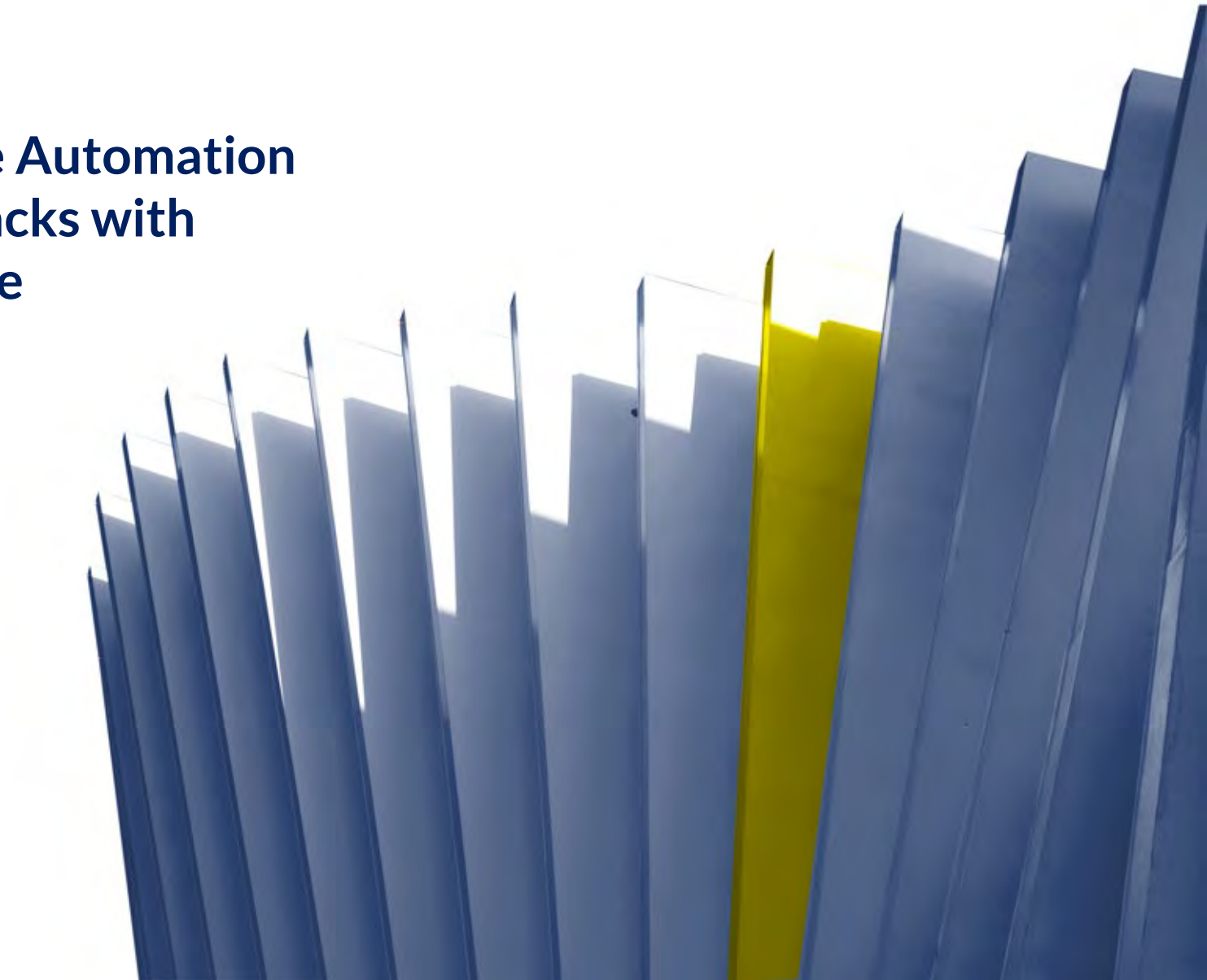
DeepSec 2022, Wien, 17/11/2022

DeepSec 2022 Talk: Towards the Automation of Highly Targeted Phishing Attacks with Adversarial Artificial Intelligence

Francesco Morano and Enrico Frumento

V2



Cefriel
POLITECNICO DI MILANO

Social Engineering never disappears

Emotet is again active in Italy (01/11/2022)

A new campaign with Italian targets aimed at conveying via email a password-protected ZIP attachment containing an XLS equipped with malicious macros.

To get infected, a user needs: to open the email, open the attached zip, enter the password, open the excel file inside the zip, enable Office macros, ignore all the warnings that Office shows and forget that you have done all this.

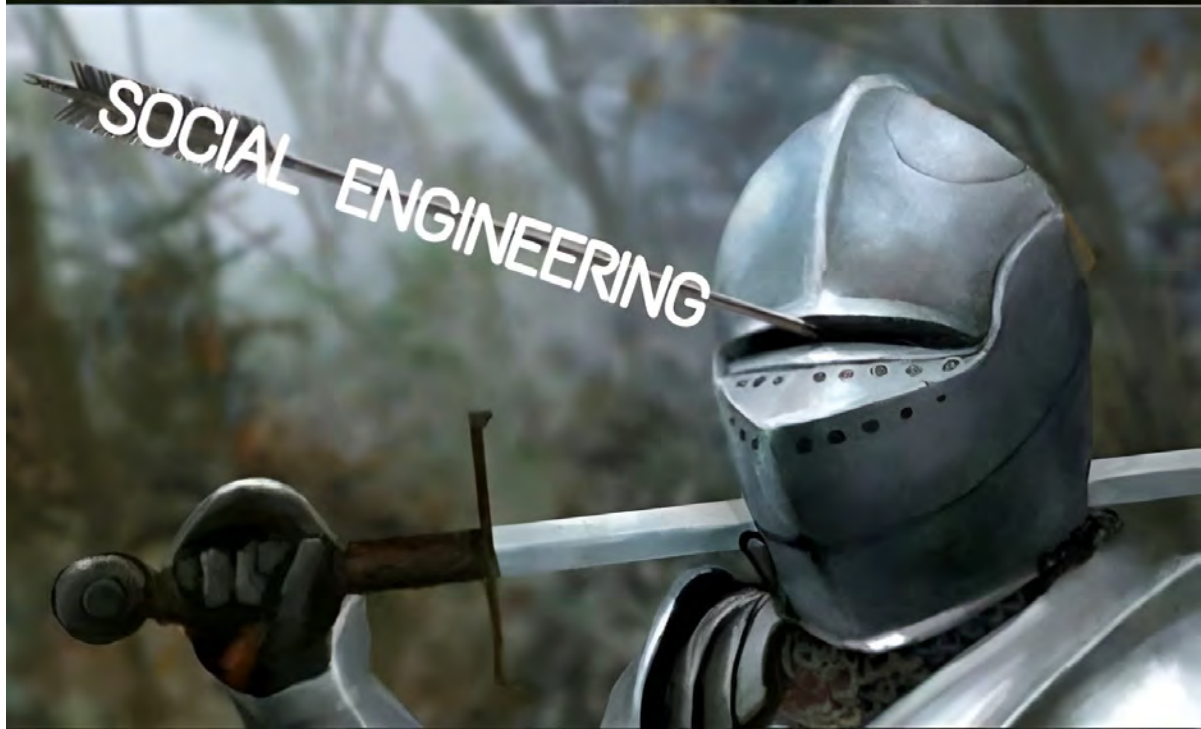
Only then the macro executes, which downloads and then executes the malware.

Social Engineering never disappears

REALLY basic Social Engineering works exceptionally well, so hackers do not need to improve sophistication.

New AI-enabled anti-deception detection systems.

Hence it's the right timing to explore what could come after…



When has Social Engineering begun?

HUMAN

+

General Model of Social Engineering attacks

MALWARE ECOSYSTEM

MODERN OSINT

(AB)USE OF PSYCHOLOGY AND COGNITIVE SCIENCE

EVOLUTION OF THE ATTACK VECTORS

AUTOMATIC SOCIAL ENGINEERING ATTACKS (ASE)

ECONOMIC DRIVERS

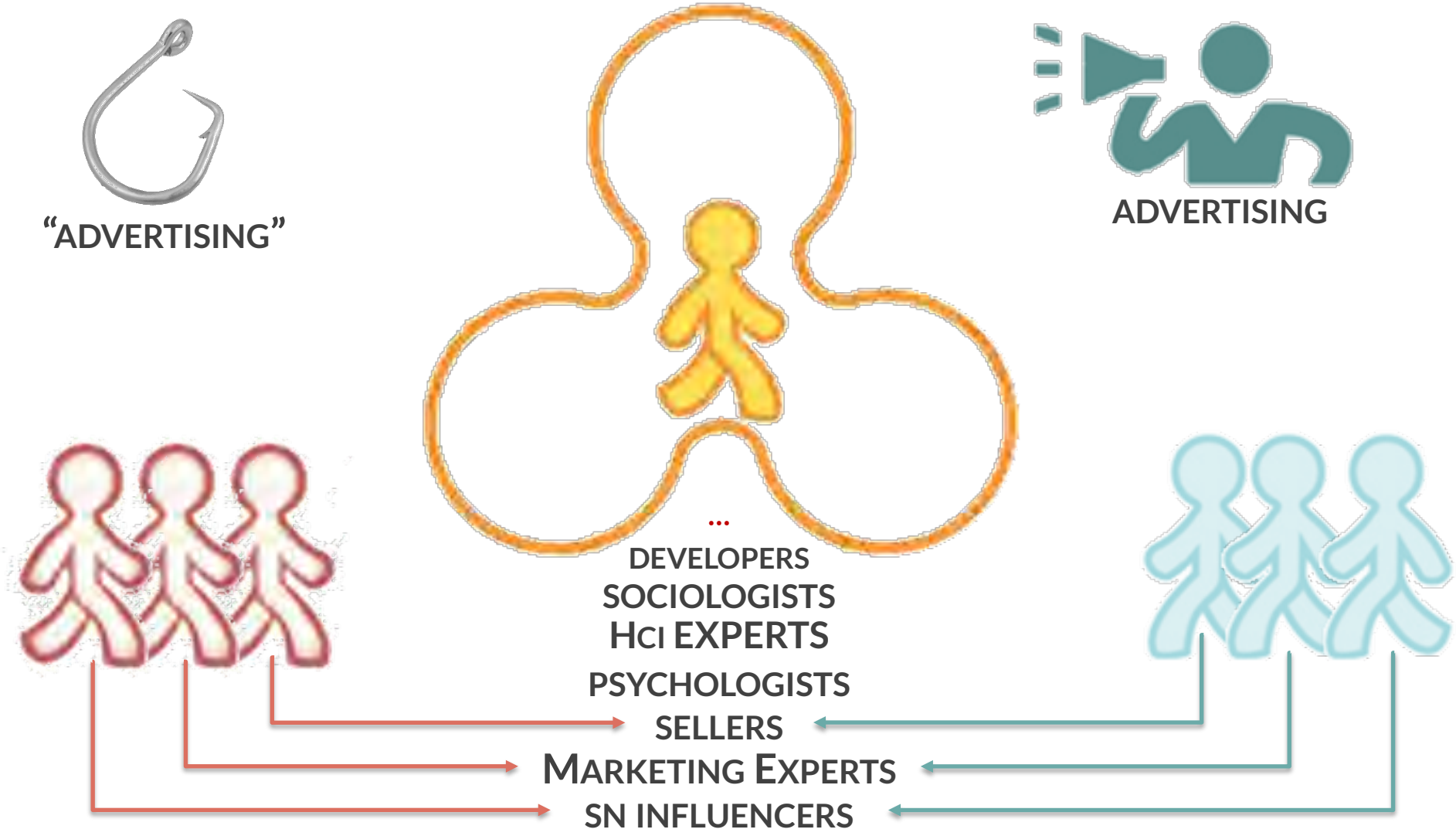
Cybersecurity of the human element

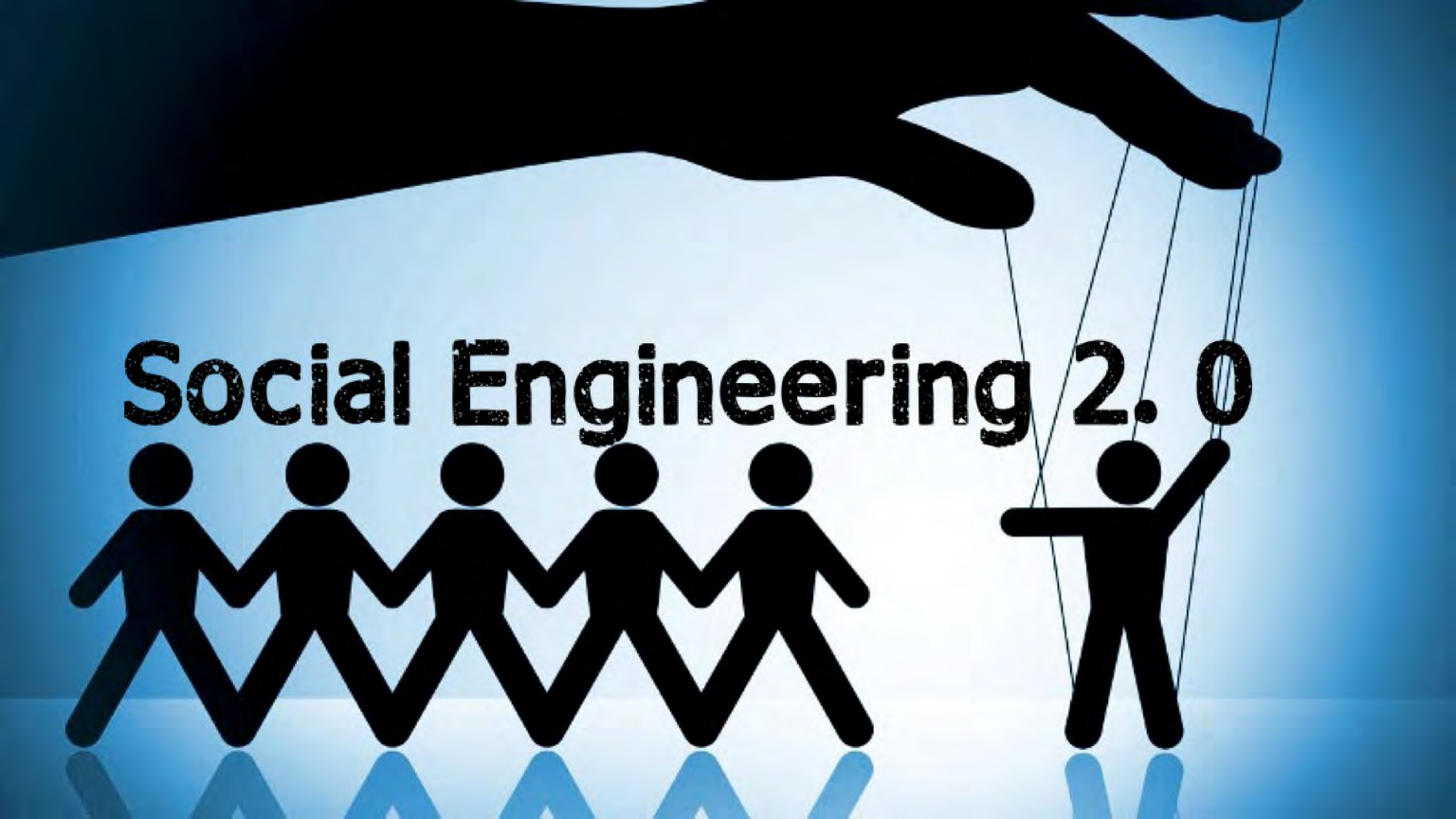
WEF/IBM (2022), **95%** of cybersecurity breaches result from human errors.

Much of the cybersecurity market instead concentrates on the technical side of an attack (IT or OT).

Organisations spend less than **5%** of their IT security budget to combat **95%** of risks (e.g., social engineering).

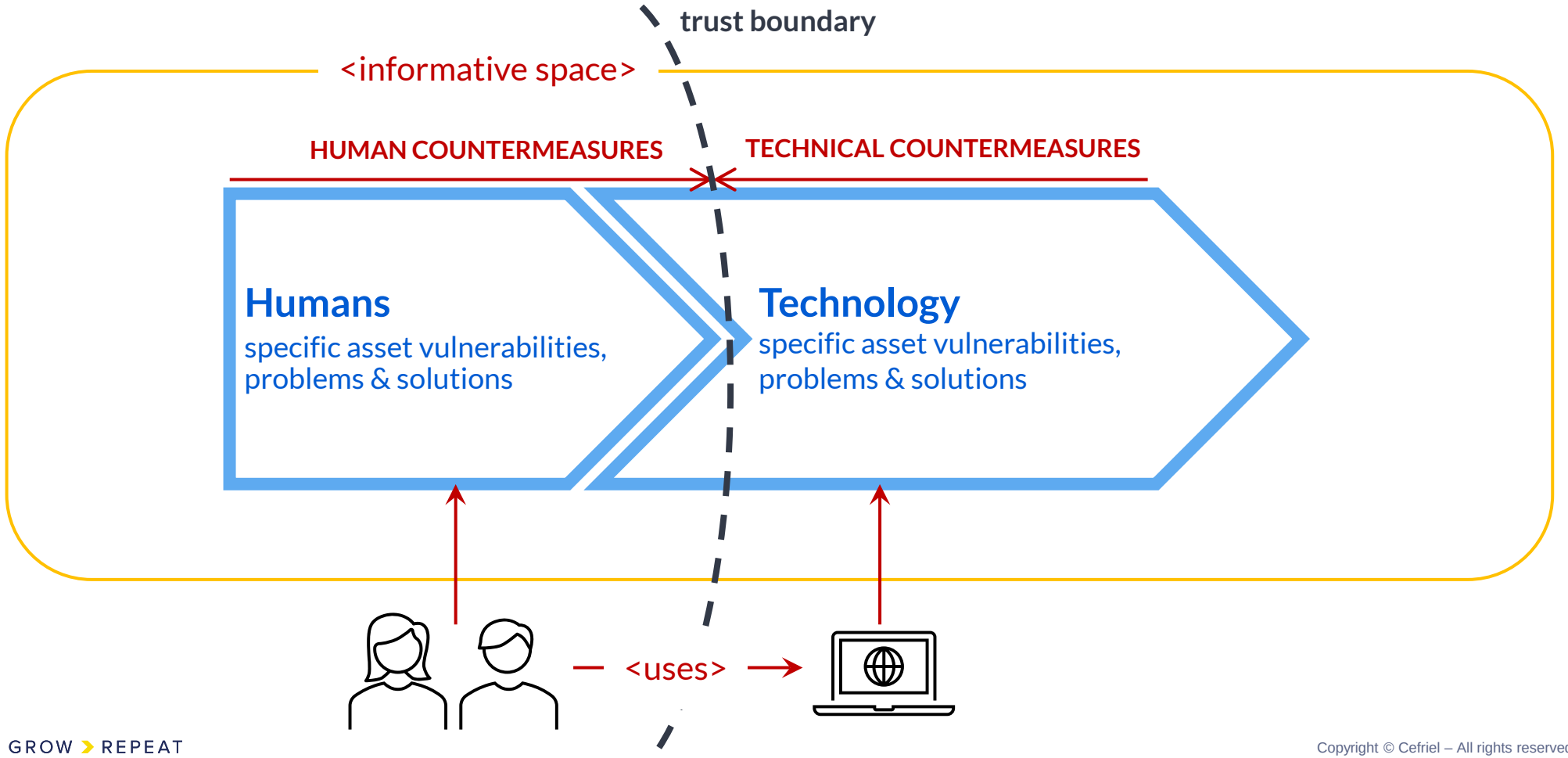
The composition of cybercrime teams is evolving




Social Engineering 2. 0

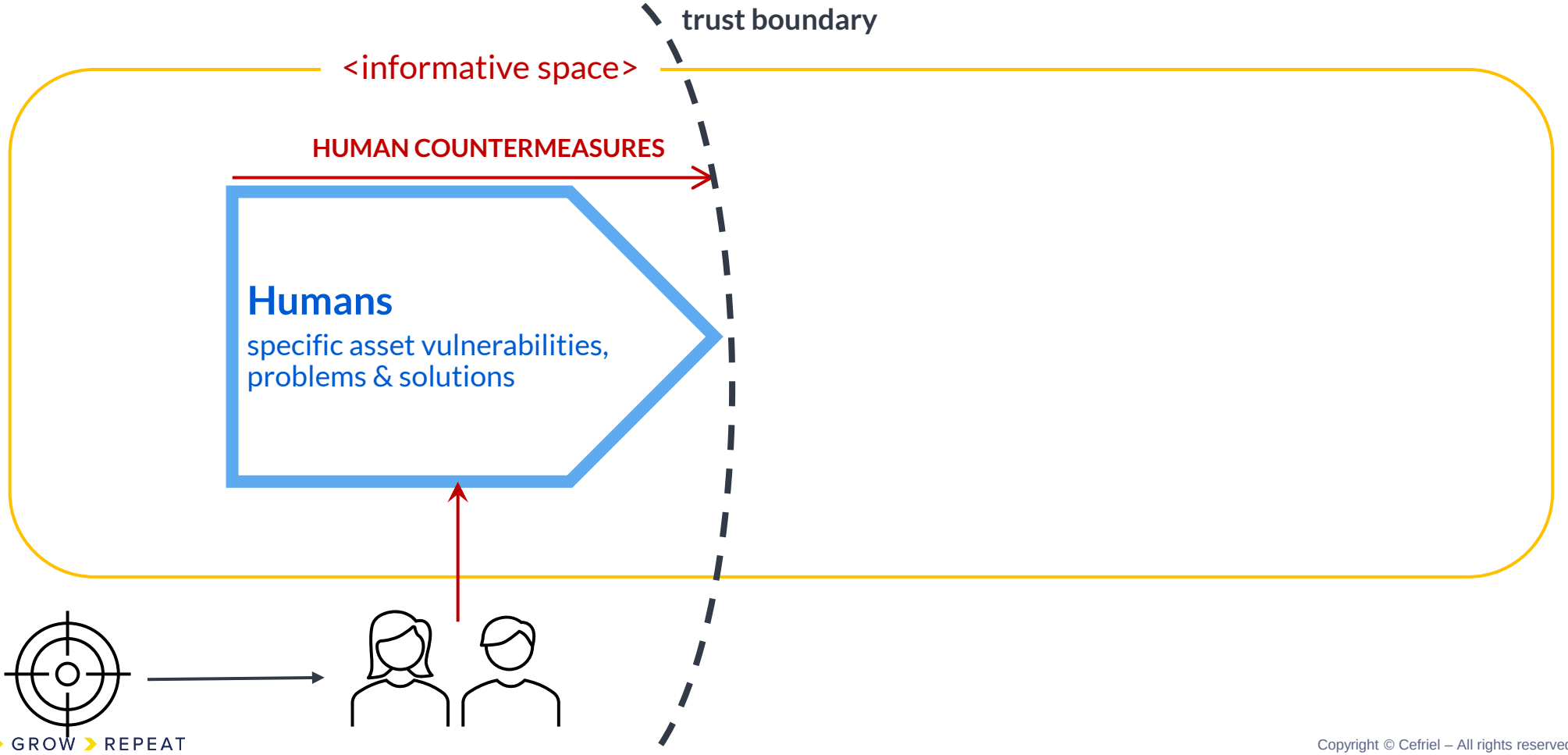
A conceptual schema of the threat

- An Asset ∈ Informative Space
- The Informative Space must be protected



A conceptual schema of the threat

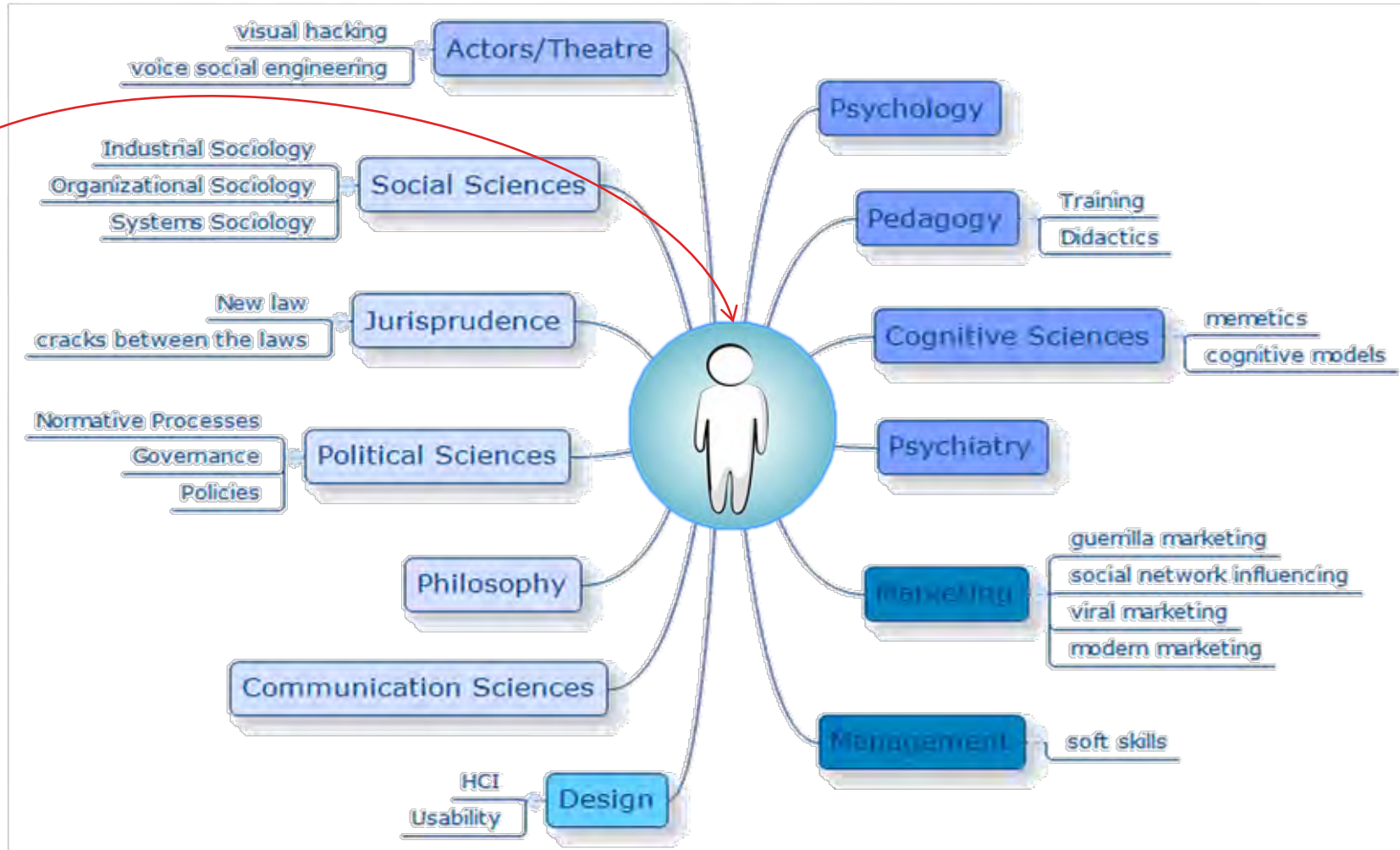
- An Asset ∈ Informative Space
- The Informative Space must be protected



The real nature of the problem

- The Human **IS** the "system" under attack
- *Which sciences contribute to modelling the attacked target?*

- A model of the attacked target defines the vulnerabilities which can be exploited through a threat
- It's a multidisciplinary problem by definition!



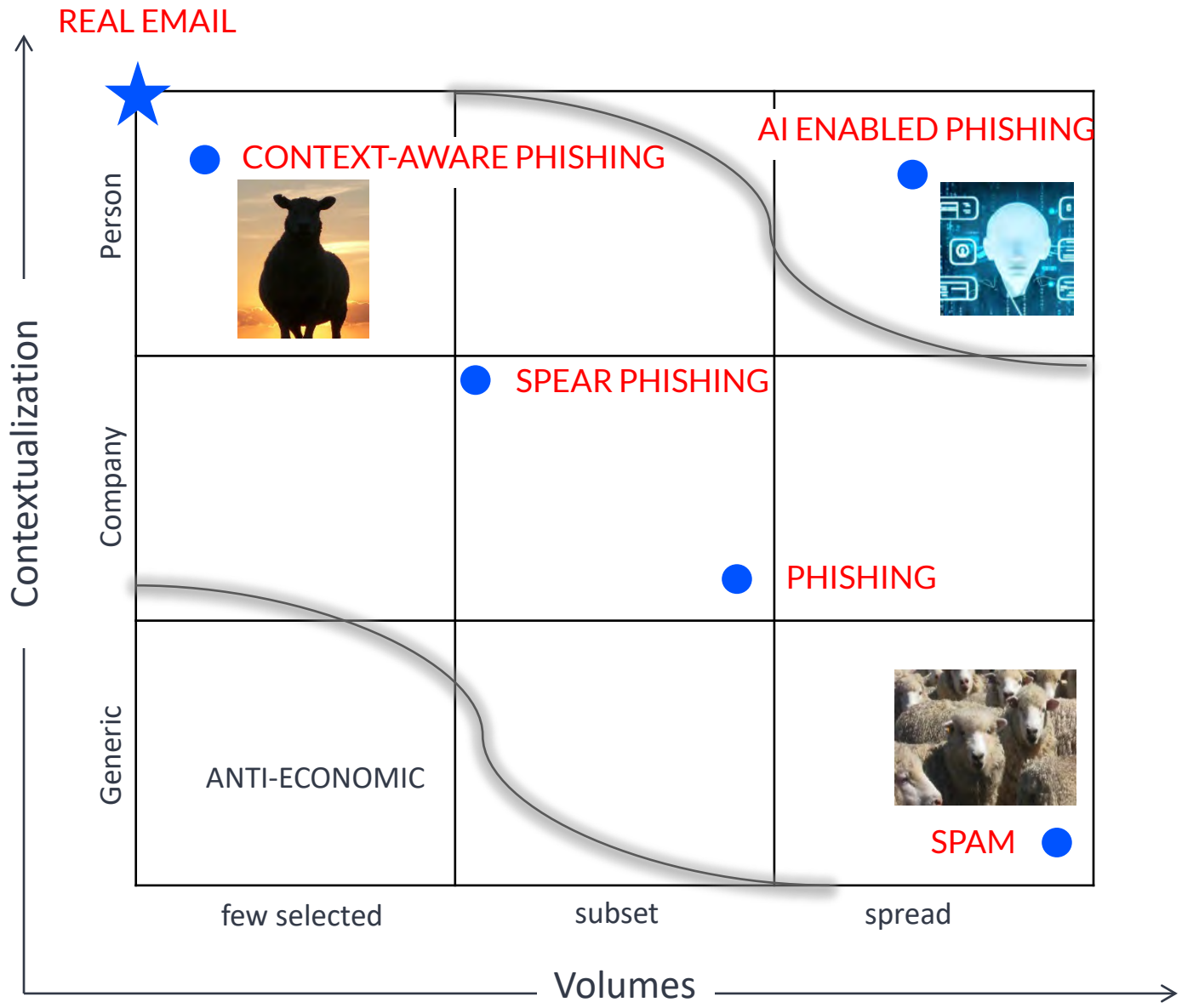
Social Engineering is the best in class

- Most remunerative TTP
- Easier than any other
- It’s enough to be smarter than your victims
- A lot of low-hanging fruits
- Humans cannot be permanently patched
- The same old tricks always work
- Very little code to write
- Very few working defences
- …

Mapping 4 levels

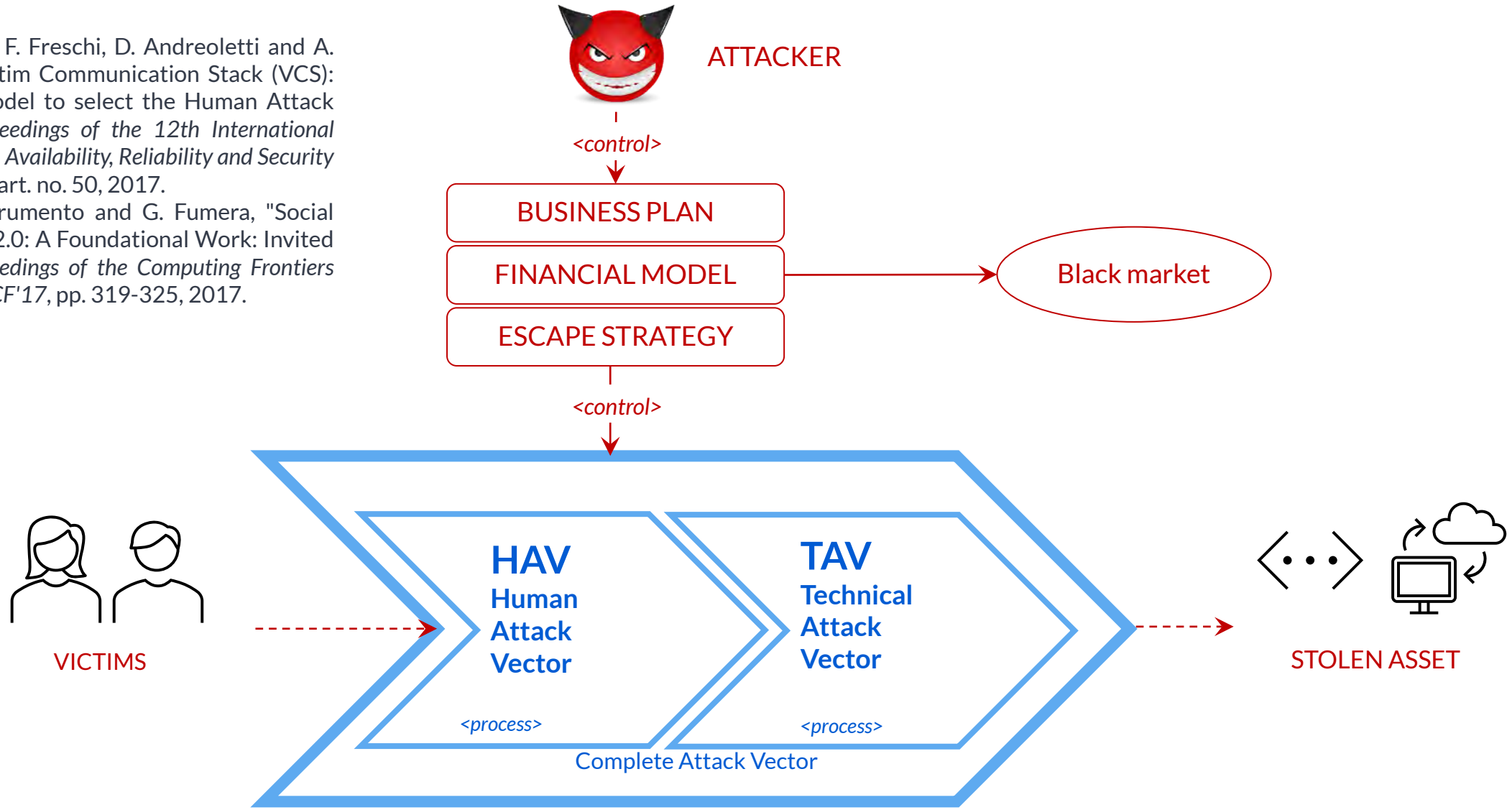
A model to categorise different kinds of phishing.

All types of phishing could be mapped on this plane.

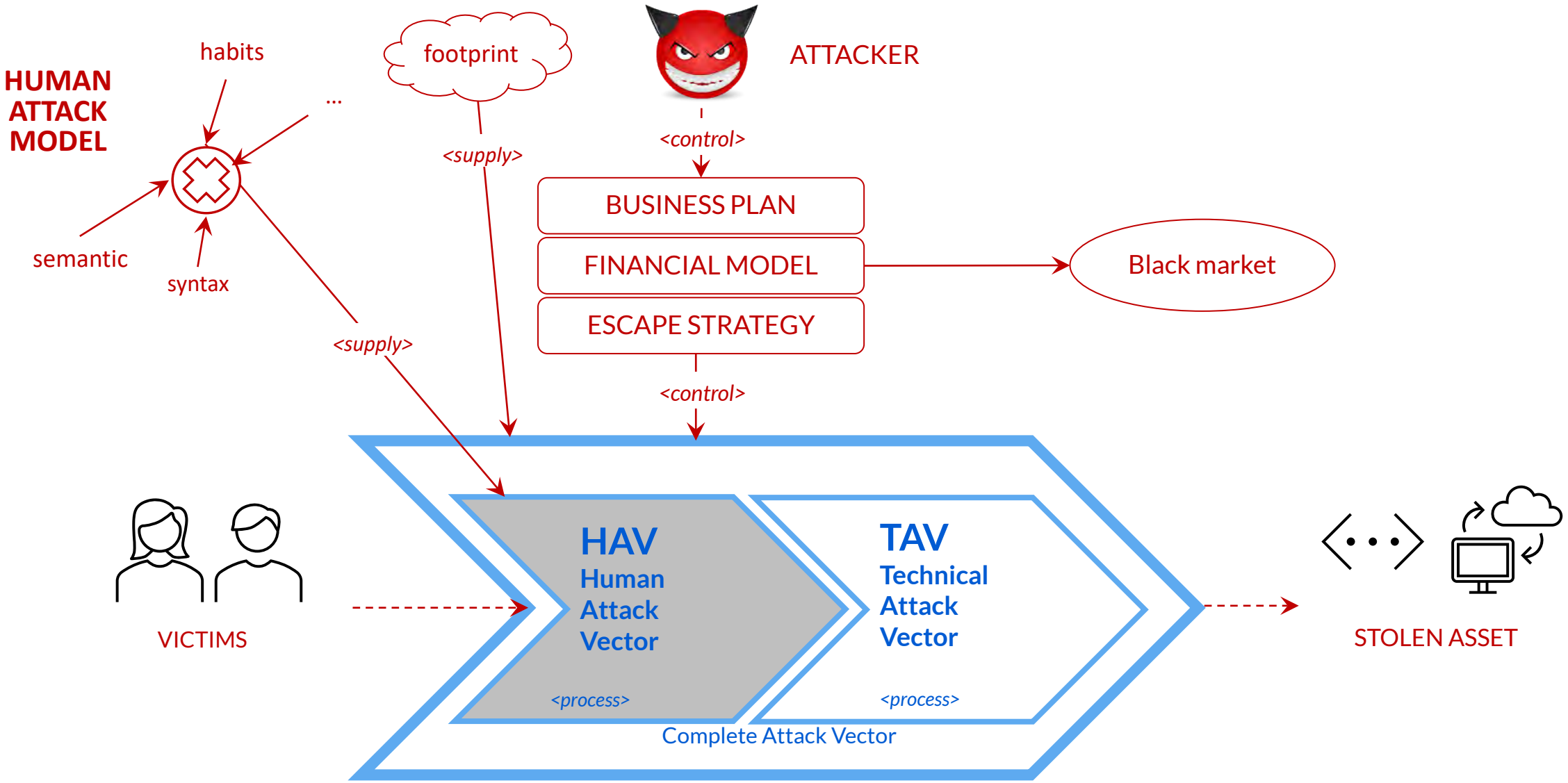


Phases of a Social Engineering Attack

- E. Frumento, F. Freschi, D. Andreoletti and A. Consoli, "Victim Communication Stack (VCS): A flexible model to select the Human Attack Vector", *Proceedings of the 12th International Conference on Availability, Reliability and Security - ARES '17*, p. art. no. 50, 2017.
- D. Ariu, E. Frumento and G. Fumera, "Social Engineering 2.0: A Foundational Work: Invited Paper", *Proceedings of the Computing Frontiers Conference - CF'17*, pp. 319-325, 2017.



Phases of a Social Engineering Attack



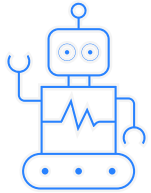
Attack based on personality

- Have a business plan
- Understand **who** is the handler of the asset to steal
- Understand "**how**" is the handler

- **If it is a human:**
  - Create a personalised Human Attack Vector
  - Create a Technical Attack Vector (subordinate)
- **If it is an IT system:**
  - Create a Technical Attack Vector (a different one)
  - Deliver the attack

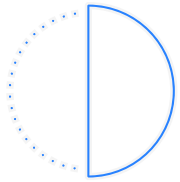
Use AI to automate the human-related threats

CHALLENGE



- **Automate SE attacks against humans and improve the reach of attacks.**
- **Improve Red Team efficiency**

SCOPE



AI can be used:

- **against humans**
- **to assist humans** (anti-deception detection systems)
- to **improve VA**

What exists

- **Knowbe4**: **AIDA** stands for Artificial Intelligence Driven Agent and uses artificial intelligence to dynamically create integrated campaigns that send emails, text and voicemail to an employee (private tool)

- **SNAP_R** for Twitter (abandoned SW)
  - automated spear phishing framework had between 30% and 66% success rate
  - the result is comparable to the 45% reported for largescale manual spear phishing efforts
  - computational and theoretical complexities are very low

- **BlackHat 21**: E. Lim et al., Turing in a Box: Applying Artificial Intelligence as a Service to Targeted Phishing.
  - They didn't cover OSINT+AI and did very few tests

Work of a professional designer without any former experience with spear phishing

WITHOUT AI TOOLS

- 1 hour to define the correct text
- 4 hours for a decent logo
- 2 hours for pagination and site creation
- **1 overall day for a similar result**

VS

WITH AI TOOLS

- **Total working time of 30 min for generating the email and the companion site!**

Work of a professional designer without any former experience with spear phishing

- With **Namelix**, we can generate a name of a realistic company (also used for the phishy URL) and with **LogoAI**, a good logo
- With **Stable Diffusion**, we created a nice set of images to be used in the phishing email
- With "**This person does not exist**", we created a credible profile for the email.
- With **Rytr**, which supports many languages and is based on GPT-3, we created the email text, starting from a simple seed, *"here is the invoice. Check it out,"* using different communication templates: business pitch, communication and email.
- Chosen a standard font
- Manually add a **Call-To-Action** button (register/discover more)
- Added **fingerprinting** and malware

Caffelab | Look who just opened! hello@caffelab.com

to me

Tue, 25 Oct, 22:55 (15 hours ago)

CAFFELAB

Hi Friend,

We're excited to share with you a new place in the neighborhood to get caffeinated and caffeinate your ideas. We want to make it easy for you to find a space for collaboration or just a quiet place to focus.

CaffeLab is a coffee shop, workspace and event venue all rolled into one!

You'll be able to work alone or with a group of people on laptops, enjoy a cup of our carefully roasted coffee, or take part in one of our workshops. We also have plenty of events coming up that are open to the public.

Check out our website for more details and discount!

Register

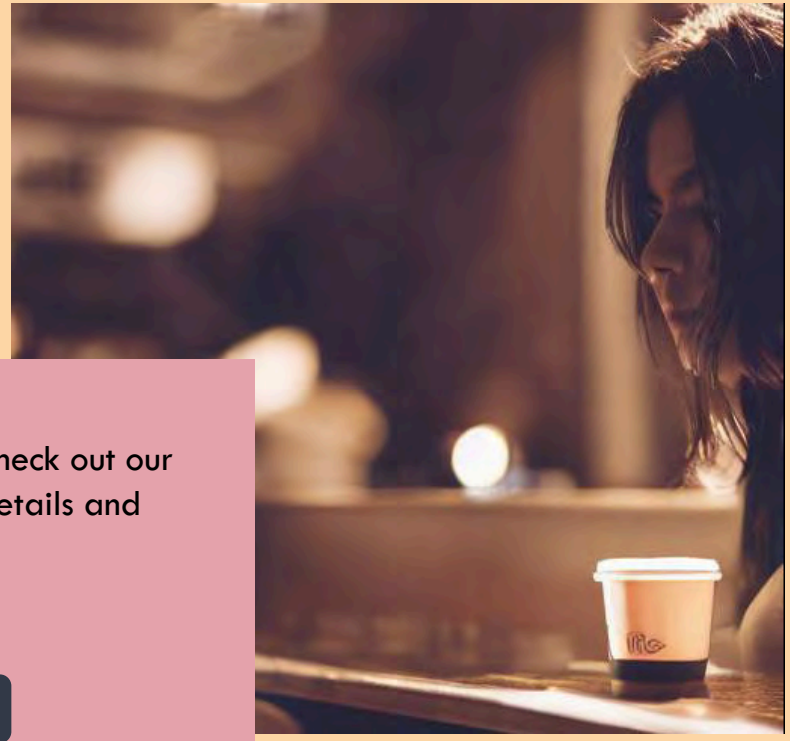
New Tab

Search Google or type a URL

CAFFeLAB

About Services Login **Register**

CaffeLab is a coffee shop, workspace and event venue all rolled into one!!



Register now and check out our website for more details and discount!

Register now

Tanger | Look who just opened! hello@tanger.com

to me

Tue, 25 Oct, 22:55 (15 hours ago)

Hi Jane,

We're so glad to have you with us! You'll be getting **20% off** your first order today.

Bistrot Gourmet is an online Korean restaurant that offers a diverse selection of Korean dishes at affordable prices.

From spicy and savory to sweet and tangy, our menu has something for everyone. Pay attention to our special deals, we update them regularly, so you never know what could be coming up next.

We hope you enjoy your dining experience with us!

View our menu

Cheers, Nina

New Tab

Search Google or type a URL

Tanger

About Menu Services **Order**

Welcome to Tanger!

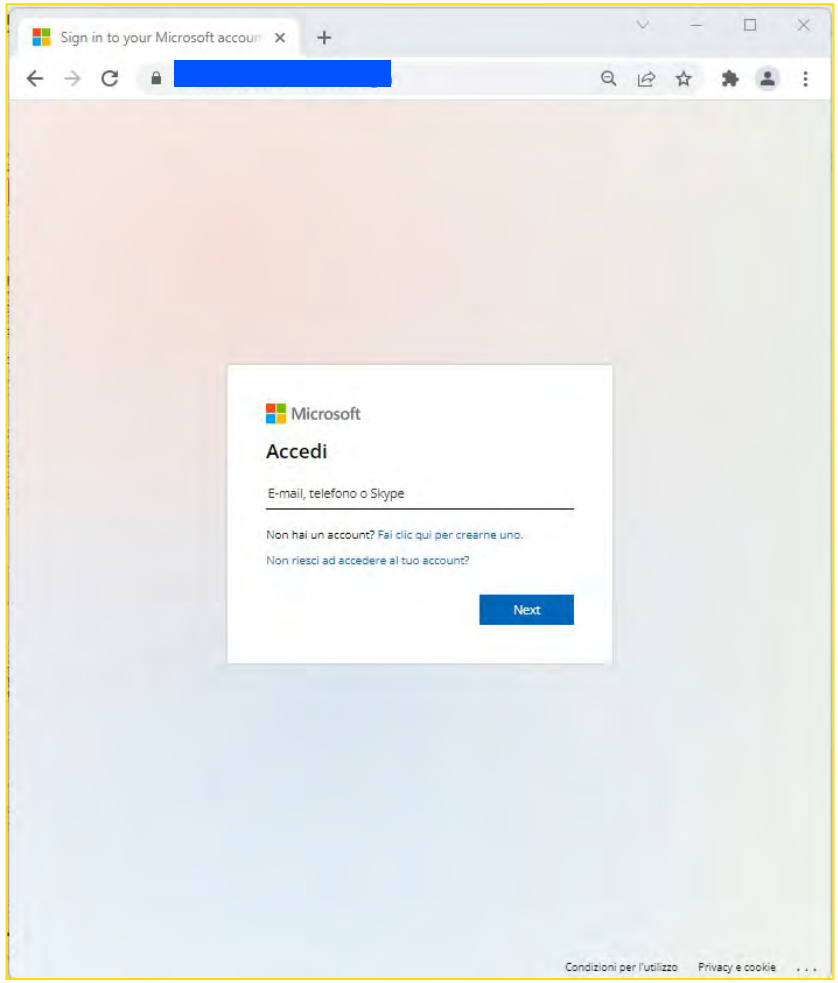
No matter what kind of cuisine you're in the mood for, our bistrot will satisfy your cravings.

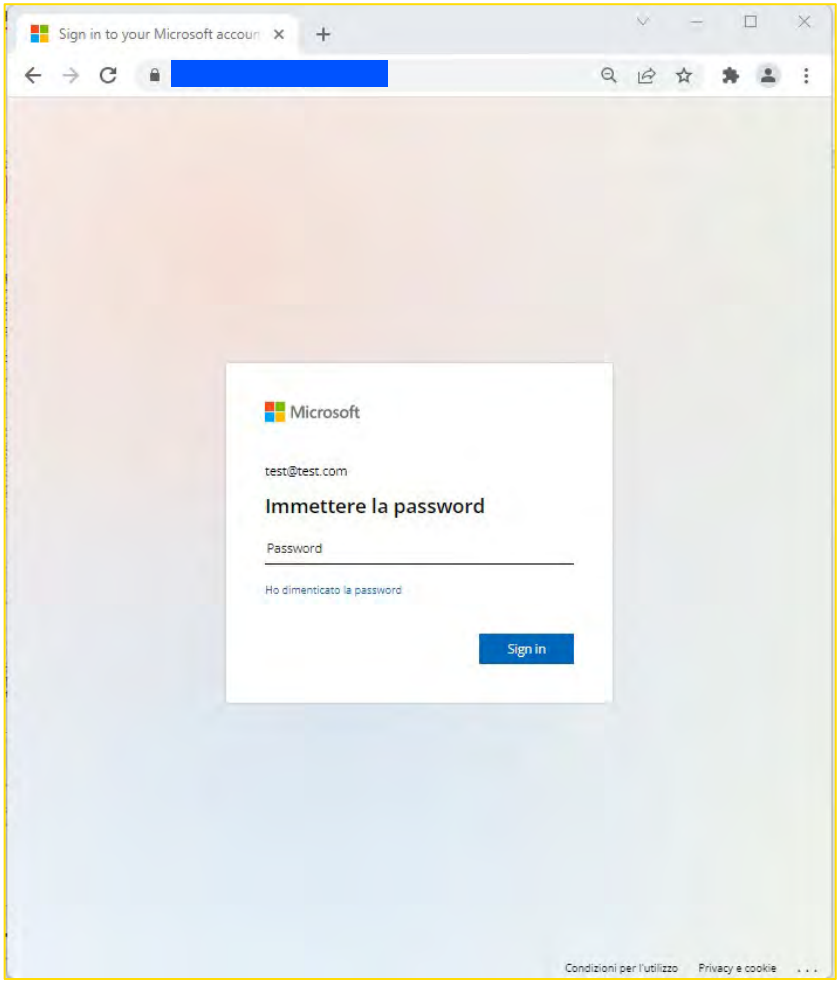
View our menu

Fake log in page

After the CTA, we used a clone of the Microsoft Office 365 login authentication portal.

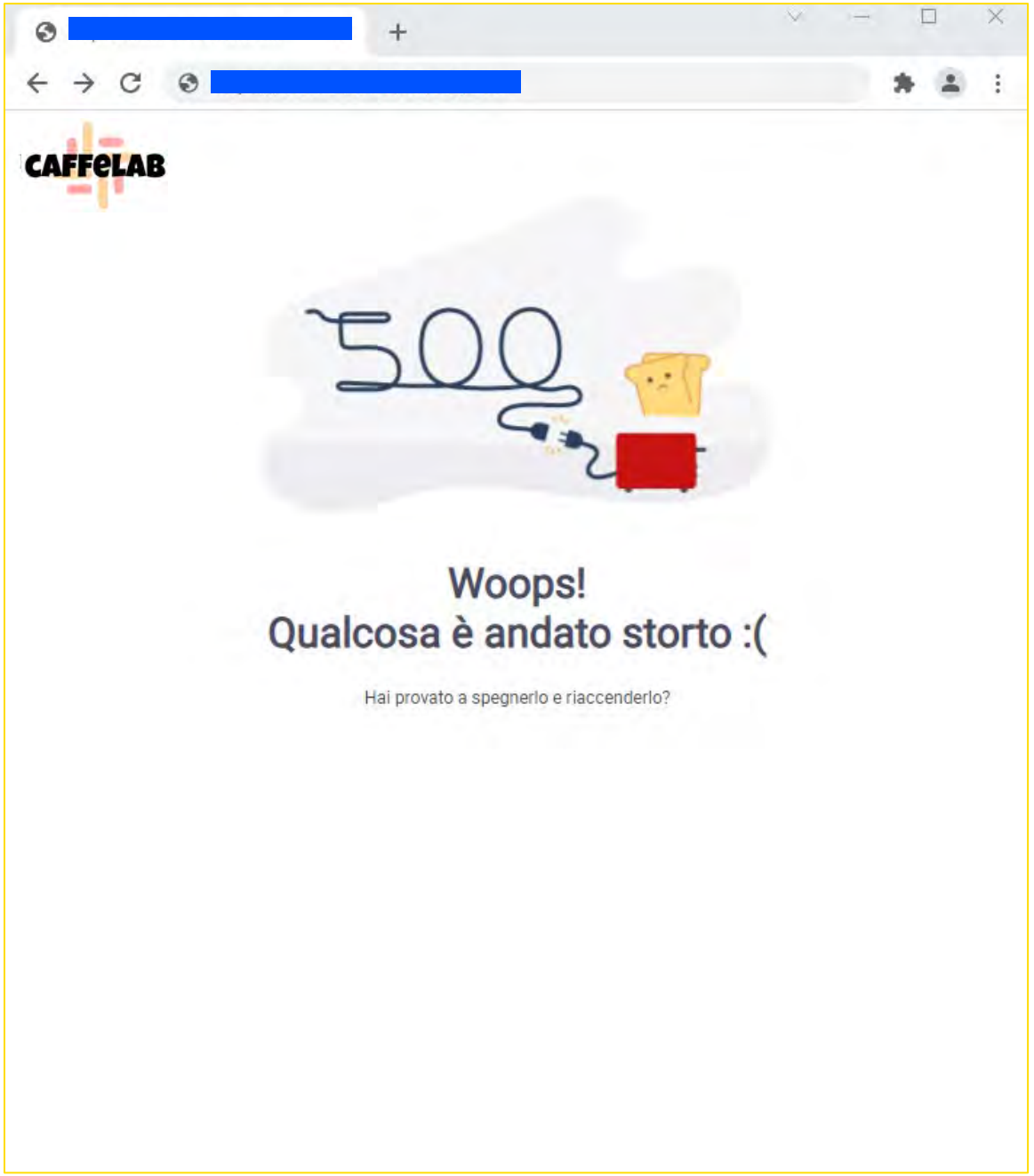
The victim enters their email address and password in two independent steps. In this way, you can study the behaviour of each field.





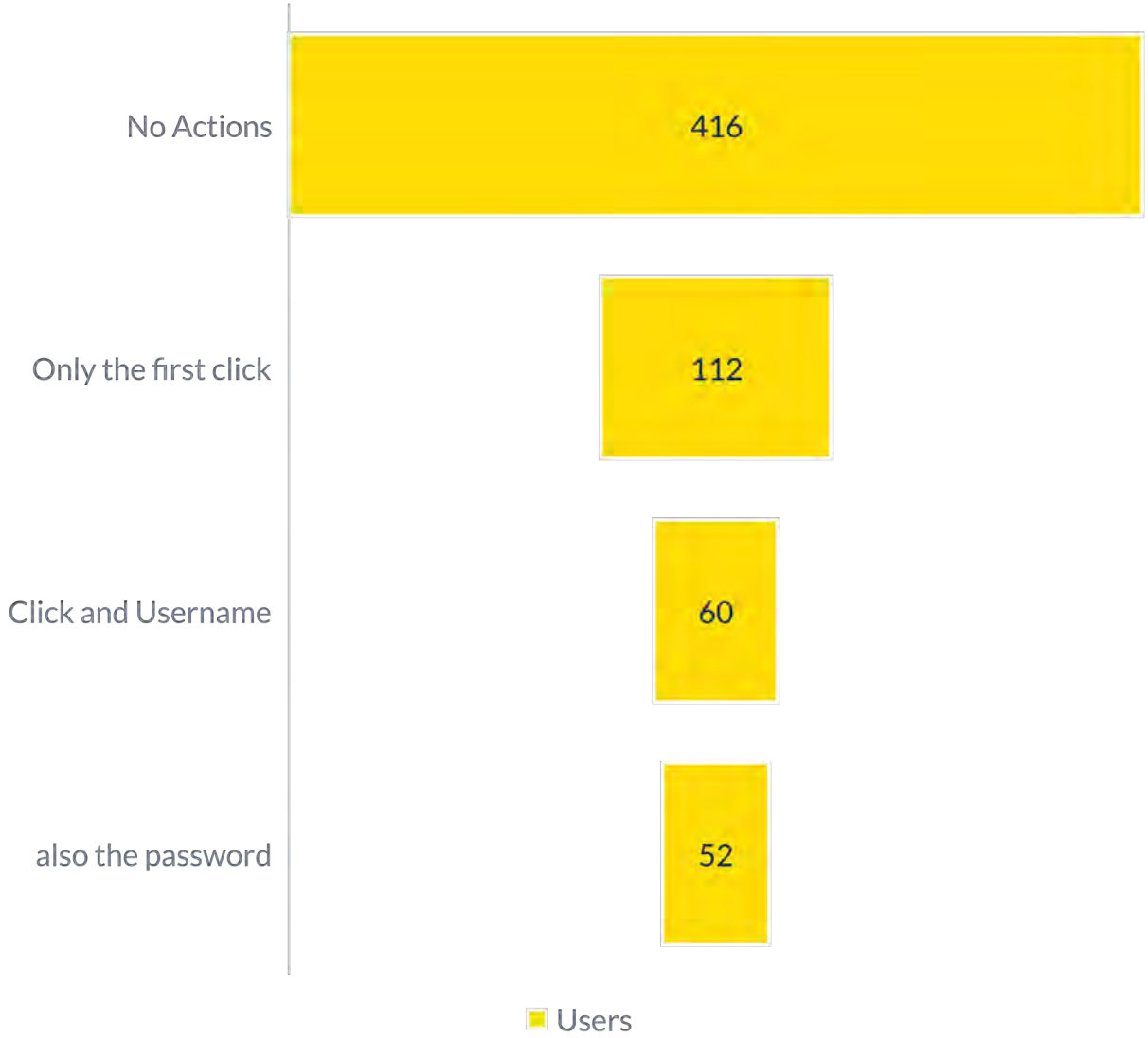
Third Step

- If the victim enters his/her credentials, is redirected to an error page.
- The error page helps to evaluate the behaviour of people, leveraging on a situation of uncertainty (e.g., someone reloads the page or re-enters his credentials).

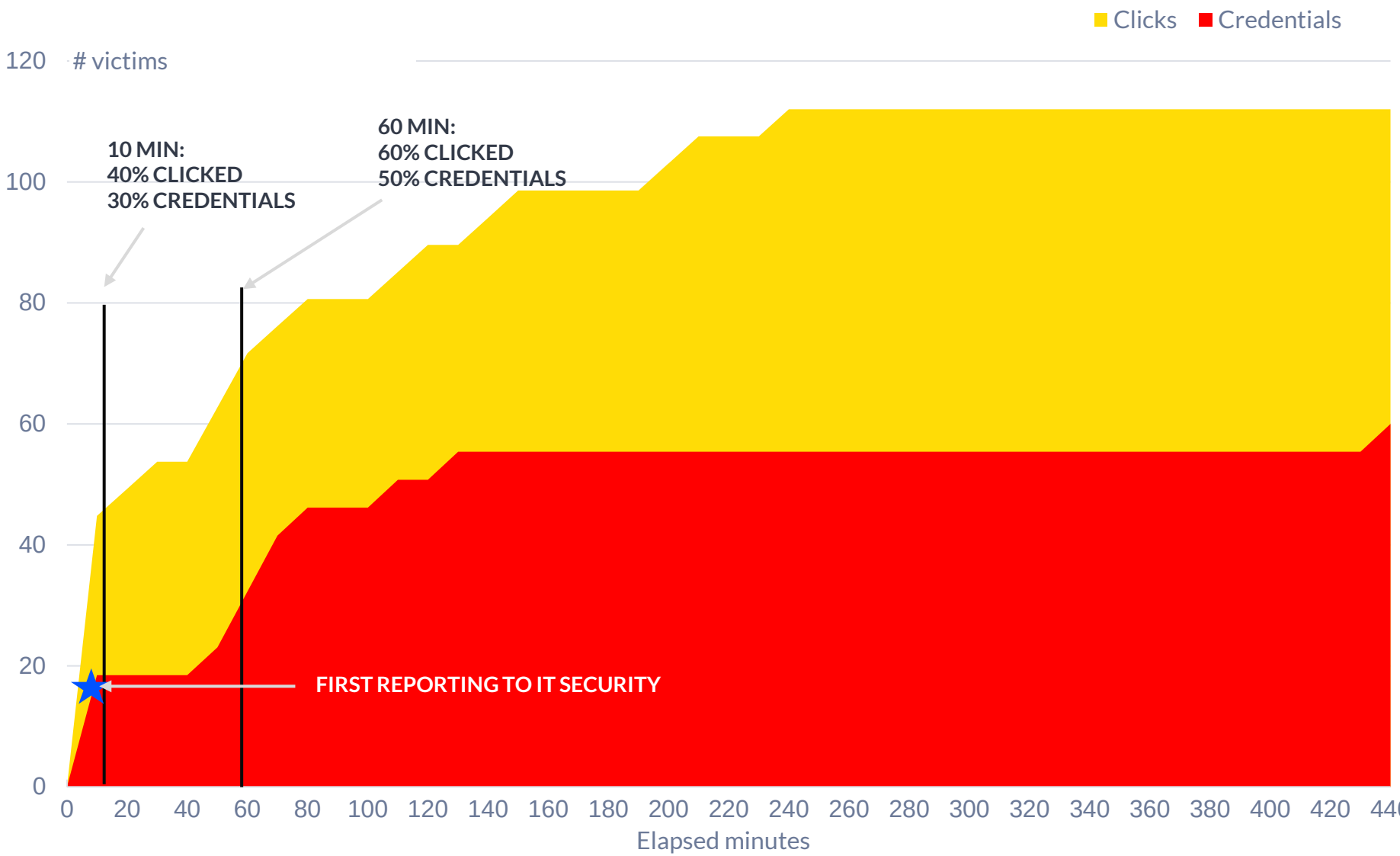


Results of a real test

- The campaign's effectiveness can be estimated at around **20%:** 112 ppl of the sample (~530 ppl) clicked on the malicious link.
- The effectiveness of the second step, the insertion of usernames, is about **11%** (60 ppl).
- **9%** also entered the password (52 ppl).
- Most users did not report receiving a suspicious or phishing email, approx**. 93%**.



Results in time



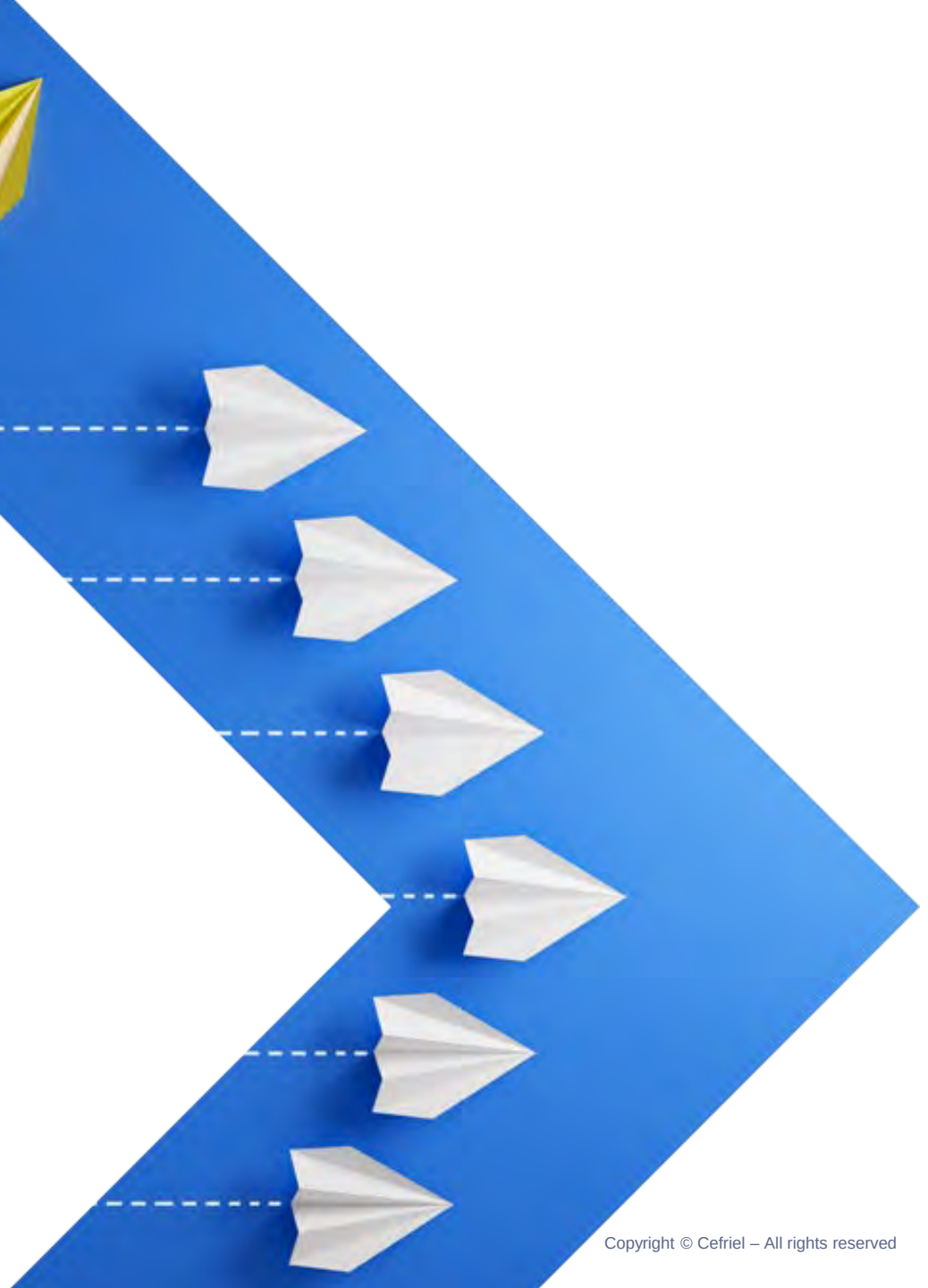
The first user to report, somehow, the attack was after 5 minutes, when there were already approx. 20 victims.

What we are missing

- Selection of the suitable victim
- More targeted selection of the impersonated brand
- CSS
- AI-generated logos aren’t easy
- One-click automation of the entire process
- Avoid anti-spam filters
- Semantics of the attack

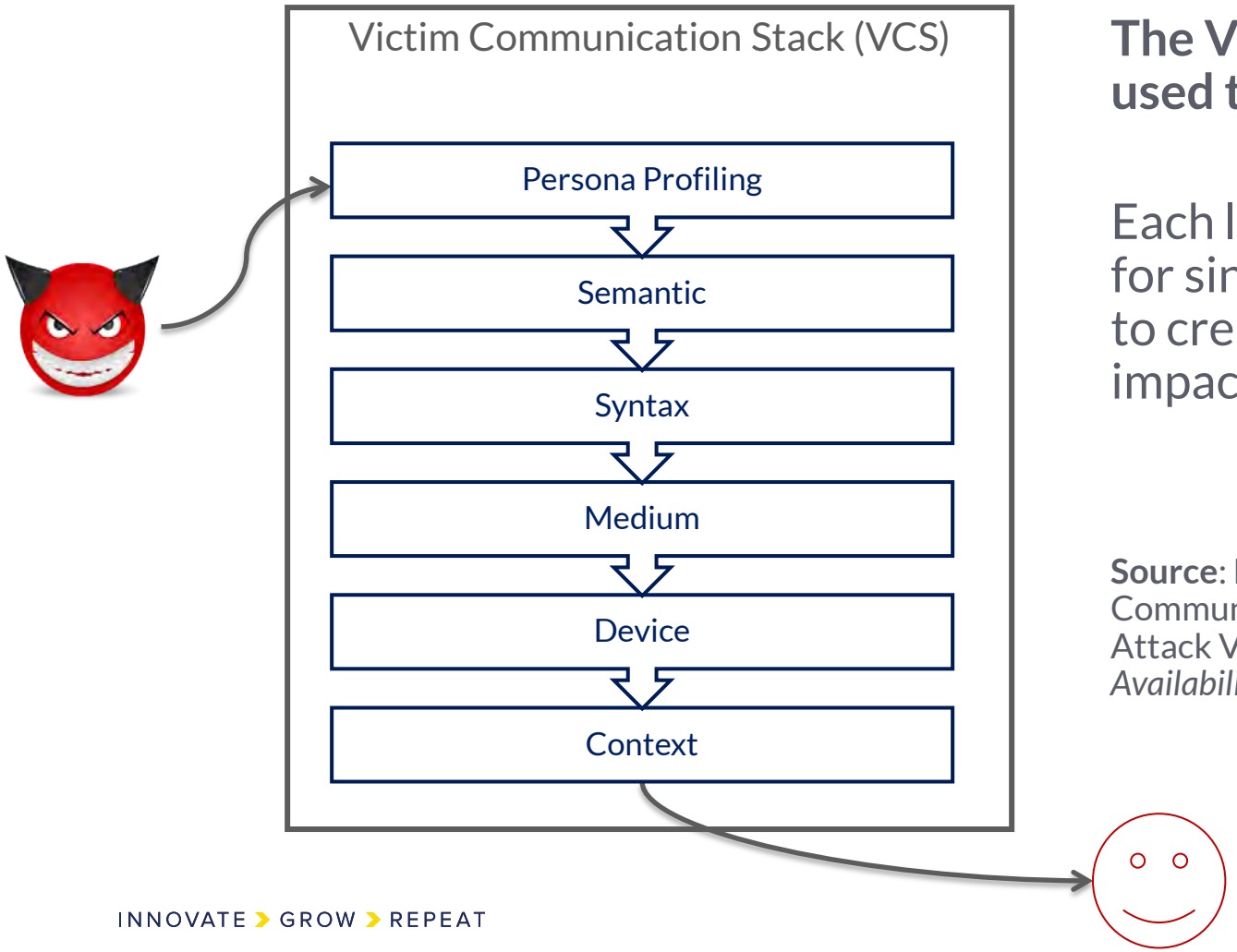
WHY WE DO THIS?

- More efficient Red Teaming
- Automation of phishing simulations
- Better understanding of human-related risks
- Link with Training programs (e.g., interleaving of automated phish simulations as exams)



Human Attack Vector

We require a model before, the Victim Communication Stack (VCS).

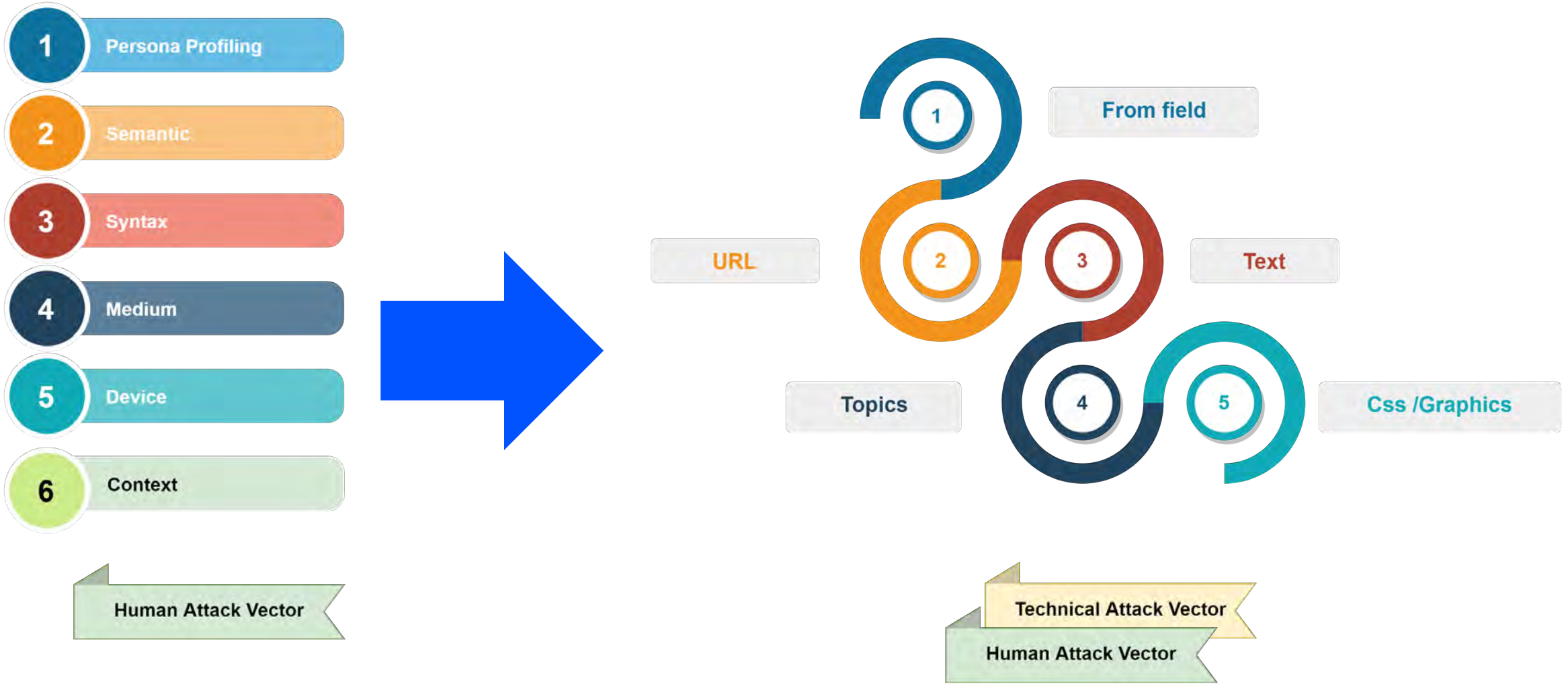


The VCS is the theoretical communication stack used to build an HAV.

Each layer has different possible choices, either for single victims or groups of victims. The aim is to create a model for attacking and reducing the impact of “creativity”.

Source: E. Frumento, F. Freschi, D. Andreoletti and A. Consoli, "Victim Communication Stack (VCS): A flexible model to select the Human Attack Vector", *Proceedings of the 12th International Conference on Availability, Reliability and Security - ARES '17*, p. art. no. 50, 2017.

Human Attack Vector and Technical Attack Vector



HAV and TAV example

Emotet is again active in Italy (01/11/2022)

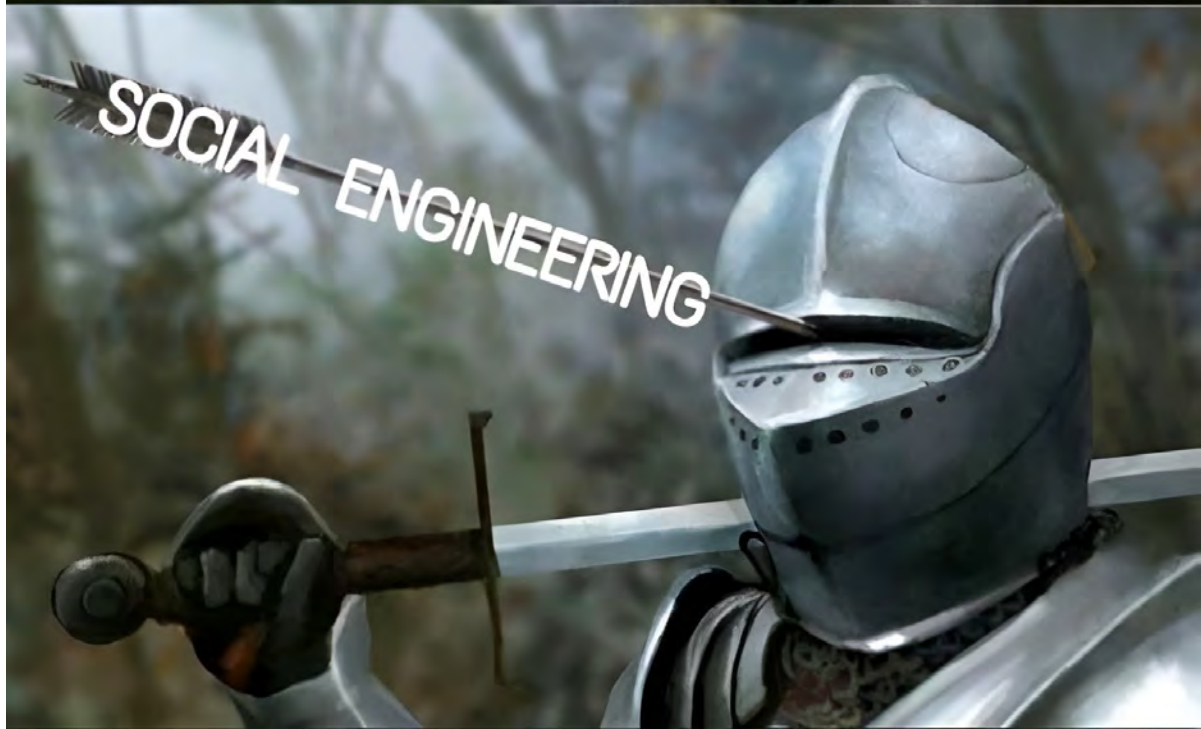
HAV

A new campaign with Italian targets aimed at conveying via email a password-protected ZIP attachment containing an XLS equipped with malicious macros.

To get infected, a user needs: to open the email, open the attached zip, enter the password, open the excel file inside the zip, enable Office macros, ignore all the warnings that Office shows and forget that you have done all this.

TAV

Only then the macro executes, which downloads and then executes the malware.



Phishing ≠ Spear phishing

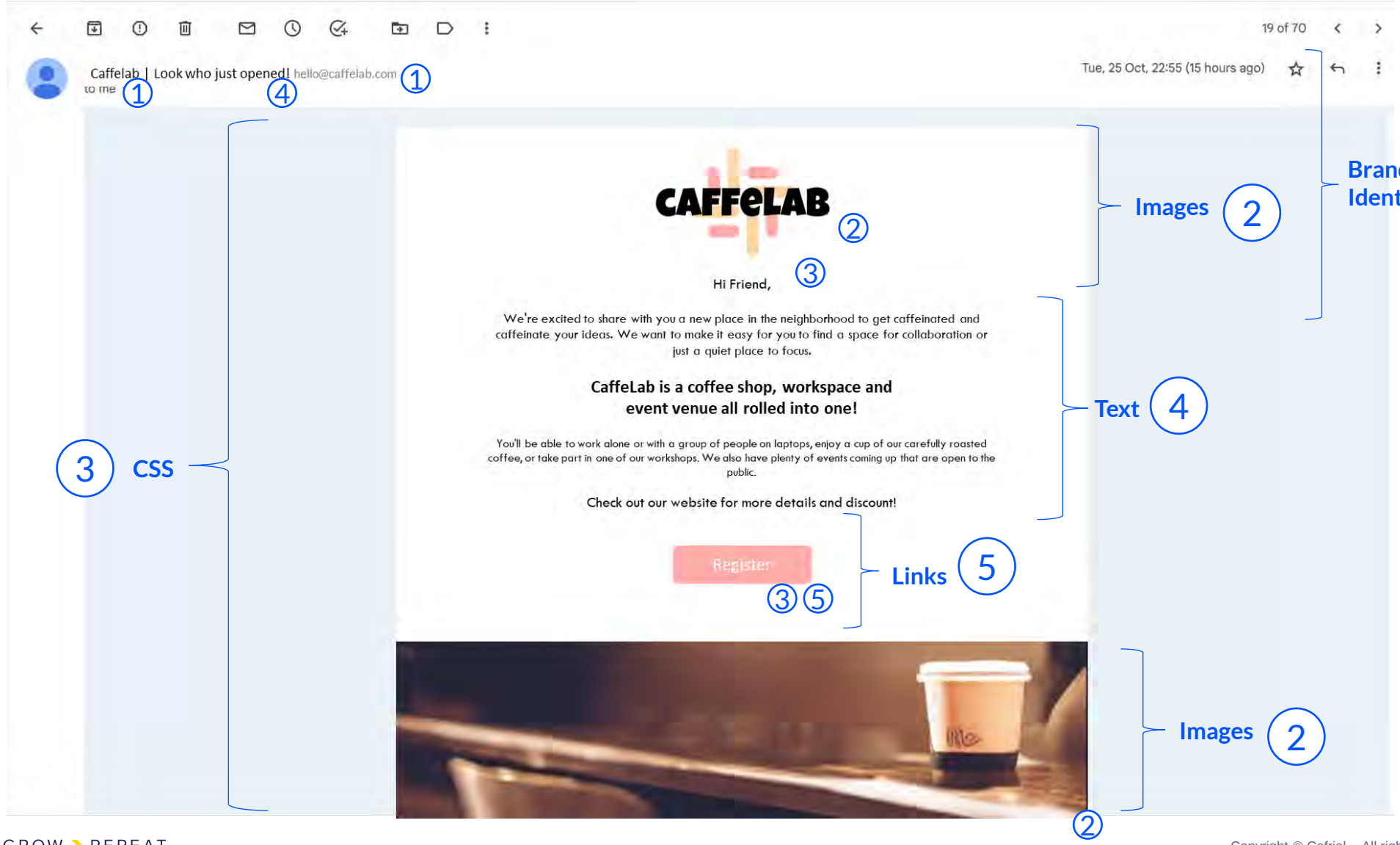
Phishing

- Is a more sophisticated form of SPAM that, thanks to graphics, delivers a more sophisticated hook, specialised for a subsample of users belonging to the targeted company (e.g., targeted Bank customers are falling more into this hook than those who are not). The business model is not flat. It is usually sent to fewer people as SPAM but also to people not belonging to the users' category chosen, supposing that for them, the hook does not work.

Spear Phishing

- A specialised form of phishing sent only to the company's customers, which the mail pretends to come (in this sample, eBay). The return of this type of phishing is more significant. The victims are selected because they are customers of the targeted organisation. Victims are chosen on the Social Networks using OSINT techniques or setting up customers' assistance unofficial pages on the social networks.

Phishing mail template



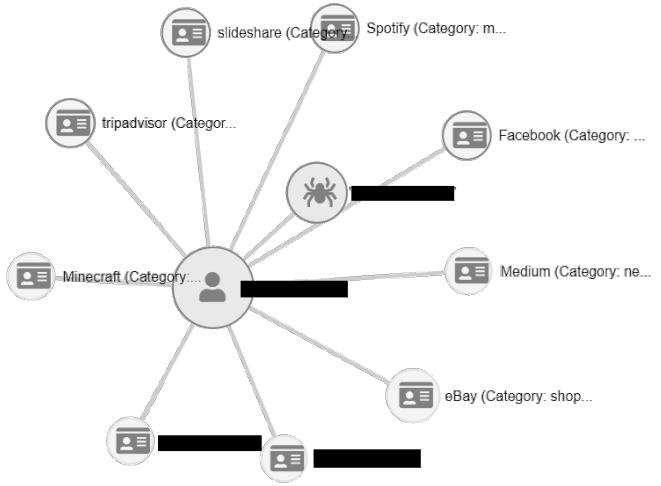
1 - Brand Identity

1. Persona Profiling
2. Semantic
3. Syntax
4. Medium
5. Device
6. Context

Human Attack Vector

OSINT Research over the chosen victim.

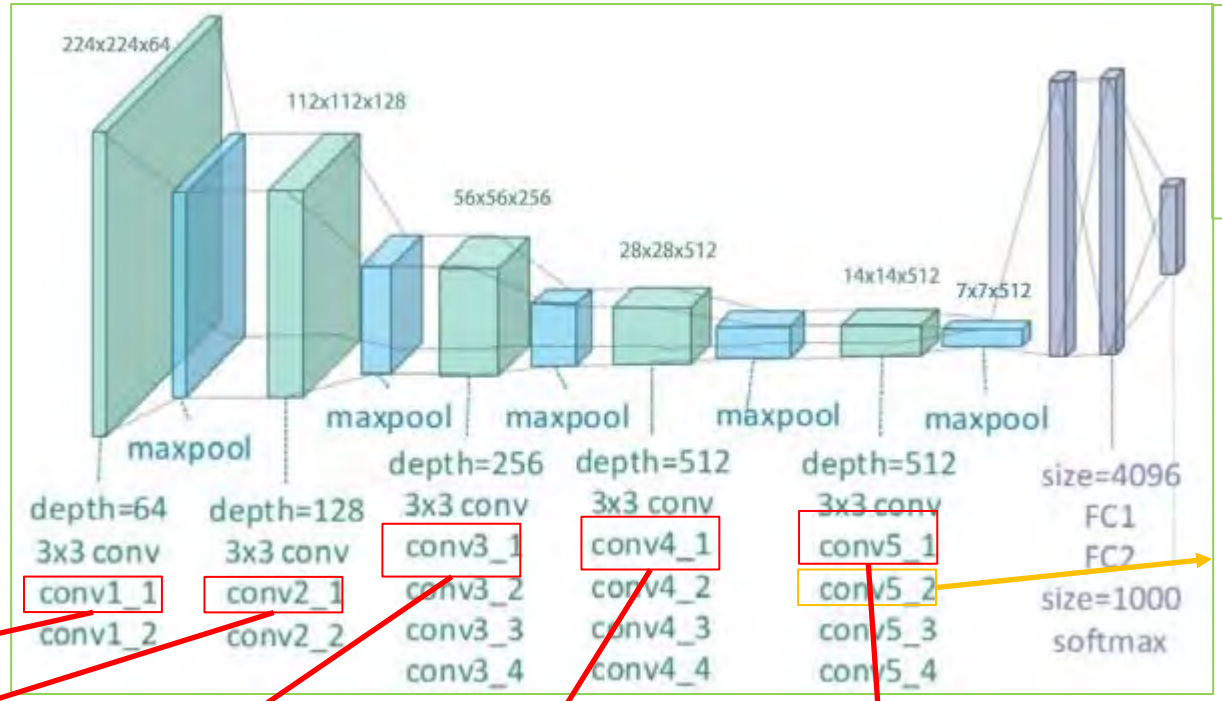
Spiderfoot Output



	Spotify	Facebook	Ebay	Tripadvisor
Ebay	0.4303366	0.35032713	**1.0**	0.26906437
Google	0.6327295	0.50059026	0.45825684	0.3524051
Microsoft	0.6098279	0.16725439	0.34018078	0.1796725
Outlook	0.6098279	0.000841757	0.03805846	0.05037361
Amazon	0.6098279	0.29049876	0.46660656	0.27714825
LinkedIn	0.45898113	0.6356593	0.36349726	0.37835854
Paypal	0.46584445	0.3903677	0.45583403	0.25550362
Facebook	0.6058922	**1.0**	0.35032713	0.28615004
Netflix	0.713966	0.34745294	0.42040065	0.2508172

Cosine similarity matrix between the output of OSINT searches and the brands most used for phishing attacks, according to the Word2Vec English GoogleNews Negative300 module.

2 - Images



VGG-19 Imagenet Weights

Chosen Brand Identity Logo

Content

block5_conv2
shape: (1, 26, 32, 512)
min: 0.0 max: 2410.8796
mean: 13.764149

Style

block1_conv1
shape: (1, 336, 512, 64)
min: 0.0
max: 835.5256
mean: 33.97525

block2_conv1
shape: (1, 168, 256, 128)
min: 0.0
max: 4625.8857
mean: 199.82687

block3_conv1
shape: (1, 84, 128, 256)
min: 0.0
max: 8789.239
mean: 230.78099

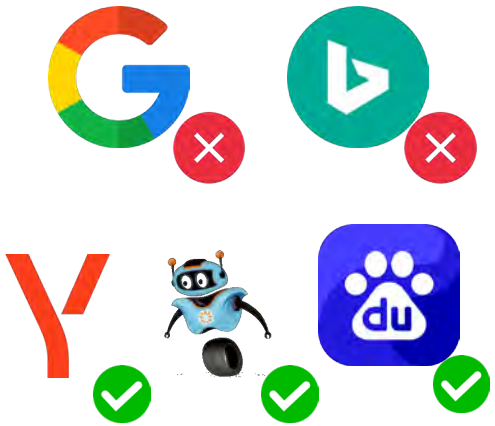
block4_conv1
shape: (1, 42, 64, 512)
min: 0.0
max: 21566.135
mean: 791.24005

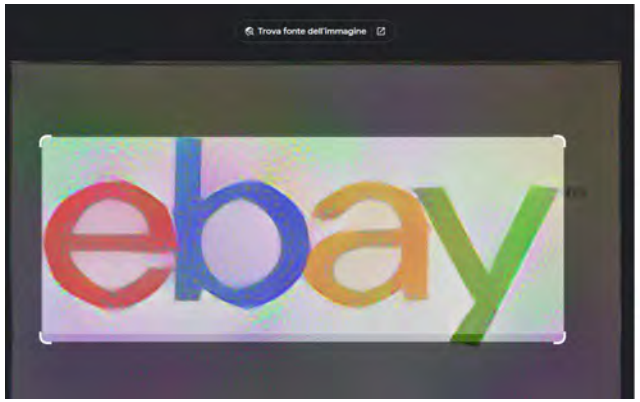
block5_conv1
shape: (1, 21, 32, 512)
min: 0.0
max: 3189.2542
mean: 59.179478

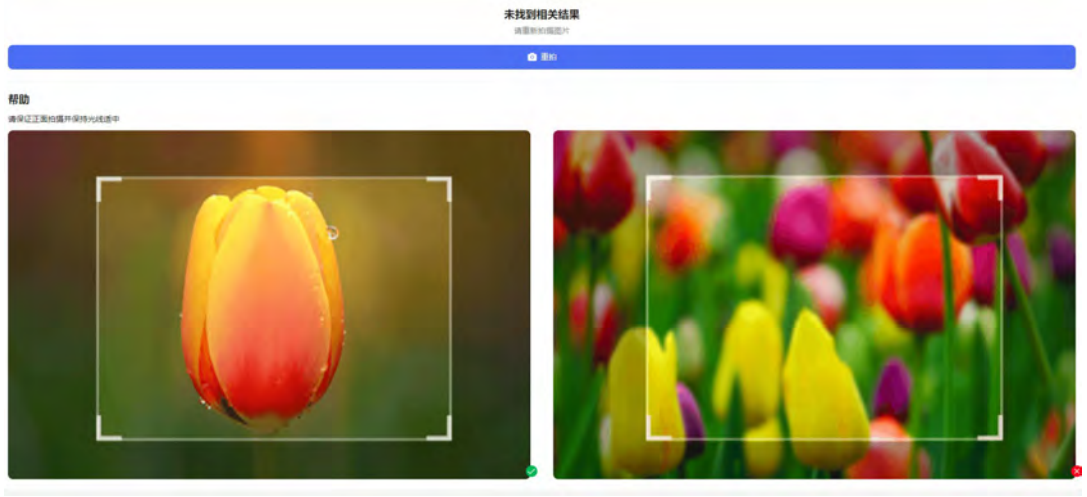
$$G^l_{cd} = \frac{\sum_{ij} F^l_{ijc}(x) F^l_{ijd}(x)}{IJ}$$

Gram Matrix

2 - Images

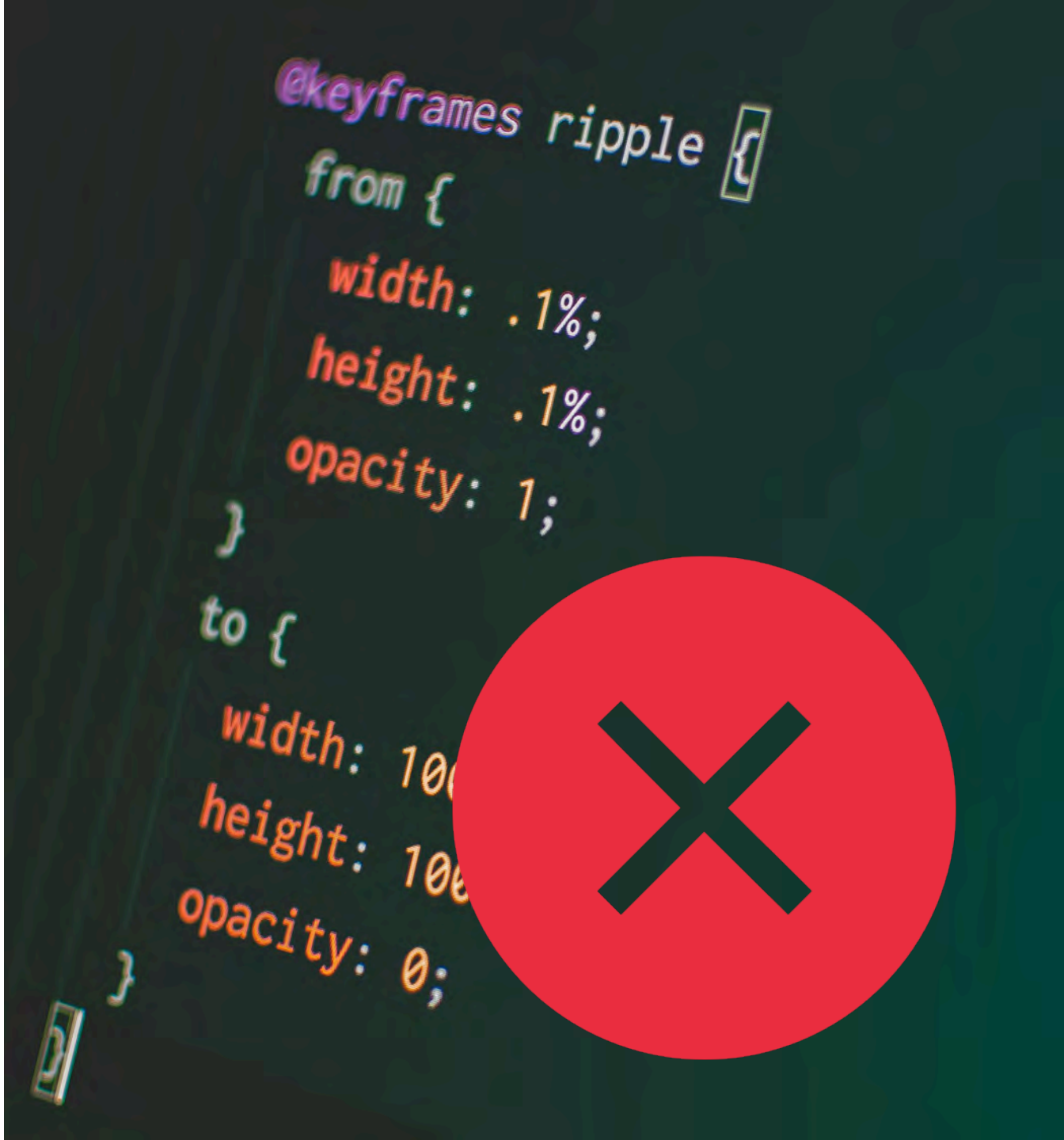






3 - CSS

- **Targeted and automated CSS/Style** is the only wholly **uncovered area**
- This is due to several factors:
  - lack of a suitable AI model
  - lack of a precise dataset due to high heterogeneity
  - results can be obtained, but none of them is satisfactory.



4 - Text

A 6 billion parameter, autoregressive text generation model trained on The Pile.

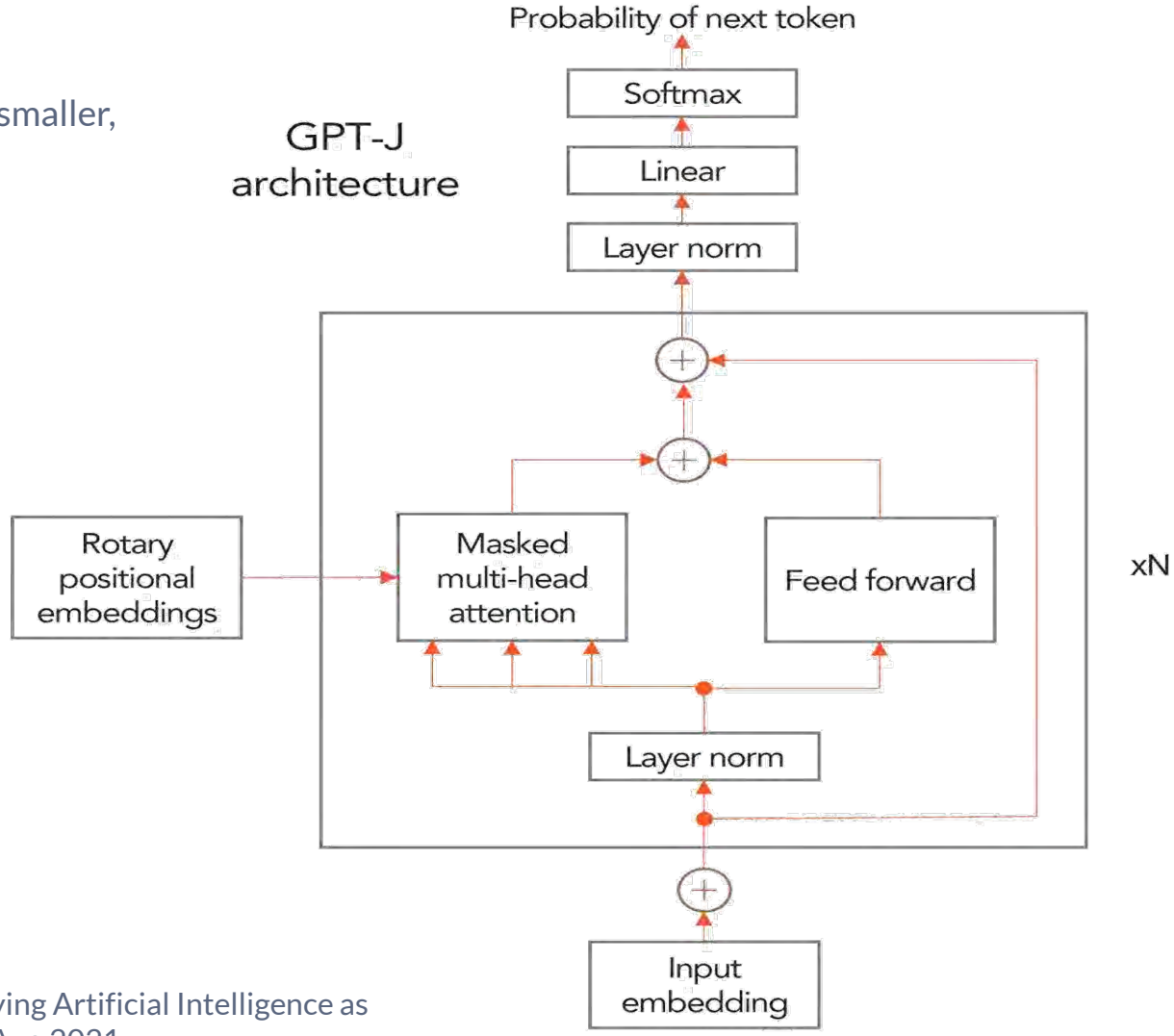
The Pile is an 825 GiB diverse, open-source language modelling data set comprising 22 smaller, high-quality datasets combined.

Model Details

Hyperparameter	Value
n_parameters	6,053,381,344
n_layers	28*
d_model	4,096
d_ff	16,384
n_heads	16
d_head	256
n_ctx	2,048
n_vocab	50,257 (same tokenizer as GPT-2/3)
position encoding	Rotary position encodings (RoPE)
RoPE dimensions	64

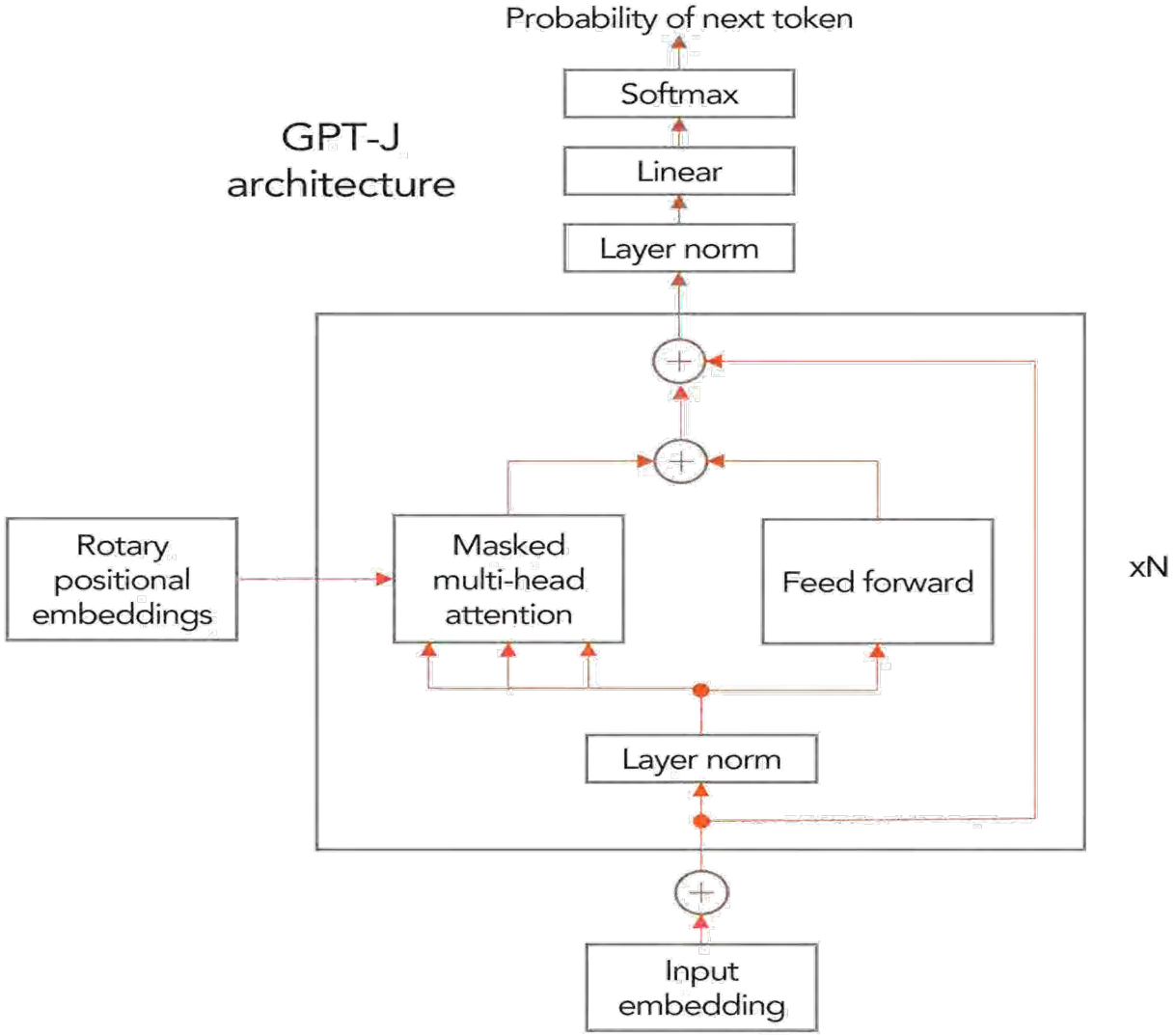
* each layer consists of one feedforward block and one self attention block

The model consists of 28 layers with a model dimension of 4096, and a feedforward dimension of 16384. The model dimension is split into 16 heads, each with a dimension of 256. Rotary position encodings (RoPE) was applied to 64 dimensions of each head. The model is trained with a tokenization vocabulary of 50257, using the same set of BPEs as GPT-2/GPT-3.



Similar study: E. Lim et al., Turing in a Box: Applying Artificial Intelligence as a Service to Targeted Phishing, Blackhat 2021, Aug 2021, https://tinyurl.com/23gjc3e3

5 – Links (typosquatting)

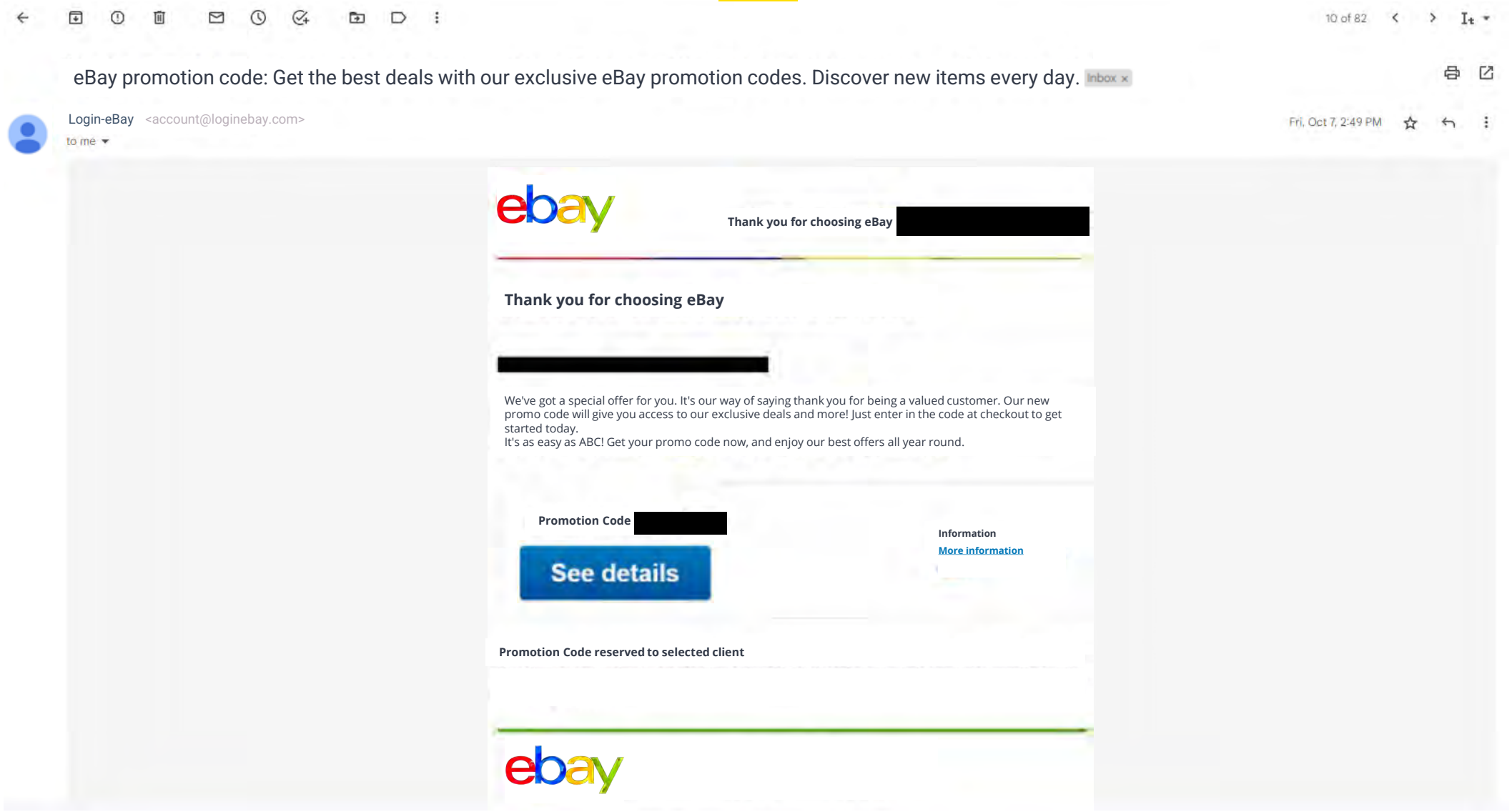


FROM:
account@LogInEbay.com
noreply@SecureEbay.com
Signin@ebay.com
LogInEbay@EbayLogin.com

Object:
Ebay promotion code: Get the best deals with our exclusive ebay promotion codes. Discover new items every day.

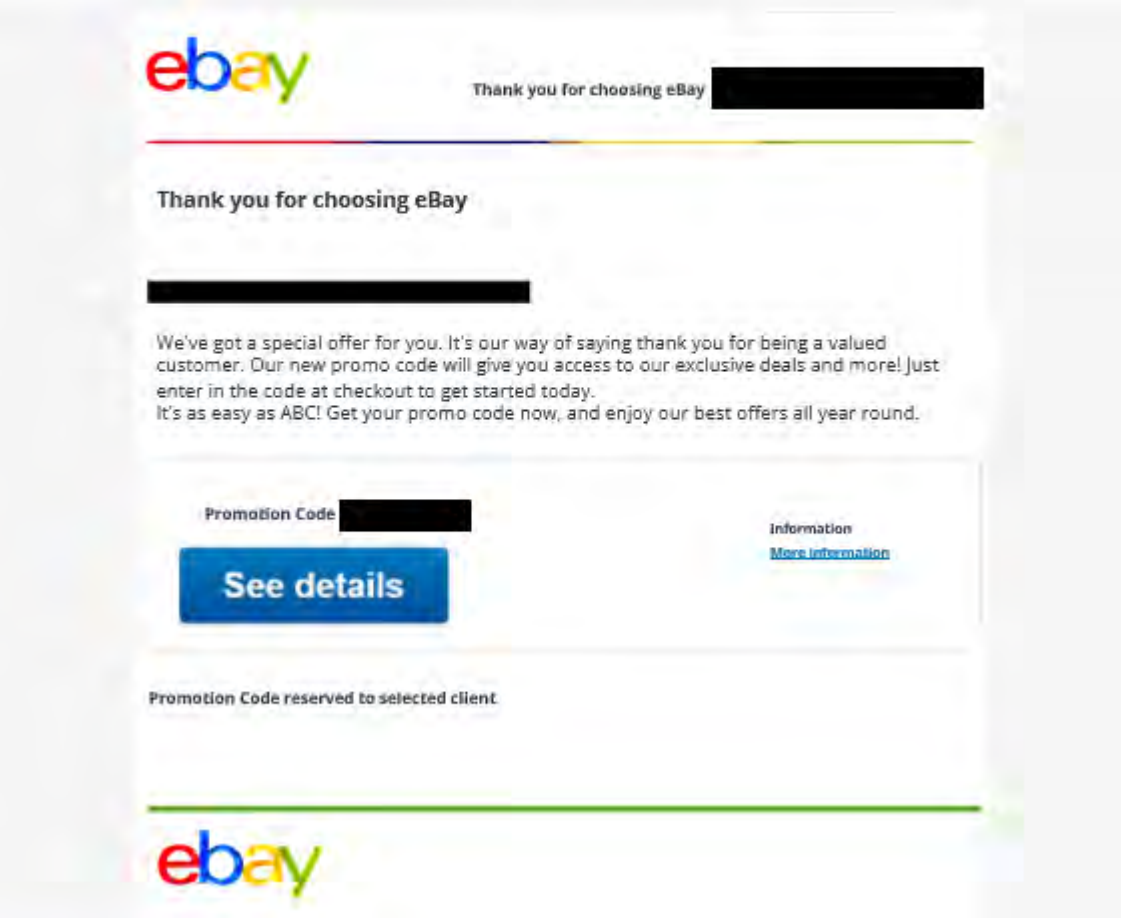
URL:
LogInEbay.com
SecureEbay.com
EbayLogin.com
Ebay-SignIn.com

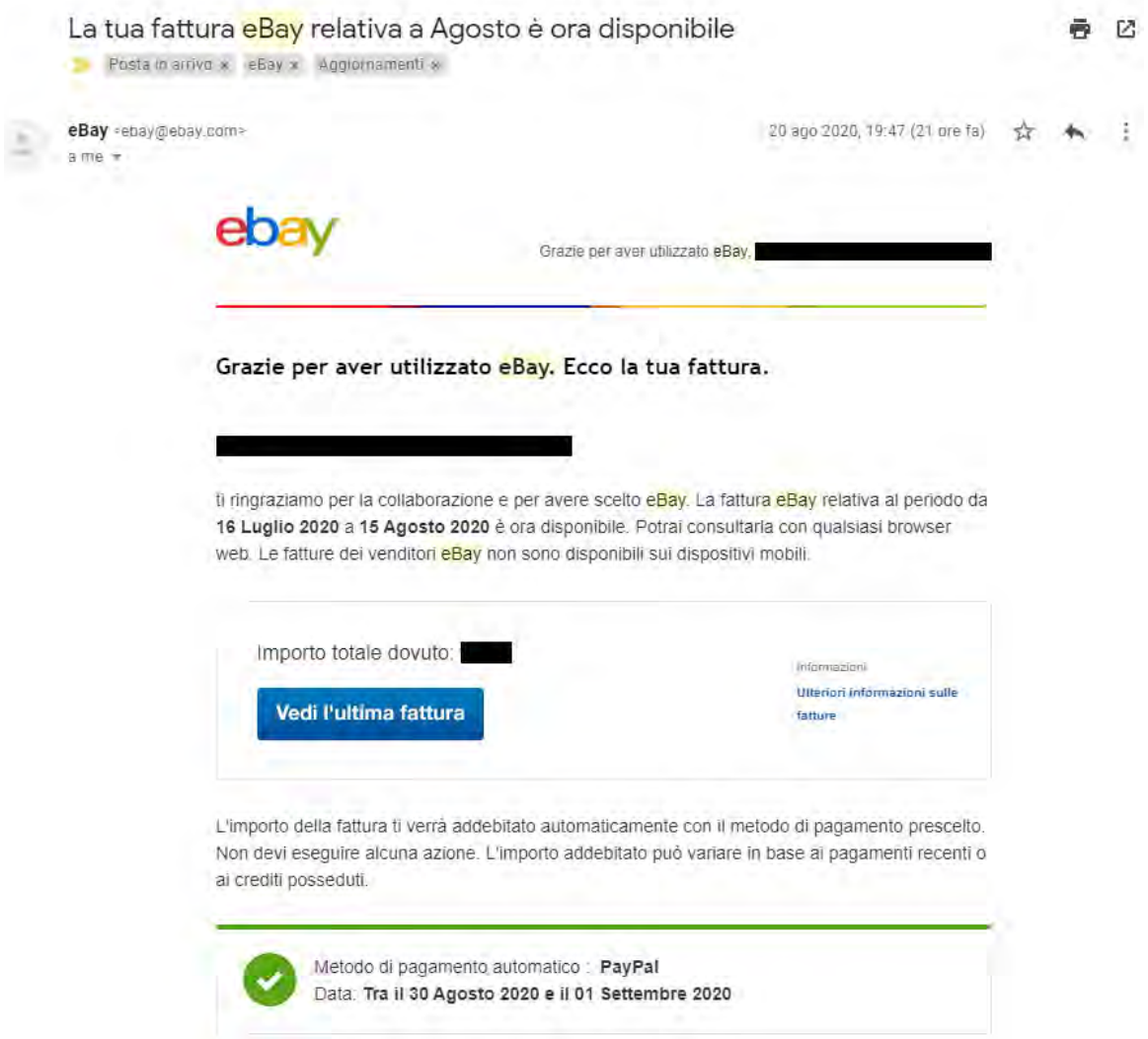
Final Result



Ebay-ish vs Ebay

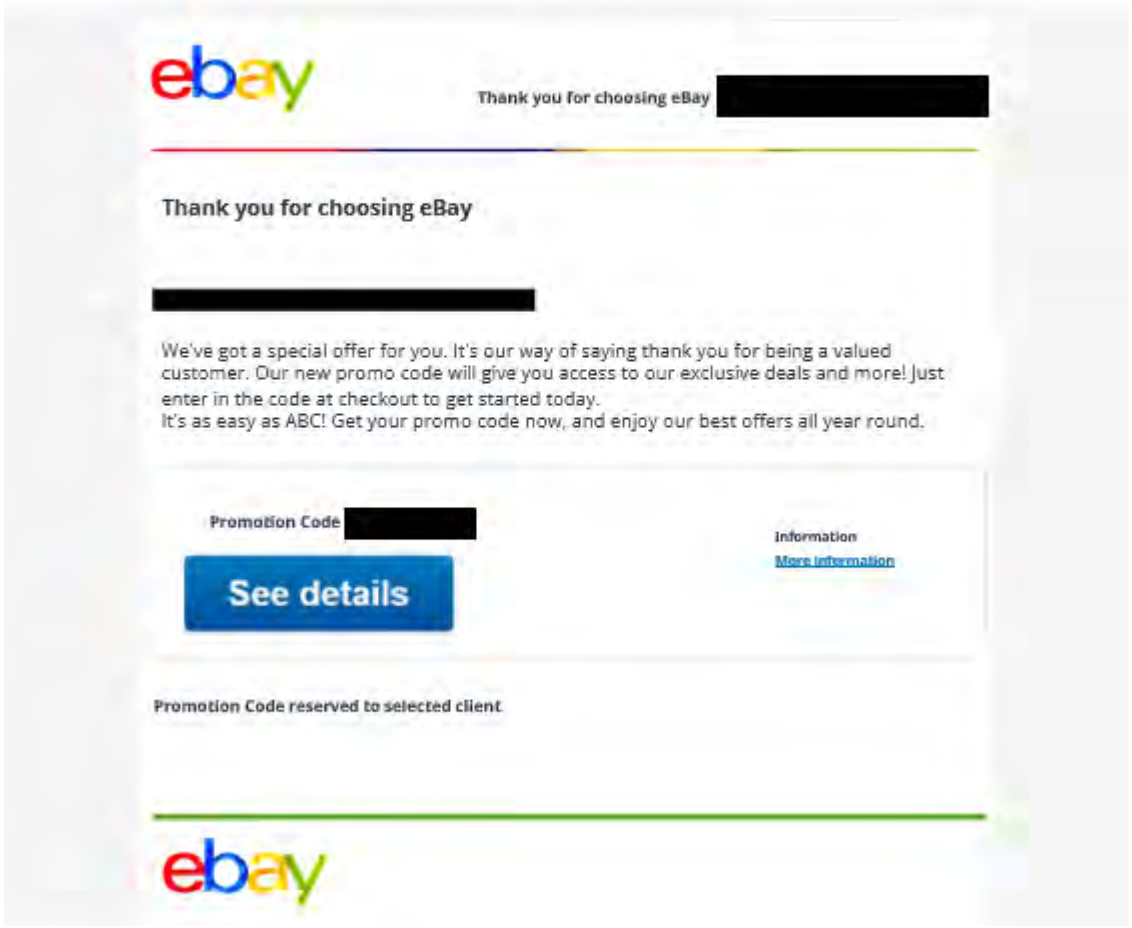
leals with our exclusive eBay promotion codes. Discover new items every day.



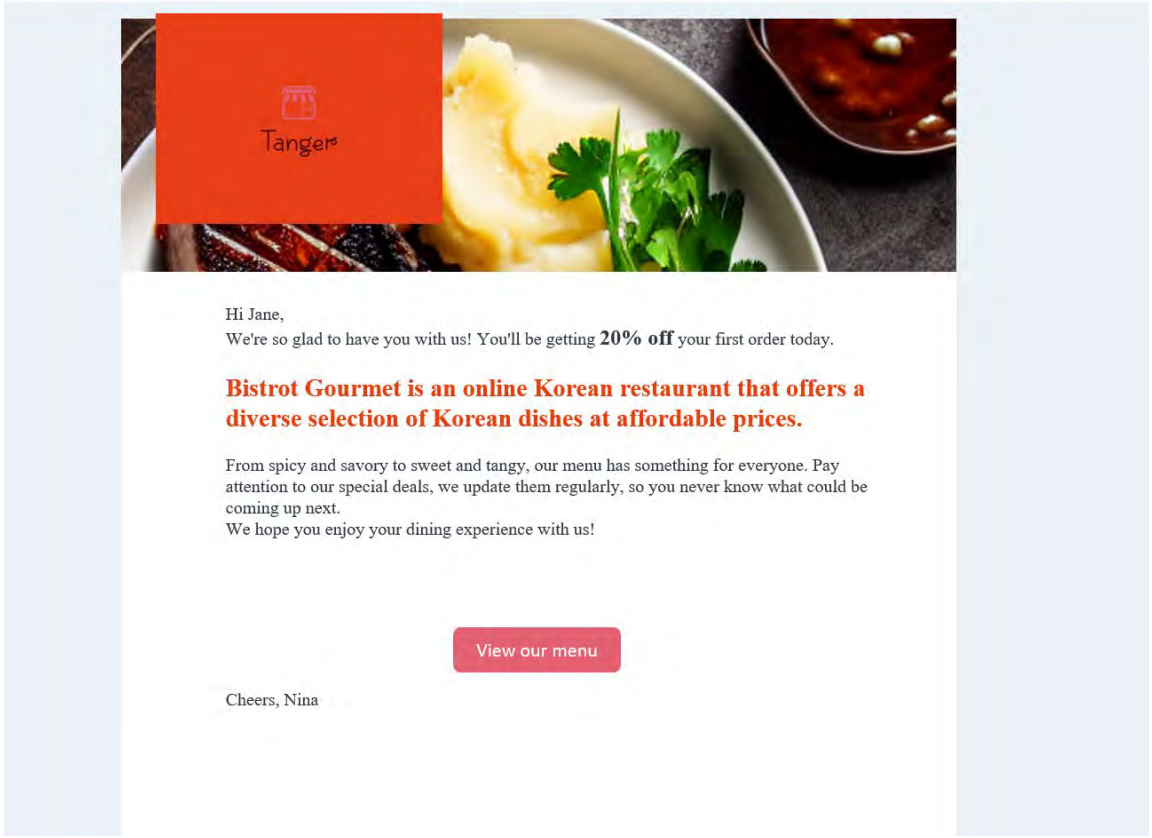


Confrontation

leals with our exclusive eBay promotion codes. Discover new items every day.



o@tanger.com Tue, 25 Oct, 22:55 (15 hours ag



Conclusions & Future works

- ***AI-Generation world runs FAAAST!***

- Improve automatic OSINT exploration of potential victims
- Include automation of uncovered aspects (e.g., CSS and assembling of email elements)
- Automate other mediums of communication (e.g., SMS or IM)
- Use new learning models (e.g., GPT-4)

Cefriel

POLITECNICO DI MILANO

WWW.CEFRIEL.COM

CONTACT

Enrico Frumento
Cybersecurity Senior Domain Expert

Cefriel - Politecnico di Milano
www.cefriel.com
Viale Sarca 226 – 20126 Milano – IT

https://www.linkedin.com/in/enricofrumento/

INNOVATE › GROW › REPEAT