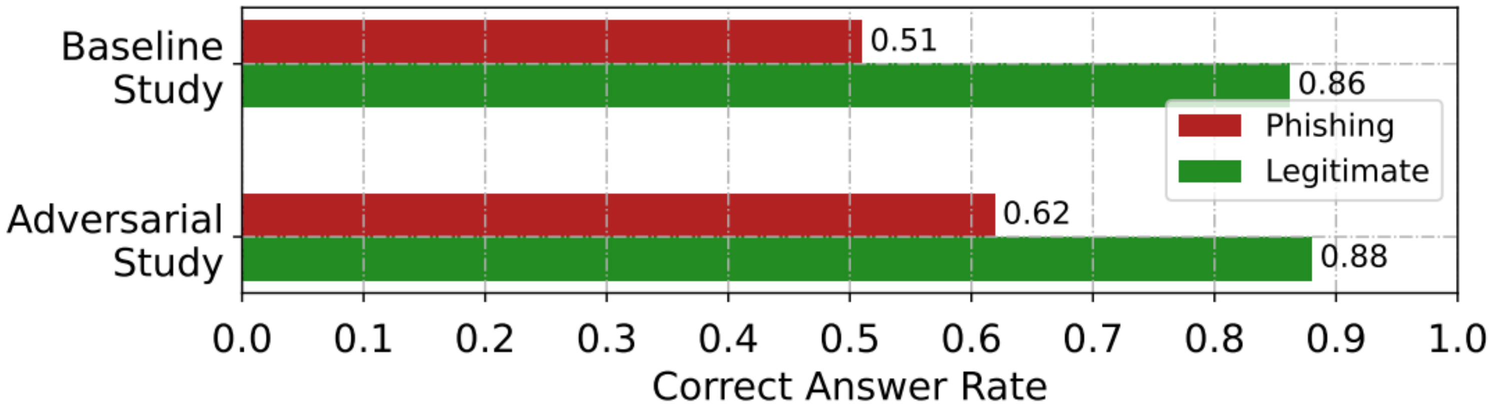


Fig. 2: Overview of baseline and adversarial study (7, 050 responses)

Our sample is deceived by phishing webpages
(even adversarial ones, to a lesser degree)

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

What did we find? (2) [ACSAC'22]

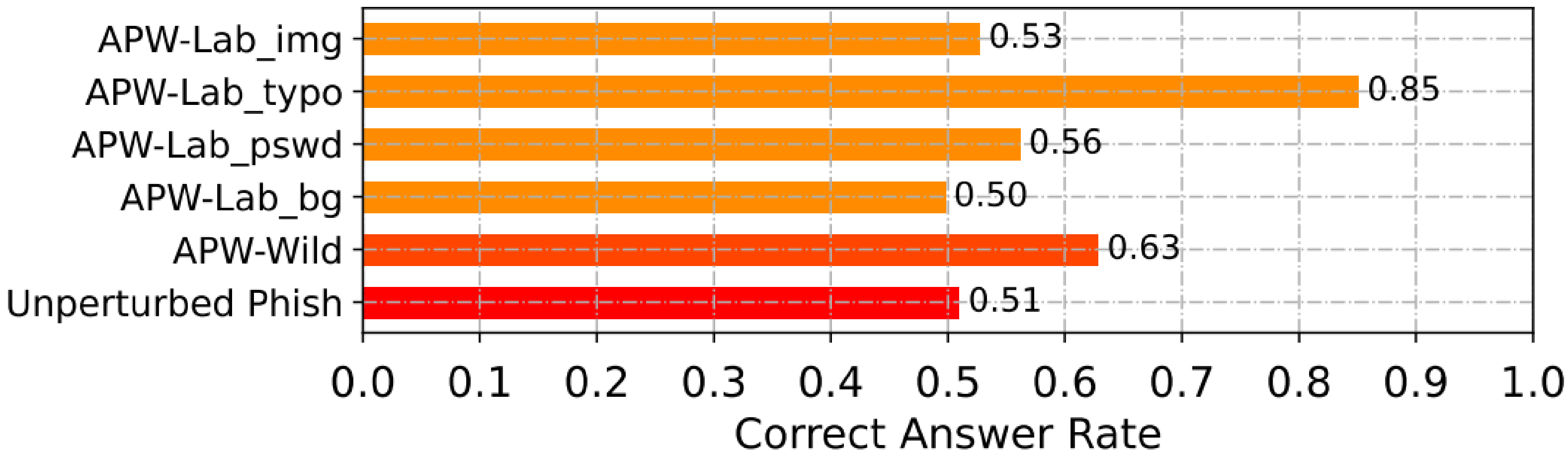


Fig. 3: Detection rate for different types of phishing webpages.

Some perturbations are easier to spot than others
(Typos make users suspicious, but changing the entire background does not!)

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

What did we find? (3) [ACSAC'22]

We also asked users to explain why they deemed any webpage to be benign or phishing.

APW-Lab. We recall (cf. Fig. 3) that our participants performed very well on *APW-Lab_typo*, for which we coded 93 responses. Among these, a large majority (69, 74%) mentioned "typo" (after making a correct detection). Intriguingly, 15% (14) provided reasons that have nothing to do with *APW-Lab_typo* (despite still rating them as phishing). E.g., P668 stated: "*figures do not look normal*". The remaining 11% incorrectly labeled the webpage as legitimate (e.g., "*Everrything looks normal*" [P621]).

Even though participants can recognize an adversarial phishing webpage as "phishing", they rarely pinpoint the perturbation that makes the webpage "adversarial" (as long as it is not text-based)

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

What did we find? (3) [ACSAC'22]

We also asked users to explain why they deemed any webpage to be benign or phishing.

Concerning *APW-Lab_img*, we have coded 61 responses. Notably, only 13% (8) pointed out the 'correct' adversarial perturbation (i.e., images on footer). E.g., P544 stated: "*low quality and strange icons at the bottom, which a legit site would not have*". In contrast, 48% (29) mentioned other reasons. E.g., P210 stated: "*Adobe doesn't require logging in to view something in it to my knowledge*". The remaining 39% incorrectly labeled the webpage as legitimate (e.g., "*norton certificate makes me think it's more legit than not.*" [242]).

Even though participants can recognize an adversarial phishing webpage as "phishing", they rarely pinpoint the perturbation that makes the webpage "adversarial" (as long as it is not text-based)

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

What did we find? (3) [ACSAC'22]

We also asked users to explain why they deemed any webpage to be benign or phishing.

For *APW-Lab_pswd*, we coded 137 responses. The majority (70, 51%), despite stemming from a correct detection, have nothing to do with our perturbation: only 8% (11) pointed out the visible password as a potential phishing indicator (e.g., "*password field is plain text*" [P1306]; or "*the password is not hidden*" [P937]). The rest 41% incorrectly labeled the webpage as legitimate (e.g., "*As a Wells Fargo customer who was literally just checking their account before starting this study I can assure you this is legitimately legit*" [P86]).

Even though participants can recognize an adversarial phishing webpage as "phishing", they rarely pinpoint the perturbation that makes the webpage "adversarial" (as long as it is not text-based)

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

What did we find? (3) [ACSAC'22]

We also asked users to explain why they deemed any webpage to be benign or phishing.

We coded 89 responses for *APW-Lab_bg*. Surprisingly, only 4% (3) of responses mention our inserted perturbation. In contrast, 48% (43) justify their (correct) phishing detection by mentioning unrelated factors. E.g., P971 stated: "*too many big competing brands at the top*". The rest 49% incorrectly labeled the page as legitimate (e.g., P321 stated: "*good grammar, good syntax, appropriate colors, logo*").

Even though participants can recognize an adversarial phishing webpage as "phishing", they rarely pinpoint the perturbation that makes the webpage "adversarial" (as long as it is not text-based)

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

What do we do? [ESORICS'23]

RQ: 'Are users suspicious of the logos generated via the generative adversarial perturbation?'

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

How did we do it? [ESORICS'23]

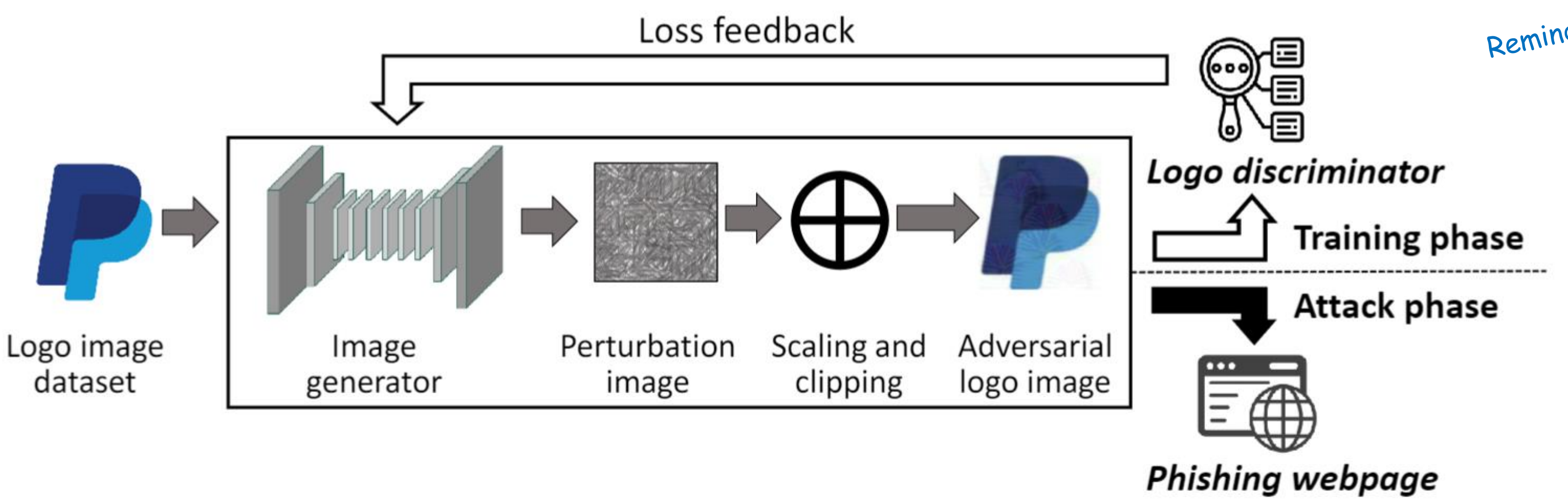
1. We take the adversarial logos generated for the ESORICS'23 paper
2. We use them to carry out two user study with the same goal: *given an "original" logo and an "adversarial" logo, can the human spot any difference?* (no priming)
   a. large set of different logos for a "vertical" study with 30 students
   b. smaller set of 21 logos for a "horizontal" study with 287 participants

UNIVERSITÄT LIECHTENSTEIN

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

How did we do it? [ESORICS'23]

1. We take the adversarial logos generated for the ESORICS'23 paper
2. We use them to carry out two user study with the same goal: *given an "original" logo and an "adversarial" logo, can the human spot any difference?* (no priming)
   a. large set of different logos for a "vertical" study with 30 students
   b. smaller set of 21 logos for a "horizontal" study with 287 participants



Reminder

Fig. 4: Generative adversarial perturbation workflow

UNIVERSITÄT LIECHTENSTEIN

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

How did we do it? [ESORICS'23]

Look at these two images for no more than 5 seconds, and then answer the similarity question.

Logo A

Logo B

On a scale from 1 to 5, how similar do you think these two logos are? *

	1	2	3	4	5	
Very Different	○	○	○	○	○	Very Similar

Look at these two images for no more than 5 seconds, and then answer the similarity question.

Logo A

Logo B

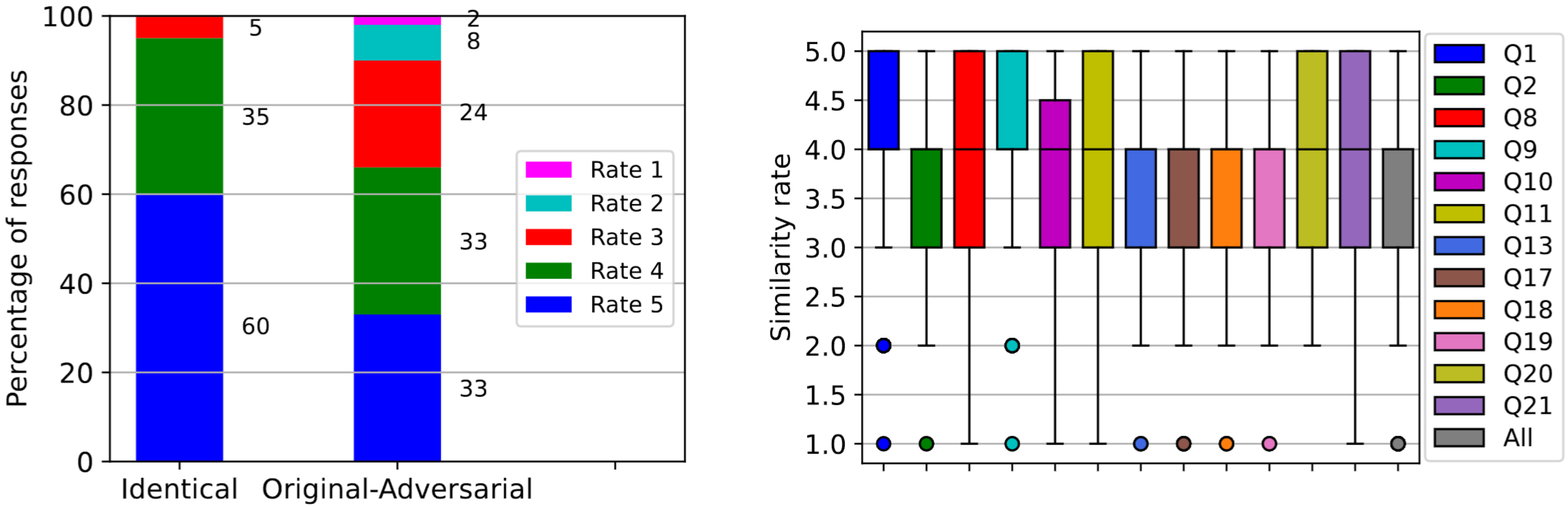
On a scale from 1 to 5, how similar do you think these two logos are? *

	1	2	3	4	5	
Very Different	○	○	○	○	○	Very Similar

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

What did we find? [ESORICS'23]

- For every question, users had to say how "similar" the two logos were (5= very similar, 1= not similar at all)



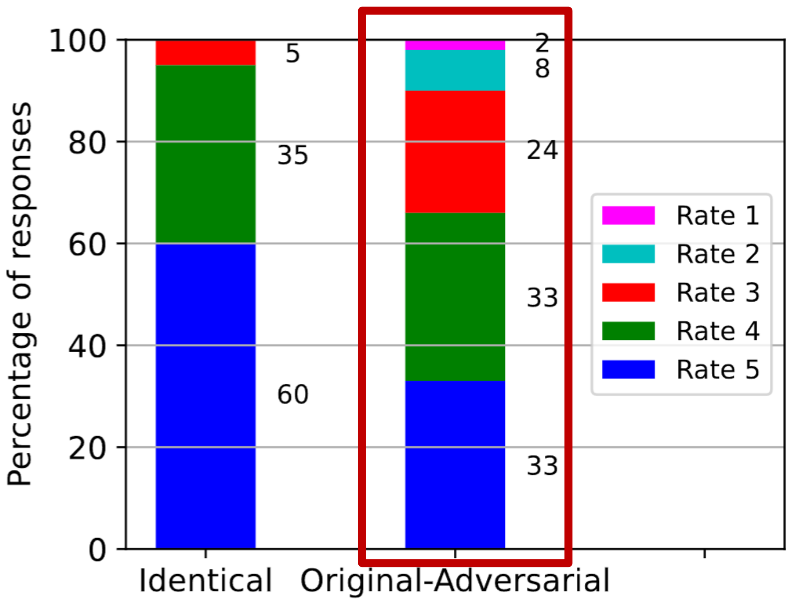
(a) Vertical Study **(b)** Horizontal Study

UNIVERSITÄT LIECHTENSTEIN

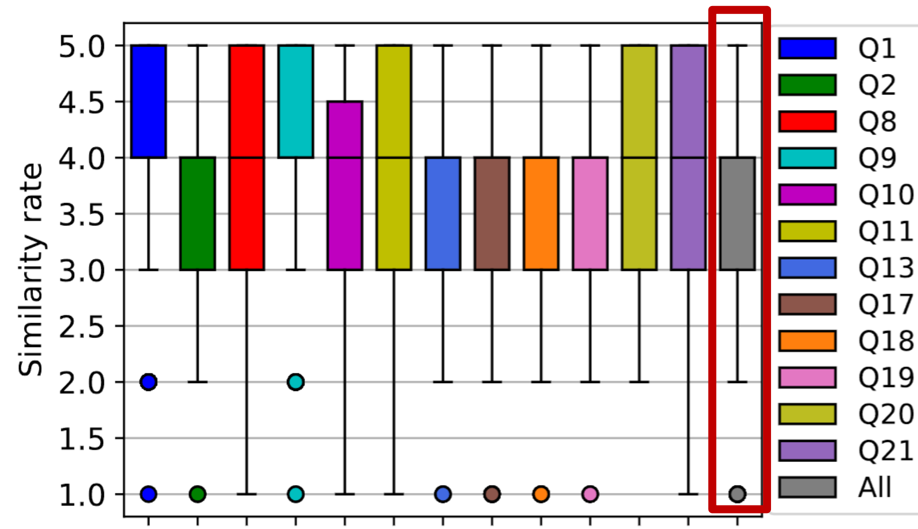
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What did we find? [ESORICS'23]

- For every question, users had to say how "similar" the two logos were (5= very similar, 1= not similar at all)



(a) Vertical Study



(b) Horizontal Study

Takeaways:

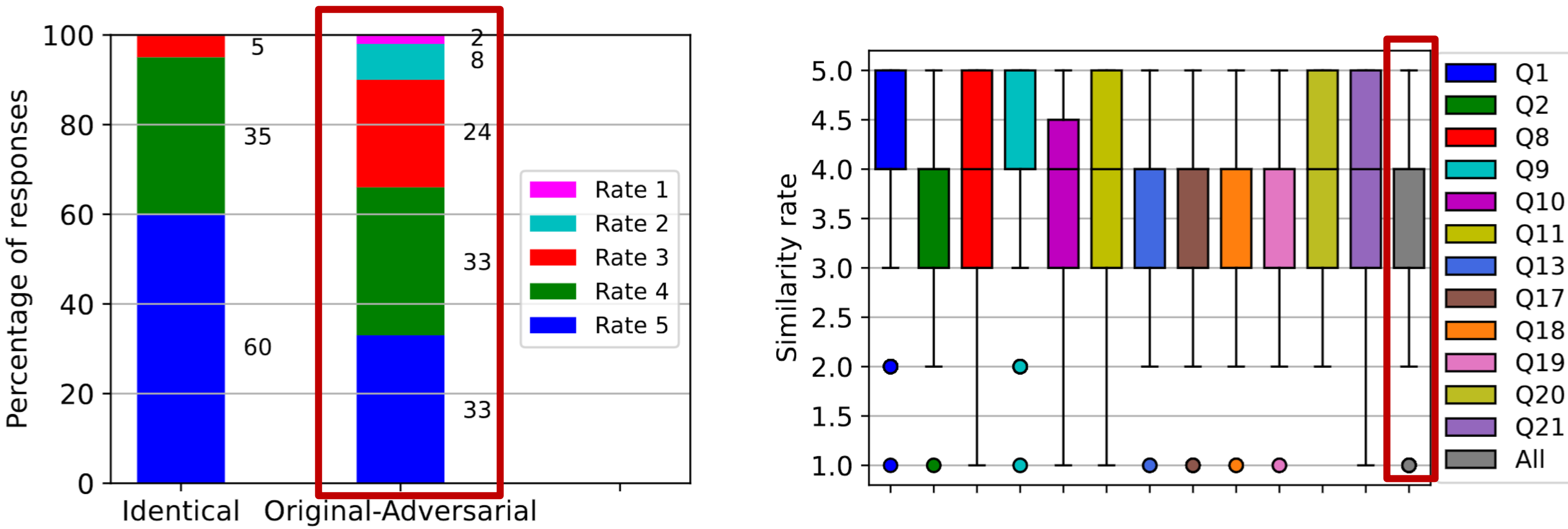
1. Vertical Study: over 85% of participants rated >=3 similarity
2. Horizontal Study: the average similarity per question was >=3

UNIVERSITÄT LIECHTENSTEIN

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

What did we find? [ESORICS'23]

- For every question, users had to say how "similar" the two logos were (5= very similar, 1= not similar at all)



(a) Vertical Study **(b)** Horizontal Study

Takeaways:

1. Vertical Study: over 85% of participants rated >=3 similarity
2. Horizontal Study: the average similarity per question was >=3

UNIVERSITÄT LIECHTENSTEIN

Humans are (likely to be) deceived

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

What do we do? [USENIX Sec’24]

RQ: ‘Does LogoMorph deceive humans, *too*?’

UNIVERSITÄT
LIECHTENSTEIN

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

How did we do it? [USENIX Sec'24]

1. We take the adversarial webpages (not just logos!) generated in the USENIX Sec'24 paper *which bypassed PhishIntention* (the target system)
2. We use them to carry out a user study (N=150): *can users identify a phishing webpage* (half of the webpages are benign)*?* <u>(priming)</u>
   a. First, we do this with "non-adversarial" logos
   b. Then, we do this with "adversarial" logos generated via LogoMorph

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

How did we do it? [USENIX Sec'24]

1. We take the adversarial webpages (not just logos!) generated in the USENIX Sec'24 paper *which bypassed PhishIntention* (the target system)
2. We use them to carry out a user study (N=150): *can users identify a phishing webpage* (half of the webpages are benign)*?* (priming)
   a. First, we do this with "non-adversarial" logos
   b. Then, we do this with "adversarial" logos generated via LogoMorph

How did LogoMorph work?

Figure 1: **Adversarial Logo Examples**—We show the original logo and two attack examples generated by our `LogoMorph`.

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

How did we do it? [USENIX Sec’24]

1. We take the adversarial webpages (not just logos!) generated in the USENIX Sec’24 paper *which bypassed PhishIntention* (the target system)
2. We use them to carry out a user study (N=150): *can users identify a phishing webpage* (half of the webpages are benign)*?* (priming)
   a. First, we do this with “non-adversarial” logos
   b. Then, we do this with “adversarial” logos generated via LogoMorph

Generated with the diffusion model

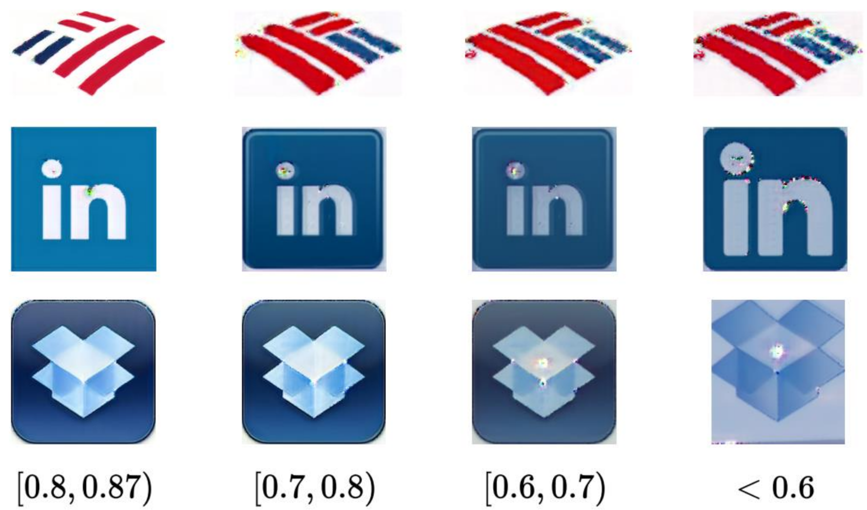


Figure 5: **Image Logo Attack Examples**—We show example logo images of different similarity levels compared with the original logos. All of them are below the detection threshold of 0.87.

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

How did we do it? [USENIX Sec’24]

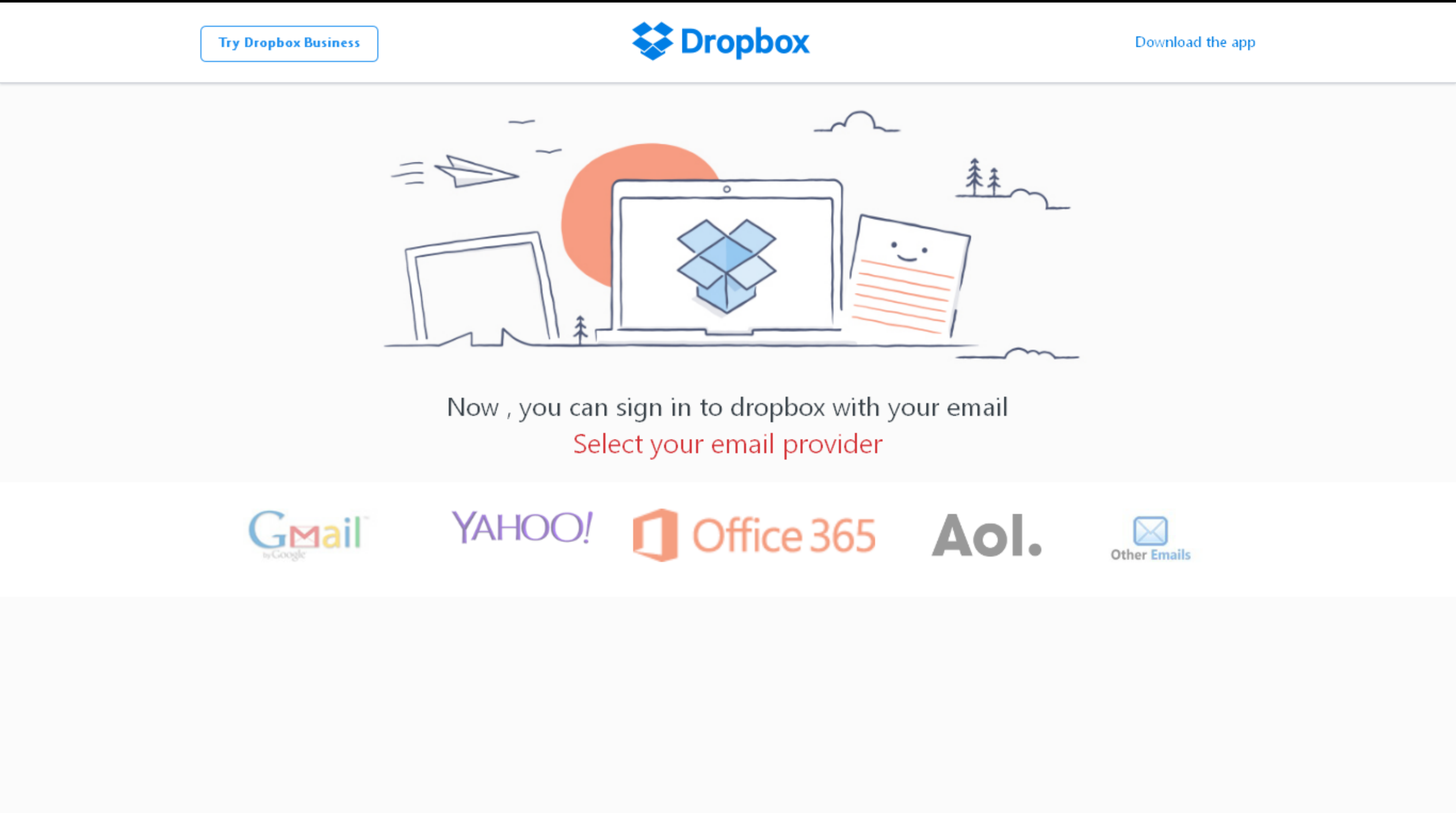
1. We take the adversarial webpages (not just logos!) generated in the USENIX Sec’24 paper *which bypassed PhishIntention* (the target system)
2. We use them to carry out a user study (N=150): *can users identify a phishing webpage* (half of the webpages are benign)*?* <u>(priming)</u>
   a. First, we do this with “non-adversarial” logos
   b. Then, we do this with “adversarial” logos generated via LogoMorph

Generated via brute-force search

Figure 4: **Text Logo Attack Examples**—The first row displays the brand’s original logo. The second row shows attack fonts with cosine similarity (about 0.86) that is slightly below the detection threshold. The third row exhibits adversarial logos with a lower cosine similarity (about 0.79). All these fonts can bypass detection.

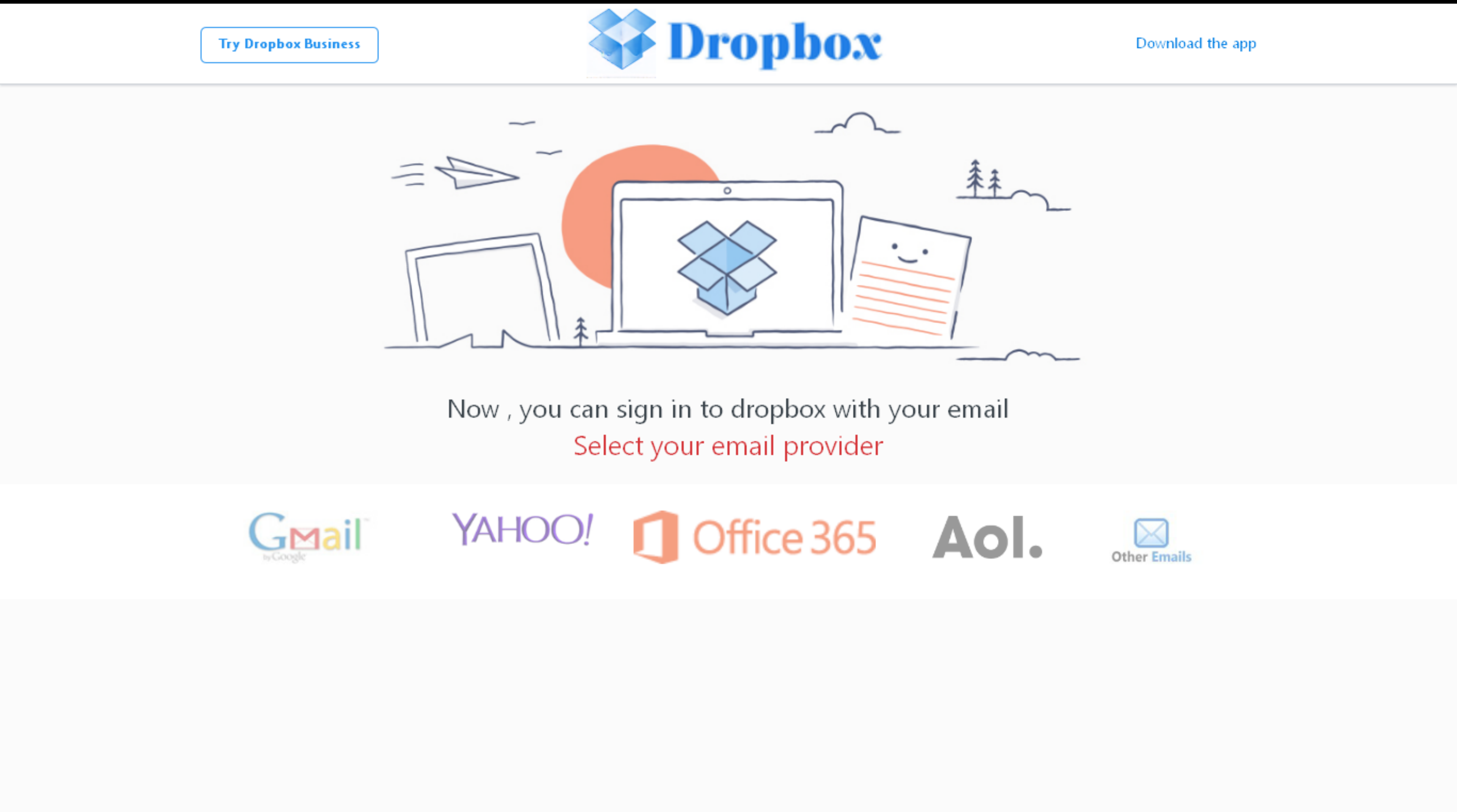
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

How did we do it? [USENIX Sec’24]



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

How did we do it? [USENIX Sec’24]



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

What did we find? [USENIX Sec'24]

- The impression is that users can recognize adversarial-phishing webpages slightly better…

Study	Accuracy	TPR	TNR
Adversarial	0.69	0.59	0.79
Baseline	0.60	0.45	0.75

Table 9: **Users Study Results**—The adversarial study uses phishing webpages with our adversarial logos. The baseline study uses original phishing pages. We report the overall accuracy, true positive rate (TPR), and true negative rate (TNR).

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

What did we find? [USENIX Sec'24]

- …however, when asked "what influenced your decision?", participants provide reasons that have nothing to do with the logo! (which was the only thing we changed)
  - Only 23% of the participants who correctly identified a webpage to be phishing mentioned "logo" in their responses.

Takeaway. Despite users recognizing adversarial phishing webpages slightly better than the original ones, it remains difficult for users to recognize adversarial phishing pages accurately (TPR=0.59). Also, most of the provided explanations are not related to our `LogoMorph` attack.

Conclusions

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

Outline of Today

- Using Machine Learning (ML) for Phishing Website Detection
- "Trivially" evading ML-based Phishing Website Detectors
- Using ML to evade ML-based Phishing Website Detectors
- The viewpoint of human users in the above

UNIVERSITÄT LIECHTENSTEIN

Two goals:

- Inspire you (to do/consider doing research in computer security)
- Entertain you (research should be fun)

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

Outline of Today – Takeaways

- Using Machine Learning (ML) for Phishing Website Detection
  - **Many ways exist, which are far from perfect (but they're the best we have) → Lots of room for improvement**
- "Trivially" evading ML-based Phishing Website Detectors
- Using ML to evade ML-based Phishing Website Detectors
- The viewpoint of human users in the above

UNIVERSITÄT LIECHTENSTEIN

Two goals:

- Inspire you (to do/consider doing research in computer security)
- Entertain you (research should be fun)

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

Outline of Today – Takeaways

- Using Machine Learning (ML) for Phishing Website Detection
  - **Many ways exist, which are far from perfect (but they're the best we have) → Lots of room for improvement**
- "Trivially" evading ML-based Phishing Website Detectors
  - **Real attackers favor cheap tactics, which are often effective (hard to convince reviewers that these "cheap tactics" are interesting…)**
- Using ML to evade ML-based Phishing Website Detectors
- The viewpoint of human users in the above

UNIVERSITÄT LIECHTENSTEIN

Two goals:

- Inspire you (to do/consider doing research in computer security)
- Entertain you (research should be fun)

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

Outline of Today – Takeaways

- Using Machine Learning (ML) for Phishing Website Detection
  - **Many ways exist, which are far from perfect (but they're the best we have) → Lots of room for improvement**
- "Trivially" evading ML-based Phishing Website Detectors
  - **Real attackers favor cheap tactics, which are often effective (hard to convince reviewers that these "cheap tactics" are interesting…)**
- Using ML to evade ML-based Phishing Website Detectors
  - **You can go crazy with sophisticated techniques to bypass state-of-the-art systems (but always consider how expensive they are…)**
- The viewpoint of human users in the above

UNIVERSITÄT LIECHTENSTEIN

Two goals:

- Inspire you (to do/consider doing research in computer security)
- Entertain you (research should be fun)

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

Outline of Today – Takeaways

- Using Machine Learning (ML) for Phishing Website Detection
  - **Many ways exist, which are far from perfect (but they're the best we have) → Lots of room for improvement**
- "Trivially" evading ML-based Phishing Website Detectors
  - **Real attackers favor cheap tactics, which are often effective (hard to convince reviewers that these "cheap tactics" are interesting…)**
- Using ML to evade ML-based Phishing Website Detectors
  - **You can go crazy with sophisticated techniques to bypass state-of-the-art systems (but always consider how expensive they are…)**
- The viewpoint of human users in the above
  - **ALWAYS consider that humans are the ultimate target of phishing websites (attackers want to phish people–not evade systems!)**

Two goals:

- Inspire you (to do/consider doing research in computer security)
- Entertain you (research should be fun)

The many faces of AI in the Phishing-website landscape

Giovanni Apruzzese

University of St. Gallen – November 28th, 2024

UNIVERSITÄT
LIECHTENSTEIN