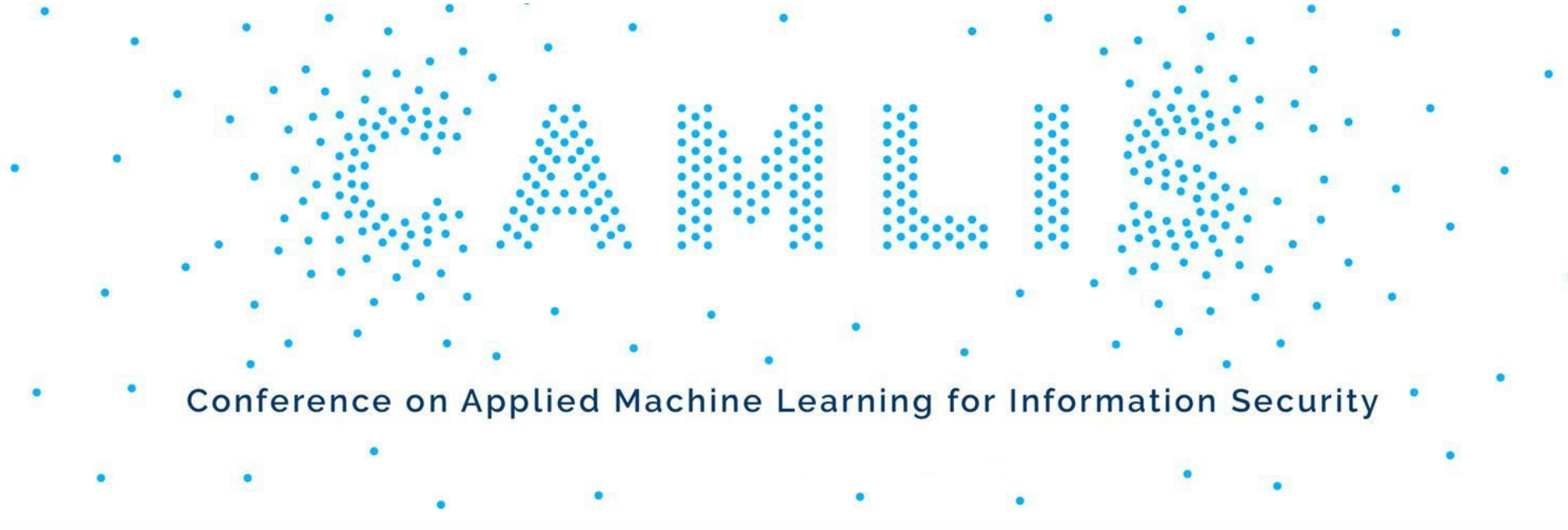


Machine Learning Security Evasion Competition (MLSEC) 2022

Anti-Phishing challenge solution

Biagio Montaruli – biagio.montaruli@sap.com
Solution: github.com/biagiom/mlsec_2022

Roadmap

- Bio
- Introduction
- Solution
- Closing remarks

Bio

- AI security researcher @ SAP Labs France
- PhD candidate @ EURECOM (France)
- Involved in the TESTABLE EU project

Introduction to the Anti-phishing challenge

GOAL: Manipulate 10 phishing web pages in order to evade 8 ML models.

- **Evade:** the returned score must be < 0.1.

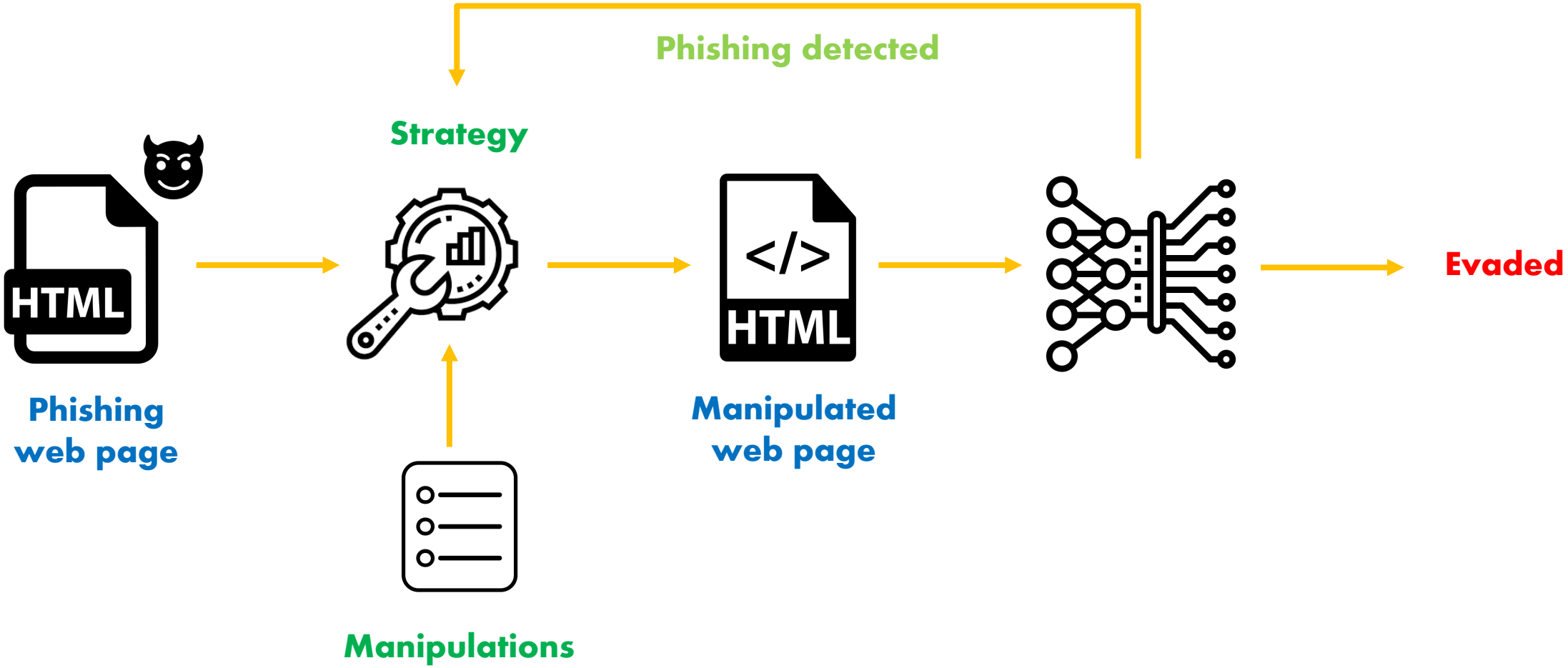
CONSTRAINT: the manipulated and original web pages must have the same rendering.

- The screenshots of the rendered manipulated and original web page must have the same SHA-256 hash.

BLACK-BOX SCENARIO: attack only requires querying the target models and observing their outputs.

- No need to access their internal parameters or knowing how the models work.

Solution: adversarial attacks against anti-phishing ML models



Manipulations – Injection of <input> tags

Injection of **<input>** tags embedded into **<noscript></noscript>**

- Different types: **text**, **submit**, **radio**, **search** and **button**

```
<noscript>
  <input id="continue" tabindex="5" class="a-button-input" type="submit" aria-labelledby="continue-announce"/>
  <input id="continue" tabindex="5" class="a-button-input" type="submit" aria-labelledby="continue-announce"/>
  <input id="continue" tabindex="5" class="a-button-input" type="submit" aria-labelledby="continue-announce"/>
  <input id="continue" tabindex="5" class="a-button-input" type="submit" aria-labelledby="continue-announce"/>
  <input id="continue" tabindex="5" class="a-button-input" type="submit" aria-labelledby="continue-announce"/>
</noscript>
```

```
<noscript>
  <input id="sayhello" tabindex="7" class="a-search-input" type="search" aria-labelledby="say-hello"/>
  <input id="sayhello" tabindex="7" class="a-search-input" type="search" aria-labelledby="say-hello"/>
  <input id="sayhello" tabindex="7" class="a-search-input" type="search" aria-labelledby="say-hello"/>
  <input id="sayhello" tabindex="7" class="a-search-input" type="search" aria-labelledby="say-hello"/>
  <input id="sayhello" tabindex="7" class="a-search-input" type="search" aria-labelledby="say-hello"/>
  <input id="sayhello" tabindex="7" class="a-search-input" type="search" aria-labelledby="say-hello"/>
  <input id="sayhello" tabindex="7" class="a-search-input" type="search" aria-labelledby="say-hello"/>
  <input id="sayhello" tabindex="7" class="a-search-input" type="search" aria-labelledby="say-hello"/>
</noscript>
```

Manipulations – Change type of `password` `<input>` tags

Change the type of **`password` `<input>`** tags to **`text`**

- The presence of **`password` `<input>`** tag is a relevant feature for the ML models since it is very common in phishing web pages
- Changing it to **`text`** allows to reduce the confidence score and at the same time achieving the same rendering

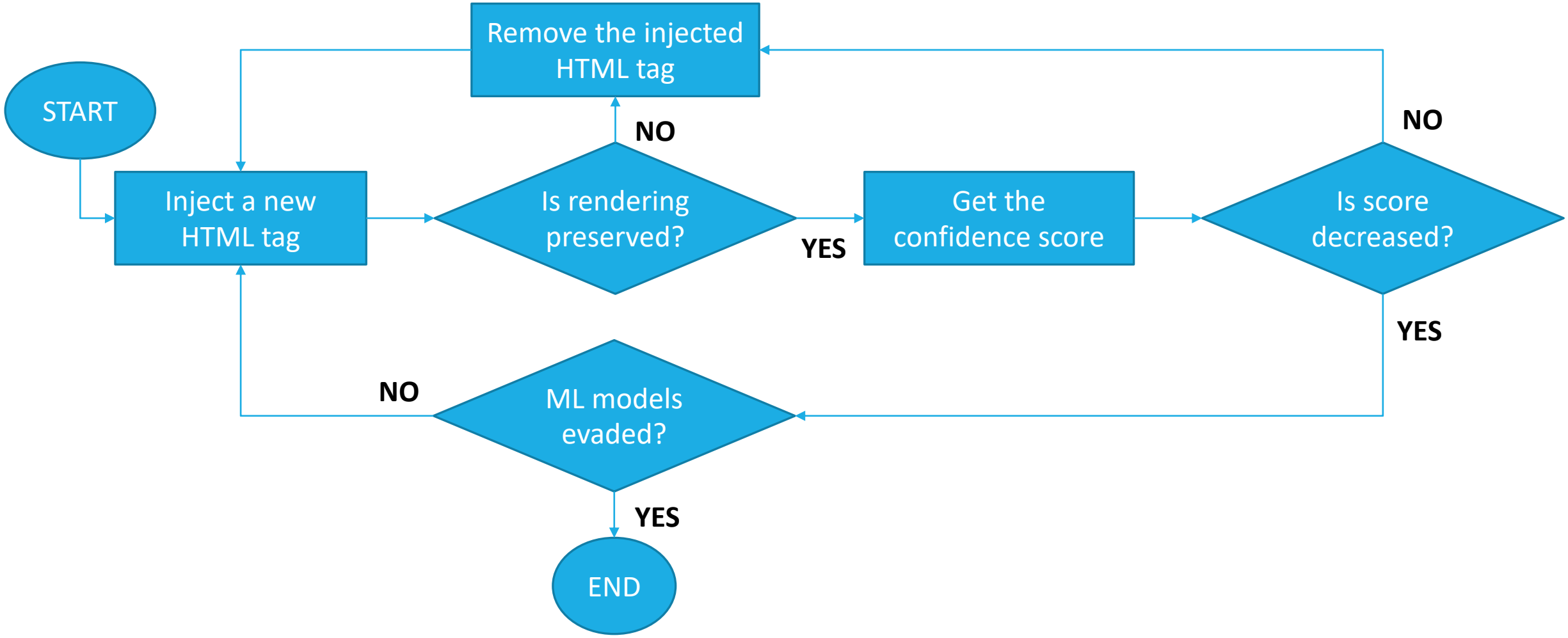
```
<div class="a-row a-spacing-base">
  Email or mobile phone number
  <input type="email" maxlength="128" id="ap_email" name="email" tabindex="1" class="a-input-text a-span12 auth-autofocus
  Password
  <input type="password" maxlength="1024" id="ap-credential-autofill-hint" name="password" class="a-input-text hide"/>
  <input type="text" maxlength="1024" id="ap-credential-autofill-hint" name="password" class="a-input-text hide"/>
```

Manipulations – Add unused JS code

Add random JS code nested in `<noscript></noscript>`

```
<noscript>
    <script type='text/javascript'>var ue_t0=ue_t0||+new Date();</script>
    <!-- sadfnjdsa374fdsdsagds6tewddstw2w7fhagsafd -->
    <script>var aPageStart = (new Date()).getTime();</script>
    <script type='text/javascript'>var ue_t0=ue_t0||+new Date();</script>
    <!-- sadfnjdsa374fdsdsagds6tewddstw2w7fhagsafd -->
    <script>var aPageStart = (new Date()).getTime();</script>
    <script type='text/javascript'>var ue_t0=ue_t0||+new Date();</script>
    <!-- sadfnjdsa374fdsdsagds6tewddstw2w7fhagsafd -->
    <script>var aPageStart = (new Date()).getTime();</script>
    <script type='text/javascript'>var ue_t0=ue_t0||+new Date();</script>
    <!-- sadfnjdsa374fdsdsagds6tewddstw2w7fhagsafd -->
    <script>var aPageStart = (new Date()).getTime();</script>
    <script type='text/javascript'>var ue_t0=ue_t0||+new Date();</script>
    <!-- sadfnjdsa374fdsdsagds6tewddstw2w7fhagsafd -->
    <script>var aPageStart = (new Date()).getTime();</script>
    <script type='text/javascript'>var ue_t0=ue_t0||+new Date();</script>
</noscript>
```

Strategy

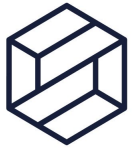

START
Inject a new HTML tag
Is rendering preserved?
NO
Remove the injected HTML tag
YES
Get the confidence score
Is score decreased?
NO
YES
ML models evaded?
NO
YES
END

Closing remarks

- Wrap-up: generation of adversarial exaples against ML models for classification of phishing web pages

- Special thanks to all the organizers:
  - Dr. Hyrum Anderson (Robust Intelligence)
  - Zoltan Balazs (CUJO AI),
  - Eugene Neelou (Adversa AI)

- Congratulations to all the participants!



ROBUST INTELLIGENCE

CUJOAI

ADVERSA