Luigi Gallo

Tutor: Prof. Alessio Botta
XXXIV Cycle
III year presentation

UNIVERSITÀ DEGLI STUDI DI NAPOLI
FEDERICO II

A Machine and Human Learning approach for Phishing Defense in a large company

Background

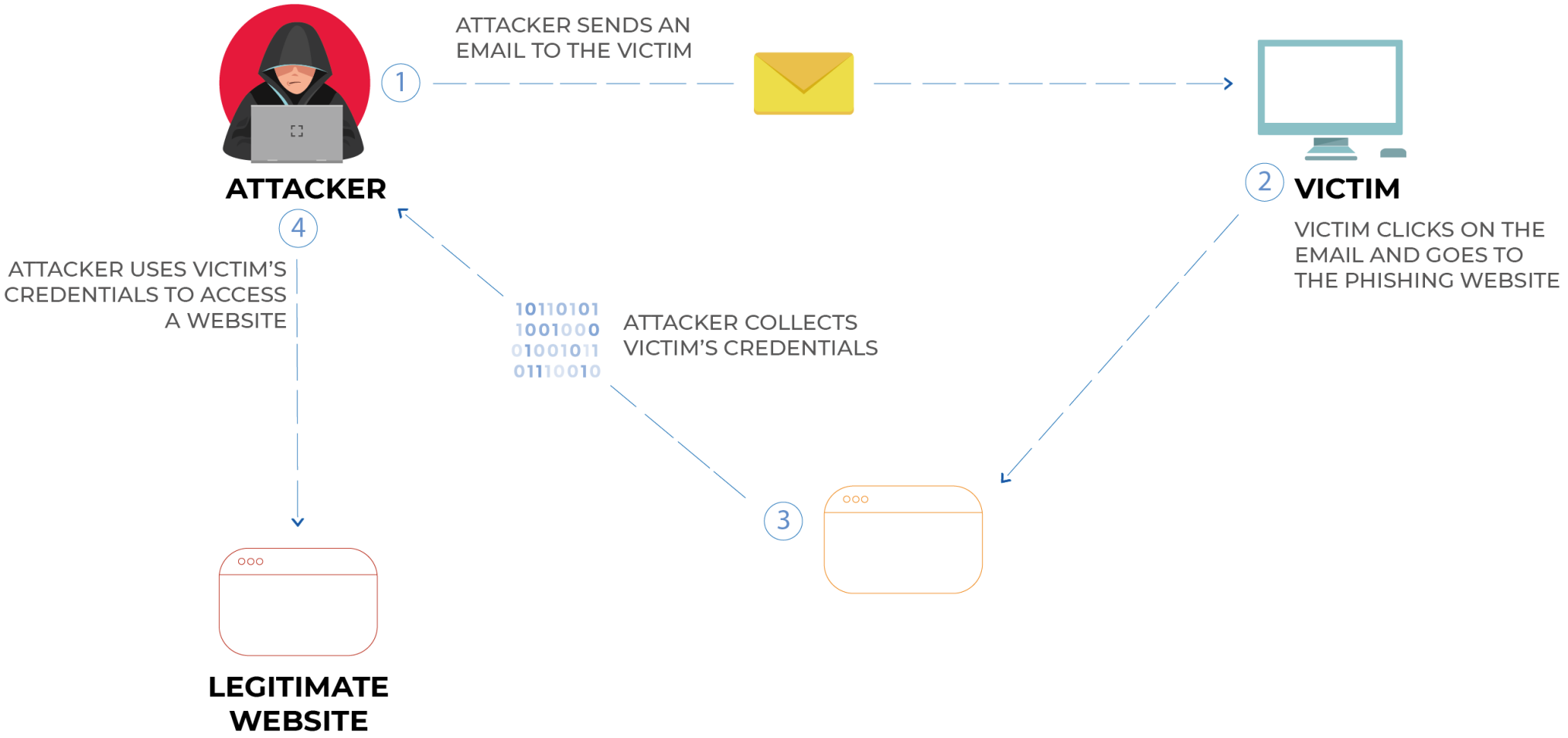
- Master's Degree in 2018
  - Anomaly Detection in traffic traces with Big Data Analytics
- Internship at ArcLab Dieti – Comics Research group
  - Anomaly Detection in traffic traces with Big Data Analytics
  - Cloud Robotics Architectures
- Cyber Security Lab (Telecom Italia Lab)
  - The use of Machine Learning technologies for Security purposes
  - Security in 5G Networks (3GPP)
  - Scouting and Testing of novel security solutions

Context & Contribution

- **Context**
  - Phishing email attacks
  - Human factor in Cyber Security
  - Machine Learning for Cyber Security purposes

- **Contribution**
  - An email threat management system based on
    - the use of automatic machine learning classifiers to anticipate security incidents, trained with real data collected on the field
    - the analysis of both the technical and cognitive characteristics of a phishing attack
    - the best defence methodologies found during the previous analyses, validated by large-scale social experiments

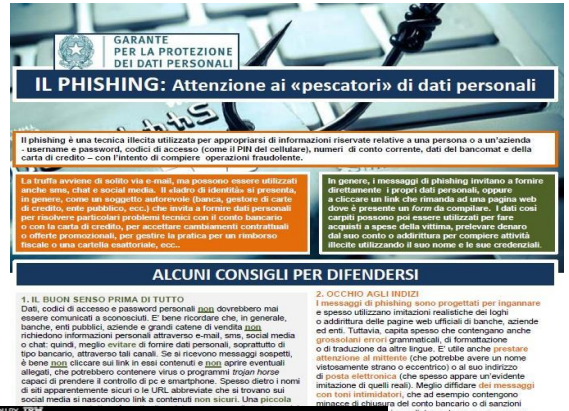
Phishing Attacks

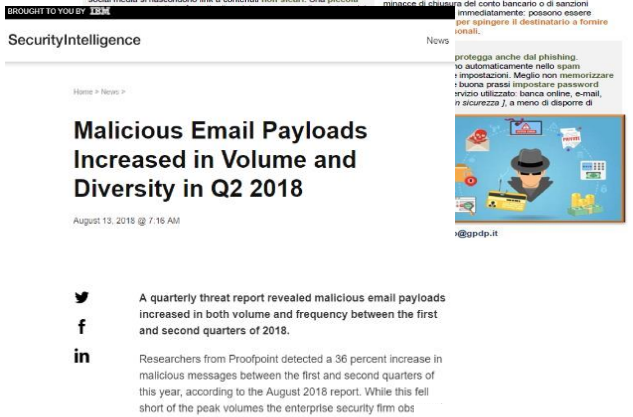


Motivation (1/3): phishing as major security issue

Email is currently one of the most used channels for making (starting) cyber attacks, **increasing in number and in malignance** ($1.8 billions of monetary losses in USA in 2019).

Several security and data protection agencies have raised alarms





CEO Fraud e CyberSecurity

Un'azienda aerospaziale austriaca ha licenziato di recente il suo presidente e il suo Cfo dopo aver perso quasi 50 milioni di dollari per una cyber frode.

Secondo uno studio internazionale oltre 400 imprese ogni giorno sono vittime di frodi finanziarie attraverso tecniche di ingegneria sociale ed in particolare di spear phishing.

Anche in Italia nell'ultimo anno si è registrata una recrudescenza degli attacchi mirati ai vertici aziendali, l'ultimo eclatante caso ha riguardato un dirigente di Confindustria che, con una banalissima mail, è stato spinto a fare un bonifico di mezzo milione di euro.

In passato economia e cyber space interagivano poco, ora invece i due piani non sono più separabili e garantire un cyber space sicuro per fare business, diventa un vantaggio competitivo.

Gerardo Costabile, tra i più autorevoli esperti italiani di cybersecurity, ci spiega perché rischi che possono essere noti alla maggior riportati o gestiti in modo efficace e quali sono i controlli essenziali che i vertici di tutte le aziende dovrebbero rispettare e impleme loro strutture e le filiere produttive in cui operano.

Attraverso una rassegna della casistica internazionale, Costabile illustra, in particolare, che cos'è e come funziona la BEC (Business quali sono le vittime predilette, i punti critici che impattano sul fattore umano e sulle infrastrutture tecnologiche e soprattutto gli s

Luigi Gallo

Motivation (2/3): even a “simple” bullet can hurt

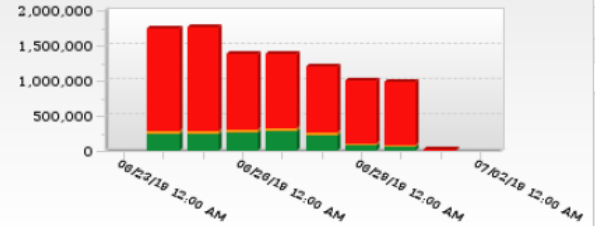
- People increasingly publish **personal information**, typically used to make email attacks **trustworthy and captivating** (social engineering techniques)
- The main problem is that these attacks are very sophisticated and mingle with a lot of noise (marketing, advertising, errors, newsletters, sex photos etc.)
- The problem with large companies is that the number of employees who may **fall victim of phishing or download malware** is considerable

Executive Summary (Inbound)

Monday, Jun 24, 2019 01:00 AM to Monday, Jul 01, 2019 01:00 AM CEST

Message and Threat Summary

Messages



Message Types	Count	%
Single threat messages:	7,878,141	82.6
Multiple threat messages:	201,162	2.1
Total Threat Messages:	8,079,303	84.7
Clean messages:	1,458,656	15.3
Total Messages:	9,537,959	100

Motivation (3/3): the need for automatic tools

SMTP was not designed with a *security-by-design* approach:
the recipient of an email cannot authenticate the sender!

Anti-Spam filters help to mitigate the problems of STMP and Spam Emails

✓ Network overload
✓ Loss of time and productivity
✓ Irritation and discontent

but

X Vechicle for attacks

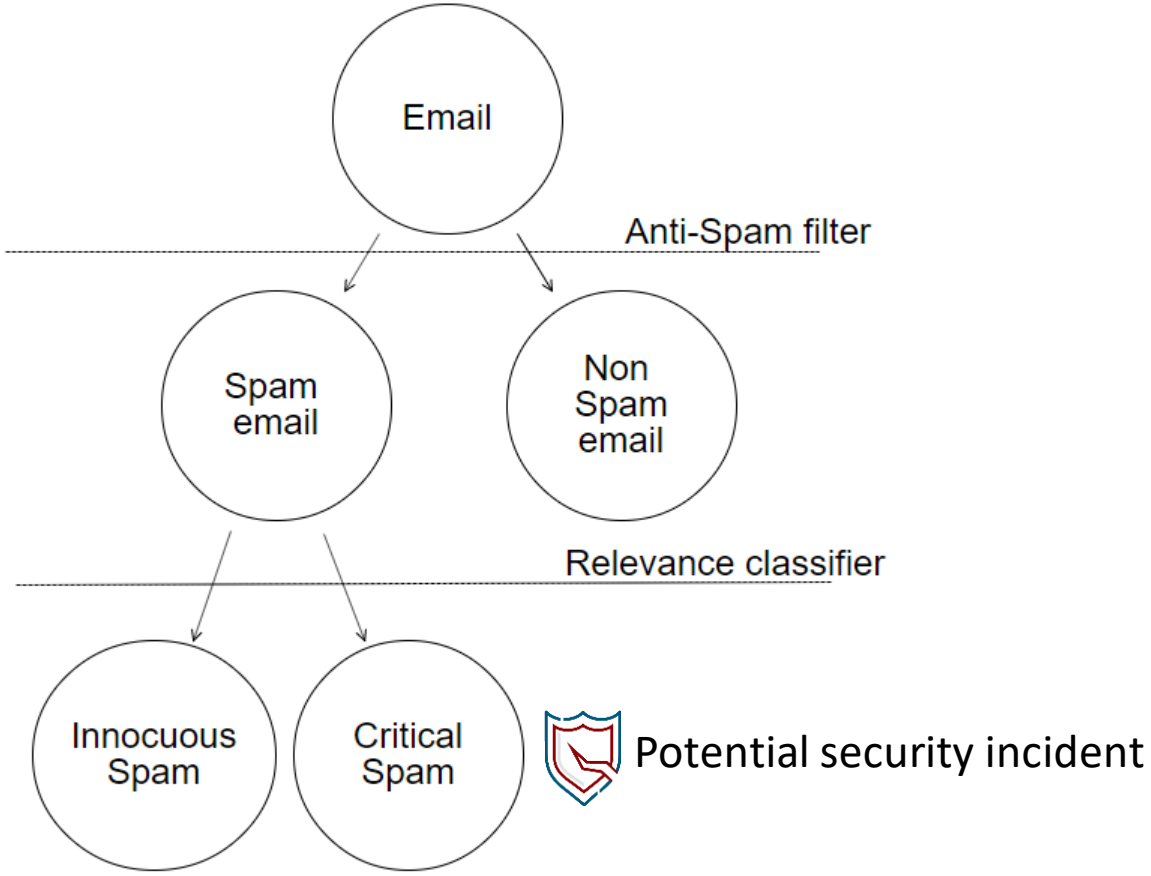
(miss-classifications, evasion and poisoning attacks)

outcome →

- Identity Theft
- Direct economic losses
- Serious reputational consequences
- Denial of access to services

Entire anti-phishing groups in SOCs to check for security incidents among the user-reported emails (mostly noise)

IDEA : highlighting “the needle in the haystack”

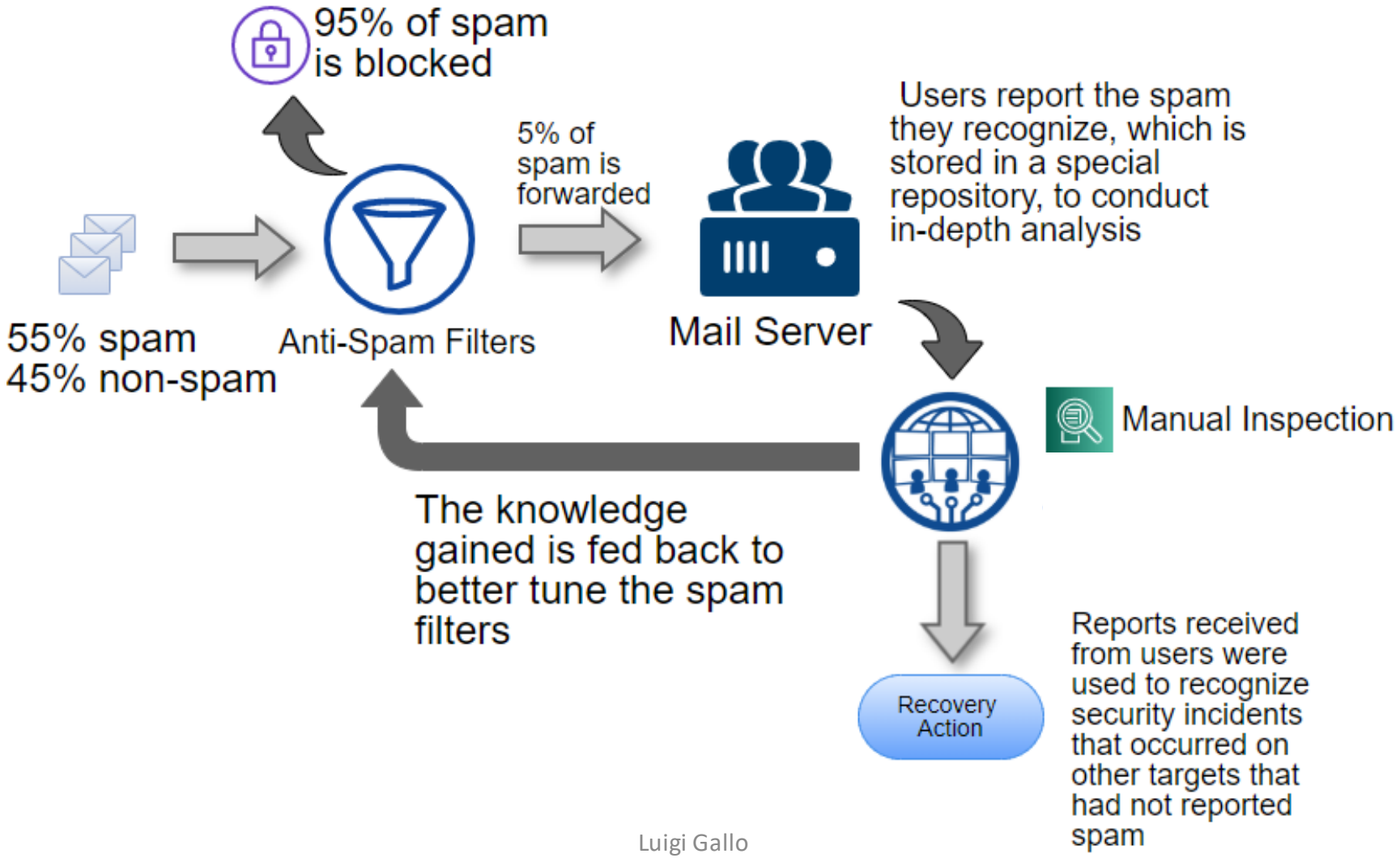


Security incident defined as
“a security-relevant system event in which the system’s security policy is disobeyed or otherwise breached” [RFC4949]

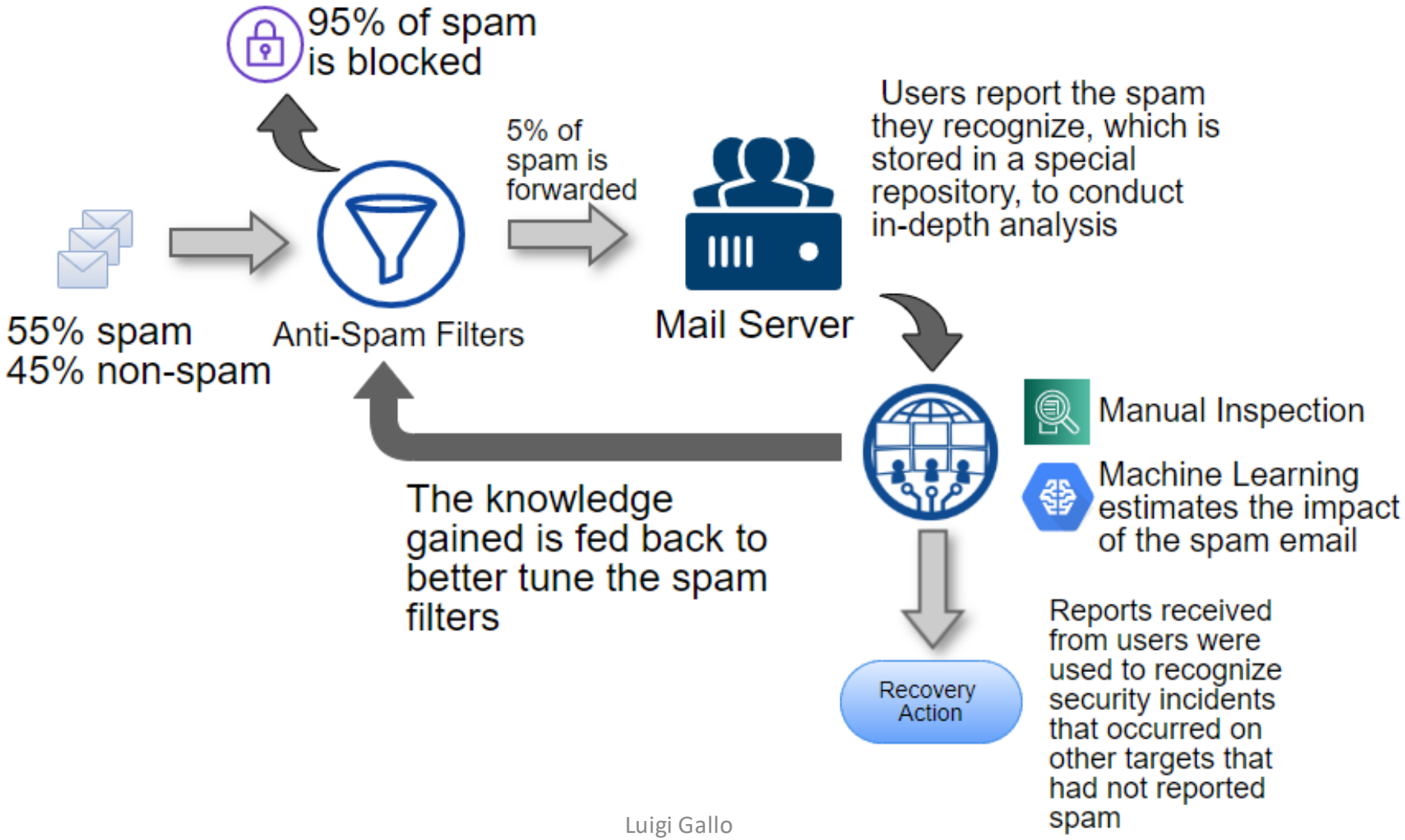
Two necessary conditions as long as a security incident occurs:

- The recipient is deceived by the email
- The “payload” of deception is not trivial

Scenario


95% of spam is blocked
5% of spam is forwarded
Users report the spam they recognize, which is stored in a special repository, to conduct in-depth analysis
55% spam
45% non-spam
Anti-Spam Filters
Mail Server
Manual Inspection
The knowledge gained is fed back to better tune the spam filters
Recovery Action
Reports received from users were used to recognize security incidents that occurred on other targets that had not reported spam

Scenario


95% of spam is blocked
5% of spam is forwarded
Users report the spam they recognize, which is stored in a special repository, to conduct in-depth analysis
55% spam
45% non-spam
Anti-Spam Filters
Mail Server
Manual Inspection
Machine Learning estimates the impact of the spam email
The knowledge gained is fed back to better tune the spam filters
Recovery Action
Reports received from users were used to recognize security incidents that occurred on other targets that had not reported spam

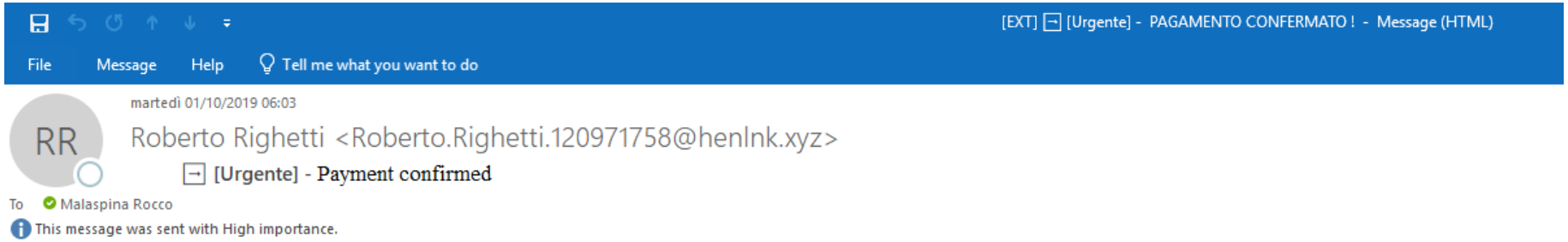
Dataset collected

22K (unique) spam emails reported by employees of the company (every day the new ones come in), labelled as one of the following classes:

- **Critical spam - Class 1, Positive**: spam emails that have created a security incident or at least required a defensive action to prevent future infections
- **Not relevant spam - Class 0, Negative**: spam emails with low or no degree of danger, and did not require any recovery action.

Examples


Deletion of Account - Message (HTML)
File
Message
Help
Tell me what you want to do
TA
mercoledì 14/08/2019 07:15
TIM Admin <sbpupil.ias@eircom.net>
Deletion of Account
To
No Reply
Dear TIM user,
We are deleting all emails that are no longer used by our database, in order to reduce congestion, please follow the link below to confirm that your email is in use.
Click here to update and prevent deletion.
If you do not do this within 24 hours, your email will be removed from our database.
Sorry about the inconvenience
Copyright © 2019 TIM Corporation. All rights reserved.



Hi there,
Look at the clock... What time is it?
Within a few hours from now, you should see your first ones credited thousand euros on your account.
Everything happens quickly, and in a big way.

Yes,
your life is about to come to a turning point, a positive turn...
You have to follow the
following steps to ensure that you are on board when this happens.

Start now by clicking right here

A pension? A trip? The
money to take care of your parents or send your kids to college?
Or all of it...?
All this is about to become a reality...

You and your family deserve it.
Come on
take a look now and you'll believe me.
Anyway, you don't have to pay for anything.

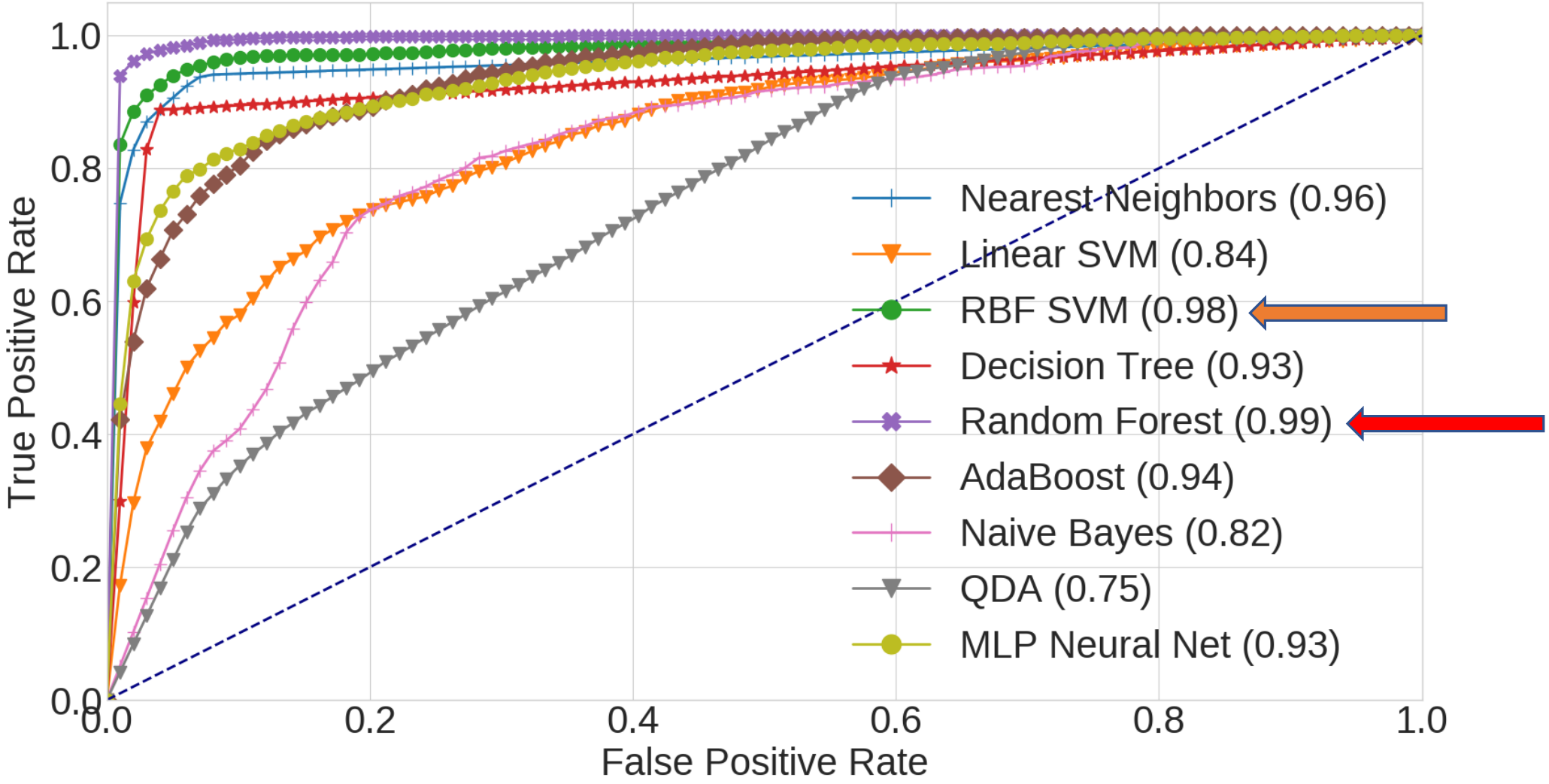
It's
one of those opportunities you really can't let slip through your fingers.
What are you waiting for?

Feature set

- 79 features coming from 8 different ***feature fields***

Field	Description
General	General information, mostly extracted from the smtp headers: if any smtp server is blacklisted, size of the mail, number of recipients etc, plus all those features that give us information about the email's origin and destination.
Content*	Features extracted from the text in the content of the email: language, number of words, number of deceiving words, number of disguised words, readability indexes, simplicity and correctness of the text etc. All the *Content* features have been calculated also on the text extracted with an Optical Character Recognition tool, generating the **Content_View** features (as described in the next feature field).
View	Features extracted from the screenshot of the email as it is displayed to the recipient: height and width of the screenshot, number of images, amount of text within the content but not read by the recipient etc.
Subject	Features extracted from the subject of the email: number of words, number of characters, if there are non-ASCII characters, if the email is forwarded or answered.
Links	Features about the links in the email: number, number of link domains, information from URL analysis service, etc.
Attachments	Features about the email attachments: number, type, size, information from sandboxes and antivirus, etc.
Other	Other types of information not in the previous fields: number of malicious entities known thanks to Threat Intelligence activities, role in the company of recipients, etc.

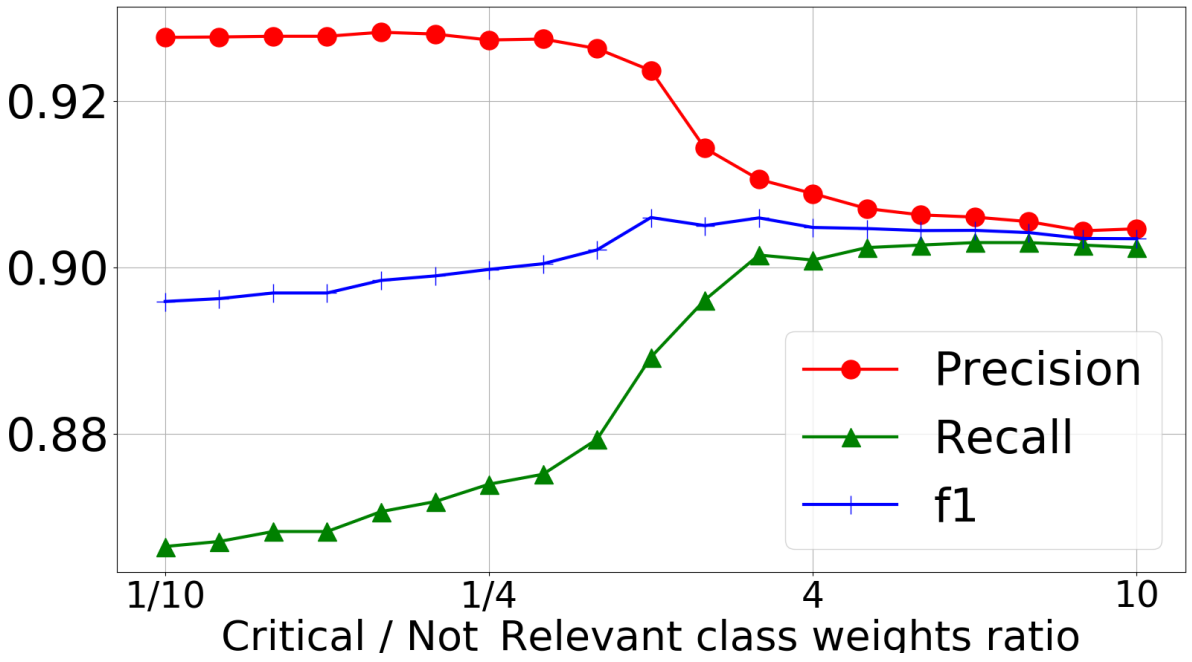
Selecting supervised Machine Learning models



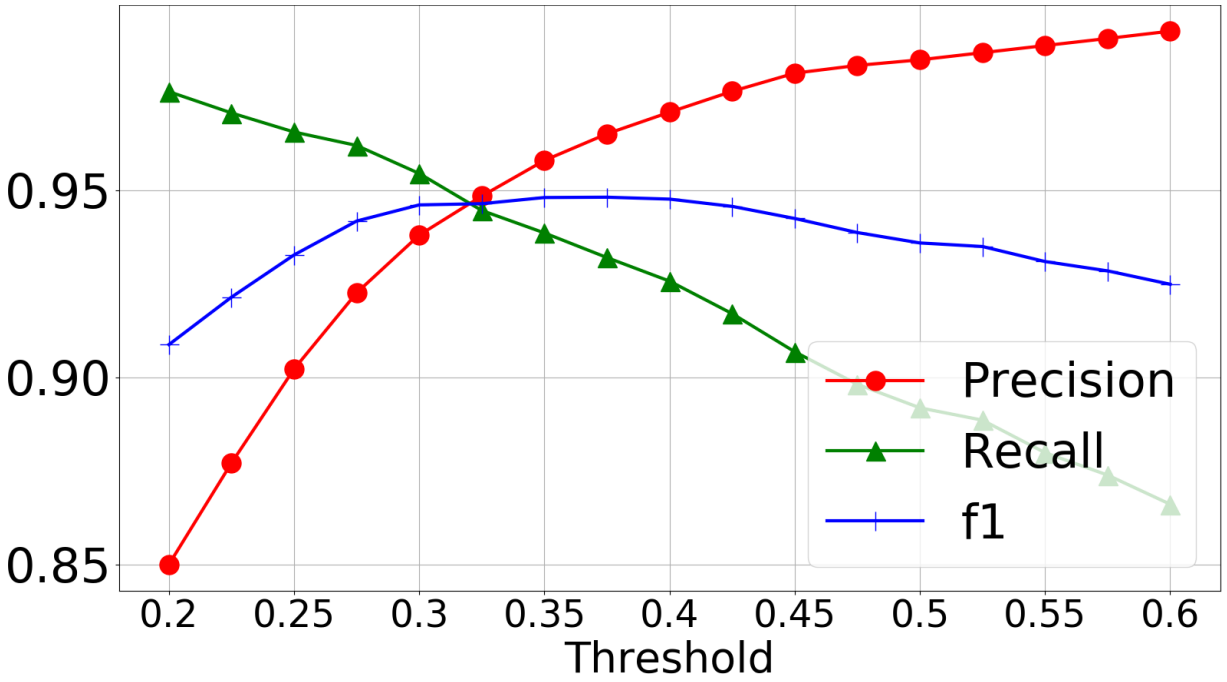
Support Vector Machine (RBF Kernel) and Random Forest have the best performance.

RBF SVM vs Random Forest

Support Vector Machines (Kernel RBF)



Random Forest

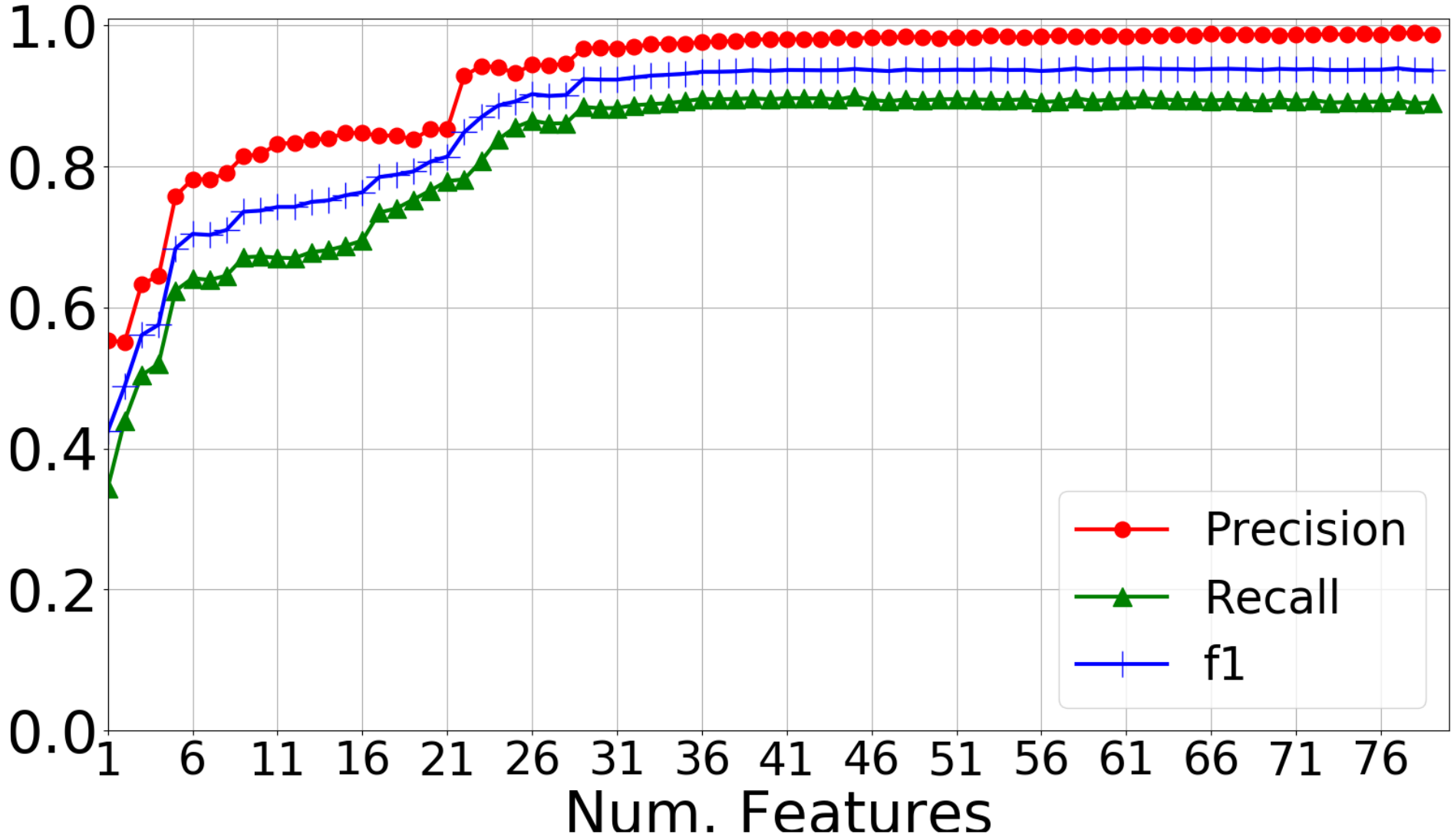


Random Forest shows better performance than RBF SVM

(after both class weights and threshold tuning process)

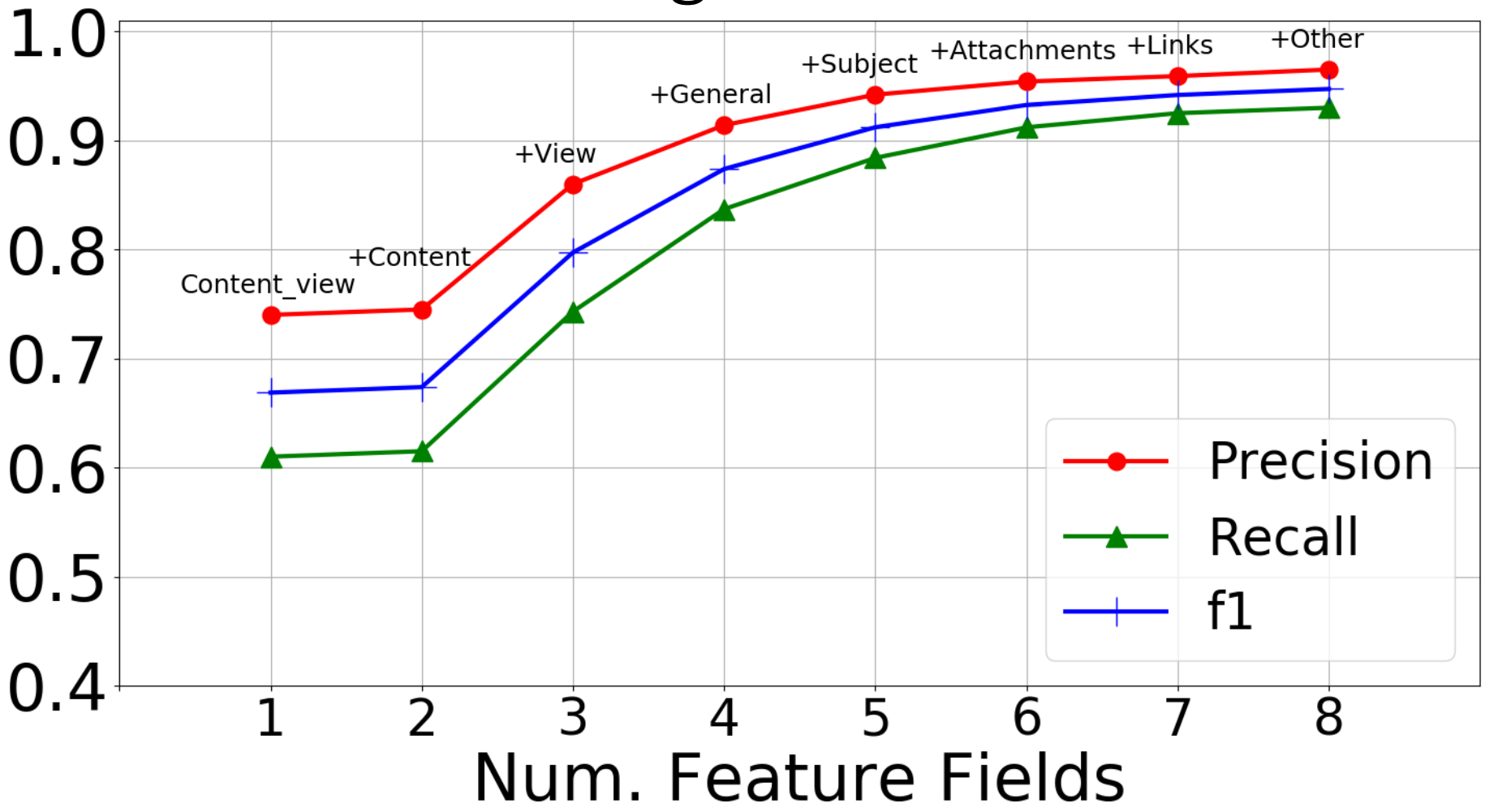
Which features really worked?

Performance using the best X features



36 features are enough to capture the problem.

Performance using the best X features



The view aspects and the content of the email are the most important traits to focus on (in order to create/detect an effective phishing email)

Evading the detector with adversarial samples

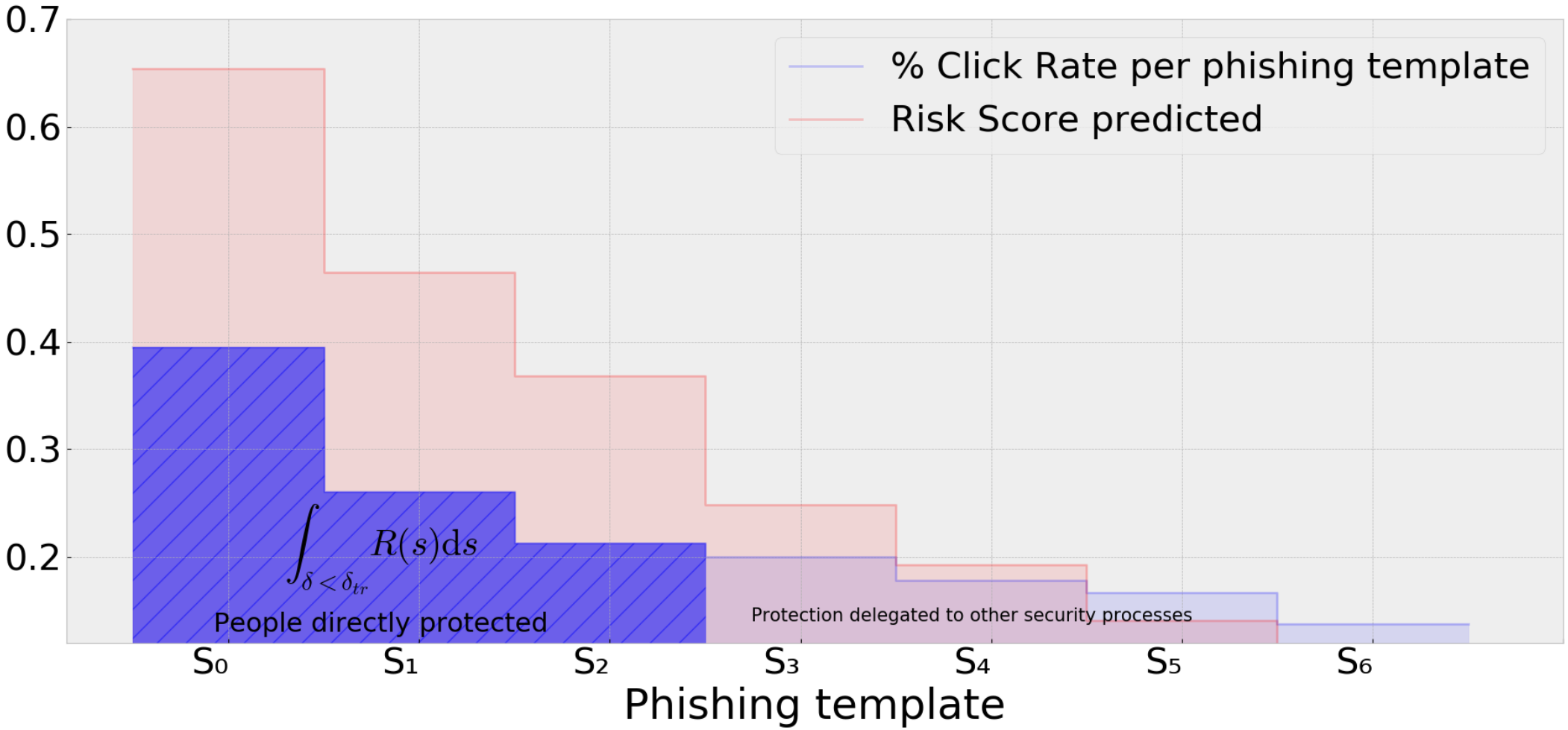
Data-driven security awareness

- Clustering of positive samples
- Starting from the most suited centroid, altering some of the features with a perturbation δ
- Seven adversarial samples representing the phishing templates used in our experiment have been obtained
- Such synthetic emails have been sent to a total of 41,154 people, of all levels of expertise, education and age. Each phishing template reached 5879 random people. The purpose is to measure the degree of success of each template.

Table – Manipulations performed to generate adversarial samples.

#	Manipulation
δ^0	No manipulation
δ^1	Alteration of the readability of the content by smudging the punctuation
δ^2	Alteration of the correctness of the content by injecting typing errors
δ^3	Deletion of hidden text (white text on white background)
δ^4	Remotion of deceiving words from the subject
δ^5	Dispersion of the deceiving message by adding a long block of text at the bottom of the content and words in the subject line
δ^6	Insertion of multiple points where to click by adding clickable images

Results of the awareness campaign experiment



Ongoing and future work

SPAMLEY

A system that allows the collection of data, useful data to understand which individuals are most vulnerable to which phishing emails

AIMS AND HYPOTHESES

1 EMAIL FEATURES

Email technical features and *persuasion principles* leverage user decision-making.

2 COGNITIVE TRAITS

The human being is considered "the weakest link". *Big-Five personality traits* are one way to explain why some people are more susceptible than others to phishing attacks.

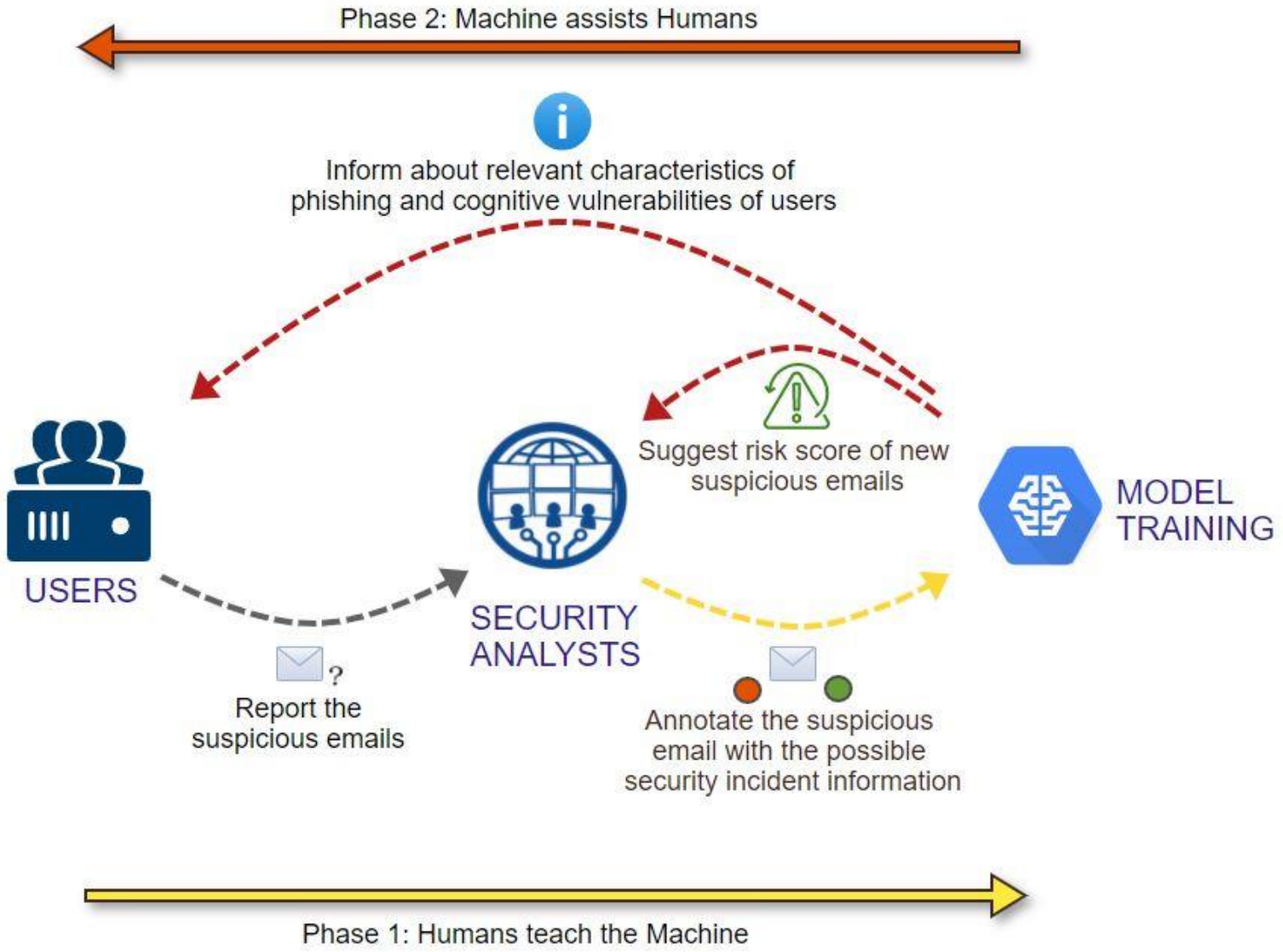
3 THEIR INTERACTION

Understanding the relationship between them could be the missing link to implementing adhoc anti-phishing courses to reduce users' susceptibility to phishing.

https://spamley.comics.unina.it/

Scan the QRcode to enjoy the email test

Machine and Human Learning collaborative approach



Conclusions

Support Vector Machine and Random Forest classifiers achieve the best performance

The full feature set considered allows to obtain up to 91,6% of recall and up to 95,2% of precision with supervised approaches

Highly dangerous spam emails can be detected with (only) 36 features

A large scale social experiment confirms the above points

This system has been integrated to help the Threat Intelligence processes of the partner company

A new dataset is about to be released to allow the scientific community to conduct technical-cognitive studies on phishing

Publications

- **Luigi Gallo**, Alessandro Maiello, Alessio Botta, Giorgio Ventre, 2 Years in the anti-phishing group of a large company, Computers & Security, Volume 105, 2021, 102259, ISSN 0167-4048, https://doi.org/10.1016/j.cose.2021.102259.
- **Luigi Gallo**, Alessio Botta, and Giorgio Ventre. 2019. Identifying threats in a large company's inbox. In Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks (Big-DAMA '19). Association for Computing Machinery, New York, NY, USA, 1–7. DOI:https://doi.org/10.1145/3359992.3366637
- Antonia Affinito, Alessio Botta, **Luigi Gallo**, Mauro Garofalo, and Giorgio Ventre. 2020. Spark-based port and net scan detection. Proceedings of the 35th Annual ACM Symposium on Applied Computing. Association for Computing Machinery, New York, NY, USA, 1172–1179. DOI:https://doi.org/10.1145/3341105.3373970
- A. Botta, **L. Gallo** and G. Ventre, "Cloud, Fog, and Dew Robotics: Architectures for Next Generation Applications," 2019 7th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud), 2019, pp. 16-23, doi: 10.1109/MobileCloud.2019.00010.

In preparation

- “DewROS: a Platform for Informed Dew Robotics in ROS” (tentative title)
- “Security testing methodologies for Network Traffic Analyzers” (tentative title)
- “A game-based platform for phishing awareness testing” (tentative title)

Credits Summary

	Credits year 1								Credits year 2								Credits year 3									
		1	2	3	4	5	6			1	2	3	4	5	6			1	2	3	4	5	6			
	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Total	Check
Modules	20	1,6	0	3	0	6	5	**15,6**	14	0	3	0	9	2,8	0	**14,8**	10	0	0	3	4	0	0	**7**	**37,4**	**30-70**
Seminars	5	0,4	0	0,4	1,5	0	0	**2,3**	6	0,4	0	1,6	2,7	4	1,4	**10,1**	7	1,3	0,4	0,7	0,4	0	0	**2,8**	**15,2**	**10-30**
Research	35	8	10	6,6	8,5	4	5	**42,1**	40	9,6	7	8,4	3,3	0,7	6,1	**35,1**	43	8,7	9,6	6,3	5,6	10	10	**50,2**	**127,4**	**80-140**
	60	10	10	10	10	10	10	**60**	60	10	10	10	15	7,5	7,5	**60**	60	10	10	10	10	10	10	**60**	**180**	**180**

THANKS !

Backup Slides

Annotation consistency evaluation

		Analyst 1		Row Marginals	
		Positive	Negative		
Analyst 2	Positive	42	10	52	rm^1
	Negative	4	207	211	rm^2
Column Marginals		46	217	263	
		cm^1	cm^2		n

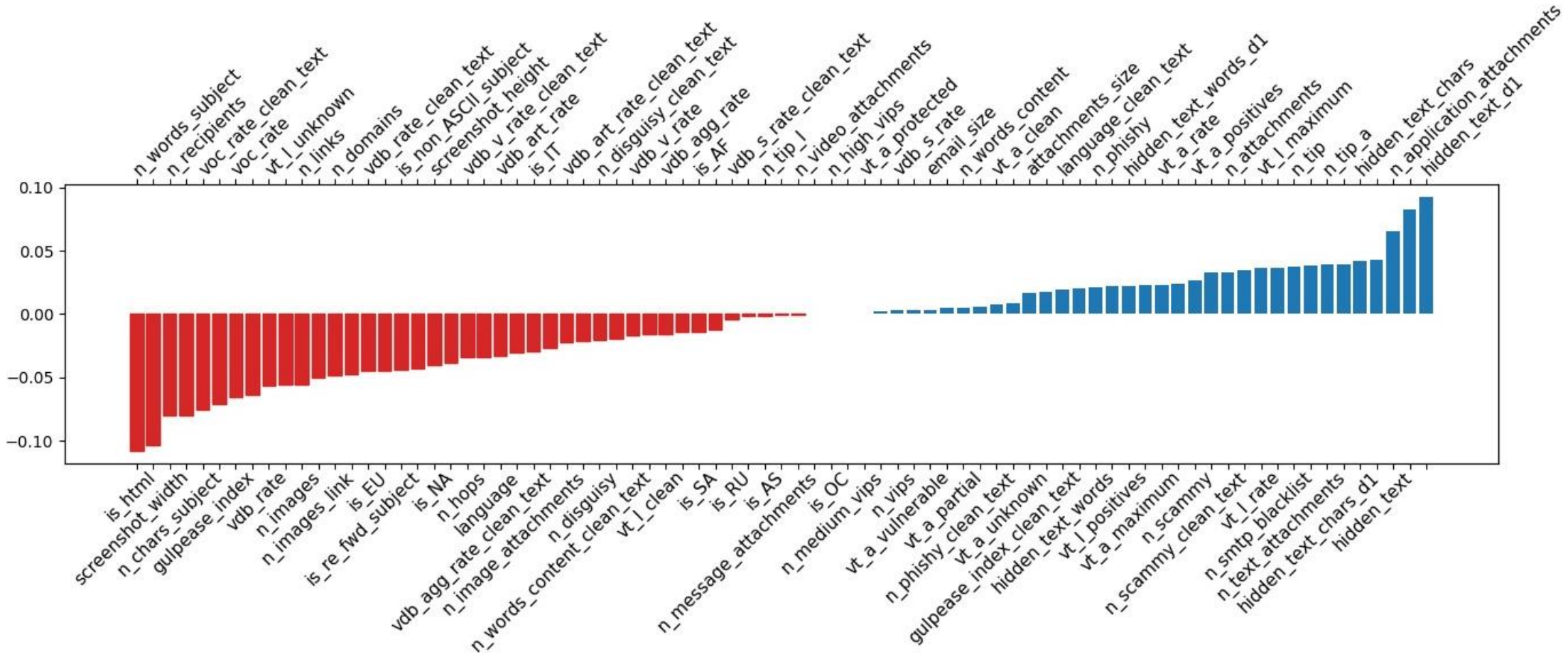
Percentage of agreement: **94,67%**

Kappa statistic: **83,3%**

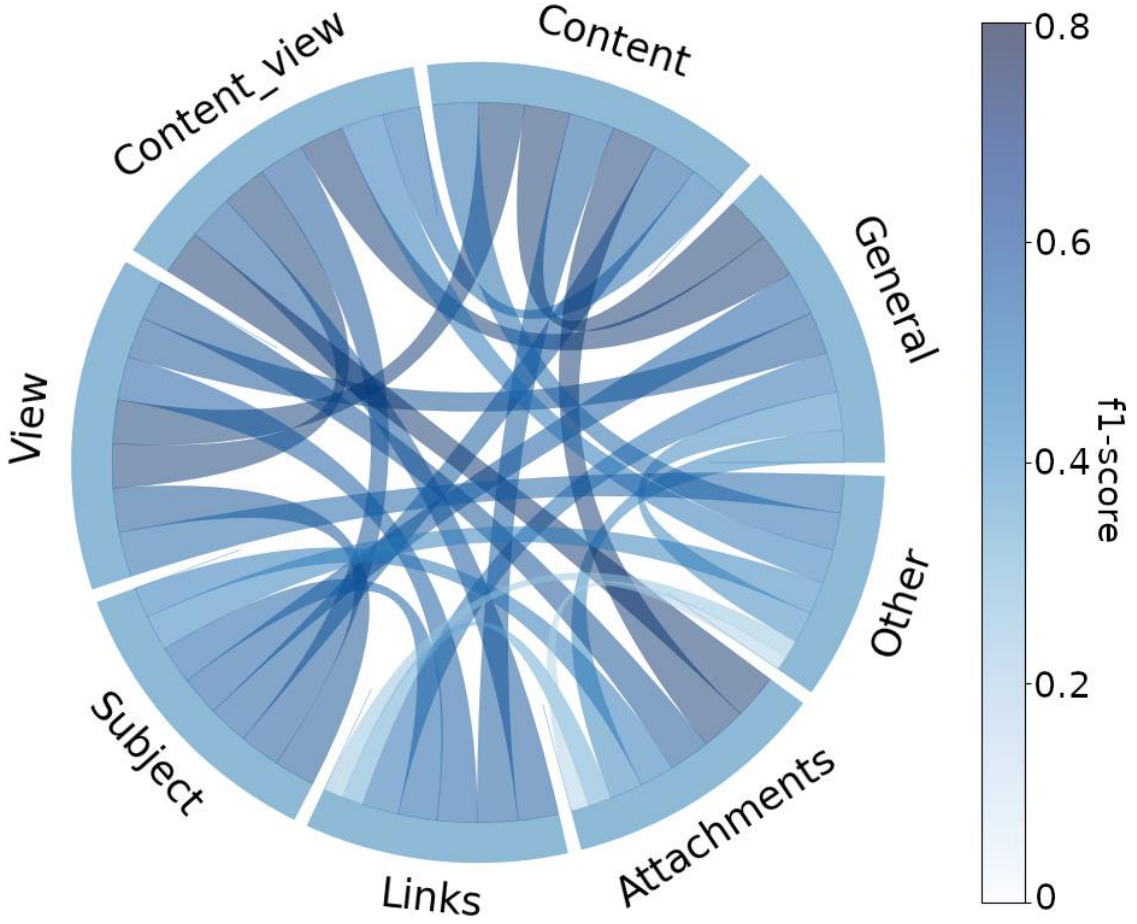
Full Feature set

Field	Feature	Description
General	is_html	if it is an html mail
	n_smtp_blackist	the number of smtp servers traversed in the blacklists
	email_size	the size of the email
	n_recipients	the number of recipients
	n_hops	the number of SMTP hops
	is_IT	if the email comes from Italy
	is_EU	if the email comes from Europe
	is_NA	if the email comes from North America
	is_SA	if the email comes from South America
	is_RU	if the email comes from Russia
	is_AS	if the email comes from Asia
	is_AF	if the email comes from Africa
	is_OC	if the email comes from Oceania
Content[3]	language[3]	the language of the mail
	voc_rate[3]	the rate of words of the content in the vocabulary
	vdb_rate[3]	the rate of words of the content within the basic vocabulary
	vdb_agg_rate[3]	the rate of adjectives within the content
	vdb_v_rate[3]	the rate of verbs within the content
	vdb_s_rate[3]	the rate of nouns within the content
	vdb_art_rate[3]	the rate of articles within the content
	gulpease_index[3]	readability index (Italian - Gulpease index [27], English - Flesch formula [15])
	n_words_content[3]	number of words in the content
	n_disguisy[3]	number of disguised words in the entire email (content, subject, address)
	n_phishy[3]	number of deceiving words, related to phishing, in the content and subject
	n_scammy[3]	number of deceiving words, related to scamming, in the content and subject
View	screenshot_width	the width of the email as it is displayed to the recipient
	screenshot_heigth	the heigth of the email as it is displayed to the recipient
	n_images	number of images
	n_images_links	number of images as links
	hidden_text[4]	percentage of text in the content not displayed to the recipient
	hidden_text_words[4]	number of words in the content not displayed to the recipient
	hidden_text_chars[4]	number of characters in the content not displayed to the recipient
Subject	n_words_subject	number of words in the subject
	n_char_subject	number of characters in the subject
	is_non_ASCII_subject	if the object contains non-ASCII characters
	is_re_fwd_subject	if the email is replied or forwarded
Links	n_links	number of links
	n_domains	number of link domains
	vt_l_rate	rate of links considered malicious by at least one engine of VirusTotal
	vt_l_maximum	maximum number of VirusTotal engines that consider a link as malicious
	vt_l_positives	number of links considered malicious by at least one engine of VirusTotal
	vt_l_clean	number of links not considered malicious by all engines VirusTotal
	vt_l_unknown	number of unknown links to VirusTotal
Attachments	n_attachments	number of attachments
	n_image_attachments	number of image type attachments
	n_application_attachments	number of application type attachments
	n_message_attachments	number of message type attachments
	n_text_attachments	number of text type attachments
	n_video_attachments	number of video type attachments
	attachments_size	average size of attachments
	vt_a_rate	rate of attachments considered malicious by at least one engine of VirusTotal
	vt_a_maximum	maximum number of VirusTotal engines that consider an attachment as malicious
	vt_a_positives	number of attachments considered malicious by at least one engine of VirusTotal
	vt_a_clean	number of attachments not considered malicious by all VirusTotal engines
	vt_a_vulnerable	number of attachments considered malicious by VirusTotal engines not including corporate antivirus
	vt_a_partial	number of attachments considered partially malicious by VirusTotal engines not including corporate antivirus
	vt_a_protected	number of attachments considered malicious by VirusTotal engines including corporate antivirus
	vt_a_unknown	number of unknown attachments to VirusTotal
Other	n_tip	number of entities in TIP
	n_tip_a	number of attachments in TIP
	n_tip_l	number of links in TIP
	n_vips	the number of vips among the recipients
	n_medium_vips	the number of managers among the recipients
	n_high_vips	the number of top managers among the recipients

Feature importance



f1-score of all possible pairs of feature fields



Clean text extraction scheme