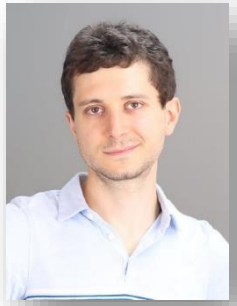
The many faces of AI in the Phishing-website Landscape

Giovanni Apruzzese

November 26th, 2025

UNIVERSITÄT LIECHTENSTEIN

whoami: Dr. **Giovanni Apruzzese**



Background:

- Assistant Professor at the University of Liechtenstein (UniLie) since Sept. 2022
- Assistant Professor at Reykjavik University (Iceland) since Aug. 2025
- Joined UniLie in 2020 as a PostDoc researcher (supervisor: Prof. Pavel Laskov)
- Studied (BSc, MSc, PhD) at the University of Modena, Italy (from 2010 to 2020)
  - BSc [2010-2013] and MSc [2013-2016] in "Computer Engineering"
  - PhD [2016-2020] in "Information and Communication Technologies"
- In 2019, spent 6 months at Dartmouth College, USA as a visiting scholar.

What do I do (and what do I like)?

- Research
  - Cybersecurity, machine learning, networked and distributed systems, intrusion detection, phishing, gaming, human factor in cybersecurity
- Reviewing
- Teaching
- Collaborating (and interacting) with experts and passionate individuals

Contact: giovannia@ru.is (further info: https://giovanniapruzzese.com)

Outline

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

Outline (truth)

- Show you how to break ML-based systems
- Show you how operational ML-based systems fail
- Show you how "easy" it is to cause harm by exploiting ML methods
- Show you (a glimpse of) the human factor in ML&Cybersecurity

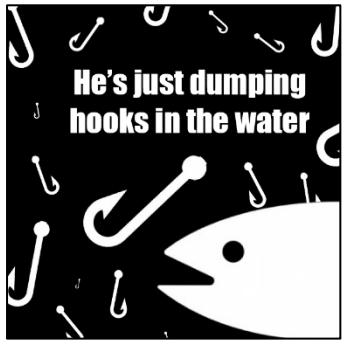
Two goals:

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

Phishing Website Detection (via ML)

Current Landscape of Phishing

- Phishing attacks are continuously increasing
- Most detection methods still rely on *blocklists* of malicious URLs



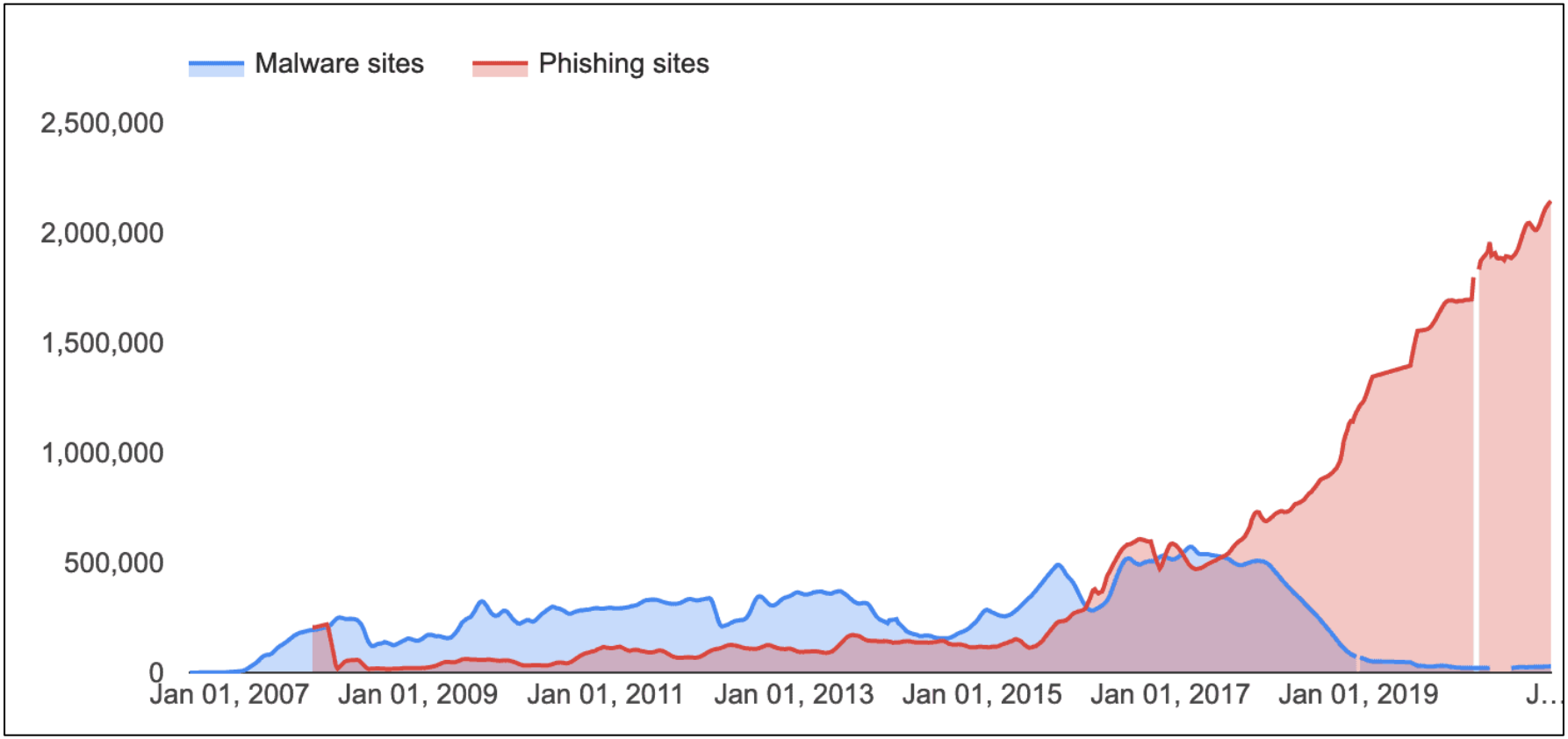
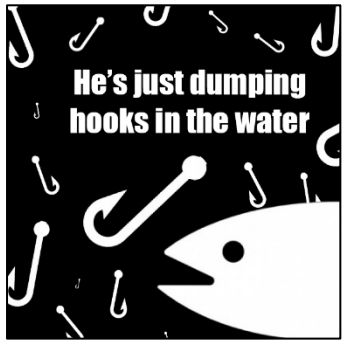


Image source: https://www.tessian.com/blog/phishing-statistics-2020/

Giovanni Apruzzese, *PhD*
https://giovanniapruzzese.com

Current Landscape of Phishing

- Phishing attacks are continuously increasing
- Most detection methods still rely on *blocklists* of malicious URLs



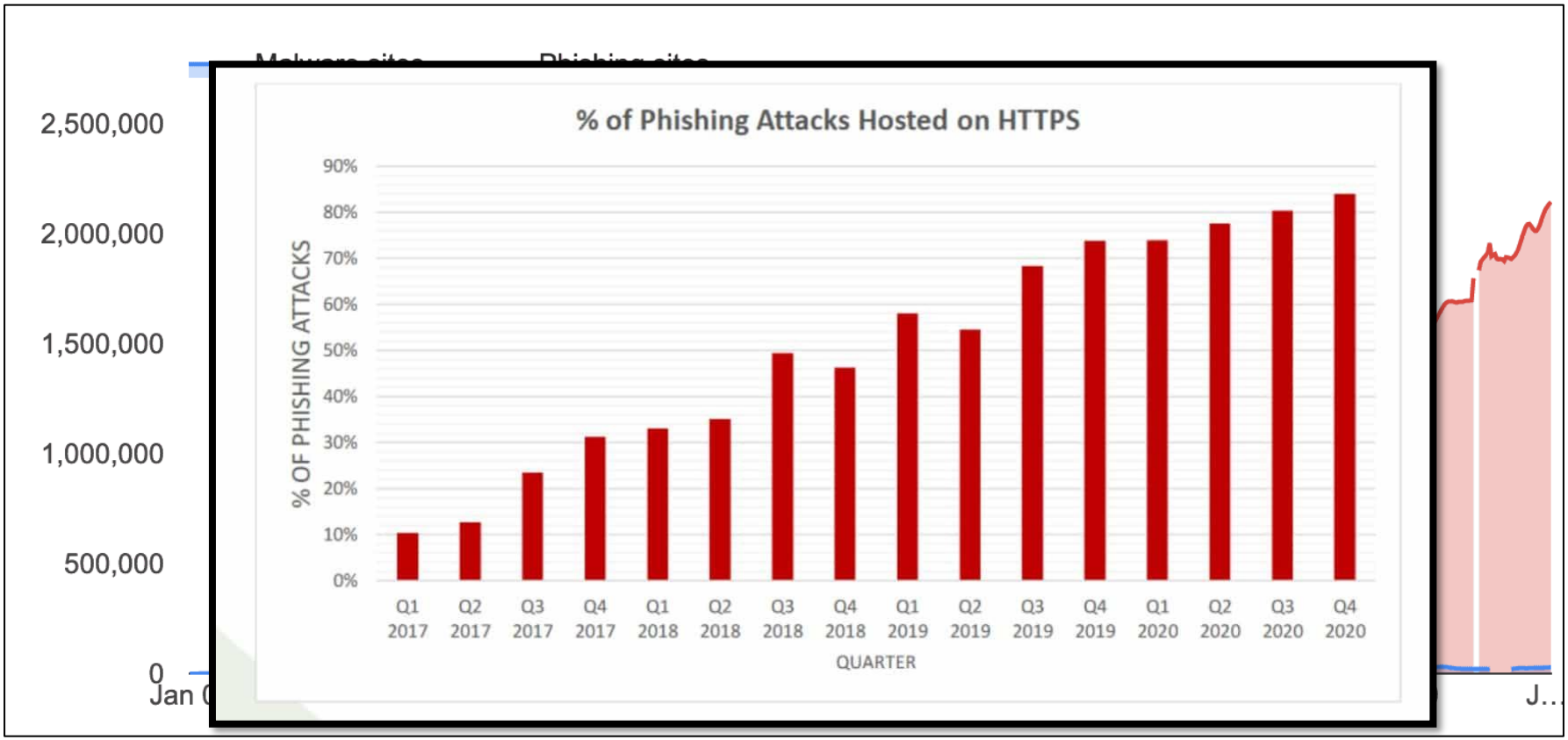


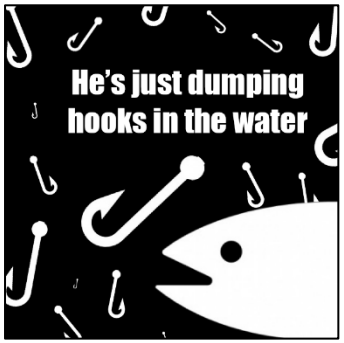
Image source: https://www.tessian.com/blog/phishing-statistics-2020/

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

Giovanni Apruzzese, *PhD*
https://giovanniapruzzese.com

Current Landscape of Phishing

- Phishing attacks are continuously increasing
- Most detection methods still rely on *blocklists* of malicious URLs



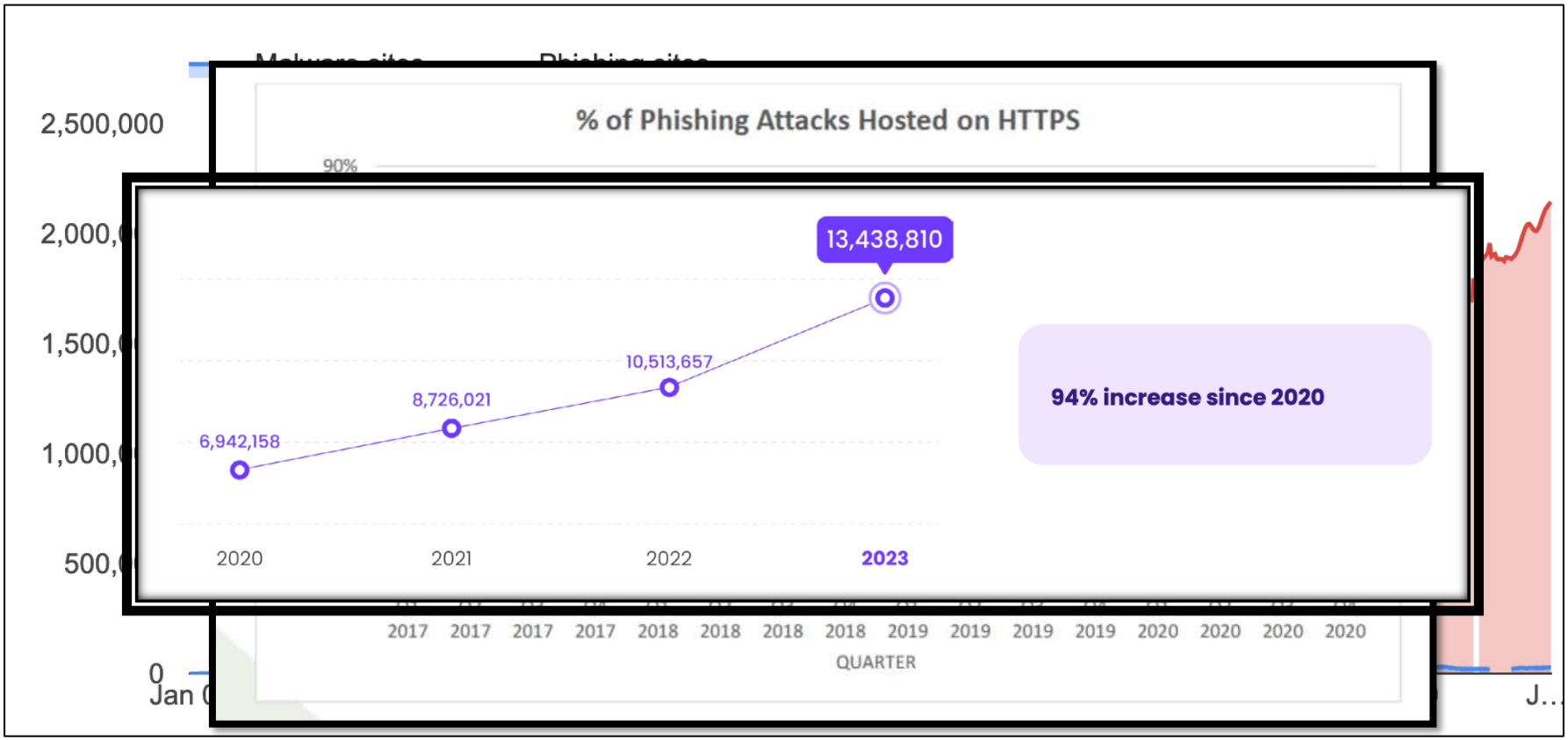


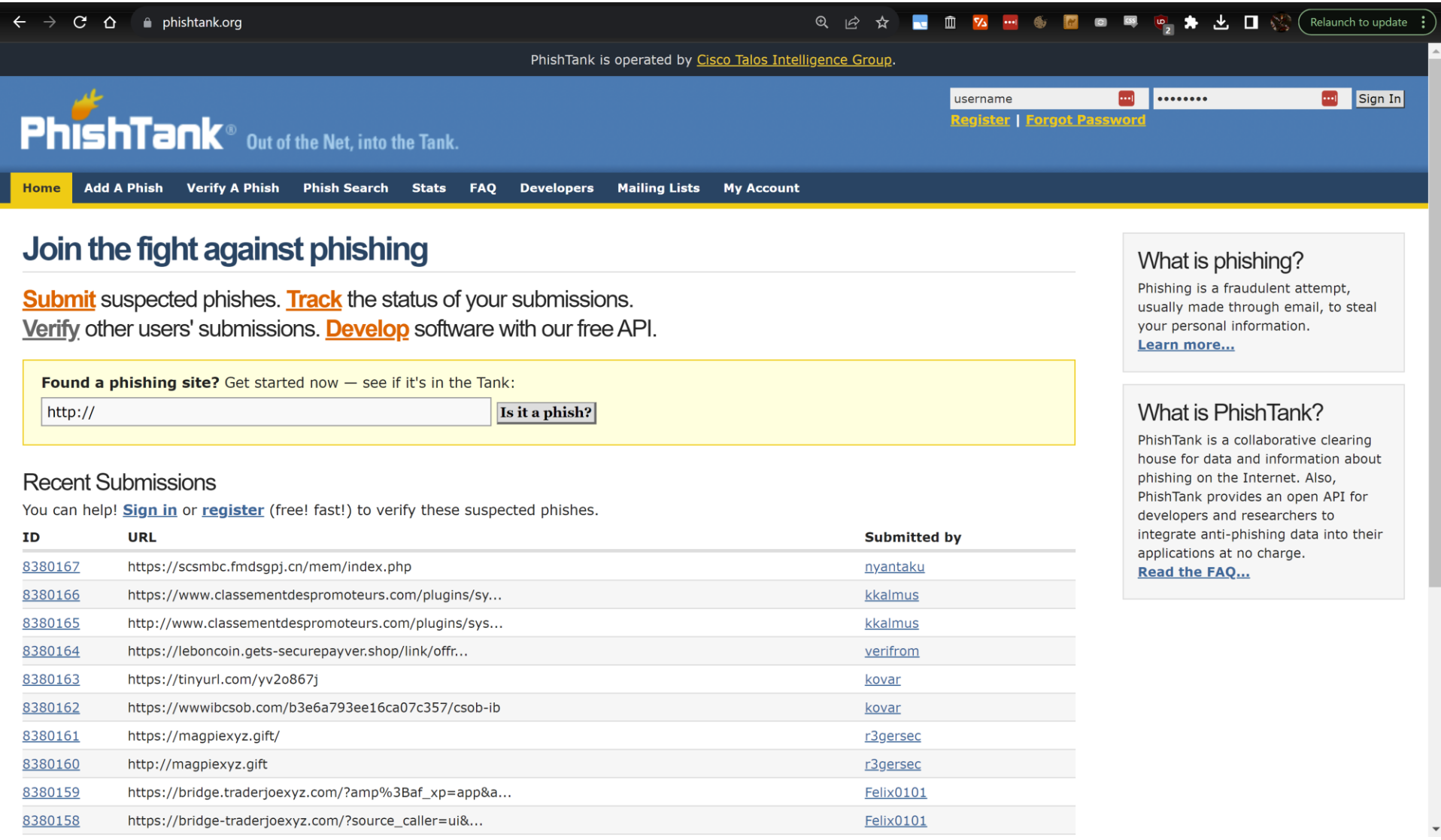
Image source: https://www.tessian.com/blog/phishing-statistics-2020/

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*
https://giovanniapruzzese.com

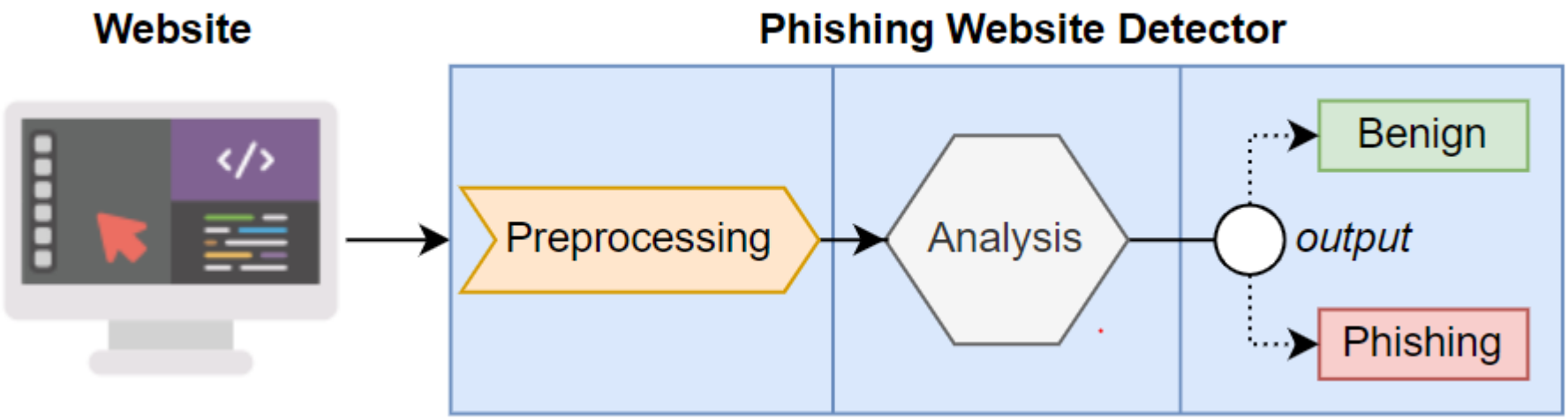
Up-to-date list of phishing URLs: PhishTank (www.phishtank.org)



Question: how do you think such blocklists are kept up to date?

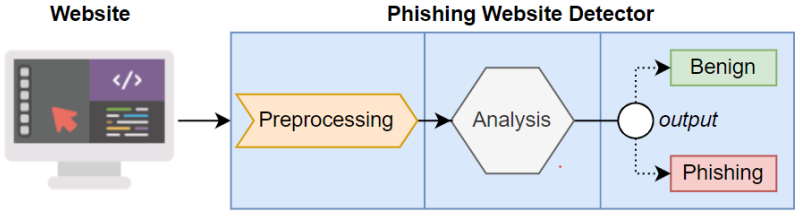
Current Landscape of Phishing – Countermeasures

- Countering phishing websites can be done via *data-driven* methods



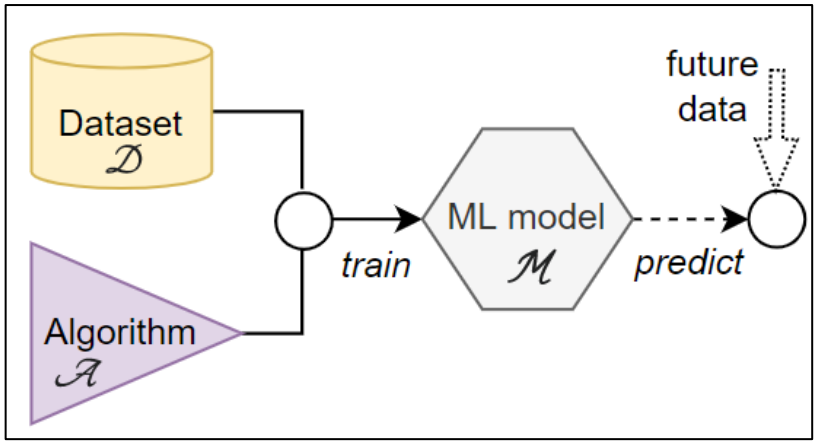
Current Landscape of Phishing – Countermeasures (ML)

- Countering phishing websites can be done via *data-driven* methods



- Such methods include (also) Machine Learning techniques:

We will focus on these for now



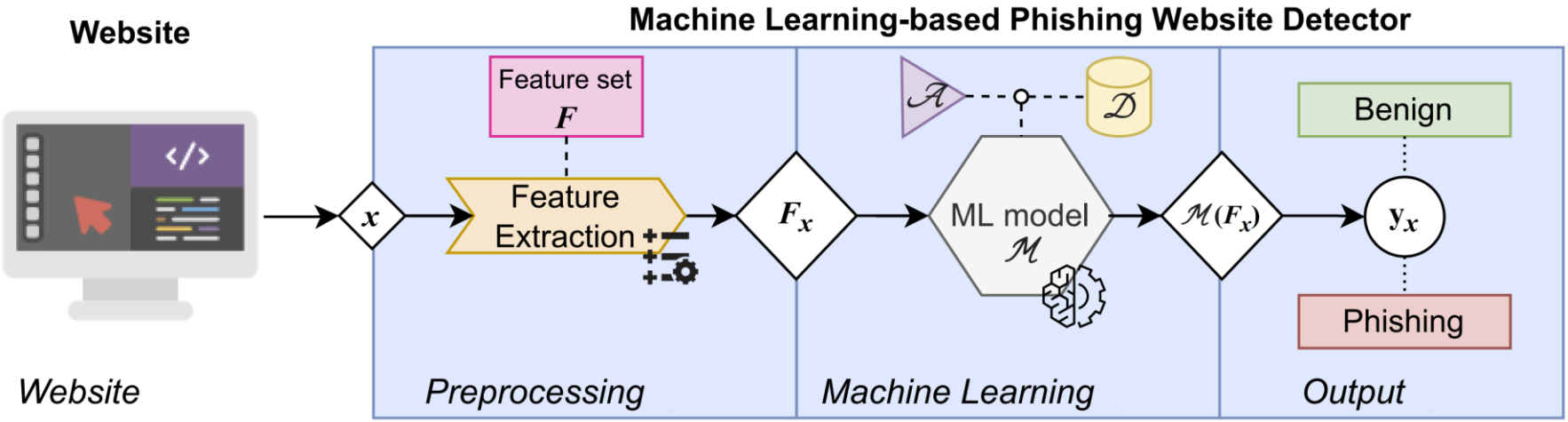
- Machine Learning-based Phishing Website Detectors (ML-PWD) are very effective [1]
  - Even popular products and web-browsers (e.g., Google Chrome) use them [2, 3]

[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*

Phishing Website Detection (via ML)

- The *detection* of a phishing webpage can entail the analysis of various elements, e.g.:
  - 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) More on these later

- These analyses can be done via ML-based classifiers [4,5]
  - To apply ML for phishing website detection, we typically need to "preprocess" the webpage and extract the "feature representation" that will be analysed by the (trained) ML



[4]: Mohammad, Rami M., Fadi Thabtah, and Lee McCluskey. "Predicting phishing websites based on self-structuring neural network." Neural Computing and Applications 25 (2014): 443-458.
[5]: **Apruzzese, Giovanni**, Mauro Conti, and Ying Yuan. "SpacePhish: The evasion-space of adversarial attacks against phishing website detectors using machine learning." ACSAC, 2022

Do feature-based detectors work? (evidence from ACSAC'22 [5])

It is possible to develop good ML-based detectors by analysing various types of "features" (URL-based, HTML-based, or a combination thereof) 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	$0.96_{\pm 0.007}$	$0.025_{\pm 0.0077}$	$0.81_{\pm 0.020}$	$0.013_{\pm 0.0037}$

[5]: **Apruzzese, Giovanni**, Mauro Conti, and Ying Yuan. "SpacePhish: The evasion-space of adversarial attacks against phishing website detectors using machine learning." ACSAC 2022

Do feature-based detectors work? (evidence from ACSAC'22 [5])

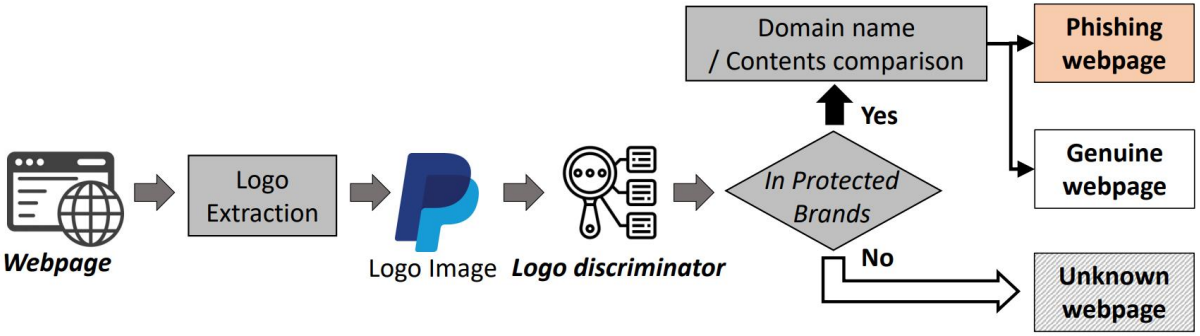
It is possible to develop good ML-based detectors by analysing various types of "features" (URL-based, HTML-based, or a combination thereof) and by using diverse types of ML algorithms, such as random forests (RF), logistic regression (LR), or convolutional neural networks (CN)

$\mathcal{A}$	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: extremely high consumption of resources

Phishing Website Detection: Reference Based (visual similarity)

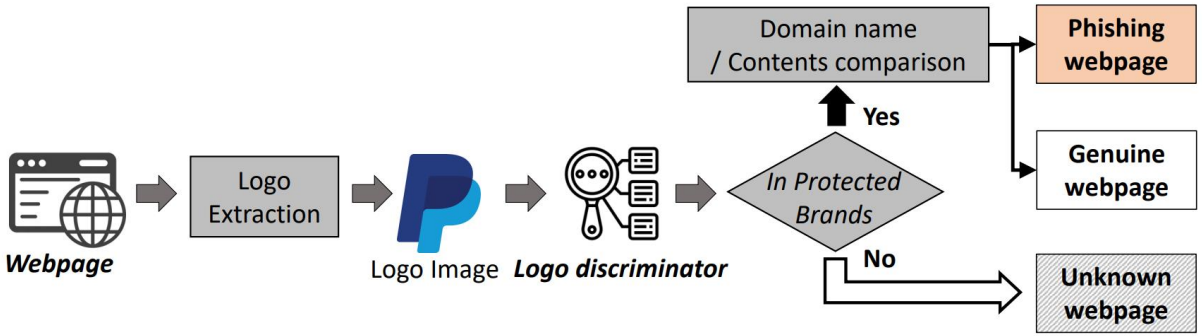
- Some detectors leverage the intuition that most phishing webpages try to mimic well-known brands, but they are hosted under a different domain (e.g., [6,7])
- These *reference based* detectors can provide some protection against phishing websites that target a restricted set of brands (e.g., PayPal, Amazon, Google).



[6]: Liu, Ruofan, et al. "Inferring phishing intention via webpage appearance and dynamics: A deep vision based approach." 31st USENIX Security Symposium (USENIX Security 22). 2022.
[7]: Lin, Yun, et al. "Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages." 30th USENIX Security Symposium (USENIX Security 21). 2021.

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 (e.g., [6,7])
- These *reference based* detectors can provide some protection against phishing websites that target a restricted set of brands (e.g., PayPal, Amazon, Google).

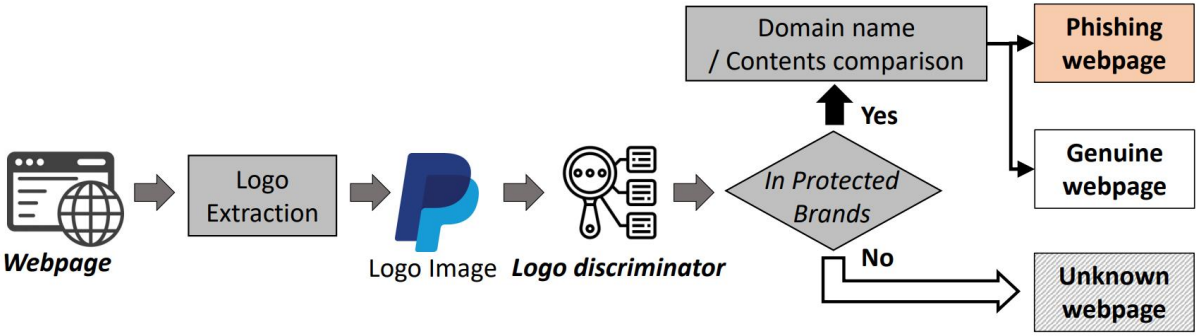


- First, they see if a webpage is visually similar to a webpage of well-known brands.
  - E.g., is this webpage similar to any webpage of PayPal, Amazon, or Google?
    - (If a match is NOT found, then the webpage is treated as benign (to avoid triggering false positives)
- Then, if a match is found, 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).

[6]: Liu, Ruofan, et al. "Inferring phishing intention via webpage appearance and dynamics: A deep vision based approach." 31st USENIX Security Symposium (USENIX Security 22). 2022.
[7]: Lin, Yun, et al. "Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages." 30th USENIX Security Symposium (USENIX Security 21). 2021.

Do logo-based detectors work? (empirical evidence from ESORICS'23 [8])

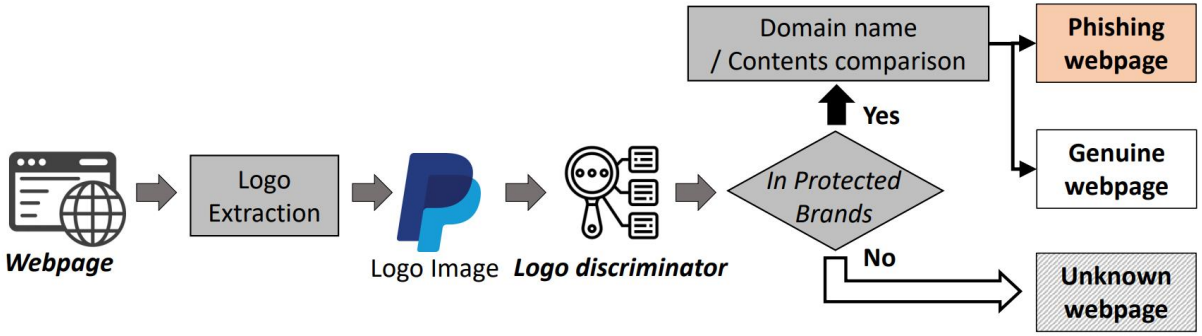
- These systems work very well when the "visual similarity" is carried out from a **logo perspective**.
- To make the procedure faster, the similarity is computed by means of discriminators reliant on deep learning models.



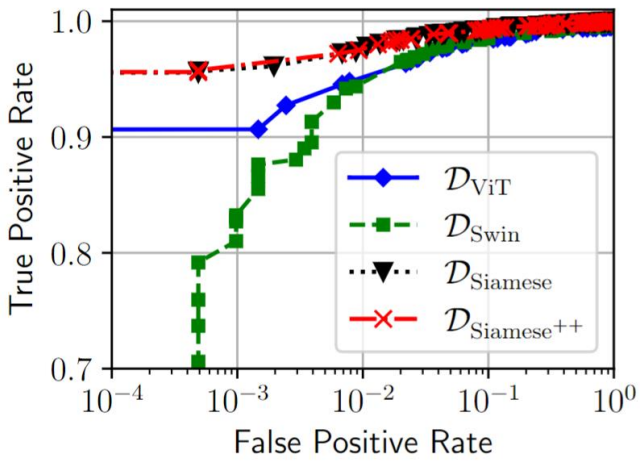
[8]: Lee, J., Xin, Z., See, M. N. P., Sabharwal, K., **Apruzzese, G.**, & Divakaran, D. M. (2023, September). Attacking logo-based phishing website detectors with adversarial perturbations. ESORICS, 2023

Do logo-based detectors work? (empirical evidence from ESORICS'23 [8])

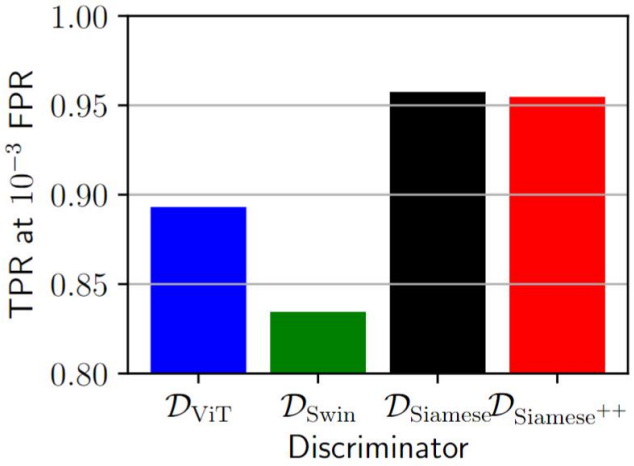
- These systems work very well when the "visual similarity" is carried out from a **logo perspective**.
- To make the procedure faster, the similarity is computed by means of discriminators reliant on deep learning models.



Code on GitHub



(a) ROC curves



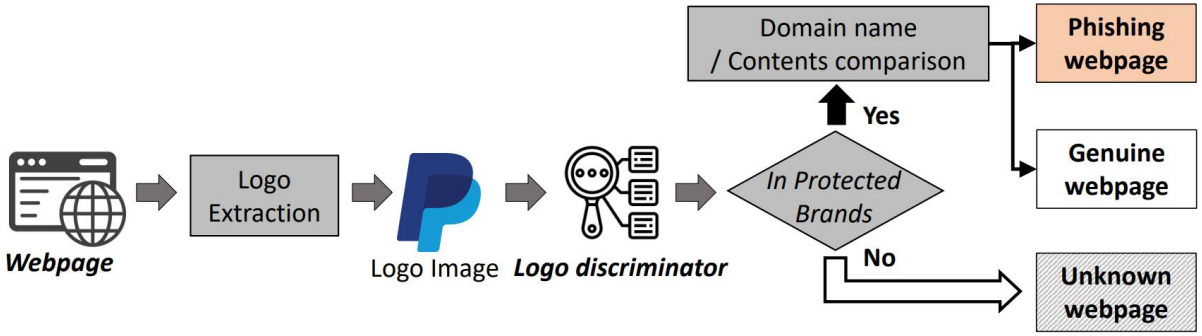
(b) TPR at 10^{-3} FPR

Models trained to identify 181 brands (~28k logos)

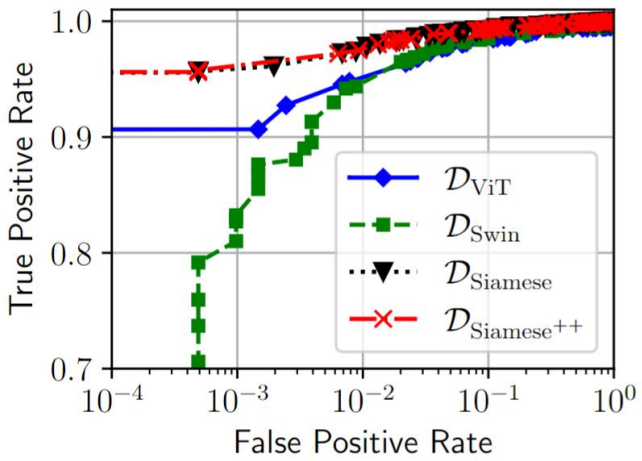
[8]: Lee, J., Xin, Z., See, M. N. P., Sabharwal, K., **Apruzzese, G.**, & Divakaran, D. M. (2023, September). Attacking logo-based phishing website detectors with adversarial perturbations. ESORICS, 2023

Do logo-based detectors work? (empirical evidence from ESORICS'23 [8])

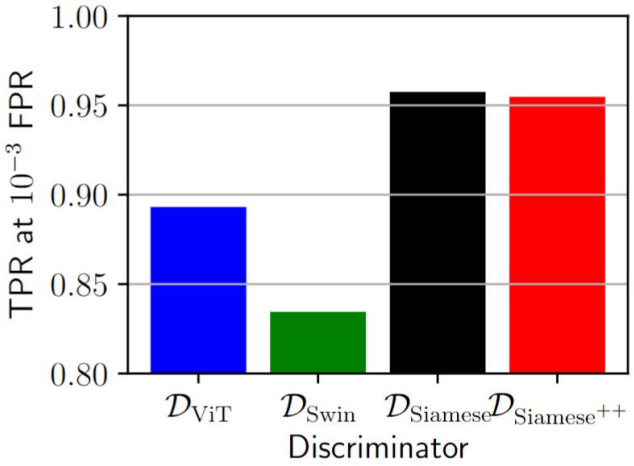
- These systems work very well when the "visual similarity" is carried out from a **logo perspective**.
- To make the procedure faster, the similarity is computed by means of discriminators reliant on deep learning models.



Code on GitHub



(a) ROC curves



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

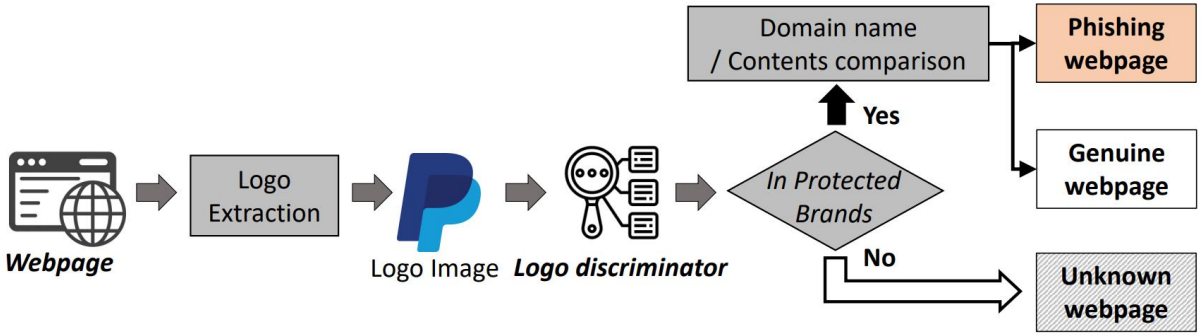
Models trained to identify 181 brands (~28k logos)

[8]: Lee, J. ... 23

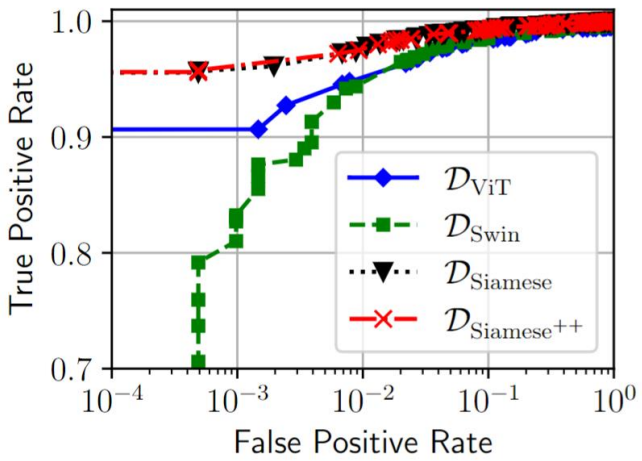
Limitation:

Do logo-based detectors work? (empirical evidence from ESORICS'23 [8])

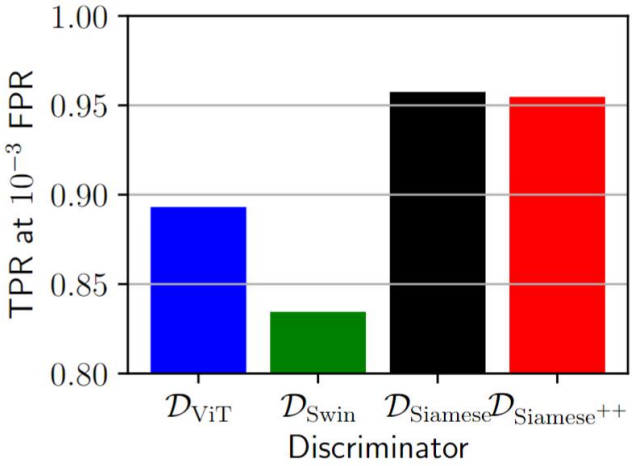
- These systems work very well when the "visual similarity" is carried out from a **logo perspective**.
- To make the procedure faster, the similarity is computed by means of discriminators reliant on deep learning models.



Code on GitHub



(a) ROC curves



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

Models trained to identify 181 brands (~28k logos)

[8]: Lee, J.

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

A deep look at feature-based PWD

Feature-based PWD

- But what are the *features* that can be used to discriminate between benign and phishing websites?

- Let's take a closer look at the (over 10-year old!) methods proposed by Mohammad et al [**An assessment of features related to phishing websites using an automated technique, 2012**]:
  - https://archive.ics.uci.edu/dataset/327/phishing+websites (shortened: https://tinyurl.com/dase25)

Feature-based PWD (cont'd)

- In one of my papers ([https://www.giovanniapruzzese.com/publications/tdsc22]), I have implemented the feature-extractor that enables one to derive a (similar) feature set by receiving a webpage as input.
  - When such features are used to develop some ML models, and such models are used to detect 'recent' webpages (collected in 2019), the performance was very high

	LNU-Phish			
Classifier	*F1-score*	*Acc*	*FPR*	*TPR*
RF	0.973	0.982	0.013	0.972
SVM	0.983	0.989	0.004	0.974
KNN	0.996	0.998	0.002	0.997
SGD	0.985	0.990	0.003	0.977
DT	0.986	0.991	0.006	0.985
LR	0.985	0.990	0.003	0.977
NB	0.955	0.971	0.010	0.932
MLP	0.990	0.994	0.001	0.983
AB	0.986	0.991	0.002	0.977
ET	0.999	0.999	0.001	0.999
GB	0.999	0.999	0.001	0.999
DnW	0.988	0.992	0.002	0.980
Bag	0.987	0.992	0.003	0.980
best	0.999	0.999	0.001	0.999

Evading Phishing Website Detectors

UNIVERSITÄT
LIECHTENSTEIN

Phishing in a nutshell

- Phishing websites are taken down quickly
  - The moment they are reported in a blocklist, they become useless

Most phishing attacks end up in failure [9]

[9] 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)

Phishing in a nutshell

- Phishing websites are taken down quickly
  - The moment they are reported in a blocklist, they become useless

Most phishing attacks end up in failure [9]

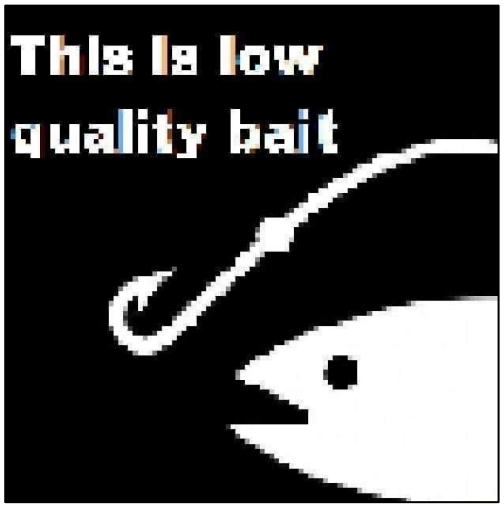
- Phishers are well aware of this fact… but they (clearly) keep doing it
  - Hence, they "have to" evade detection mechanisms…
  - *(and, as a result, we need to keep improving our systems)*
  - **…but they cannot invest "a lot of resources" for every new phishing webpage!**

This is why it is important to come up with (and assess) new "attacks" even from a defensive standpoint

[9] 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)

Phishing in a nutshell – cont'd

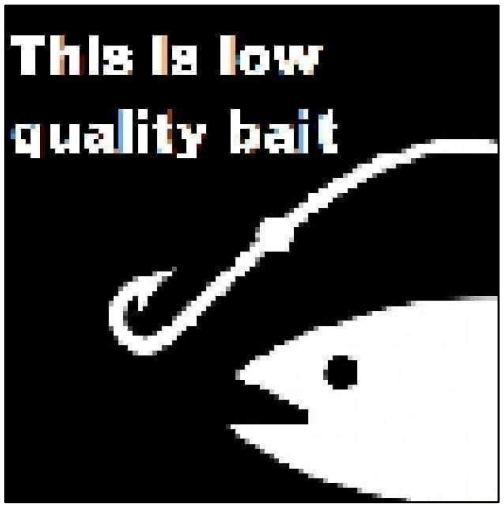
- Even when a user lands on a phishing website, the attack is not complete yet
- The user must still click on the button / submit the credentials



Mere "evasion" of a detector is not enough
(for an attacker ☺)

Phishing in a nutshell – cont'd

- Even when a user lands on a phishing website, the attack is not complete yet
- The user must still click on the button / submit the credentials

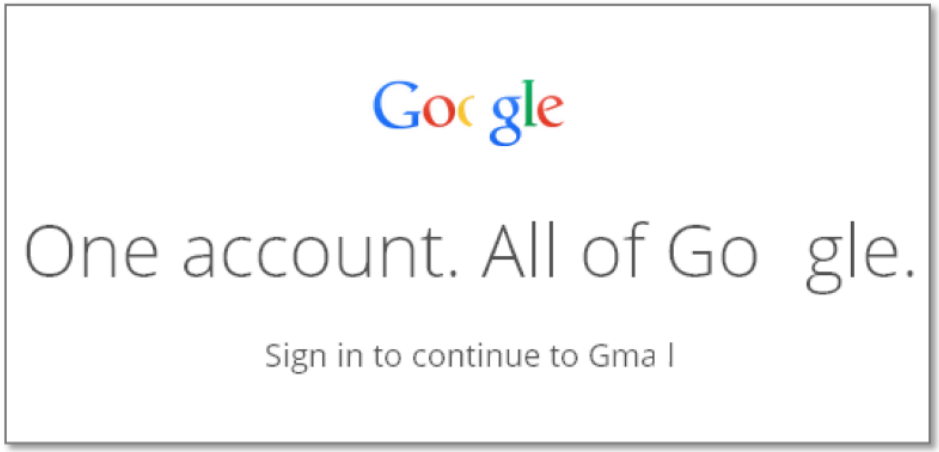


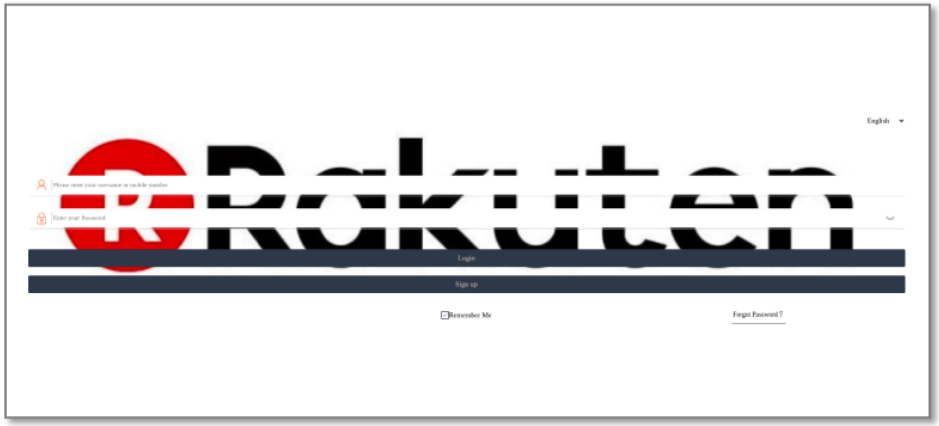
Mere "evasion" of a detector is not enough
(for an attacker ☺)

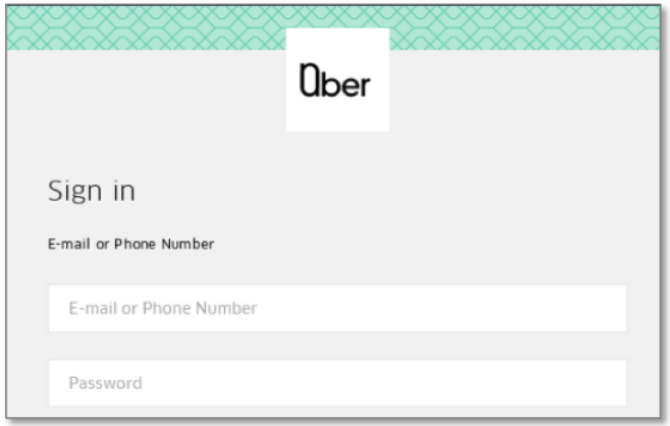
- Phishing "evasion" is a two-step process:
  - the website must bypass the detector (a "machine"), and
  - the website must be able to mislead the human user

Evidence from the real world

- The deep-learning powered detector used by a **cybersecurity company** was not able to detect webpages that looked like these ones:



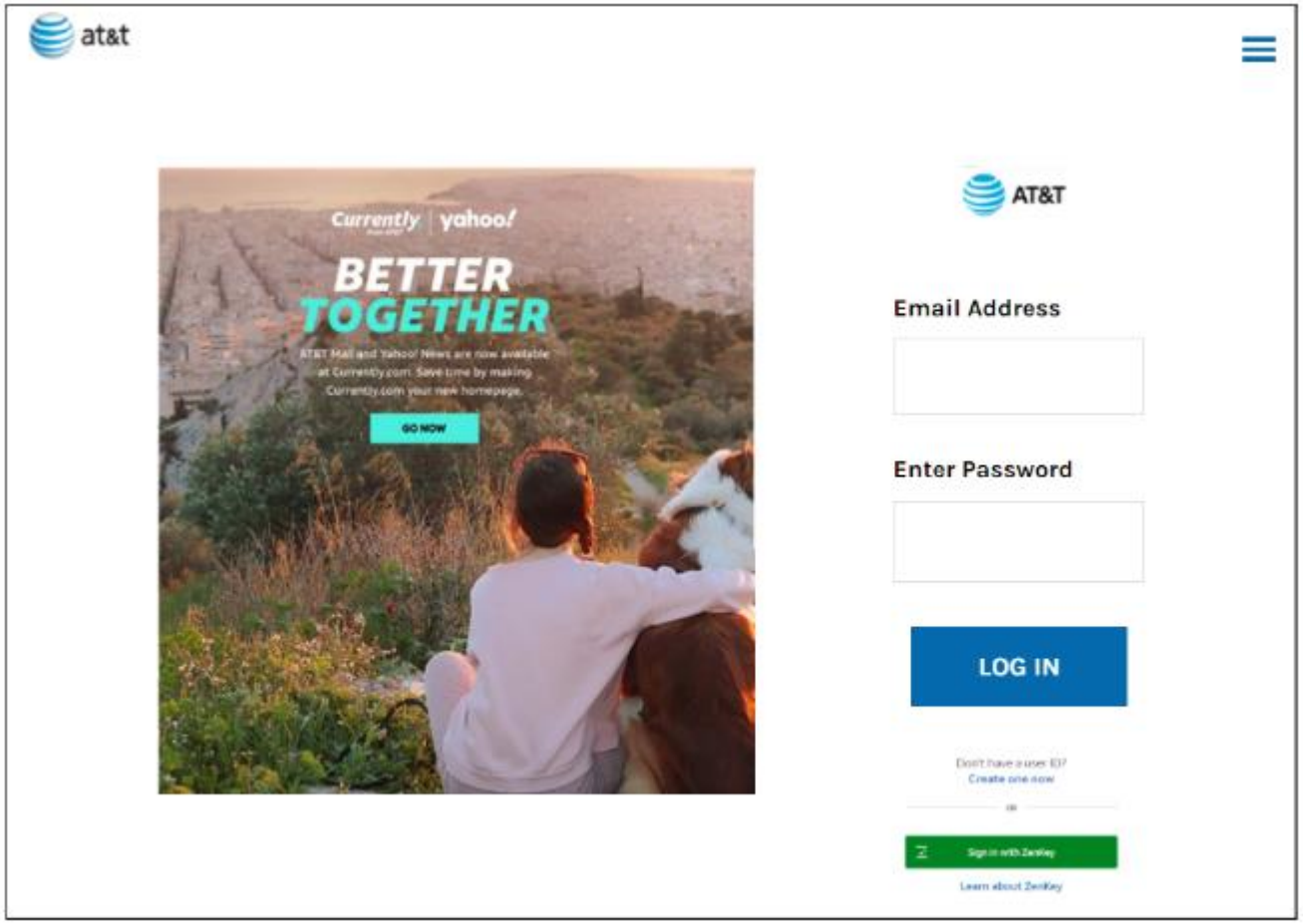




If you want to know more about this, you can ask later

Evading "feature-based" detectors – original example

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



[5]: **Apruzzese, Giovanni**, Mauro Conti, and Ying Yuan. "SpacePhish: The evasion-space of adversarial attacks against phishing website detectors using machine learning." ACSAC 2022

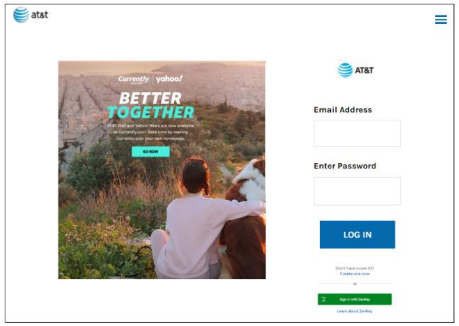
Evading “feature-based” detectors – changing the URL

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

https://www.legitimate123.weebly.com/

[5]: **Apruzzese, Giovanni**, Mauro Conti, and Ying Yuan. "SpacePhish: The evasion-space of adversarial attacks against phishing website detectors using machine learning." ACSAC 2022

Evading “feature-based” detectors – 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
```

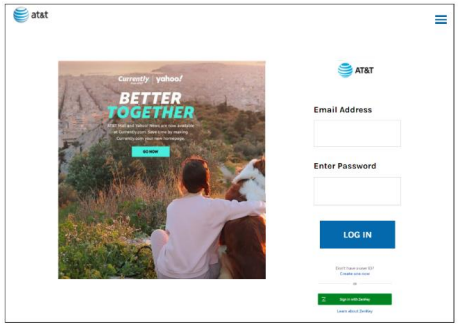
‘perturbation’

[5]: **Apruzzese, Giovanni**, Mauro Conti, and Ying Yuan. "SpacePhish: The evasion-space of adversarial attacks against phishing website detectors using machine learning." ACSAC 2022

Evading "feature-based" detectors – 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
```

'perturbation'

[5]: **Apruzzese, Giovanni**, Mauro Conti, and Ying Yuan. "SpacePhish: The evasion-space of adversarial attacks against phishing website detectors using machine learning." ACSAC 2022

Demonstration: competition-grade ML-PWD

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

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
```

```
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
}
```

Demonstration: competition-grade ML-PWD

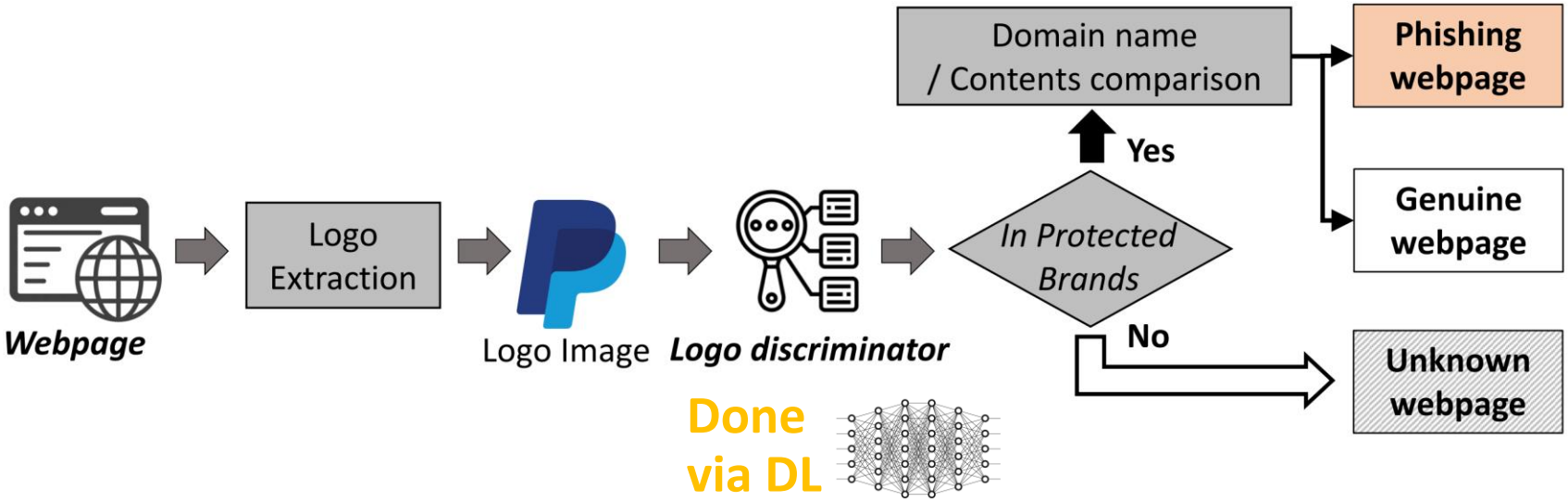
- https://spacephish.github.io (https://tinyurl.com/spacephish-demo)
- https://nbviewer.org/github/hihey54/acsac22_spacephish/blob/main/mlsec_folder/mlsec_artifact-manipulate.ipynb

```
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
        }
```

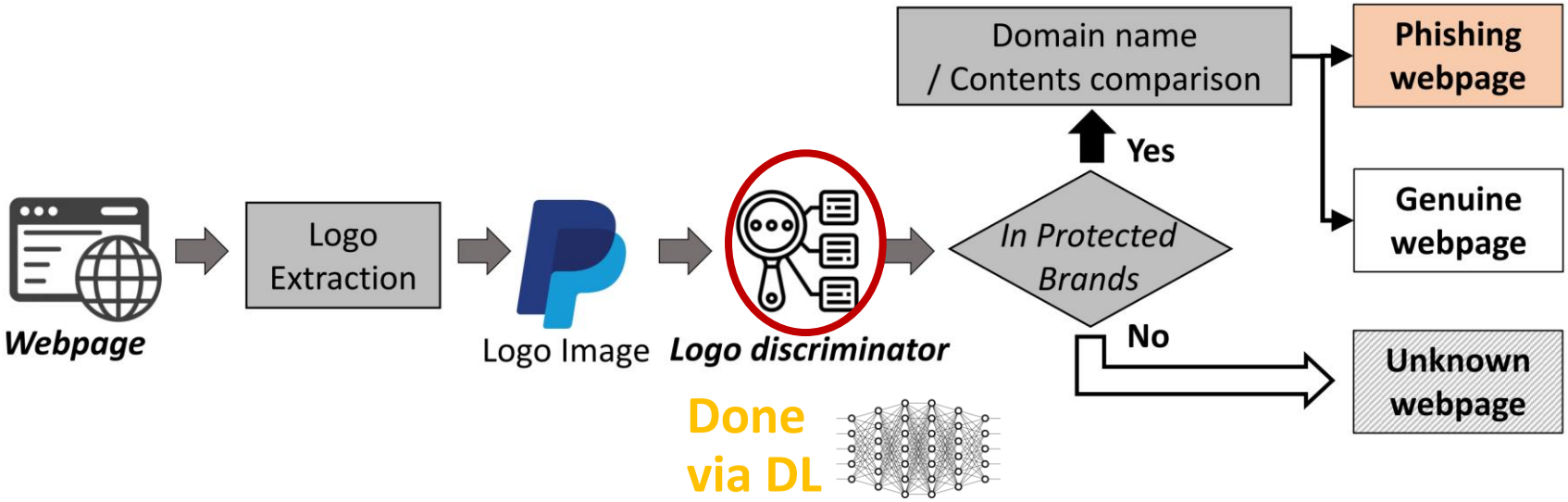
Evading "Logo-based" Phishing Website Detectors



- 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).

Evading "Logo-based" Phishing Website Detectors



Problem: these systems are tweaked to minimize false positives.

The focus is on the Logo-discriminator.

Attack: adversarial logos [8]

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

[8]: 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. ESORICS, 2023

Attack: adversarial logos [8]

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:**
   - the attacker expects the detector to have the "phished" brand(s) in the protected set (and that its logos are inspected)

No knowledge of the DL model is required!

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"

[8]: 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. ESORICS, 2023

Attack Method (how to generate adversarial logos?)

It is inspired by "GAN"

- The attack applies "Generative Adversarial Perturbations" (GAP)

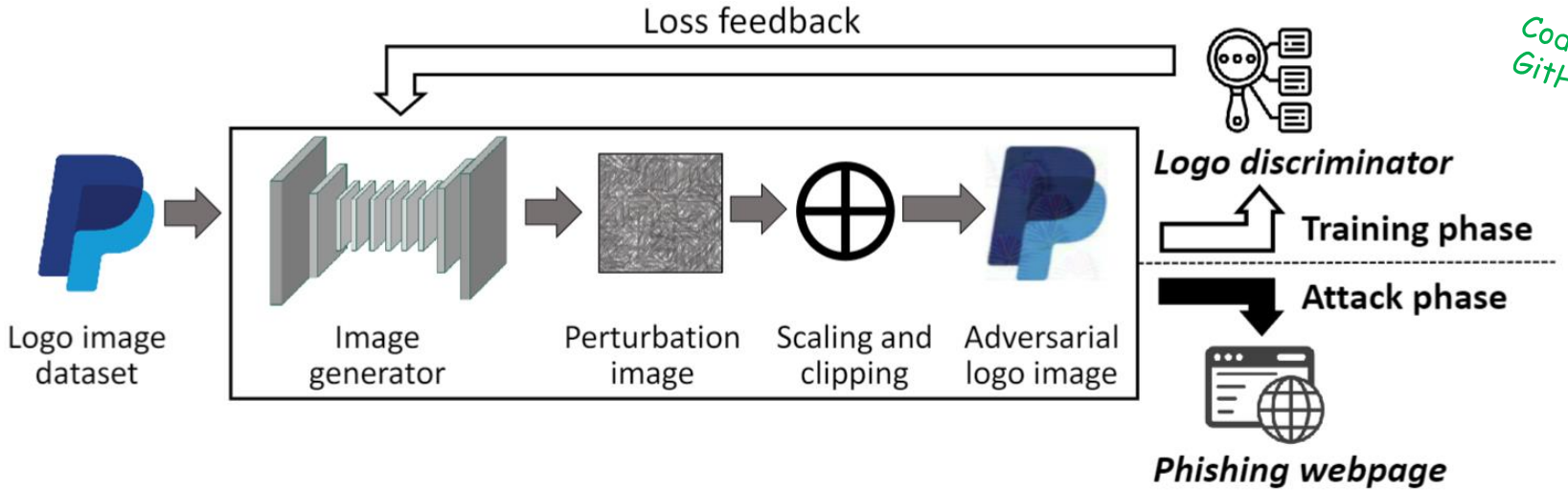


Fig. 4: Generative adversarial perturbation workflow

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

[8]: 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. ESORICS, 2023

Attack Method (how to generate adversarial logos?)

It is inspired by "GAN"

- The attack applies "Generative Adversarial Perturbations" (GAP)

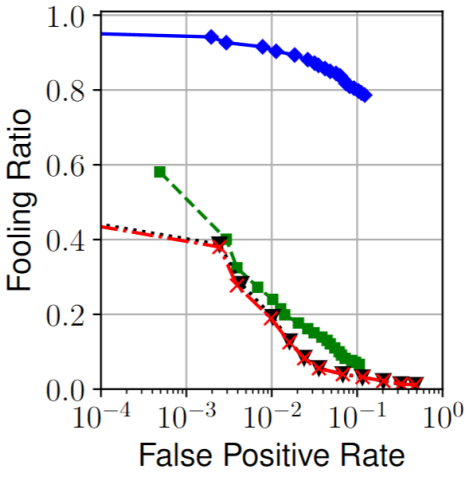
Code on GitHub

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

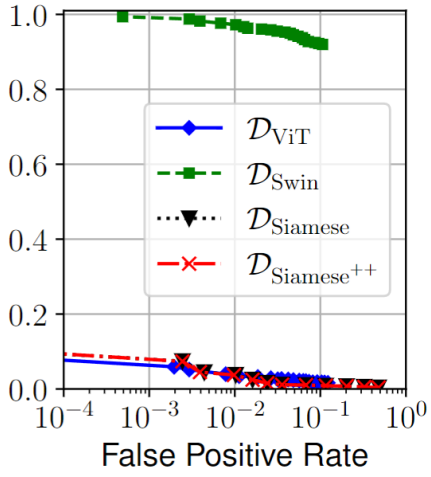
[8]: Lee, J., Xin, Z., See, M. N. P., Sabharwal, K., **Apruzzese, G.**, & Divakaran, D. M. (2023, September). Attacking logo-based phishing website detectors with adversarial perturbations. ESORICS, 2023

Results (do our adversarial logos fool the discriminators?)

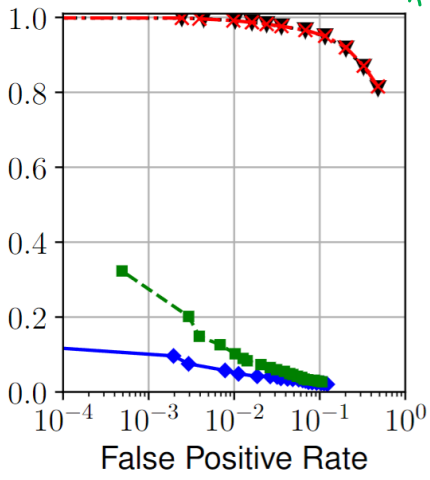
Higher = stronger attack



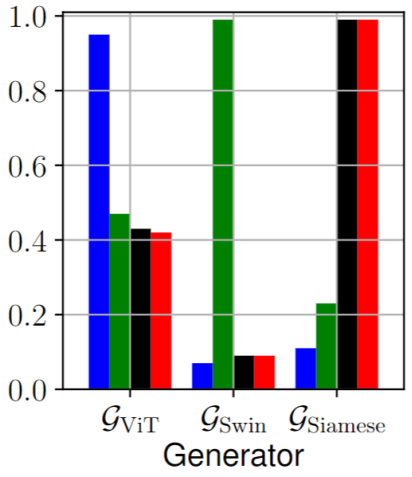
(a) $\mathcal{G}_{\text{ViT}}$



(b) $\mathcal{G}_{\text{Swin}}$



(c) $\mathcal{G}_{\text{Siamese}}$

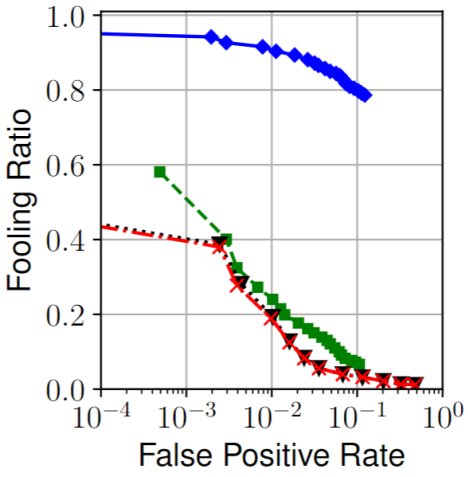


(d) at 10^{-3} FPR

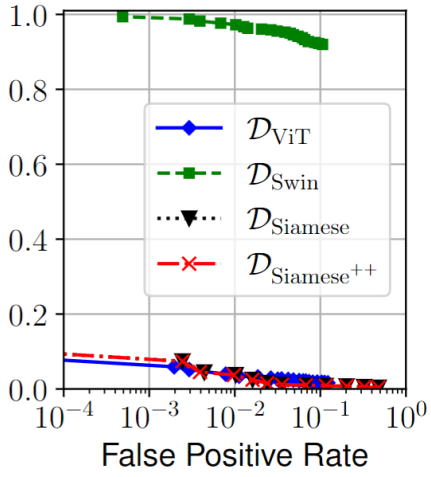
E.g.: G_{ViT} denotes the GAN trained to evade D_{ViT}

Results (do our adversarial logos fool the discriminators?)

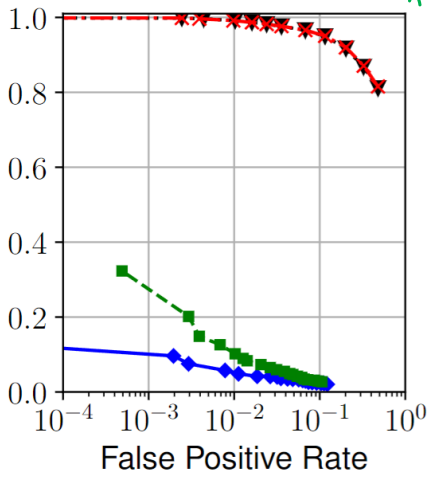
Higher = stronger attack



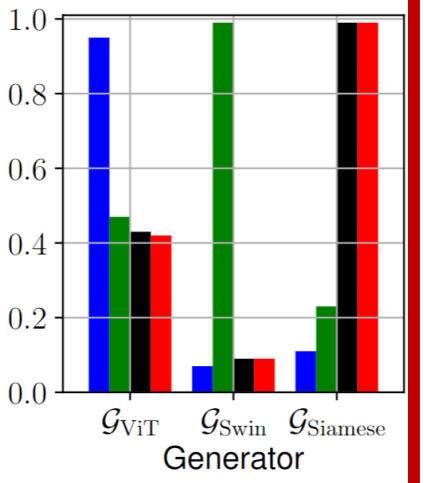
(a) $\mathcal{G}_{\text{ViT}}$



(b) $\mathcal{G}_{\text{Swin}}$



(c) $\mathcal{G}_{\text{Siamese}}$



(d) at 10^{-3} FPR

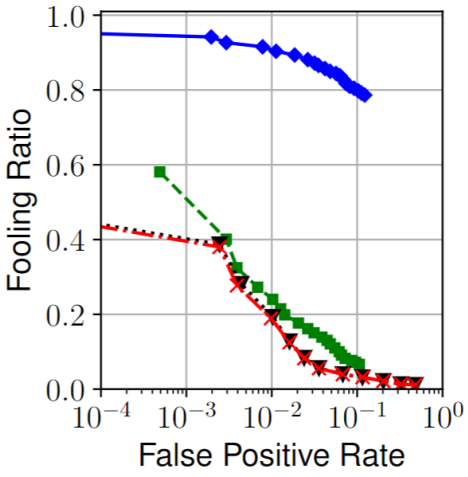
E.g.: G_{ViT} denotes the GAN trained to evade D_{ViT}

Takeaways:

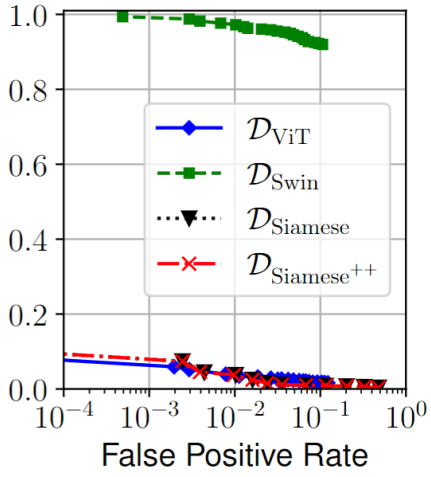
1. When the attacker and defender use the same model, the attack is ~100% effective
2. ViT is the “more robust” detector! (if the attacker is blind)

Results (do our adversarial logos fool the discriminators?)

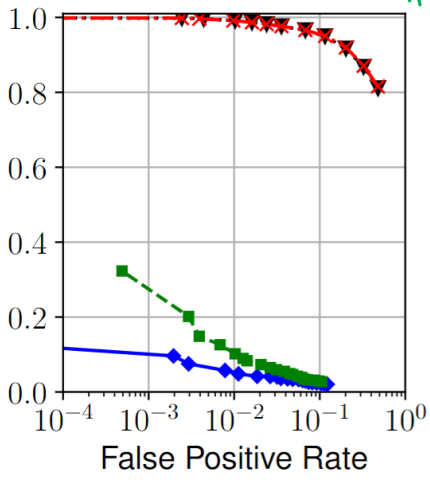
Higher = stronger attack



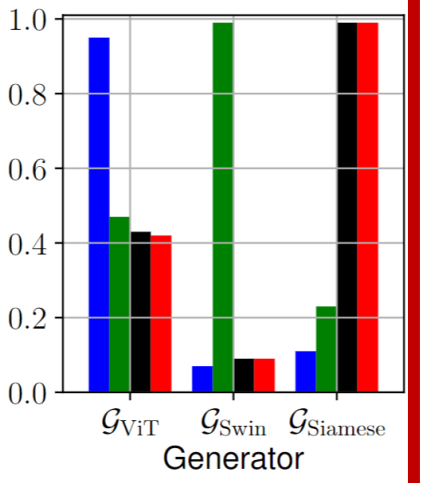
(a) $\mathcal{G}_{\text{ViT}}$



(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)

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

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.

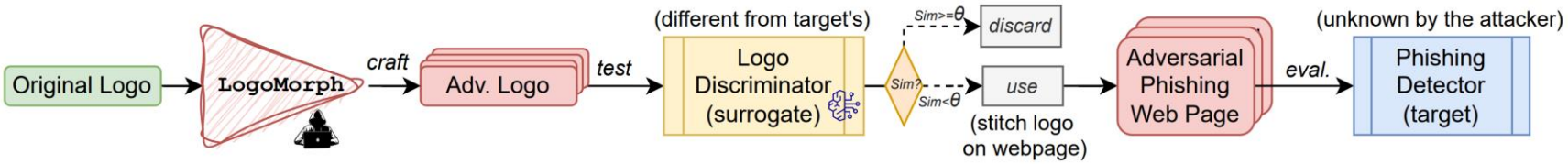


Figure 6: **Our Blackbox Experiment Setup.**—We use a surrogate logo discriminator (which is different from the one used by the target model) to generate and select adversarial logos via `LogoMorph`. Logos that bypass the surrogate discriminator (by achieving a low similarity) will be used to attack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

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.

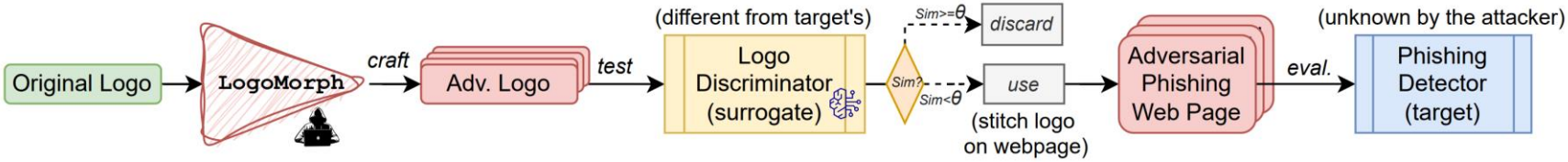


Figure 6: **Our Blackbox Experiment Setup.**—We use a surrogate logo discriminator (which is different from the one used by the target model) to generate and select adversarial logos via `LogoMorph`. Logos that bypass the surrogate discriminator (by achieving a low similarity) will be used to attack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

Some examples of LogoMorph…

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

Some examples of LogoMorph…

Generated with the diffusion model

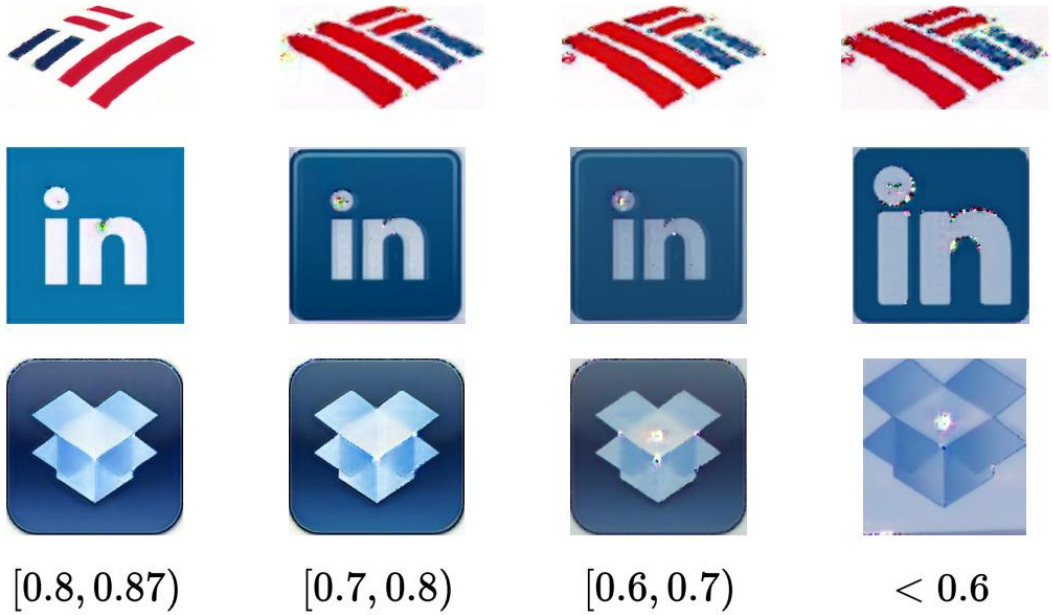


Figure 5: **Image Logo Attack Examples**—We show example logo images of different similarity levels compared with the original logos. All of them are below the detection threshold of 0.87.

Some examples of 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.

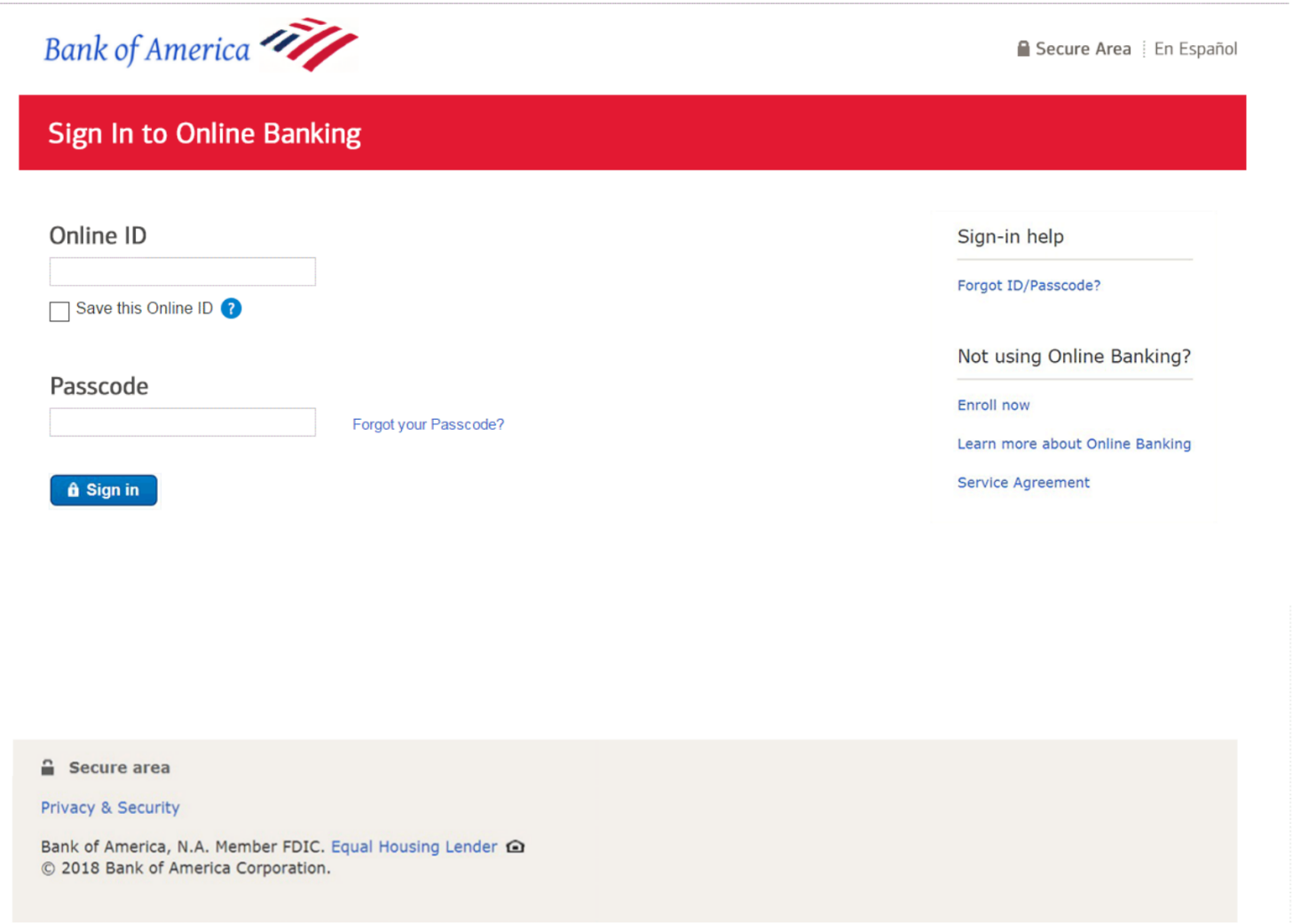


Figure 2: **Adversarial Phishing Webpage**—By using an adversarial logo crafted with `LogoMorph`, this phishing webpage bypasses detectors such as PhishIntention [32] and Phishpedia [30].

[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).

Effectiveness of LogoMorph – empirical results

- For most of the brands we considered, we crafted "adversarial logos" that, when put onto a webpage, would induce the entire system to believe the page to be benign.

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.

Effectiveness of LogoMorph – empirical 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]).

[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).

Effectiveness of LogoMorph – empirical 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).

Effectiveness of LogoMorph – empirical 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).

…what about humans?

(Phishing 101)

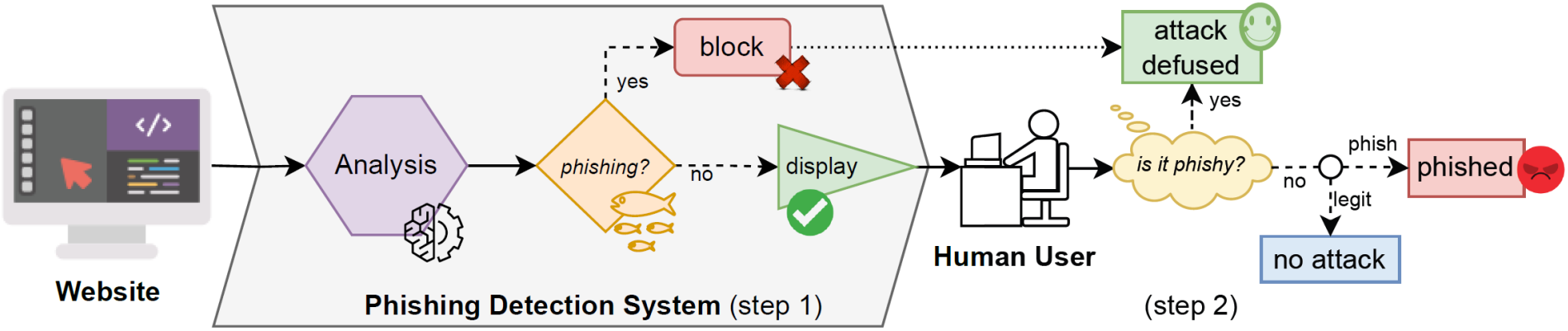


Fig. 1: Scenario: phishing detection is a two-step decision process.

Gap: Technical papers…

Typical workflow of an “adversarial machine learning” paper:

1. Propose an attack

Gap: Technical papers…

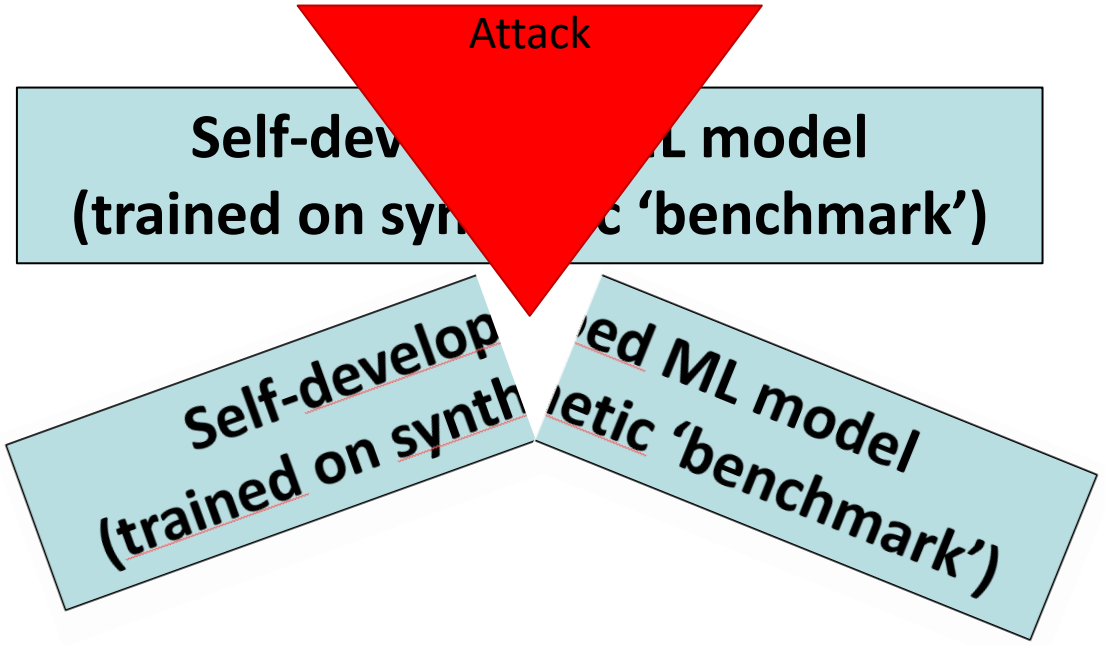
Typical workflow of an "adversarial machine learning" paper:

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

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



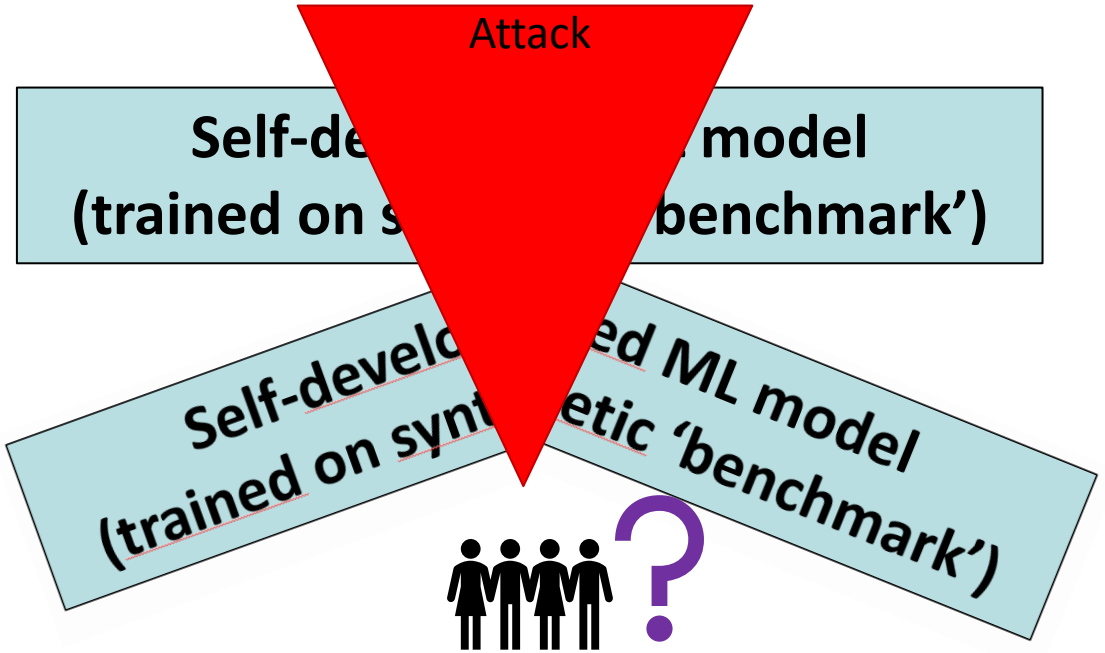
Gap: Technical papers…

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 humans (are they tricked *too*)?

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



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

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 the phishing detectors?

- Maybe the samples would be trivially blocked by the detector

What should be done:
Combine “technical” work with “user studies”

What should be done:
Combine “technical” work with “user studies”

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

How did we do it?

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 we do it?

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.

How did we do it?

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

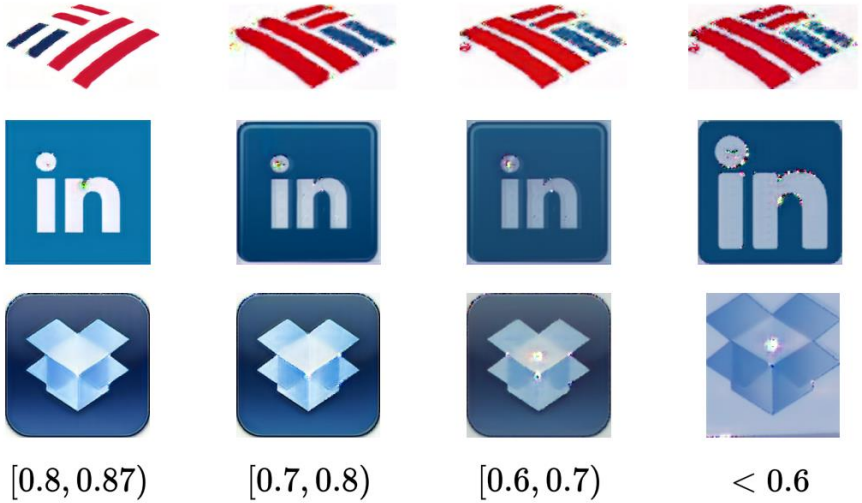


Figure 5: **Image Logo Attack Examples**—We show example logo images of different similarity levels compared with the original logos. All of them are below the detection threshold of 0.87.

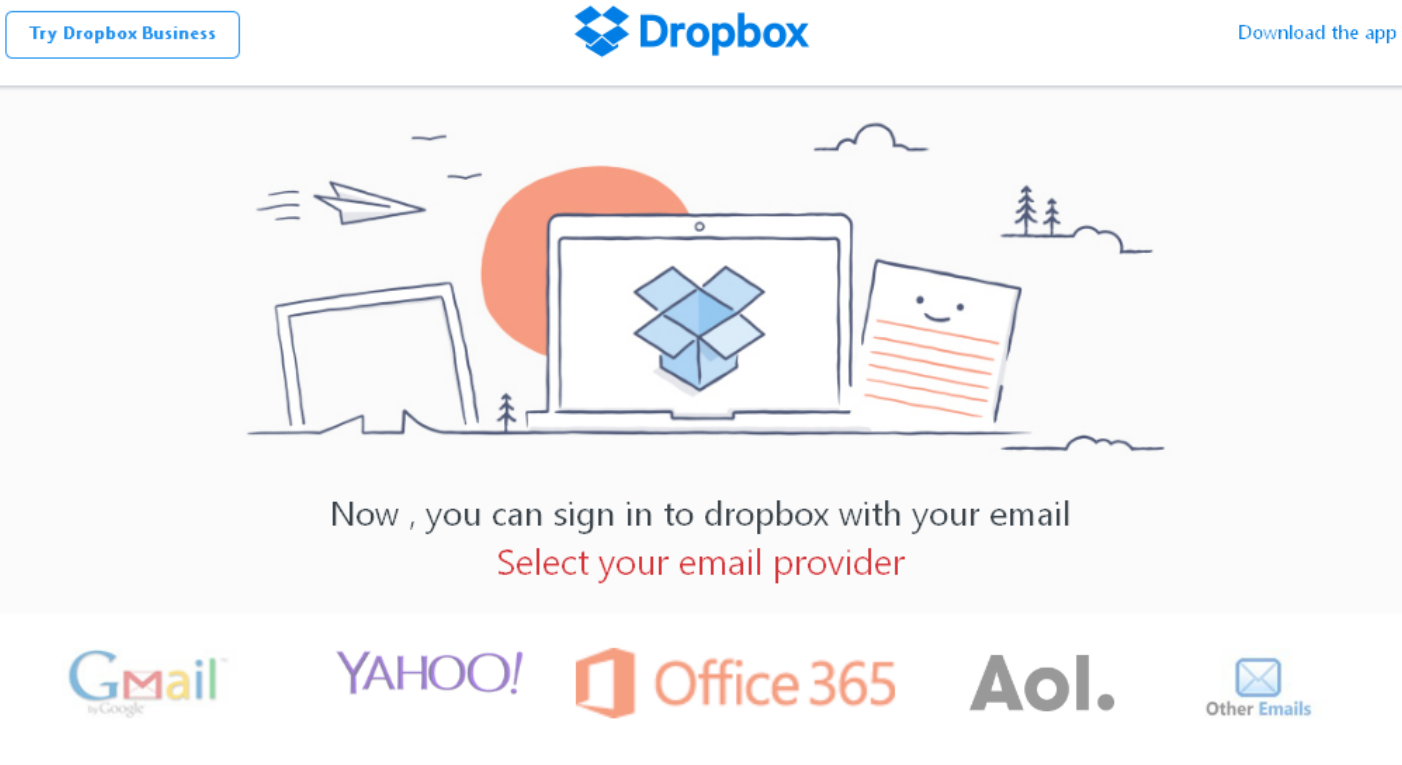
How did we do it?

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 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.

How did we do it?



How did we do it?

What did we find?

- 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).

What did we find?

- …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.

Explainable Machine Learning

UNIVERSITÄT
LIECHTENSTEIN

‘Explainable’ Machine Learning

- ML methods are typically considered as *black boxes*
  - They are so complex, that it is hard for a human to explain what led the ML model to produce any given output

‘Explainable’ Machine Learning

- ML methods are typically considered as *black boxes*
  - They are so complex, that it is hard for a human to explain what led the ML model to produce any given output
- This is a problem---especially in a cybersecurity context---because humans cannot make critical decisions without having a complete understanding of the situation

- The domain of ‘explainable machine learning’ seeks to provide a solution to this issue

'Explainable' Machine Learning

- ML methods are typically considered as *black boxes*
  - They are so complex, that it is hard for a human to explain what led the ML model to produce any given output
- This is a problem---especially in a cybersecurity context---because humans cannot make critical decisions without having a complete understanding of the situation
- The domain of 'explainable machine learning' seeks to provide a solution to this issue
- The goal is to provide an 'human-understandable *explanation'* that can assist the human analyst in understanding the output of an ML model to any given input
  - E.g., in the context of PWD, explanation methods can be used to say 'The model deemed this sample as being a phishing one because it has these properties'
- In this way, humans can make more informed decisions.
  - It can also be used to troubleshoot a given ML model: if the explanation does not make sense, then it may be a sign that the ML model is not very good

LIME

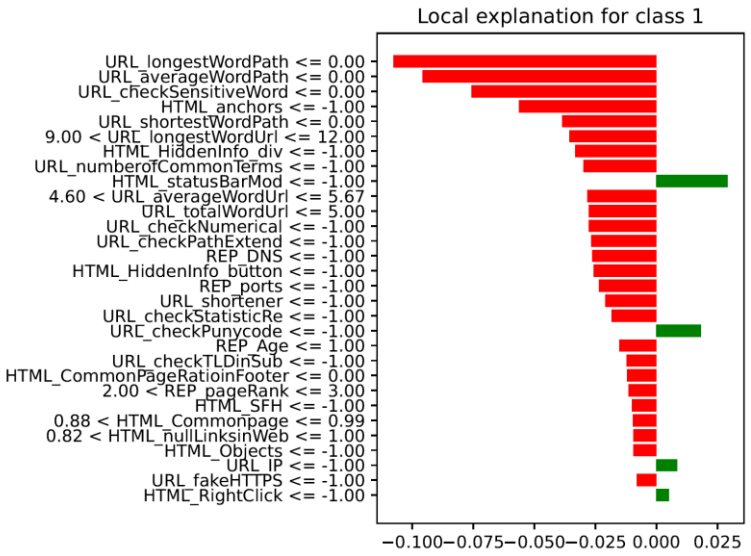
- One of the most well-known XAI techniques is *LIME* (short for 'Local Interpretable Model-agnostic Explanations') [Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.]
  - The way LIME works is very complex (and will not be covered in this lesson/course).

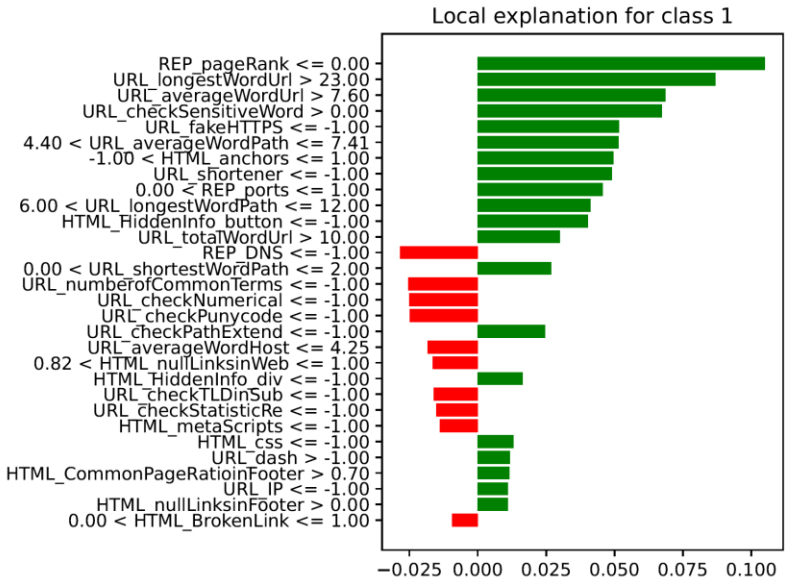
LIME

- One of the most well-known XAI techniques is *LIME* (short for 'Local Interpretable Model-agnostic Explanations') [Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.]
  - The way LIME works is very complex (and will not be covered in this lesson/course).
- At a high-level, LIME seeks to answer the question 'what makes *this sample* being of *class X*'?
  - The 'what makes' is related to, e.g., the 'features' that describe a given sample

LIME

- One of the most well-known XAI techniques is *LIME* (short for 'Local Interpretable Model-agnostic Explanations') [Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.]
  - The way LIME works is very complex (and will not be covered in this lesson/course).
- At a high-level, LIME seeks to answer the question 'what makes *this sample* being of *class X*'?
  - The 'what makes' is related to, e.g., the 'features' that describe a given sample
- For instance, in a PWD context, the output of LIME to two samples (benign, on the left; and phishing, on the right) could be the following:





‘Explainable’ Machine Learning -- problems

- Unfortunately, all that glitters is not gold!

‘Explainable’ Machine Learning -- problems

- Unfortunately, all that glitters is not gold!

- First, explanation methods can be ‘attacked’ (see [Slack, Dylan, et al. "Fooling lime and shap: Adversarial attacks on post hoc explanation methods." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020.])
  - For instance, an attacker who is aware that a given defensive system uses a XAI method, may want to fool such a method (e.g., inducing the method to provide a ‘wrong’ explanation that does not match the input, so that the human user would be fooled into believing that, e.g., a given sample is benign instead of being malicious)
  - This becomes a problem if humans trust these XAI methods too much

'Explainable' Machine Learning -- problems

- Unfortunately, all that glitters is not gold!

- First, explanation methods can be 'attacked' (see [Slack, Dylan, et al. "Fooling lime and shap: Adversarial attacks on post hoc explanation methods." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020.])
  - For instance, an attacker who is aware that a given defensive system uses a XAI method, may want to fool such a method (e.g., inducing the method to provide a 'wrong' explanation that does not match the input, so that the human user would be fooled into believing that, e.g., a given sample is benign instead of being malicious)
  - This becomes a problem if humans trust these XAI methods too much

- Second, there is (perhaps surprisingly) little evidence on how good these XAI methods actually are *for humans* (see [Nadeem, Azqa, et al. "Sok: Explainable machine learning for computer security applications." *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2023.])
  - Most existing research on these XAI methods (especially in a security context) is done by measuring their performance via metrics that do not account for human's feedback
  - Therefore, it is unclear if these methods are actually useful (today)

Conclusions

Outline of the talk

- 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

Two goals:

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

Outline of the talk – 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

Two goals:

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

Outline of the talk – 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

Two goals:

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

Outline of the talk – 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

Two goals:

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

Outline of the talk – 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 this domain)
- Entertain you (research should be fun)