#SOMOS2030

#CátedrasCiber

#ProyectosCiber

AI in Cybersecurity: Actual Applications and Future

Rafael Pastor Vargas

Esta iniciativa se realiza en el marco de los fondos del Plan de Recuperación, Transformación y Resiliencia, financiadas por la Unión Europea (Next Generation), el proyecto del Gobierno de España que traza la hoja de ruta para la modernización de la economía española, la recuperación del crecimiento económico y la creación de empleo, para la reconstrucción económica sólida, inclusiva y resiliente tras la crisis de la COVID19, y para responder a los retos de la próxima década

Financiado por la Unión Europea NextGenerationEU

GOBIERNO DE ESPAÑA | MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA | SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

Plan de Recuperación, Transformación y Resiliencia

incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD

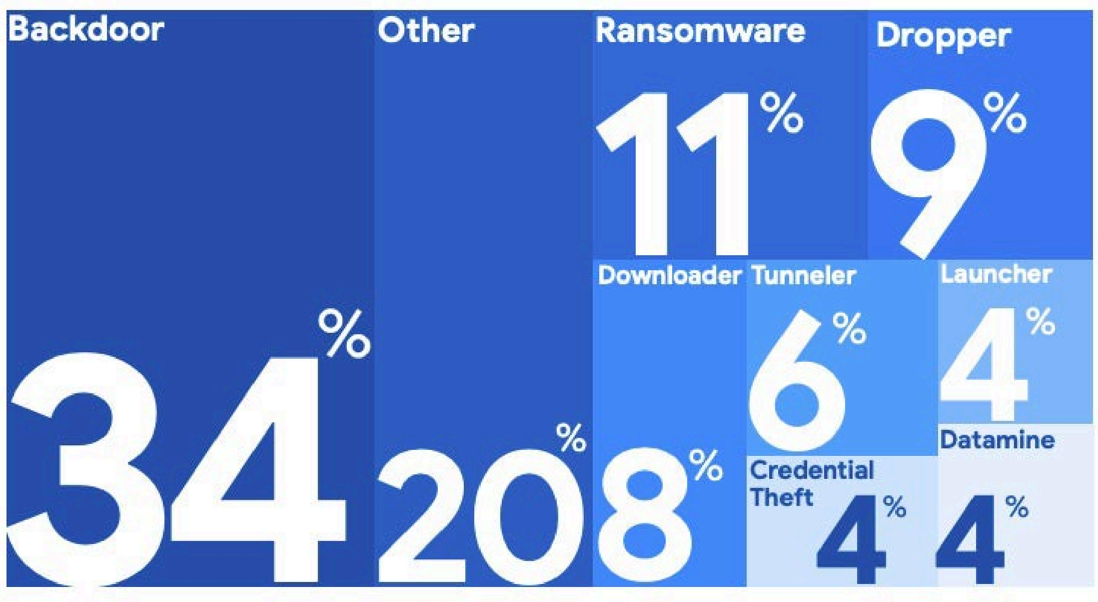
UNED

Índice

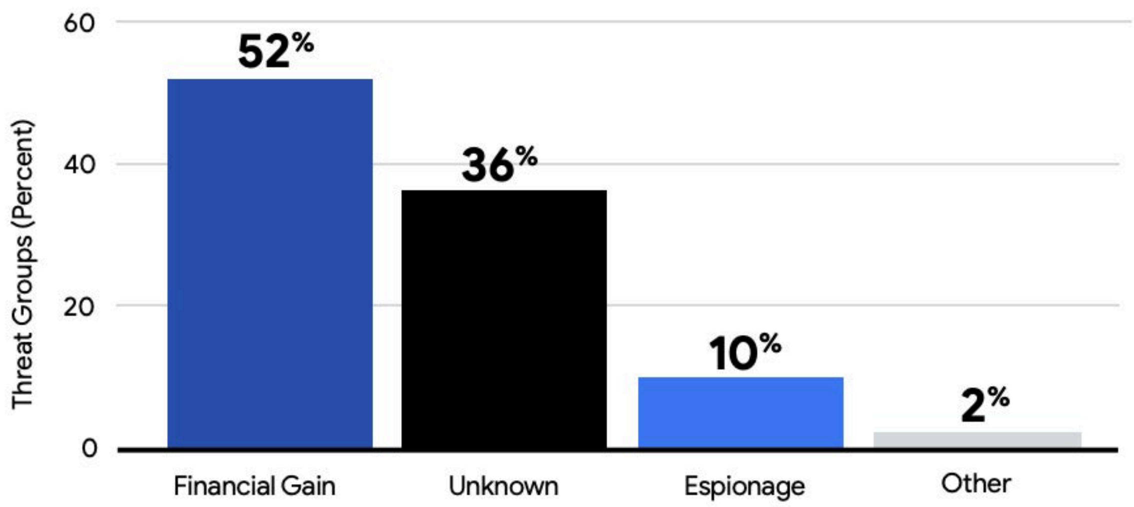
https://www.techrepublic.com/article/cyber-security-trends-google-report/

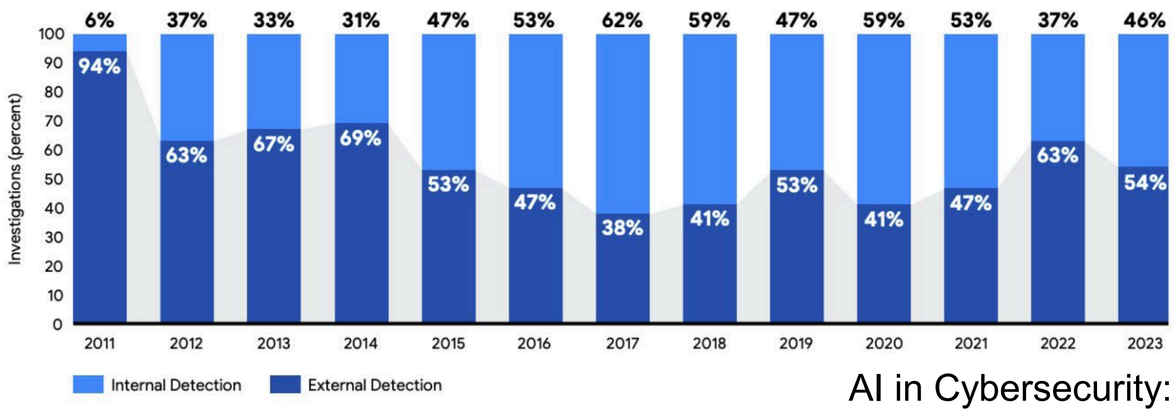
Context (I/II)

Observed Malware Families by Category, 2023



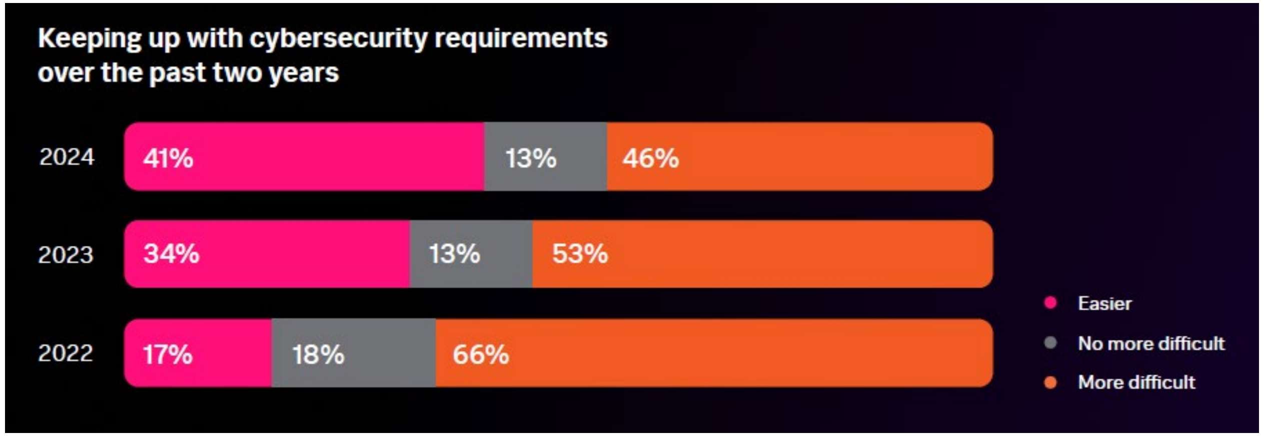
Observed Threat Groups by Goal, 2023

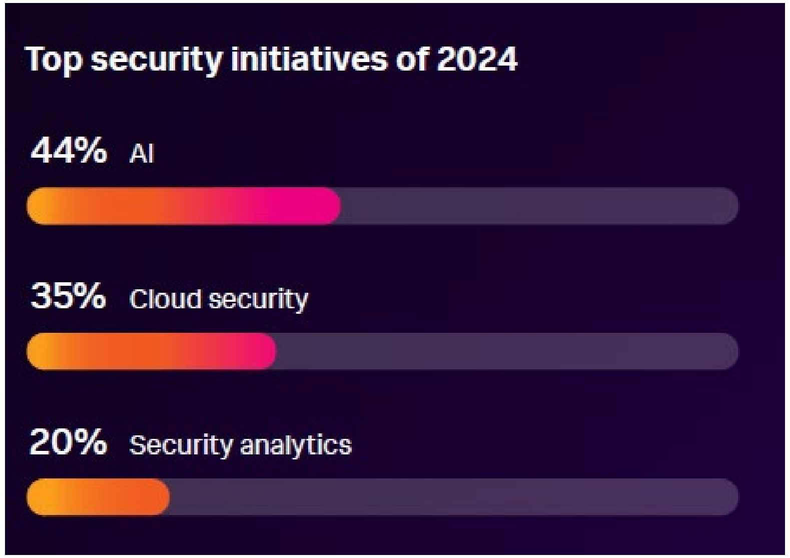


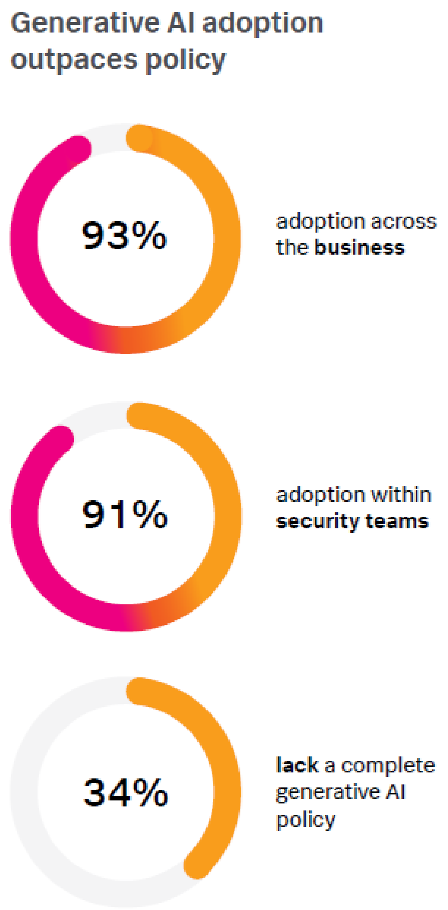


Context (II/II)

State of Security: The Race to Harness AI (Splunk)







Gen AI: Use cases in Cybersecurity (I/II)

What generative AI use cases may look like in practice

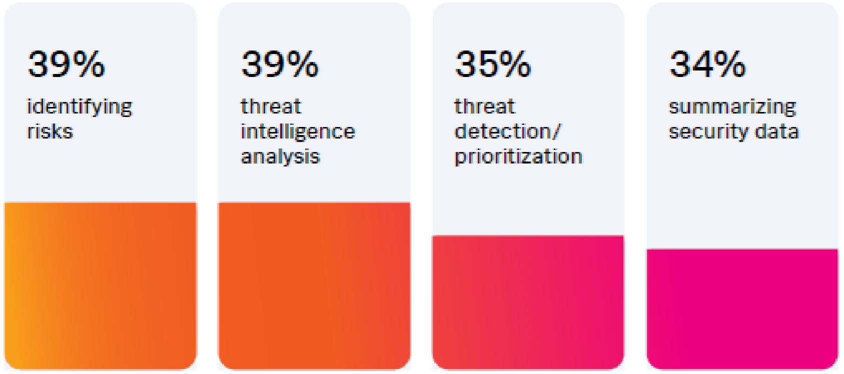
Identifying risks
Generative AI can enhance risk-based alerting by quickly aggregating diverse datasets to provide security analysts with alerts that are context-rich. Large language models (LLMs) help to deliver this information at a speed and efficiency far beyond human capability.

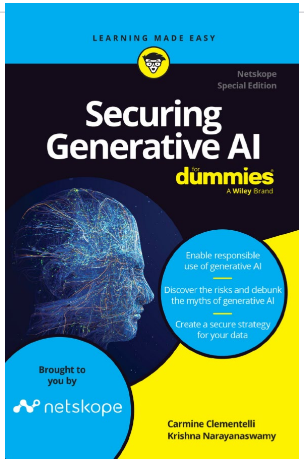
Threat intelligence analysis
LLMs can determine the indicators of compromise and MITRE ATT&CK techniques described in a threat intelligence report. This would save intelligence teams from a lot of drudgery and enable them to perform deeper analysis faster.

Threat detection and prioritization
Prioritizing and triaging alerts are tasks particularly susceptible to analyst misclassification, fatigue and human errors. Generative AI can parallel process multiple threats while improving accuracy.

Summarizing security data
Generative AI can summarize quickly, thoroughly and accurately to help security teams save time and keep up with news and information, like Biden's Executive Order on Improving the Nation's Cybersecurity.

Top generative AI cybersecurity use cases





Gen AI: Use cases in Cybersecurity (II/II)

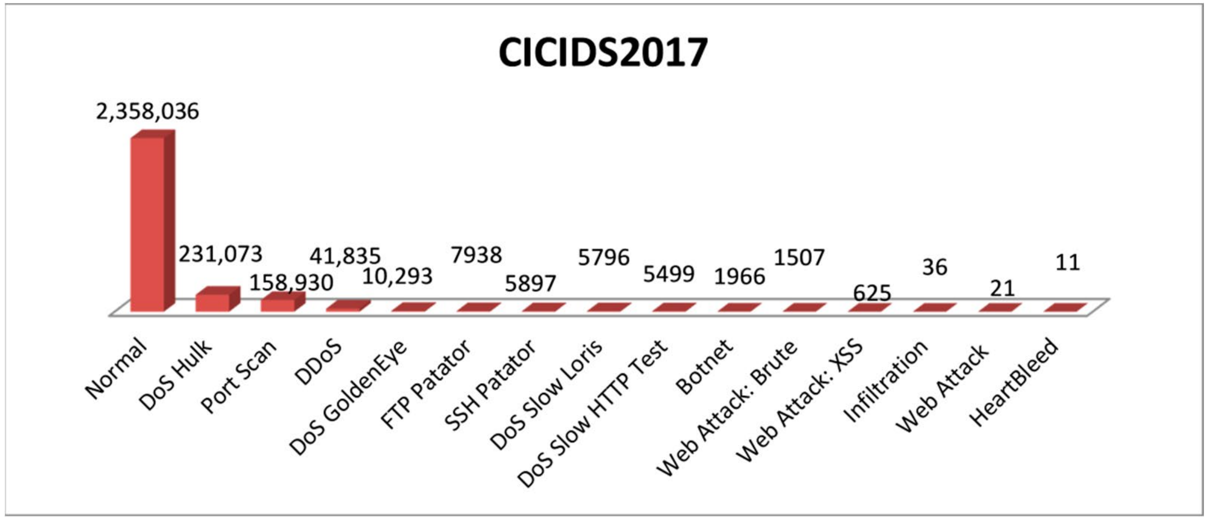
- Red teams are using AI and large language models. In 2023, Mandiant consultants used generative AI tools to speed up certain activities in red team assessments, including:
  - The creation of initial drafts of malicious emails and landing pages for faux social engineering attacks.
  - The development of custom tooling for when analysts encounter uncommon or new applications and systems.
  - The research and creation of tooling in cases where environments do not fit the operational norm that can be used again and again.

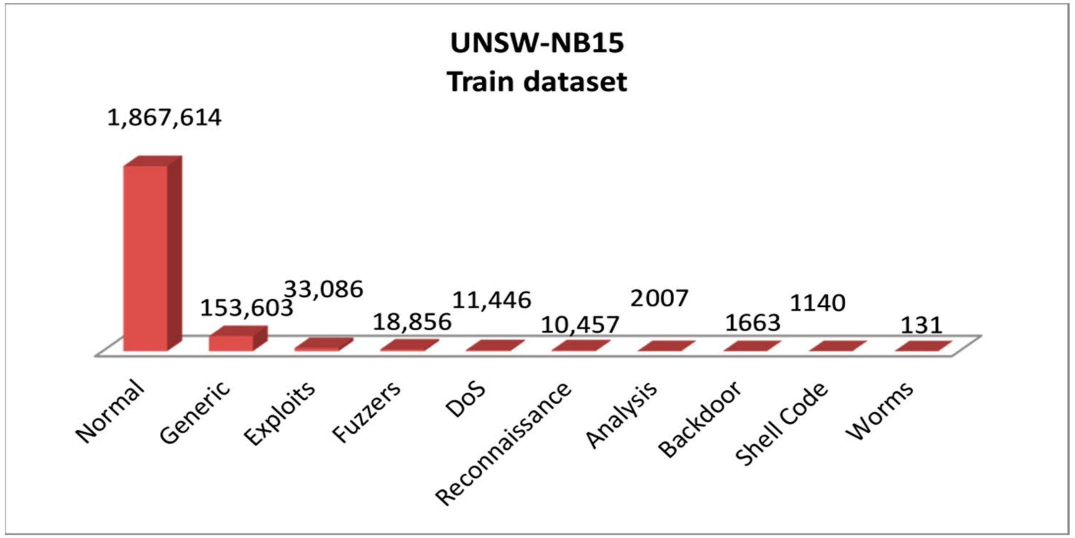
Common AI application scenarios

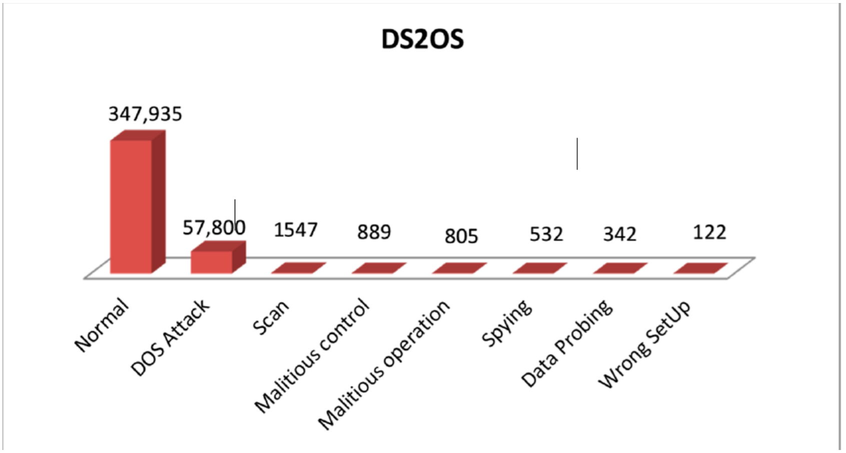
- Threat detection and response
  - Network Intrusion Detection: AI-driven intrusion detection systems can **monitor network traffic**, identify suspicious activities, and **detect intrusions** from various attack vectors like malware, phishing attempts, and **brute-force attacks**.
  - Behavioral Analysis: AI algorithms can analyze user behavior and identify deviations from normal patterns, enabling the detection of insider threats or compromised accounts.
  - Advanced Malware Detection: AI can recognize previously unknown **malware patterns and behaviors**, facilitating early detection and containment.

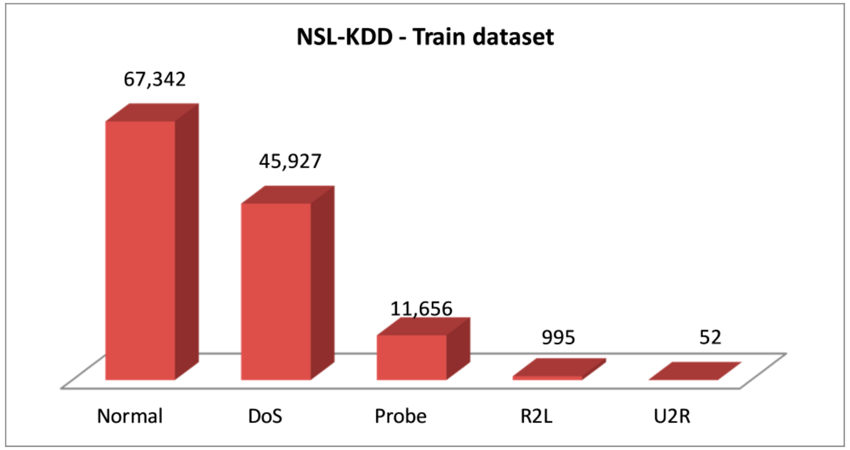
Datasets

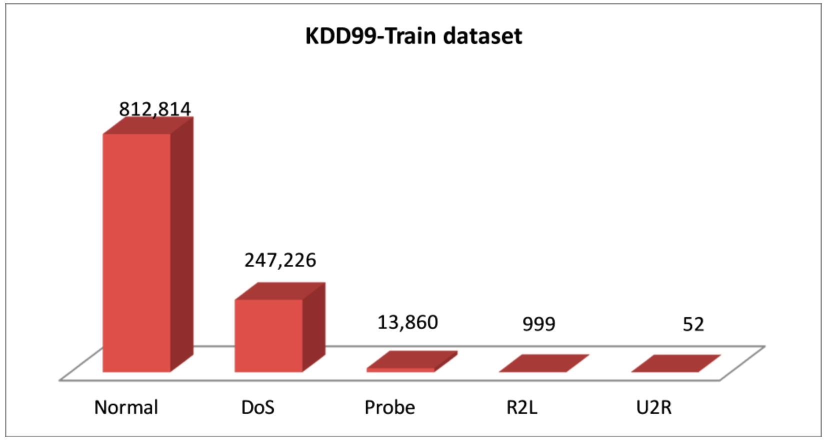
Alshaibi, A.; Al-Ani, M.; Al-Azzawi, A.; Konev, A.; Shelupanov, A. The Comparison of Cybersecurity Datasets. *Data* **2022**, *7*, 22. https://doi.org/10.3390/data7020022











IoT Attacks Identification

Alshaibi, A.; Al-Ani, M.; Al-Azzawi, A.; Konev, A.; Shelupanov, A. The Comparison of Cybersecurity Datasets. *Data* **2022**, *7*, 22. https://doi.org/10.3390/data7020022

Reference	ML Technique	IoT Attacks	Dataset	Accuracy
[46]	OS-ELM	Dataset Multiple	NSL-KDD	97.3
[47]	NN	DOS, U2R and R2L.	NSL-KDD	82.3
[35]	DT and NB.	Probing, U2R and R2L.	NSL-KDD	85.8
[23]	TAB	DoS Flooding	KDD99	99.95
[35]	DT	DOS, Reconnaissance U2R, R2L., Backdoor	KDD99	98
[26]	Ensemble Learning	Malware	AndroZoo, Drebin	94
[48]	DT.	DOS	RPL-NIDDS17	98.1
[47]	DT	DOS, Reconnaissance U2R, R2L.	UNSW-NB15	97.8
[21]	NN.	Probing, U2R and R2L	NSL-KDD	99.2
[47]	DT	DoS Reconnaissance, U2R, R2L.	NSL-KDD	98
[49]	NN.	DOS, reconnaissance and DDOS	BoT-IoT	98.26
[19]	LSSVM	Anomaly	KDD99	99.7
[27]	DFEL	Dataset Multiple	UNSW-NB15,	98.5
[21]	LSTM	DoS Flooding	ISCX2012,	99.9
[24]	Adaboost	Botnet Flooding	UNSW-NB15	99.5

A new review for DNN/Datasets

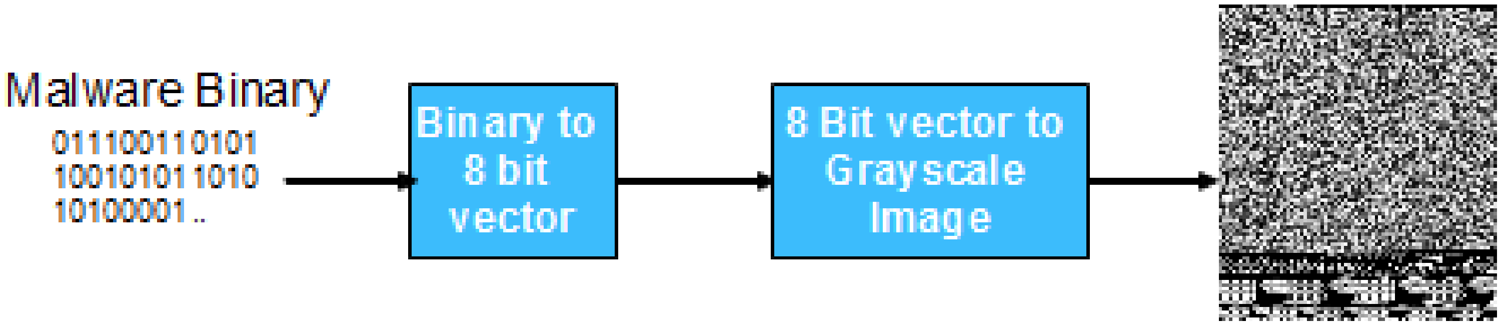
- RNN (LSTM, GRU, Transformers)

ARCHITECTURE	NUMBER OF PAPERS	PAPERS
LSTM [1]	13	[2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14]
Simple DNN ad-hoc	10	[15] [16] [17] [18] [19] [20] [21] [22] [23] [24]
CNN [25]	9	[3] [26] [8] [27] [28] [29] [30] [12] [13] [16]
GRU [31]	5	[3] [6] [8] [10] [12]
FFNN [32]	5	[7] [27] [28] [33]
RNN [34]	4	[2] [3] [8] [28]
Autoencoder [35]	4	[36] [37] [38] [39]
Federated Learning [40]	3	[41] [30] [42]
DenseNet [43]	2	[44] [27]
ResNet [45]	2	[27] [10]
HDBN [46]	1	[47]
ACID [48]	1	[49]
RandNN [50]	1	[7]
DQN [51]	1	[9]
GNN [52]	1	[53]
cGAN [54]	1	[30]
MobileNetV3 [55]	1	[56]

DATASETS	NUMBER OF PAPERS	PAPERS
UNSW-NB15 Dataset [57]	8	[44] [14] [39] [4] [18] [20] [37] [56]
TON_IOT Dataset [58]	8	[16] [14] [42] [19] [22] [26] [9] [53]
ISCX NSL-KDD 2009 [59]	7	[17] [47] [4] [20] [28] [29] [23]
CICIDS2017 Dataset [60]	6	[44] [42] [4] [10] [12] [56]
CSE-CIC-IDS2018 [60]	5	[49] [28] [29] [23] [56]
Edge-IIoTset dataset [61]	4	[17] [18] [33] [30]
Bot-IoT Dataset [62]	4	[44] [26] [53] [29]
Original Dataset	3	[8] [36] [12]
KDD Cup 1999 [63]	3	[49] [19] [37]
ISCXIDS2012 [64]	2	[49] [10]
X-IIoTID [65]	2	[19] [37]
CIC-IoT-Dataset-2022 [66]	2	[15] [7]
CIC-DDoS2019 [67]	2	[11] [13]
MalwareTextDB [68]	1	[2]
IEC 69870-5-104 Dataset [69]	1	[15]
MSU-ORNL PS Dataset [70]	1	[38]
Drebin-215 [71]	1	[17]
MQTTset dataset [72]	1	[18]
CIC IoT dataset 2023 [73]	1	[18]
IoT-ID [74]	1	[24]
Kitsune [75]	1	[5]
WSN-DS [76]	1	[6]
FLDIDSPN [77]	1	[41]
Survival [78]	1	[21]
SoReL-20M [79]	1	[27]
EMBER dataset [80]	1	[27]
NF-Ton IoT [81]	1	[53]
NF-Bot IoT [82]	1	[53]
CIRA-CIC-DoHBrw-2020 [83]	1	[10]
InSDN [84]	1	[12]
Kitsune Network Attack Dataset [85]	1	[12]

ML/DL based Malware Detection

- Datasets
  - MALWARE-TRAFFIC-ANALYSIS.NET: https://www.malware-traffic-analysis.net/
  - VIRUSTOTAL: https://www.virustotal.com
  - VirusShare: https://virusshare.com
  - theZoo: https://github.com/ytisf/theZoo (defined by the authors as a repository of live malware for your own joy and pleasure)
- Several ML algorithms
  - DNN (CNN most common)



ENISA REPORT (2021): SECURING MACHINE LEARNING ALGORITHMS
https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms

Attacks to IA: Adversarial Machine Learning

		Stage of the lifecycle									
Threats | *sub-threats*	Definition	Data Collection	Data Cleaning	Data Preprocessing	Model design	Model Training	Model Testing	Optimisation	Model Evaluation	Model Deployment	Monitoring
Evasion	A type of attack in which the attacker works on the ML algorithm's inputs to find small perturbations leading to large modification of its outputs (e.g. decision errors). It is as if the attacker created an optical illusion for the algorithm. Such modified inputs are often called adversarial examples. Example: the projection of images on a house could lead the algorithm of an autonomous car to take the decision to suddenly make it brake.										x
Use of adversarial examples crafted in white or grey box conditions (e.g. FGSM…)	In some cases, the attacker has access to information (model, model parameters, etc.) that can allow him to directly build adversarial examples. One example is to directly use the model's gradient to find the best perturbation to add to the input data to evade the model.										x

ENISA REPORT (2021): SECURING MACHINE LEARNING ALGORITHMS
https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms

Attacks to IA: Adversarial Machine Learning

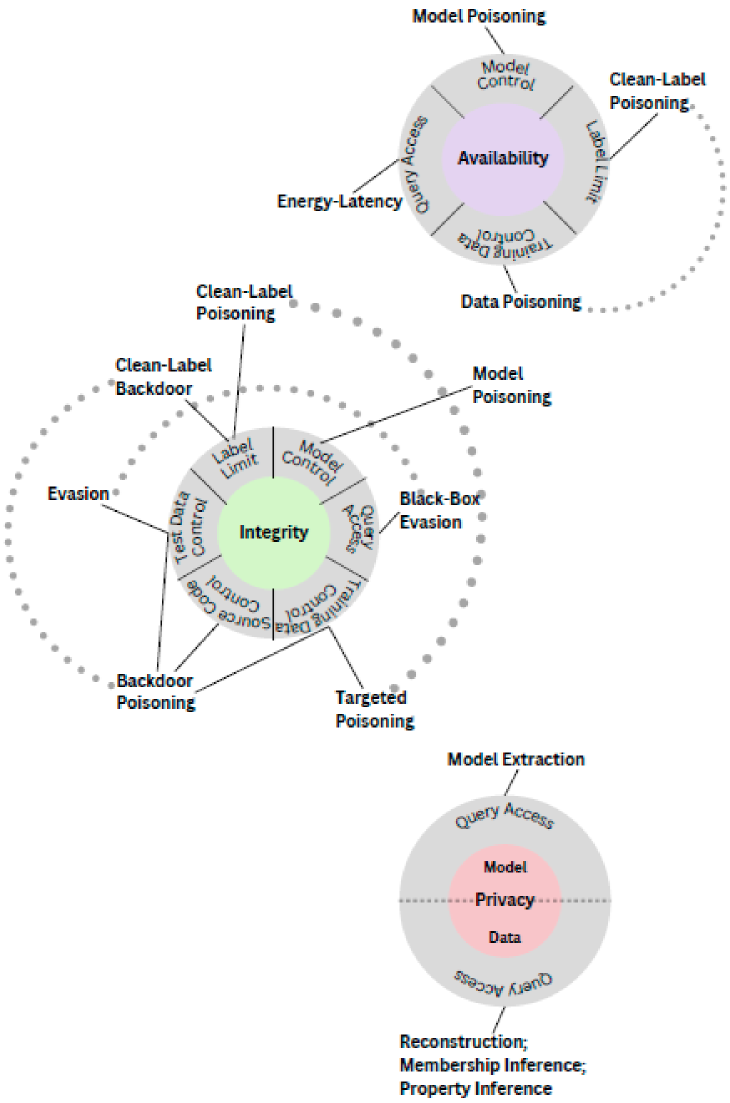
		Stage of the lifecycle									
Threats | *sub-threats*	Definition	Data Collection	Data Cleaning	Data Preprocessing	Model design	Model Training	Model Testing	Optimisation	Model Evaluation	Model Deployment	Monitoring
Oracle	A type of attack in which the attacker explores a model by providing a series of carefully crafted inputs and observing outputs. These attacks can be previous steps to more harmful types, evasion or poisoning for example. It is as if the attacker made the model talk to then better compromise it or to obtain information[204] about it (e.g. model extraction) or its training data (e.g. membership inferences attacks and Inversion attacks). Example: an attacker studies the set of input-output pairs and uses the results to retrieve training data.										x
Poisoning	A type of attack in which the attacker altered data or model to modify the ML algorithm's behavior in a chosen direction (e.g. to sabotage its results, to insert a backdoor). It is as if the attacker conditioned the algorithm according to its motivations. Such attacks are also called causative attacks. Example: massively indicating to an image recognition algorithm that images of dogs are indeed cats to lead it to interpret it this way.	x	x	x	x	x		x		x	x
Label modification	An attack in which the attacker corrupts the labels of training data. This sub-threat is specific to Supervised Learning.	x	x	x		x					

Attacks to IA: Adversarial Machine Learning

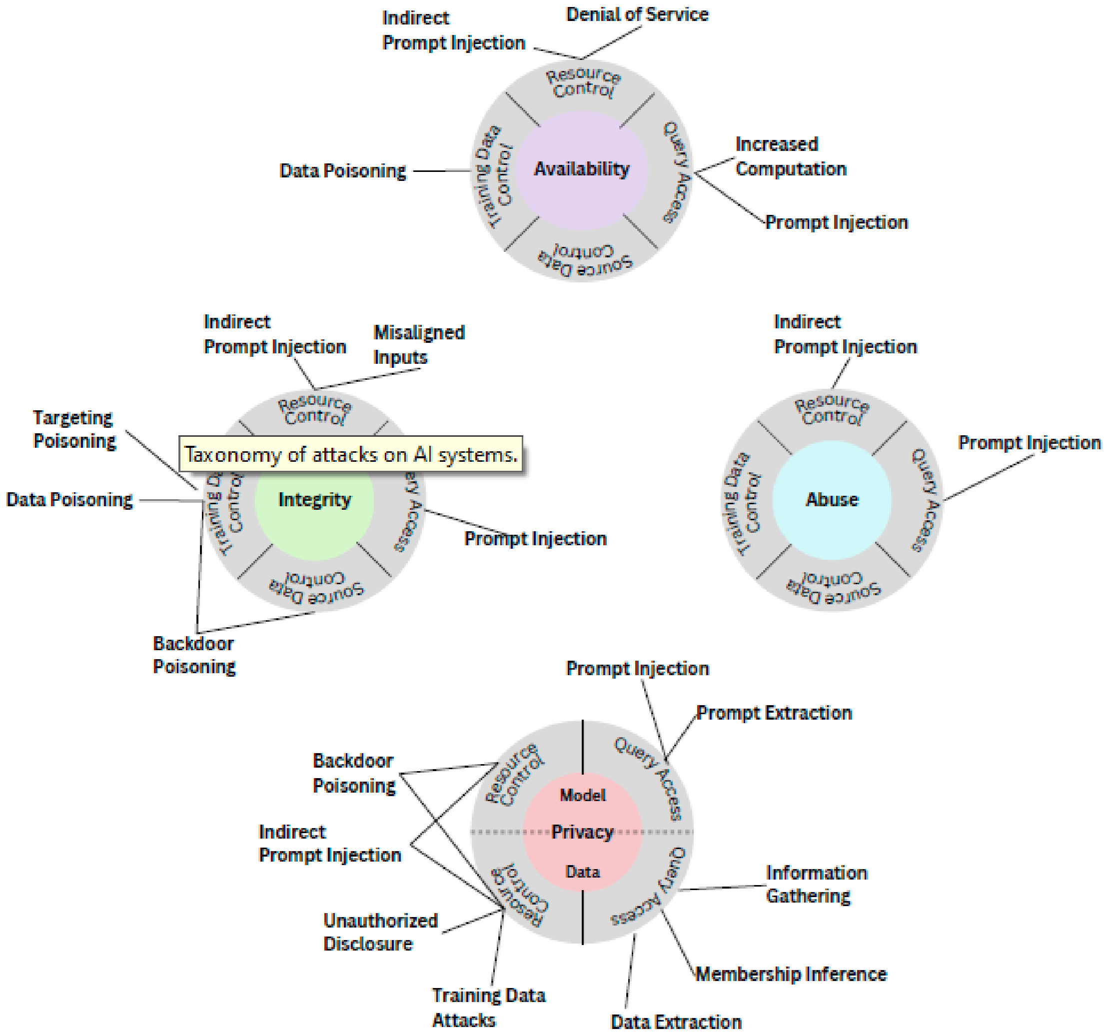
Threats | *sub-threats*	Definition	Stage of the lifecycle									
		Data Collection	Data Cleaning	Data Preprocessing	Model design	Model Training	Model Testing	Optimisation	Model Evaluation	Model Deployment	Monitoring
Model or data disclosure	This threat refers to the possibility of leakage of all or partial information about the model.[12] Example: the outputs of a ML algorithm are so verbose that they give information about its configuration (or leakage of sensitive data)	x	x	x	x	x	x	x	x	x	x
Data disclosure	This threat refers to a leak of data manipulated by ML algorithms. This data leakage can be explained by an inadequate access control, a handling error of the project team or simply because sometimes the entity that owns the model and the entity that owns the data are distinct. To train the model, it is often necessary for the data to be accessed by the model provider. This involve sharing the data and thus share sensitive data with a third party.	x	x	x	x	x	x	x	x	x	x
Model disclosure	This threat refers to a leak of the internals (i.e. parameter values) of the ML model. This model leakage could occur because of human error or contraction with a third party with a too low security level.				x	x	x	x	x	x	x

NIST Trustworthy and Responsible AI, NIST AI 100-2e2023

Adversarial Machine Learning, A Taxonomy and Terminology of Attacks and Mitigations



Taxonomy of attacks on Predictive AI systems.



Taxonomy of attacks on Generative AI systems

Some projects: Privacy and Geolocation protection

- Personal/individual behavioral modeling (Am I safe here?).
  - Geographic safe zones.
  - Anomaly detection
  - “Unsafe” geographic zones
  - Perception of insecurity and personal history (warnings).
- Grouping by zones (cities, neighborhoods, etc.)
  - Unsafe ”geographic” zones
  - Limitations/ProblemBig
  - DataVery heterogeneous trajectories
  - Preservation of anonymity
  - Non-specific data semantics

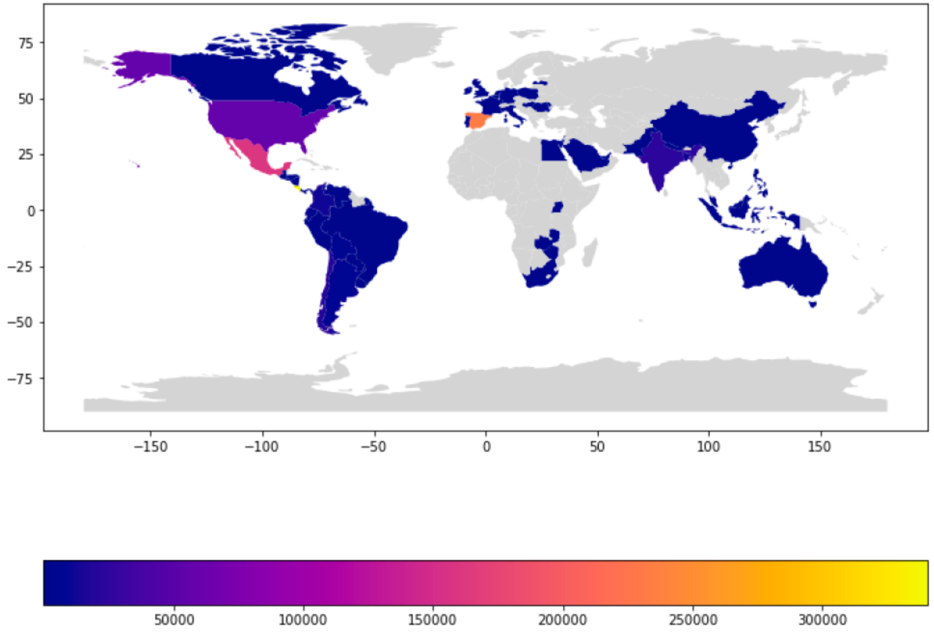
Data

- timestamp: the day, hour, minutes and seconds when the entry was created
- radius: precision in meters of the location
- speed: speed at which it has reached that point (km/h)
- altitude: altitude in meters
- isSafe: indicator indicating whether the point is safe (True) or not (False)
- monitoring: route/trajectory identifier
- userId: user identifier
- lattitude:
- latitude longitude: longitude

Folium

GeoPandas

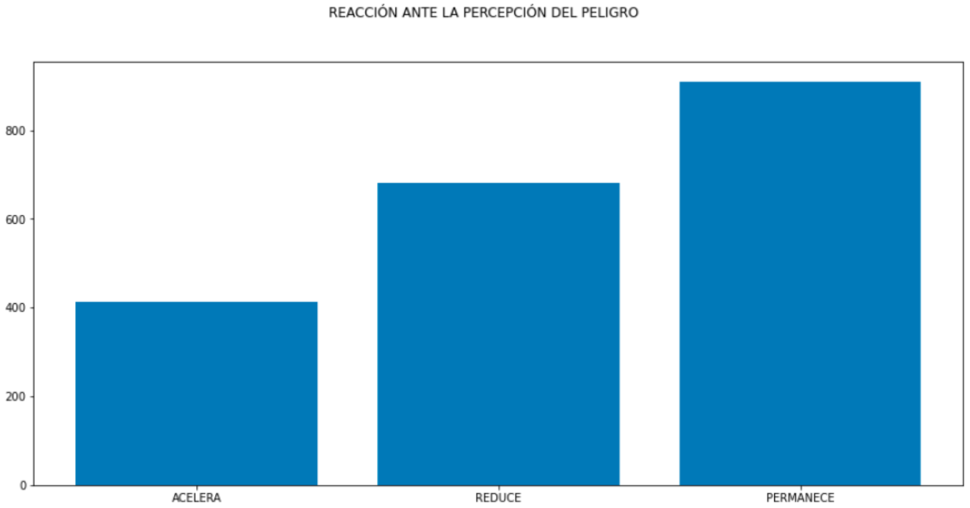
Shapely



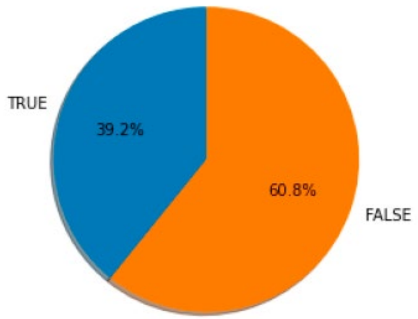
Characterization of the geo-located profile

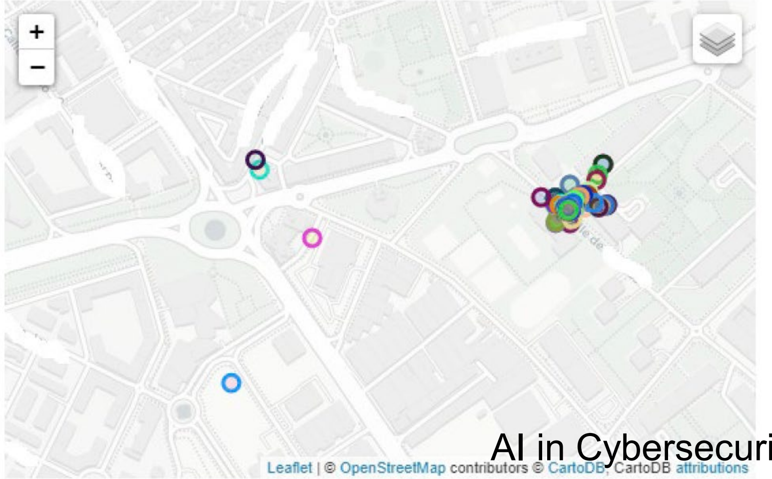
Enrichment of the original data

- distance_in_meters
  - time_in_seconds
  - isDay
  - isWeekend
- Clustering (K-Means)
  - Grouping of Unsafe Zones (N clusters by users)
  - Centroids



Cambios de velocidad en las rutas cuando el usuario indica que hay percepción de inseguridad/peligro.





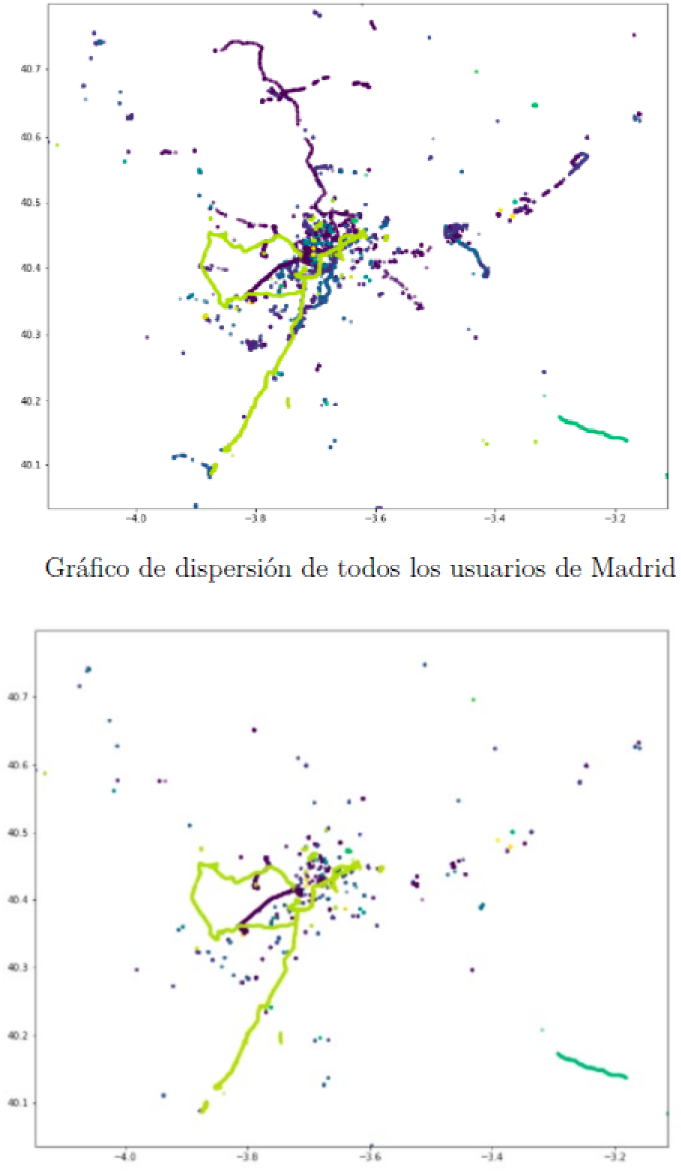
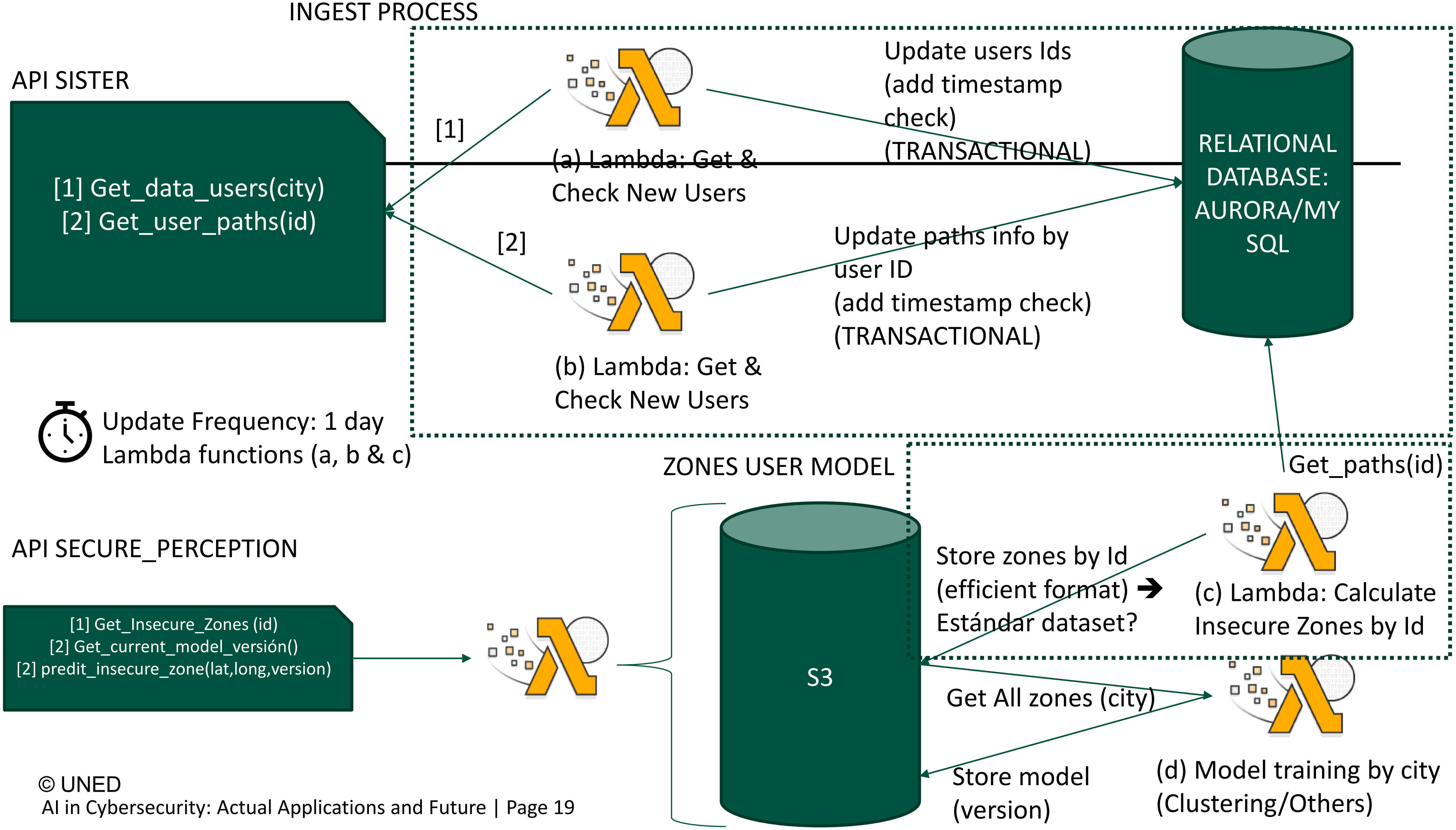


Gráfico de dispersión de todos los usuarios de Madrid

Gráfico de dispersión de zonas percibidas como inseguras de Madrid

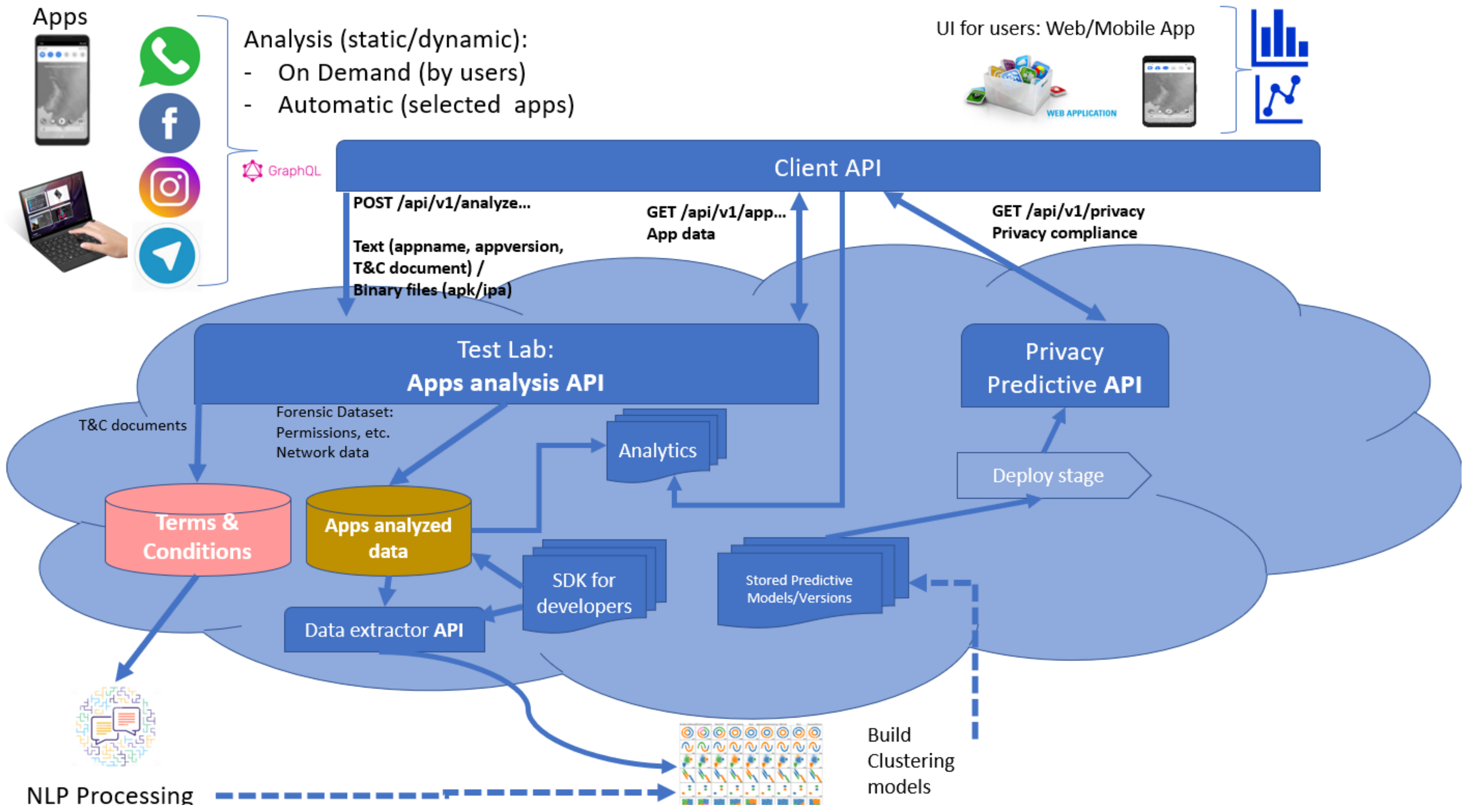

INGEST PROCESS
API SISTER
[1] Get_data_users(city)
[2] Get_user_paths(id)
[1]
[2]
(a) Lambda: Get & Check New Users
(b) Lambda: Get & Check New Users
Update users Ids (add timestamp check) (TRANSACTIONAL)
Update paths info by user ID (add timestamp check) (TRANSACTIONAL)
RELATIONAL DATABASE: AURORA/MY SQL
Update Frequency: 1 day
Lambda functions (a, b & c)
ZONES USER MODEL
Get_paths(id)
API SECURE_PERCEPTION
[1] Get_Insecure_Zones (id)
[2] Get_current_model_versión()
[2] predit_insecure_zone(lat,long,version)
S3
Store zones by Id (efficient format) ➔ Estándar dataset?
(c) Lambda: Calculate Insecure Zones by Id
Get All zones (city)
Store model (version)
(d) Model training by city (Clustering/Others)

Some projects: Smart Rural IoT and Secured Environments

- Smart Rural IoT Laboratory
  - LoT@UNED (UNED)
  - Smart Lab (NOVA)
  - Cátedra de Territorios Sostenibles y Desarrollo Local (Consorcio UNED Ponferrada)
- Within the research infrastructure, functional AI models are developed and evaluated in rural environments with computational and connectivity limitations.
- Optimization and deployment of predictive AI models on IoT devices in rural environments.

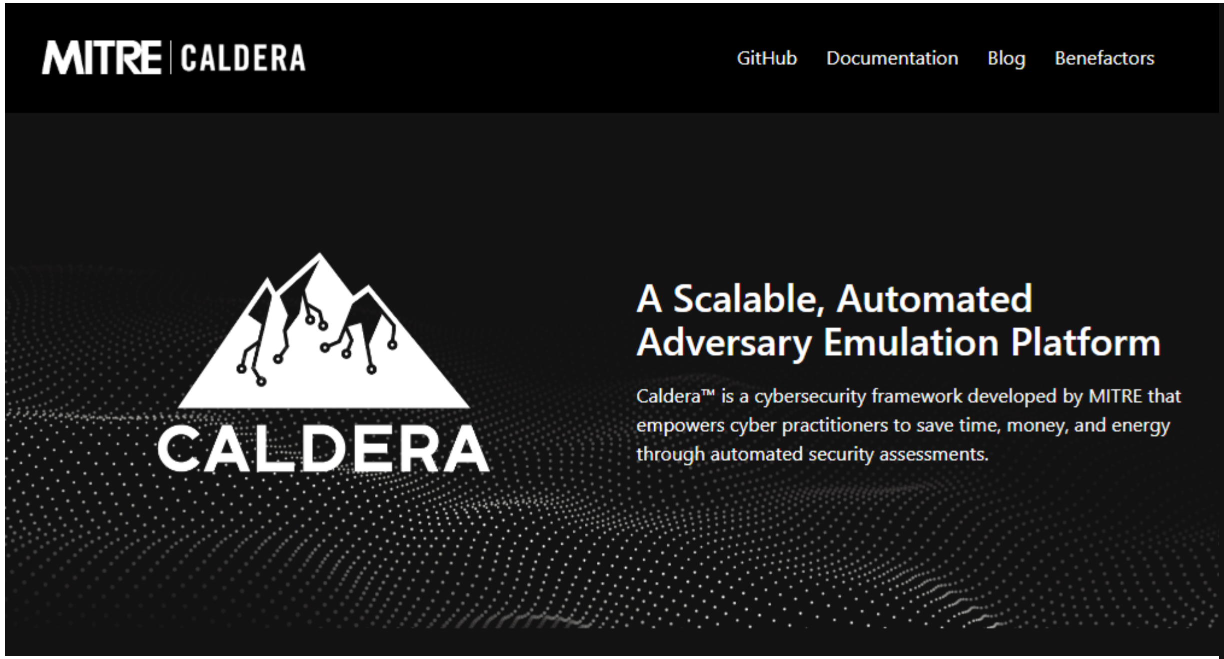
Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL
Plan de Recuperación, Transformación y Resiliencia
España | digital 2026
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD

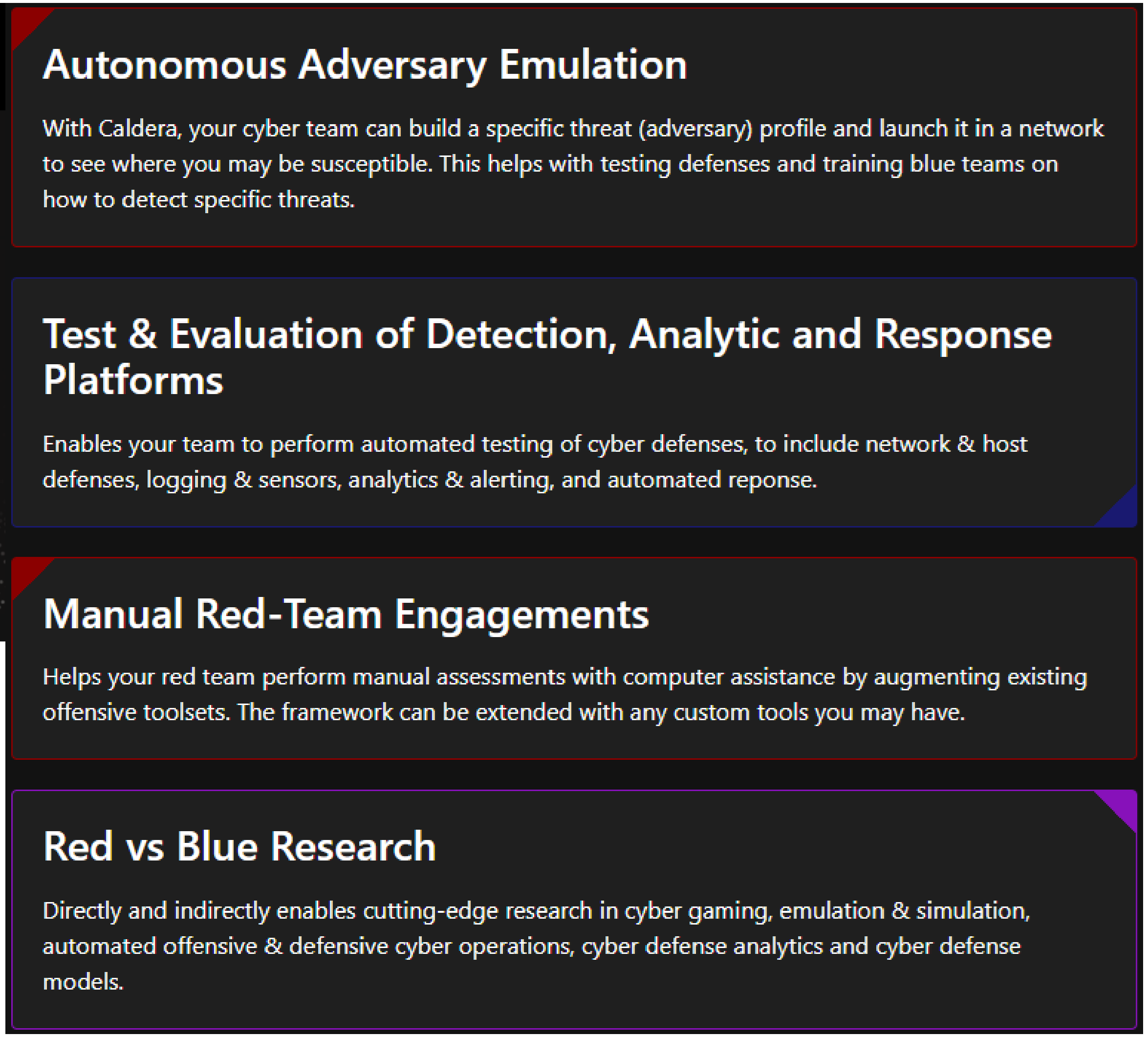
Some projects: Analysis of mobile applications from the perspective of data protection: Cyber-protection and Cyber-risks of citizen information



Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL
Plan de Recuperación, Transformación y Resiliencia
España | digital 2026
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD

Challenges and Future (I/II)





- Research on artificial intelligence
  - Caldera can be used to test artificial intelligence and other decision-making algorithms using the Mock plugin. The plugin adds simulated agents and mock ability responses, which can be used to run simulate an entire operation.

Challenges and Future (II/II)

- Foundational Time Models applied to Cybersecurity scenarios
- Reinforcement Learning in CyberSecurity
- New architectures: MamBa & Graphs Neural Networks
- Optimization/Light ML/DL models

Financiado por la Unión Europea
NextGenerationEU

GOBIERNO DE ESPAÑA
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

Plan de Recuperación, Transformación y Resiliencia

incibe_
INSTITUTO NACIONAL DE CIBERSEGURIDAD