



Security Summit

Milano 11-12-13 marzo 2025



Cybersecurity as a Service, AI e Active Adversary Come gestire una difesa proattiva ed efficace

Mauro Cicognini | Comitato Scientifico, *Clusit*

Walter Narisoni | Director Sales Engineer South EMEA, Sophos

1



Mauro Cicognini



COMITATO SCIENTIFICO



FOUNDING PARTNER



Cybersecurity as a Service, AI e Active Adversary

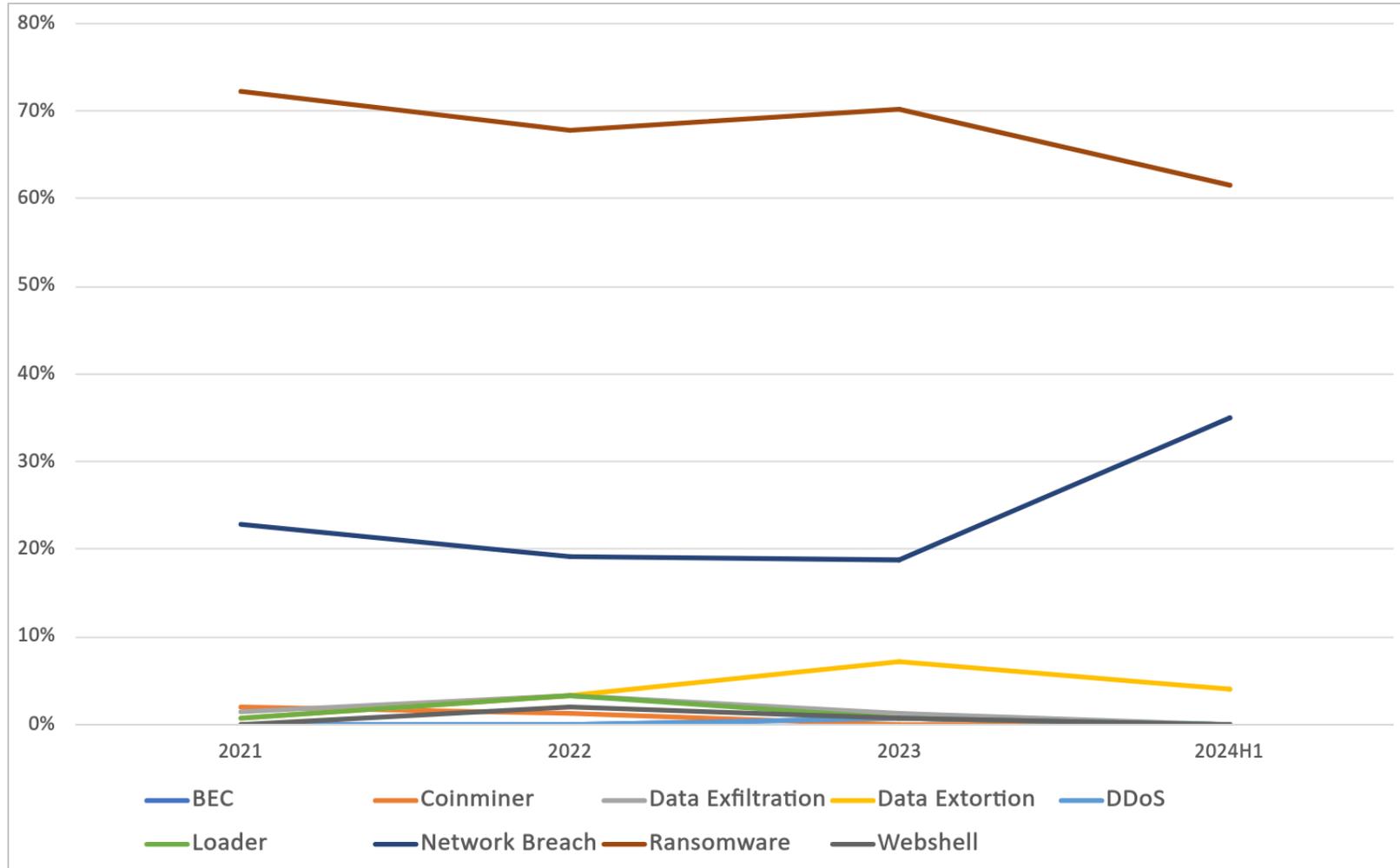
Come gestire una difesa proattiva ed efficace

SOPHOS

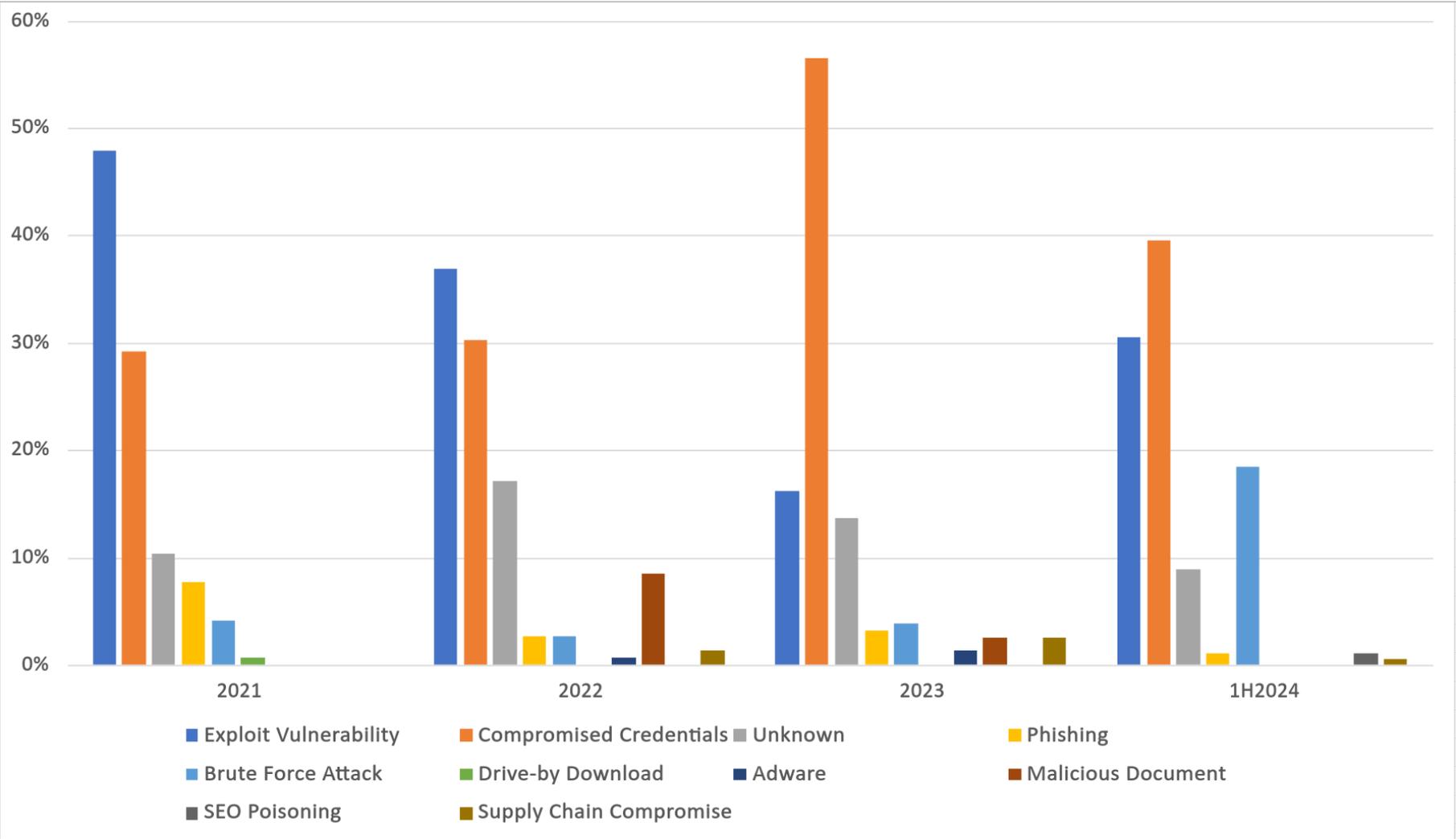


**Chiudere RDP esposti,
Utilizzare MFA e
Applicare patch ai server
vulnerabili.**

Tipi di attacco (cosa)

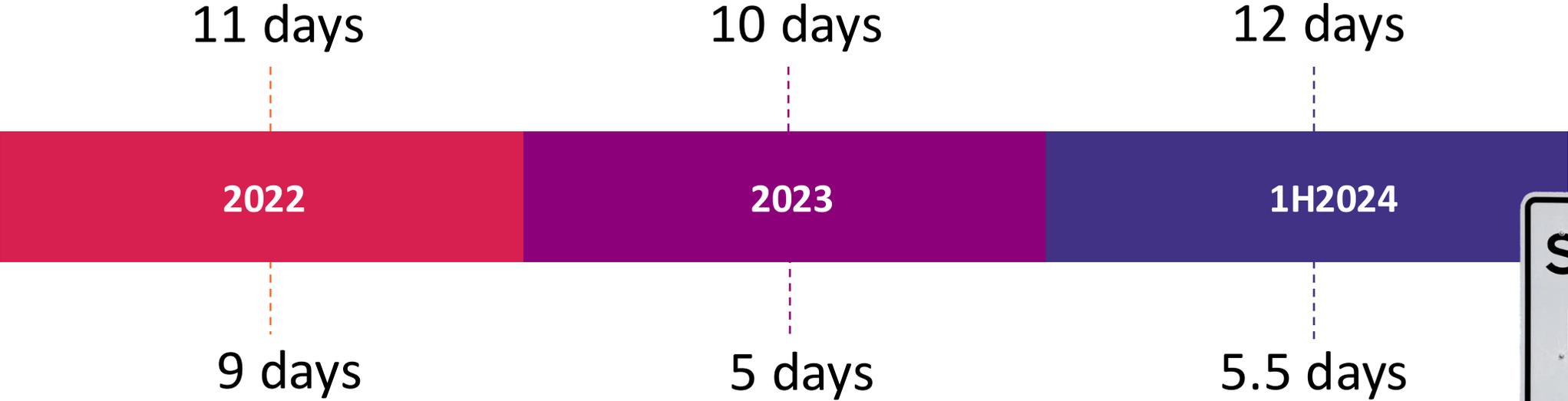


Root cause (perché)



Limite di velocità per i cattivi?

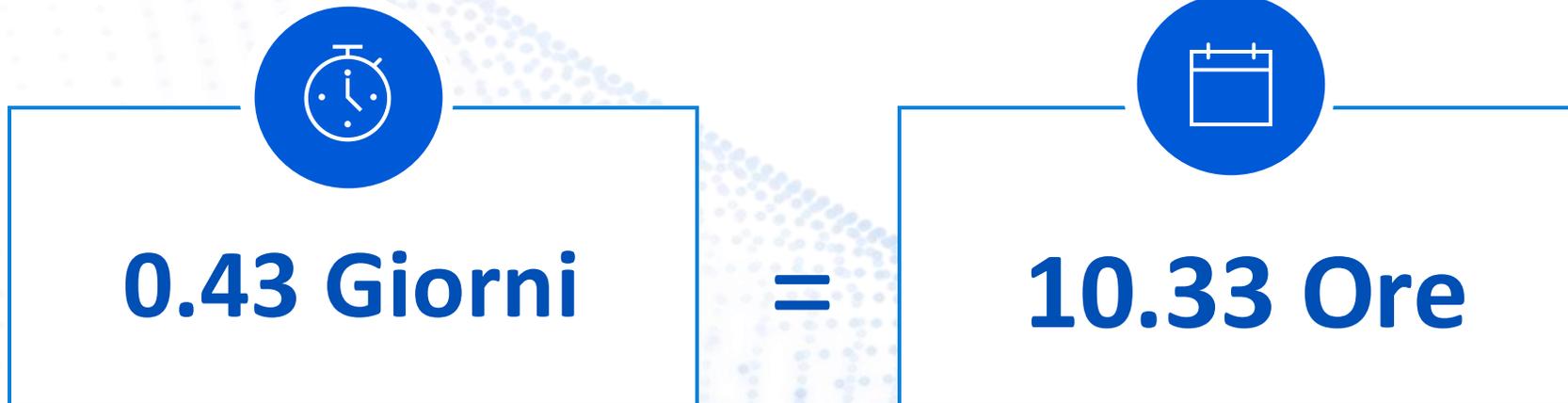
Tempo di permanenza mediano: Non-Ransomware



Tempo di permanenza mediano: Incidenti ransomware

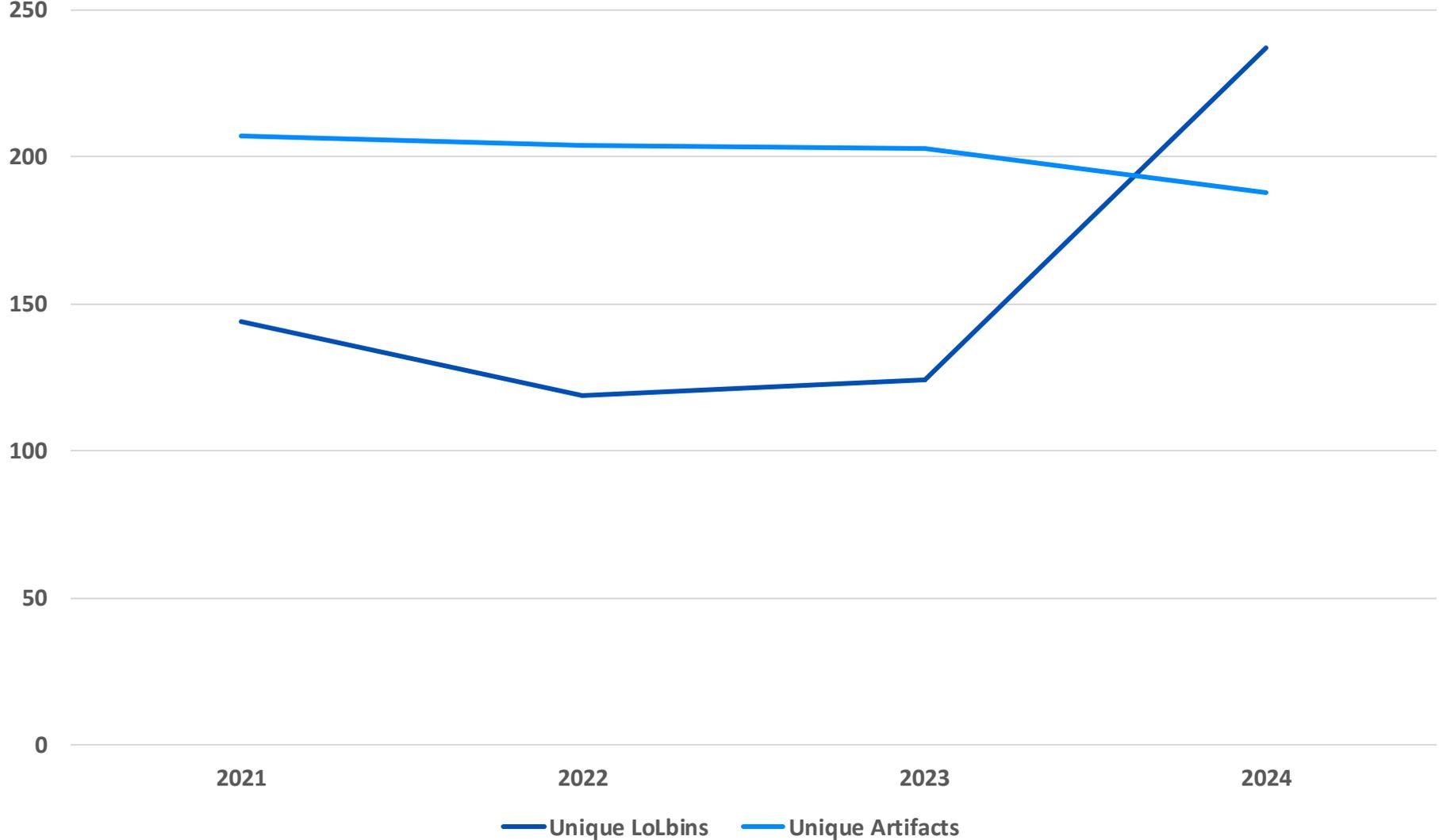


La corsa ad Active Directory



Tempo medio per raggiungere Active-Directory nel 1H2024

Nascondendosi in bella vista



Living off the Land (i.e., sfruttando strumenti informatici legittimi)

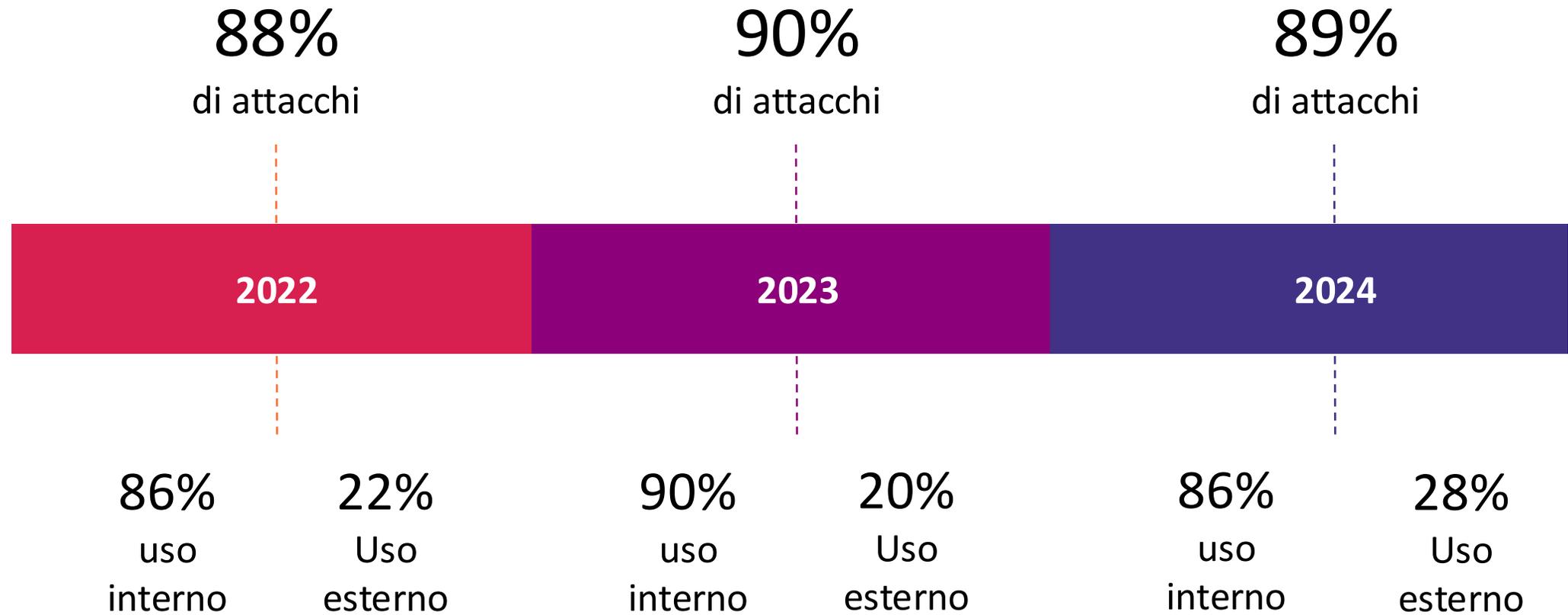
Rank	2022	2023	1H2024	Rank
1	RDP	RDP	RDP	1
2	PowerShell	PowerShell	PowerShell	2
3	cmd.exe	cmd.exe	cmd.exe	3
4	Psexec	net.exe	net.exe	4
5	Task Scheduler	Psexec	ping.exe	5
6	net.exe	Task Scheduler	nltest.exe	6
7	rundll32.exe	rundll32.exe	rundll32.exe	7
8	WMI	ping.exe	notepad.exe	8
9	whoami.exe	nltest.exe	WMI	9
10	ping.exe	reg.exe	Psexec	10
11	reg.exe	WMI	whoami.exe	11

RDP

(Remote Disaster Protocol)



Ubiquità dell'RDP negli attacchi



TOTAL RESULTS

4,160,150

TOP COUNTRIES



China	1,588,540
United States	666,832
Germany	204,481
Japan	123,558
Russian Federation	114,867

[More...](#)

TOP PORTS

3389	3,907,652
135	230,691
3388	18,466
593	2,800
5900	524

[More...](#)

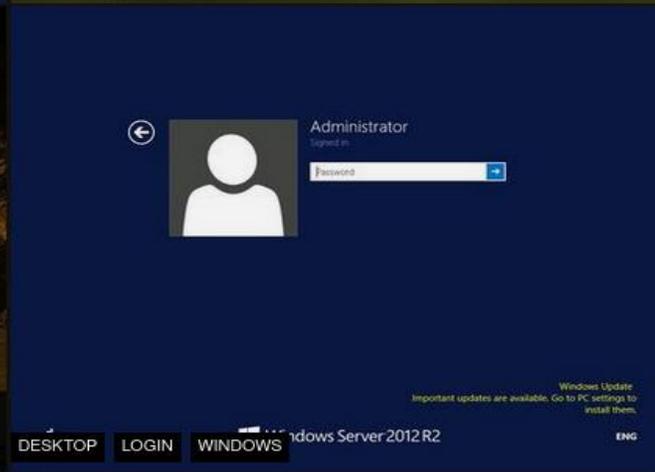
TOP ORGANIZATIONS

Tencent cloud computing (Beijing) Co., ...	429,374
Allyun Computing Co., LTD	329,382
Tencent Cloud Computing (Beijing) Co.,...	292,630
Microsoft Corporation	174,790
Amazon Technologies Inc.	105,593

[More...](#)

TOP PRODUCTS

Remote Desktop Protocol	3,902,417
Microsoft RPC Endpoint Mapper	230,691
Microsoft RPC Endpoint Mapper over HTTP	2,800



Non puoi nascondere

Host Filters

Labels:

- 832.60K remote-access
- 832.59K network-administration
- 129.51K file-sharing
- 116.45K login-page
- 94.80K jquery
- More

Autonomous System:

- 32.41K QUICKPACKET
- 31.02K MICROSOFT-CORP-MSN-AS-BLOCK
- 26.33K ALIBABA-CN-NET Hangzhou Alibaba Advertising Co.,Ltd.
- 25.91K CHINANET-BACKBONE No.31,Jin-rong Street
- 22.47K PEG-SV
- More

Location:

- 183.33K United States
- 123.99K China
- 59.23K Hong Kong
- 43.75K Russia
- 37.82K Brazil
- More

Service Filters

Service Names:

- 1.62M HTTP
- 1.09M RDP
- 462.42K UNKNOWN
- 138.83K WINRM
- 123.48K SSH
- More

Ports:

Hosts

Results: 836,349 Time: 0.20s

181.

Cordoba, Argentina

network-administration camera remote-access

- 3488/HTTP
- 4543/RTSP
- 8000/RDP
- 21901/RDP
- 21902/RDP
- 21912/RDP

184.

Microsoft Windows Arizona, United States

network-administration bootstrap jquery prototype requirejs voip remote-access email database +2

- 21/FTP
- 25/SMTP
- 53/MURMUR
- 80/HTTP
- 110/POP3
- 135/DCERPC
- 143/IMAP
- 445/SMB
- 465/SMTP
- 993/IMAP
- 995/POP3
- 1801/MSMQ
- 3306/MYSQL
- 8172/HTTP
- 8443/HTTP
- 8880/HTTP
- 50947/RDP

115.

Selangor, Malaysia

truncated

- 2/RDP
- 509/UNKNOWN

141.

Microsoft Windows Samara Oblast, Russia

network-administration remote-access

- 53/DNS
- 123/NTP
- 161/SNMP
- 500/IKE
- 1701/L2TP
- 2000/MIKROTIK_BW
- 31434/HTTP
- 32434/HTTP
- 33434/HTTP
- 34434/HTTP
- 35434/HTTP
- 36434/HTTP
- 37434/HTTP
- 38434/HTTP
- 39434/HTTP
- 42443/HTTP
- 61194/UNKNOWN
- 63389/RDP

212.

Microsoft Windows_xp Veneto, Italy

remote-access login-page network-administration email jquery

- 25/SMTP
- 110/POP3
- 143/IMAP
- 500/IKE
- 555/SSH
- 587/SMTP
- 3000/HTTP
- 4050/RDP
- 4444/RDP
- 5555/RDP

Non puoi nascondere

Host Filters

Labels:

- 832.60K remote-access
- 832.59K network-administration
- 129.51K file-sharing
- 116.45K login-page
- 94.80K jquery
- More

Autonomous System:

- 32.41K QUICKPACKET
- 31.02K MICROSOFT-CORP-MSN-AS-BLOCK
- 26.33K ALIBABA-CN-NET Hangzhou Alibaba Advertising Co.,Ltd.
- 25.91K CHINANET-BACKBONE No.31,Jin-rong Street
- 22.47K PEG-SV
- More

Location:

- 183.33K United States
- 123.99K China
- 59.23K Hong Kong
- 43.75K Russia
- 37.82K Brazil
- More

Service Filters

Service Names:

- 1.62M HTTP
- 1.09M RDP
- 462.42K UNKNOWN
- 138.83K WINRM
- 123.48K SSH
- More

Ports:

Hosts

Results: 836,349 Time: 0.20s

181. Cordoba, Argentina
network-administration camera remote-access
3488/HTTP 4543/RTSP 8000/RDP 21901/RDP 21902/RDP
21912/RDP

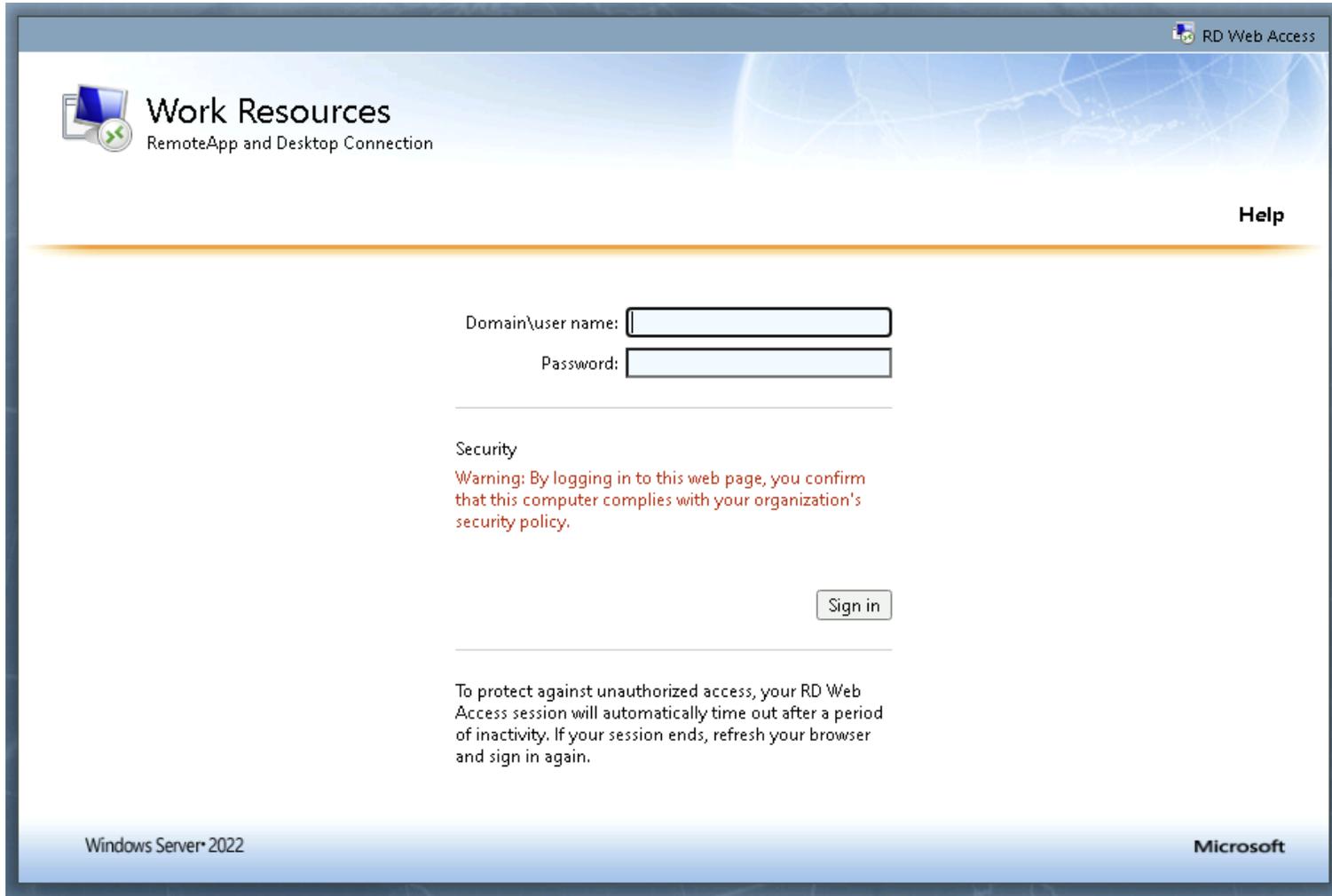
184. Microsoft Windows Arizona, United States
network-administration bootstrap jquery prototype requirejs voip remote-access email database +2
21/FTP 25/SMTP 53/MURMUR 80/HTTP 110/POP3
135/DCERPC 143/IMAP 445/SMB 465/SMTP 993/IMAP
995/POP3 1801/MSMQ 3306/MYSQL 8172/HTTP 8443/HTTP
8880/HTTP 50947/RDP

115. Selangor, Malaysia
truncated 2/RDP 509/UNKNOWN

141. Microsoft Windows Samara Oblast, Russia
network-administration remote-access
53/DNS 123/NTP 161/SNMP 500/IKE 1701/L2TP
2000/MIKROTIK_BW 31434/HTTP 32434/HTTP 33434/HTTP 34434/HTTP
35434/HTTP 36434/HTTP 37434/HTTP 38434/HTTP 39434/HTTP
42443/HTTP 61194/UNKNOWN 63389/RDP

212. Microsoft Windows_xp Veneto, Italy
remote-access login-page network-administration email jquery
25/SMTP 110/POP3 143/IMAP 500/IKE 555/SSH
587/SMTP 3000/HTTP 4050/RDP 4444/RDP 5555/RDP

Accesso Web Desktop remoto



The screenshot shows the 'Work Resources' login page for RemoteApp and Desktop Connection. The page has a blue header with the 'Work Resources' logo and text. A 'Help' link is in the top right. The main content area contains a 'Domain\user name:' field, a 'Password:' field, and a 'Sign in' button. Below the fields is a 'Security' warning in red text. At the bottom, there is a footer with 'Windows Server 2022' on the left and the 'Microsoft' logo on the right.

RD Web Access

Work Resources
RemoteApp and Desktop Connection

[Help](#)

Domain\user name:

Password:

Security
Warning: By logging in to this web page, you confirm that this computer complies with your organization's security policy.

To protect against unauthorized access, your RD Web Access session will automatically time out after a period of inactivity. If your session ends, refresh your browser and sign in again.

Windows Server 2022 Microsoft

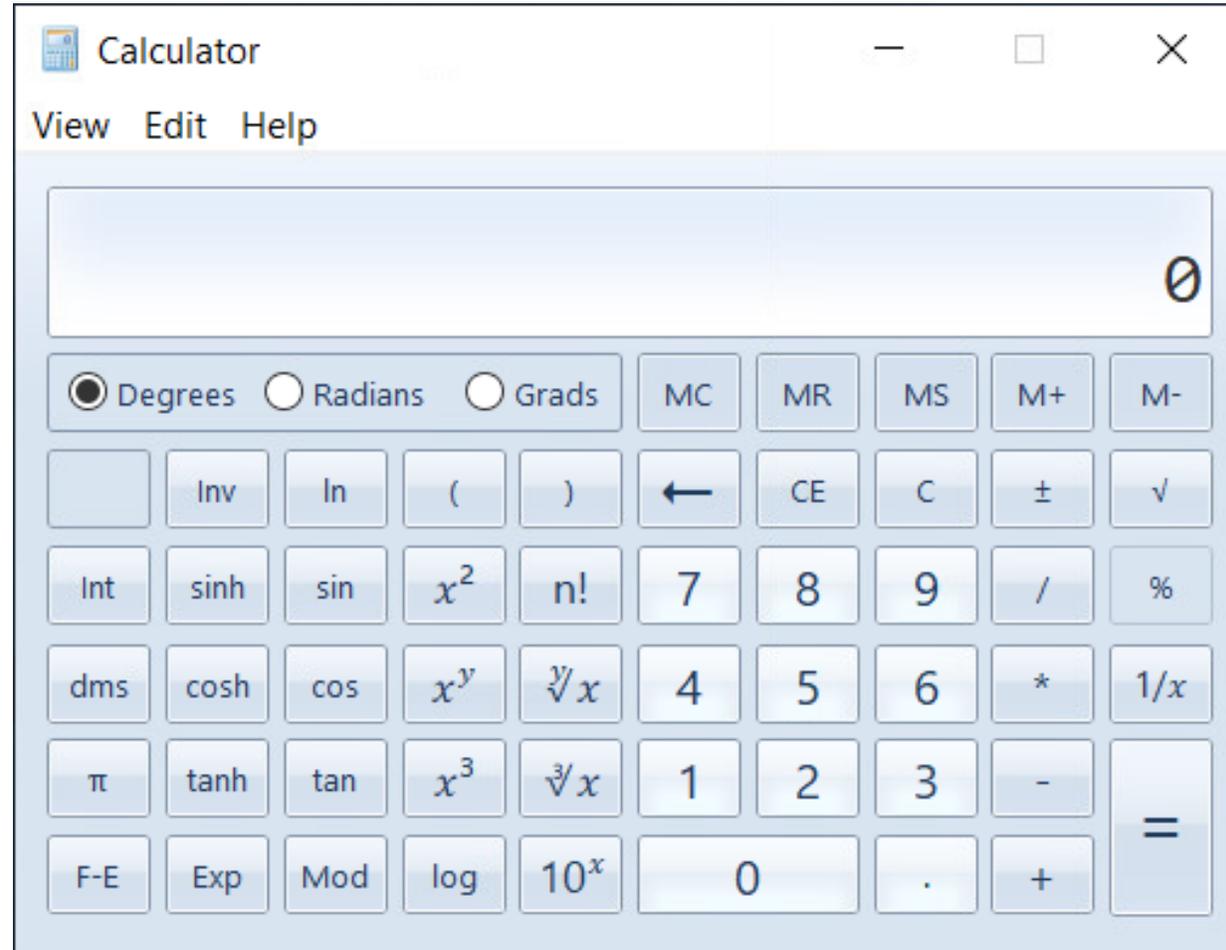
Accesso Web Desktop remoto

event_timestamp	event	subject_username	subject_domain	target_username	logon_type	name	remote_address	description
2024-02-08 04:15:53	4625	RDWebAccess	IIS APPPOOL	melissa		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:17:04	4625	RDWebAccess	IIS APPPOOL	power		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:17:18	4625	RDWebAccess	IIS APPPOOL	kimberly		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:18:22	4625	RDWebAccess	IIS APPPOOL	test.user		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:20:23	4625	RDWebAccess	IIS APPPOOL	dbaccess		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:20:28	4625	RDWebAccess	IIS APPPOOL	server		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:21:33	4625	RDWebAccess	IIS APPPOOL	client		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:21:54	4625	RDWebAccess	IIS APPPOOL	support		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:22:11	4625	RDWebAccess	IIS APPPOOL	freebox		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:22:16	4625	RDWebAccess	IIS APPPOOL	auditor		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:22:47	4625	RDWebAccess	IIS APPPOOL	steven		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:23:21	4625	RDWebAccess	IIS APPPOOL	patricia		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:24:12	4625	RDWebAccess	IIS APPPOOL	jeffrey		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:24:29	4625	RDWebAccess	IIS APPPOOL	brian		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:24:38	4625	RDWebAccess	IIS APPPOOL	daniel		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:24:44	4625	RDWebAccess	IIS APPPOOL	default		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:24:47	4625	RDWebAccess	IIS APPPOOL	test.user		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:24:54	4625	RDWebAccess	IIS APPPOOL	michelle		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:25:37	4625	RDWebAccess	IIS APPPOOL	access		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:26:32	4625	RDWebAccess	IIS APPPOOL	webadmin		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:28:10	4625	RDWebAccess	IIS APPPOOL	vendor		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:28:15	4625	RDWebAccess	IIS APPPOOL	thomas		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:28:42	4625	RDWebAccess	IIS APPPOOL	timothy		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:29:30	4625	RDWebAccess	IIS APPPOOL	care		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:29:34	4625	RDWebAccess	IIS APPPOOL	david		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:29:41	4625	RDWebAccess	IIS APPPOOL	user		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:30:07	4625	RDWebAccess	IIS APPPOOL	cookie		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:30:18	4625	RDWebAccess	IIS APPPOOL	app		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:30:22	4625	RDWebAccess	IIS APPPOOL	donald		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:30:50	4625	RDWebAccess	IIS APPPOOL	private		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:31:00	4625	RDWebAccess	IIS APPPOOL	query		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.
2024-02-08 04:31:05	4625	RDWebAccess	IIS APPPOOL	richard		3 C:\Windows\System32\inetsrv\w3wp.exe	-	An account failed to log on.

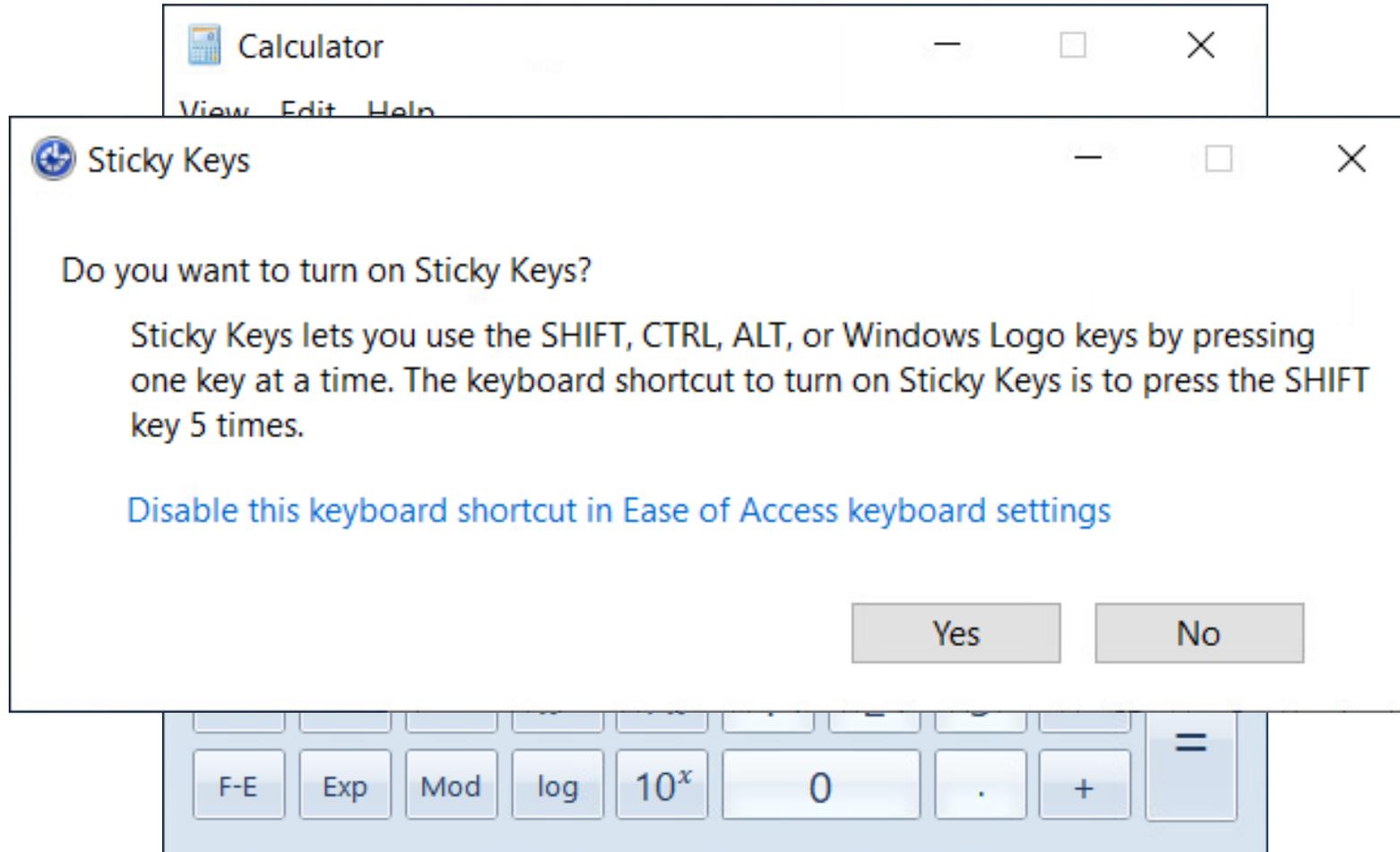
Accesso Web Desktop remoto

The image displays a screenshot of the Remote Desktop Web Access (RD Web Access) interface. The interface is presented as a window titled "RD Web Access" with a blue header bar. Inside the window, the main content area is titled "Work Resources" and "RemoteApp and Desktop Connection". Below this, there is a navigation bar with "RemoteApp and Desktops" on the left and "Help | Sign out" on the right. The main content area shows a "Current folder: /" and a single application icon labeled "Calculator". At the bottom of the window, there is a footer with "Windows Server 2022" on the left and the "Microsoft" logo on the right. A small text box at the bottom of the application area contains the following text: "This session will automatically time out after a period of inactivity. If your session ends, refresh your browser and sign in again."

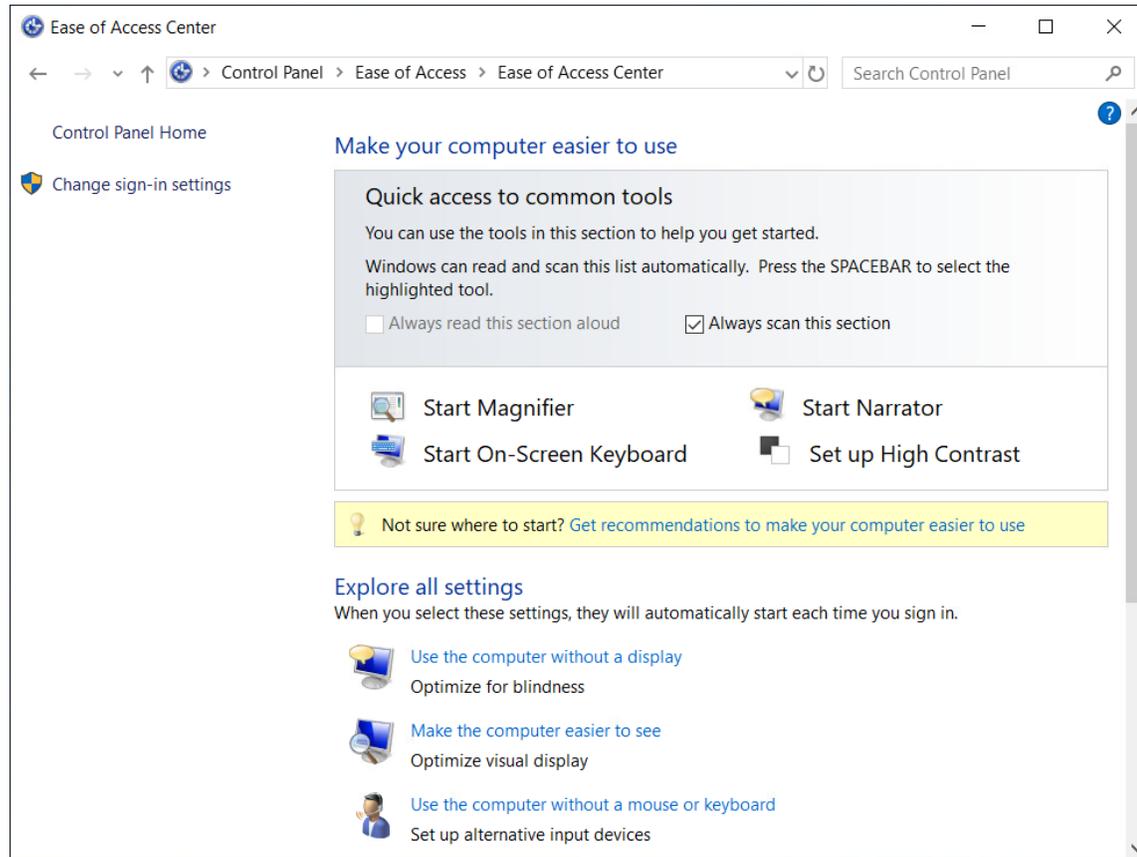
Accesso Web Desktop remoto



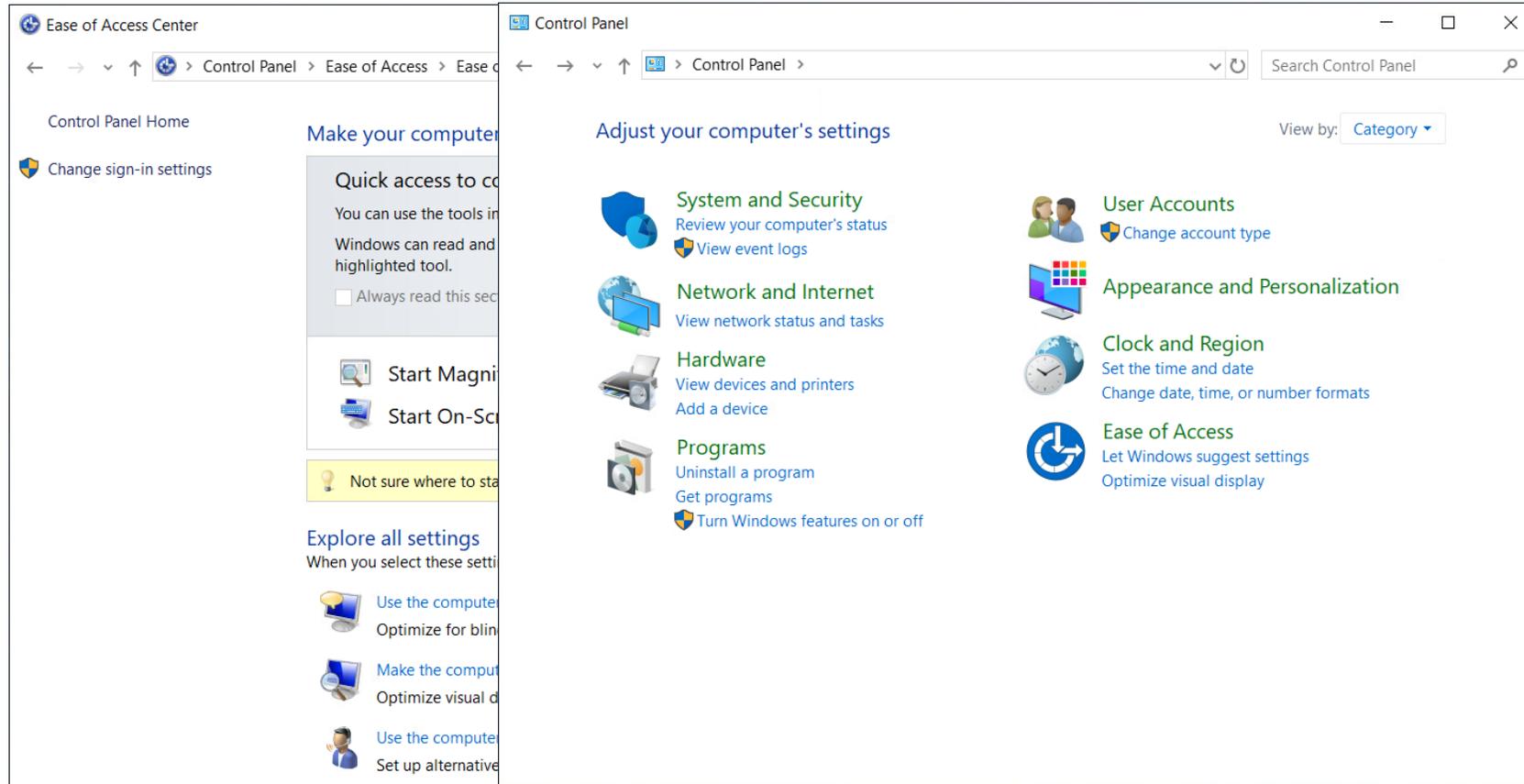
Accesso Web Desktop remoto



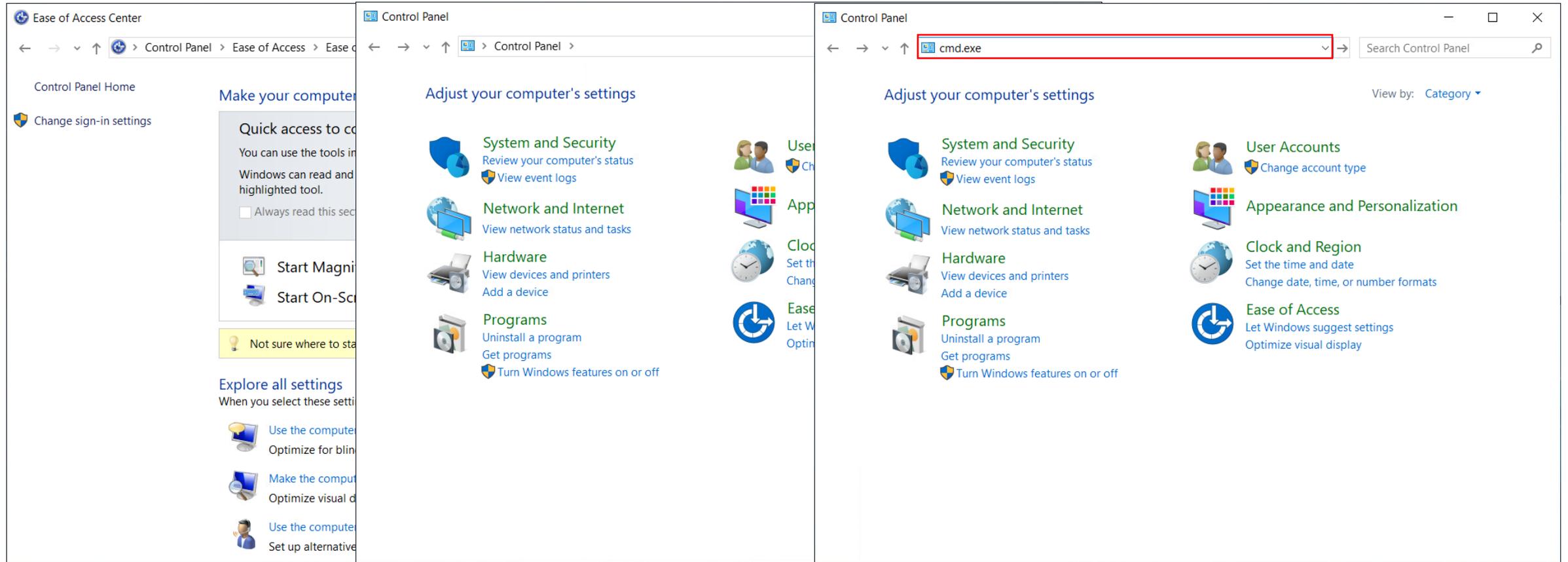
Accesso Web Desktop remoto



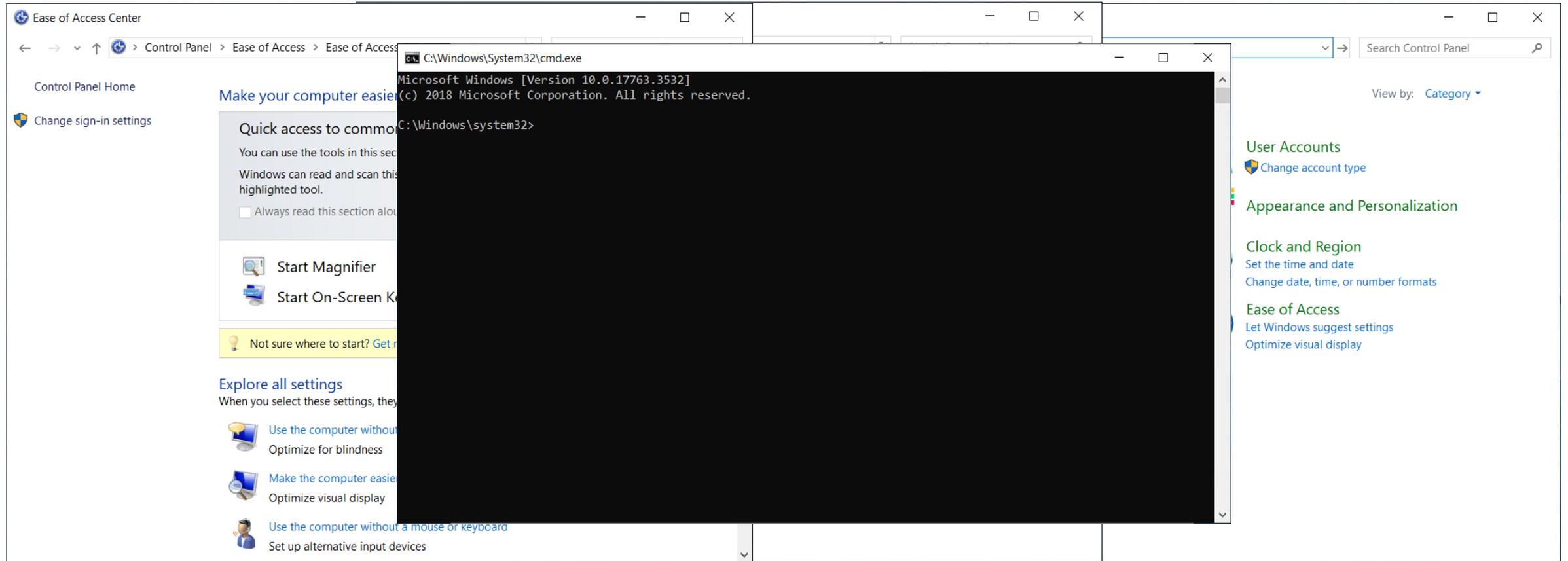
Accesso Web Desktop remoto



Accesso Web Desktop remoto



Accesso Web Desktop remoto



Accesso Web Desktop remoto

date_time	sophos_pid	cmdline
2024-01-12 15:36:53	8844:133495474138721969	"PowerShell.exe" -noexit -command Set-Location -literalPath 'C:\Users
2024-01-12 15:37:09	11168:133495474294158054	"C:\Windows\System32\cmd.exe"
2024-01-12 15:37:21	10256:133495474415113997	nltest /dclist:
2024-01-12 15:37:33	9692:133495474532985924	nltest /domain_trusts
2024-01-12 15:40:46	9848:133495476463311150	net user ██████████/domain
2024-01-12 15:40:46	9688:133495476463728999	C:\Windows\system32\net1 user ██████████/domain
2024-01-12 15:43:06	8304:133495477866076216	"C:\Windows\System32\WindowsPowerShell\v1.0\powershell.exe" -r

MFA



Vettori di attacco in evoluzione

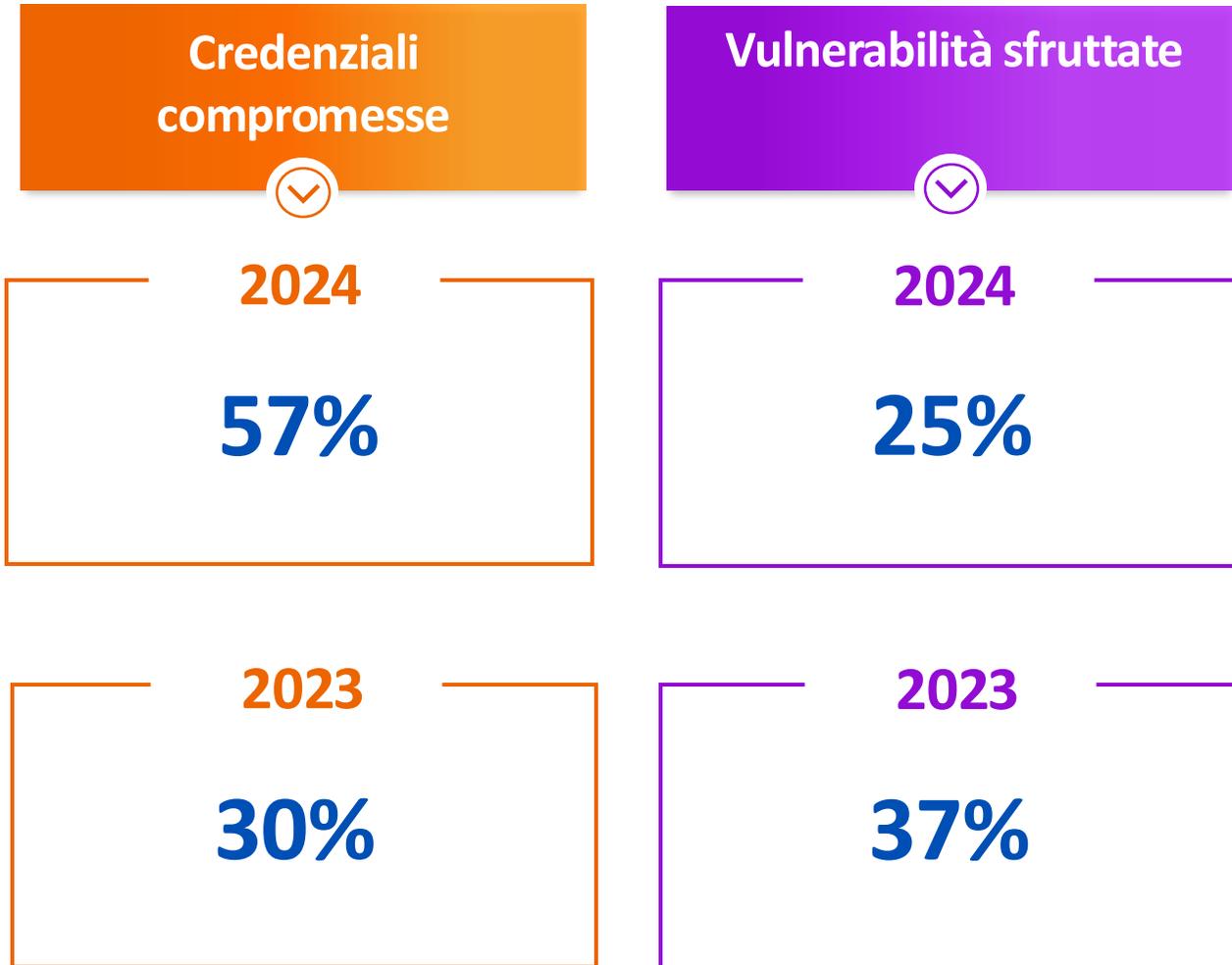
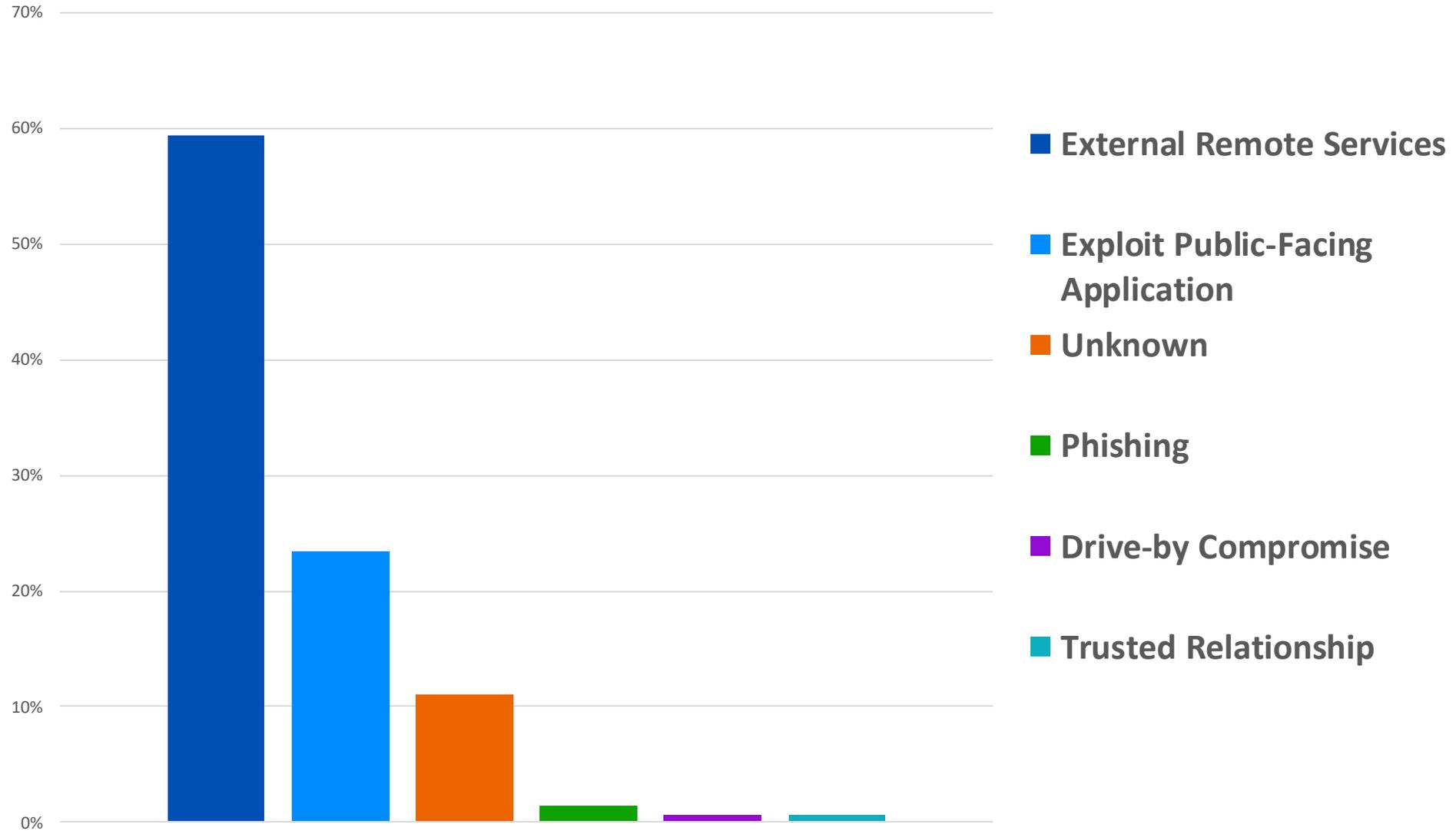


Figure 1. Select ways-in enumerations in non-Error, non-Misuse breaches (n=6,963)

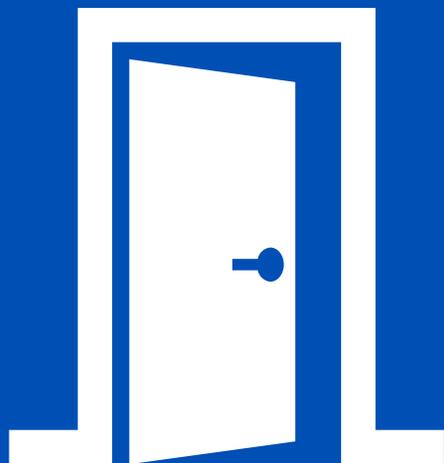
Accesso iniziale (come)



Portali pubblici sconosciuti

70%

La mancanza di MFA
lascia la porta aperta agli
avversari



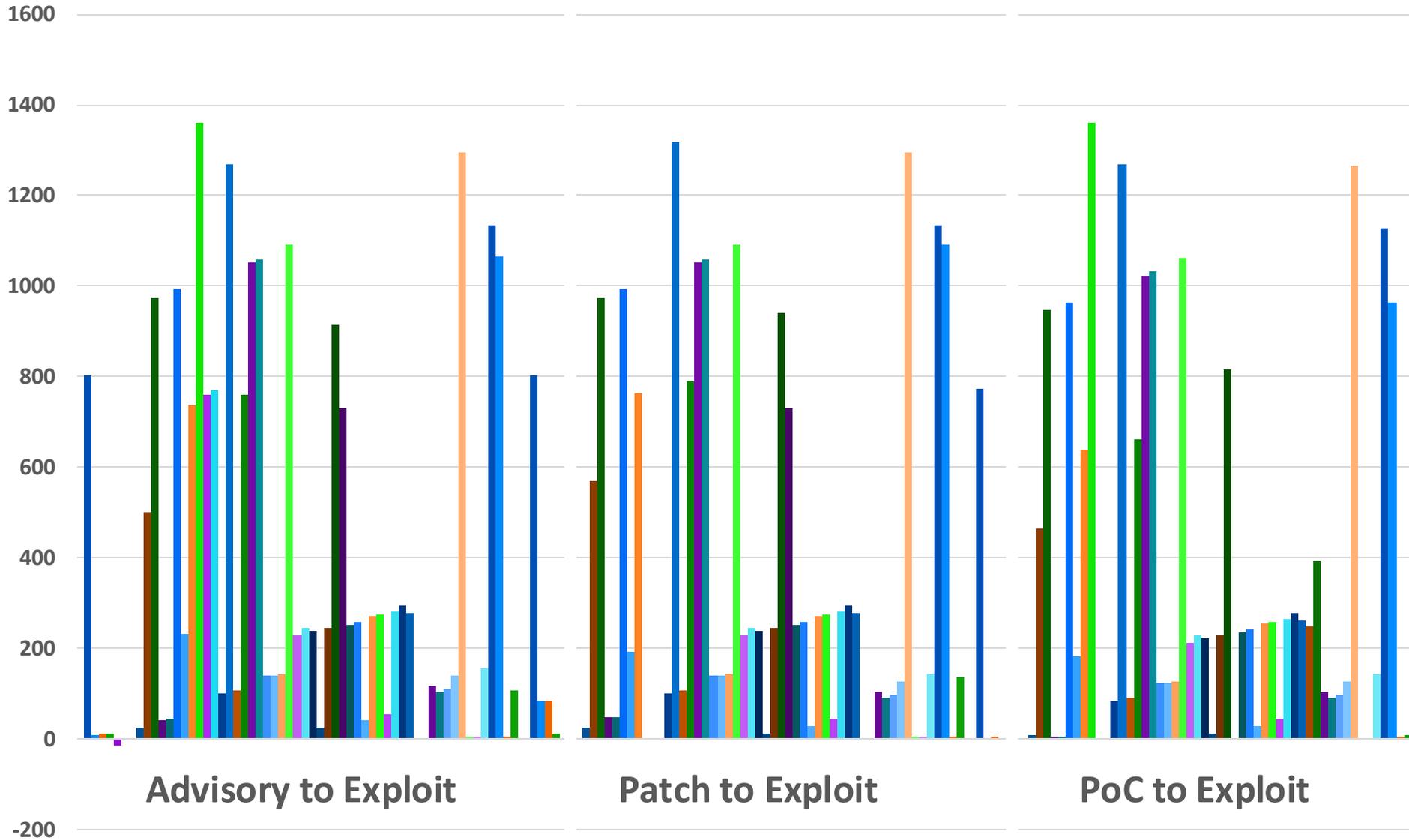
56%

Degli incidenti risolti nel 1° semestre 2024 non era configurata l'autenticazione a più fattori (MFA).

Vulnerabilità



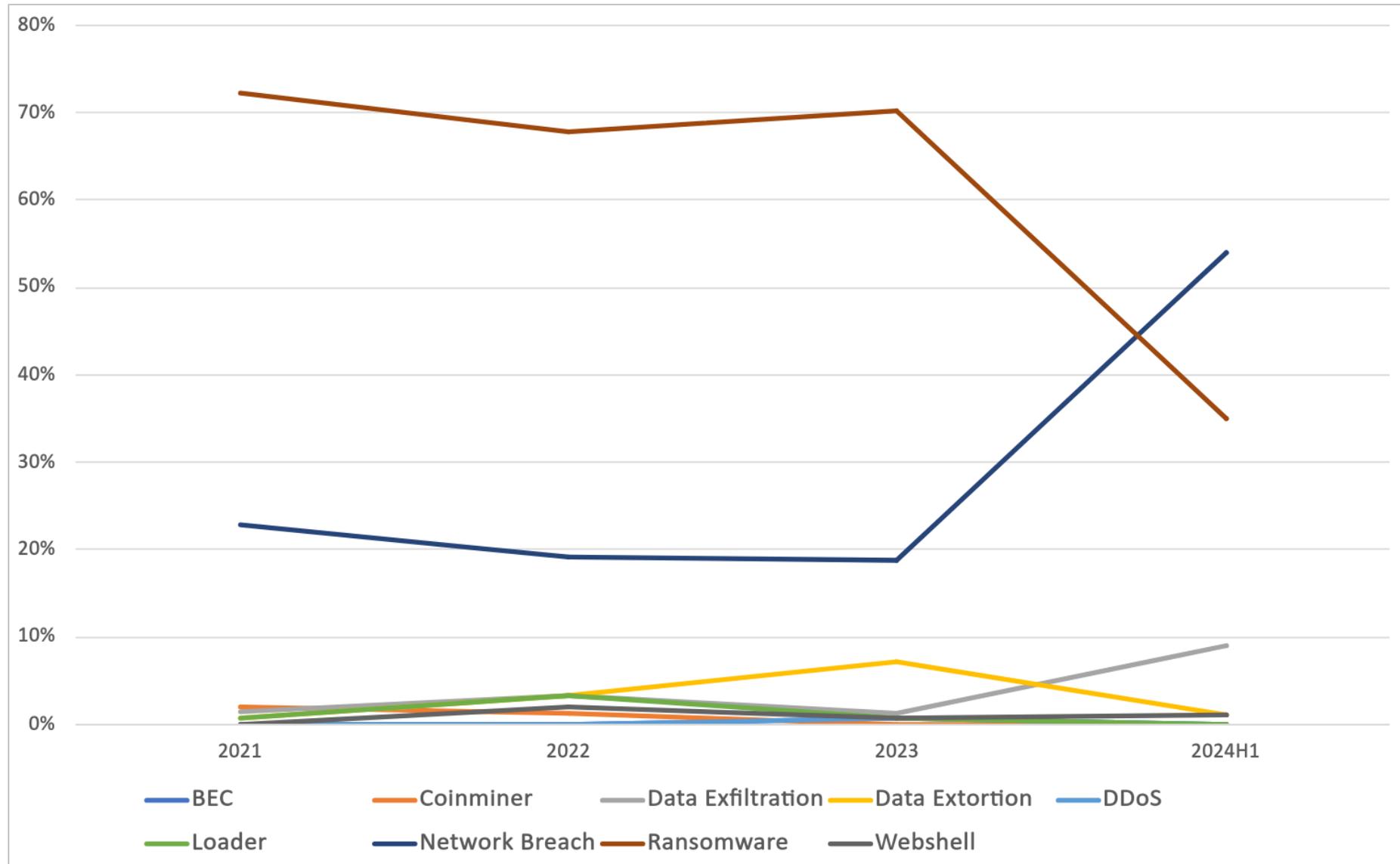
N-day, all day



Takeaways



Tipi di attacco... Rivisitato



Chiudere RDP esposti,

芭蕉 桃青



あゝあゝあゝ

あゝあゝあゝ

あゝあゝあゝ

許六





**Close exposed RDP ports,
Utilizzare MFA,**

許六

芭蕉
桃
青



**Chiudere RDP esposti,
Utilizzare MFA e
Applicare patch ai server
vulnerabili.**

I vantaggi dell'intelligenza artificiale per la sicurezza informatica



Approcci all'intelligenza artificiale

Tipo

Deep Learning AI

Apply

Utilizza reti neurali artificiali per riconoscere modelli e prendere decisioni in un modo che imita il cervello umano. APPLICA gli apprendimenti per svolgere i compiti.

Esempio: **Rileva URL dannosi**
Il modello di intelligenza artificiale viene addestrato per identificare i siti Web dannosi, consentendo ai prodotti di sicurezza di bloccarne l'accesso

Generative AI

Create

Utilizza la struttura e il modello dei dati esistenti per CREARE (generare) contenuti nuovi di zecca.

Esempio: **Riepilogo del caso di minaccia**
Il modello di intelligenza artificiale crea un riepilogo dell'attività della minaccia e fornisce agli analisti i passaggi successivi consigliati

Dimensione

Modelli di intelligenza artificiale massicci

Strumenti multiuso che vengono addestrati su grandi quantità di dati disponibili pubblicamente e possono aiutare con una vasta gamma di attività.

Esempio: **Microsoft Copilot, Google Gemini**

Piccoli modelli di intelligenza artificiale

I modelli incentrati sui risultati sono progettati, addestrati e costruiti per casi d'uso specifici.

Esempio: **Modello di rilevamento malware Android**

Deep Learning

Deep Learning | Applica le conoscenze apprese per migliorare le capacità di sicurezza informatica



Esegue attività ripetitive su vasta scala



Consente ai difensori di gestire un elevato volume di minacce



Si adatta alle minacce in evoluzione



ESEMPIO

Deep Learning in Sophos Endpoint

Diversi modelli di Deep Learning proteggono da attacchi noti e mai visti prima, comprese le minacce nelle soluzioni MS Office, nei PDF e nel formato Rich Text

Generative AI (GenAI)

GenAI | Crea contenuti per accelerare la sicurezza informatica, in particolare le operazioni di sicurezza



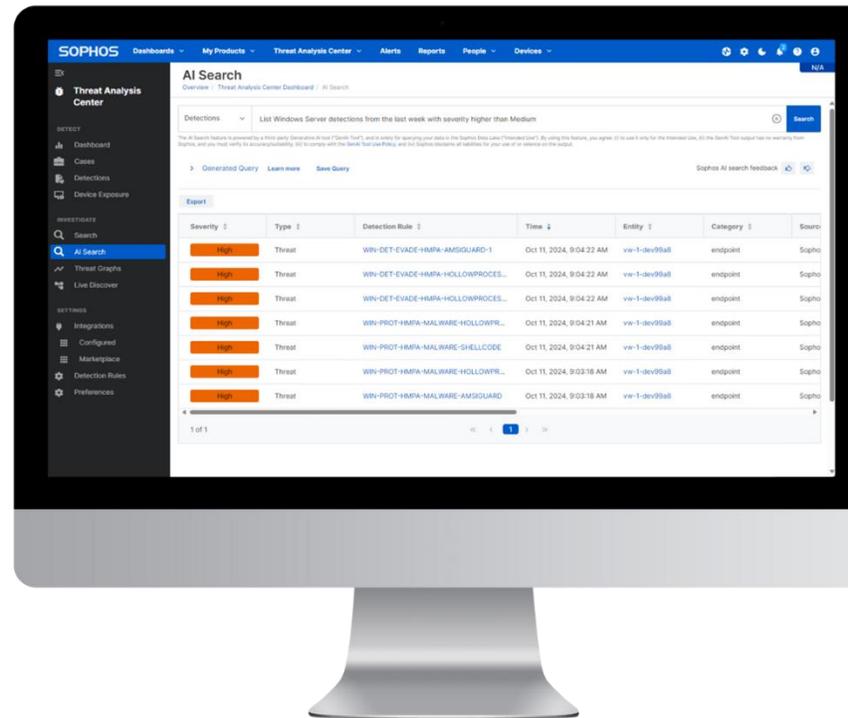
Consente agli analisti di prendere decisioni intelligenti e rapidamente



Allevia la pressione sugli analisti e previene il burnout



Riduce la barriera tecnologica alle operazioni di sicurezza



ESEMPIO Ricerca AI in Sophos XDR

Consente agli analisti della sicurezza di analizzare grandi volumi di dati di sicurezza con il linguaggio naturale, eliminando la necessità di competenze tecniche avanzate come SQL

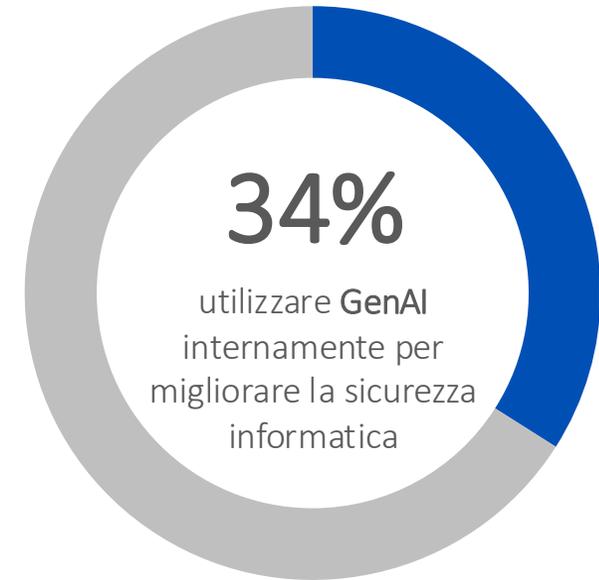
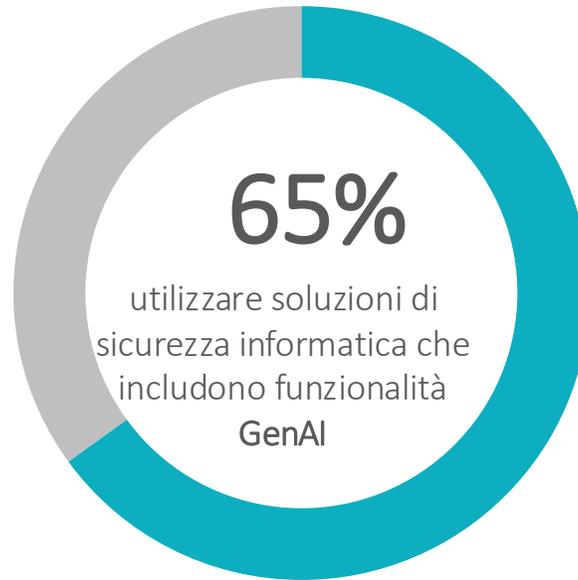
Adozione dell'intelligenza artificiale per la sicurezza informatica



AI tassi di adozione

98%

delle organizzazioni utilizza già l'intelligenza artificiale in qualche modo per la sicurezza informatica

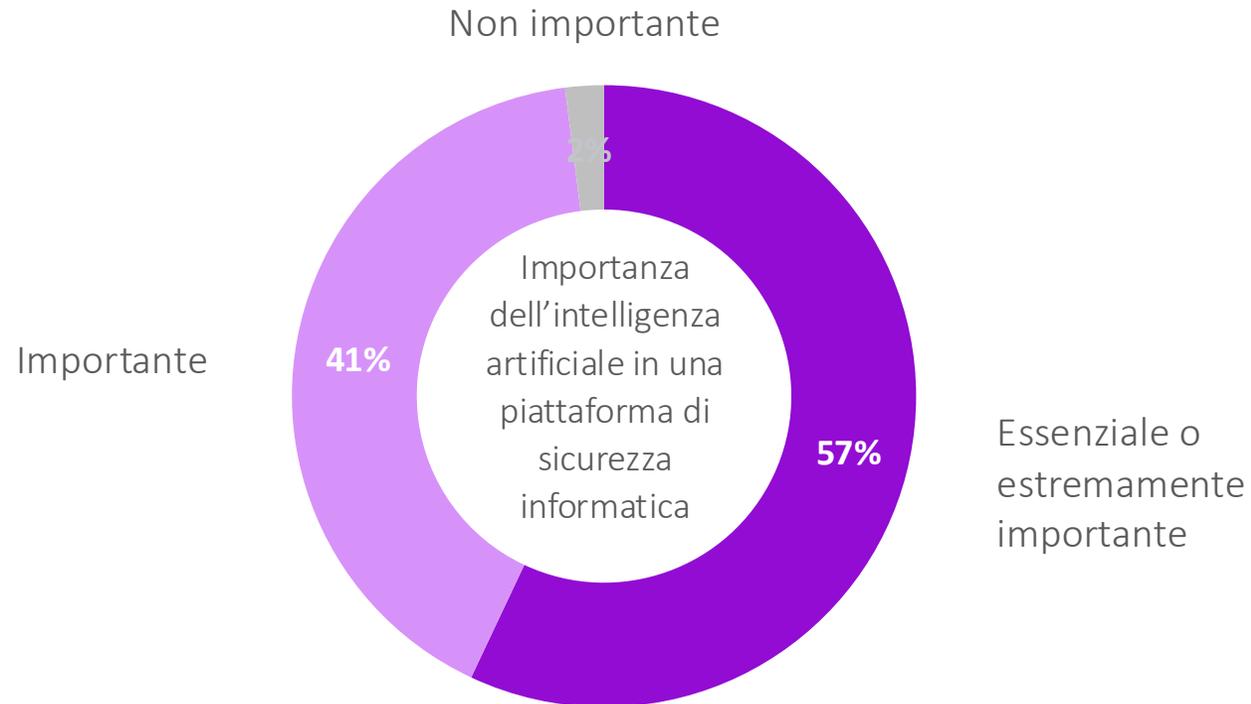


Does your organization currently use AI technologies as part of your cyber defenses? (n=400)

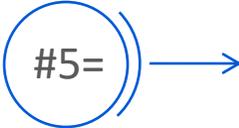
Importanza dell'intelligenza artificiale nella scelta delle soluzioni di sicurezza informatica

99%

delle organizzazioni elenca le funzionalità di intelligenza artificiale come requisito nella scelta di una piattaforma di sicurezza informatica



I vantaggi desiderati da GenAI

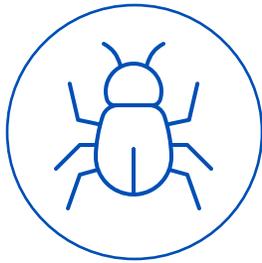
Il vantaggio più desiderato dall'IA generativa	
	Migliore protezione dalle minacce informatiche - 20% Miglioramento del ritorno dell'investimento per la sicurezza informatica (ROI) - 20%
	Aumento dell'efficienza e dell'impatto degli analisti IT - 17%
	Fiducia nel fatto che stiamo al passo con le innovazioni in materia di sicurezza informatica - 15%
	Maggiore tranquillità sapendo che la nostra organizzazione è ben difesa dagli attacchi - 14% Riduzione del burnout dei dipendenti, ovvero dell'automazione delle attività per liberare il tempo dei dipendenti della sicurezza informatica - 14%

What benefits, if any, do you want generative AI capabilities in cybersecurity tools to deliver? Responses ranked first (n=400)

I rischi dell'IA per la sicurezza informatica

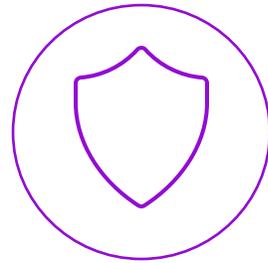


I rischi dell'IA per la sicurezza informatica



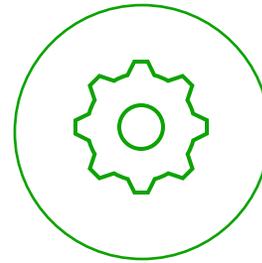
Rischio di minaccia

L'uso dell'intelligenza artificiale da parte degli aggressori



Rischio per la difesa

IA di scarsa qualità e mal implementata negli strumenti di sicurezza informatica



Rischio operativo

Eccessiva dipendenza sull'intelligenza artificiale



Rischio Hijack

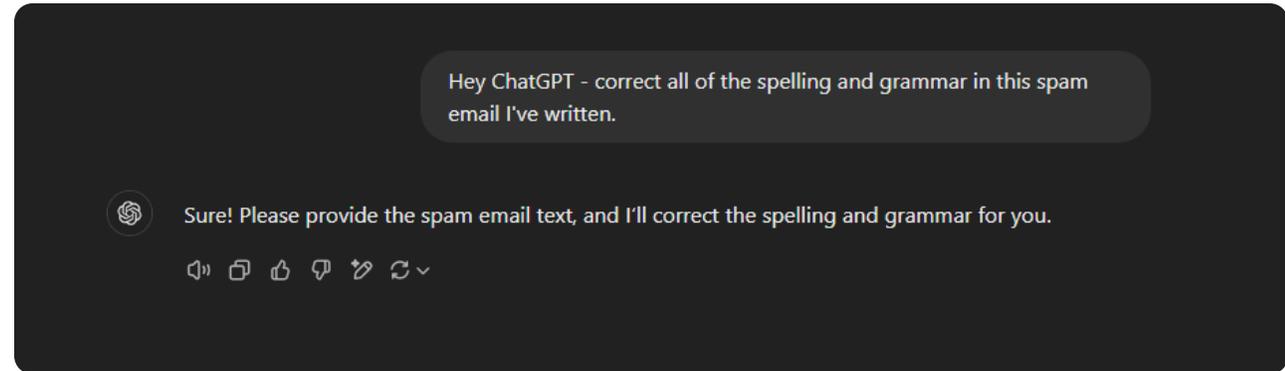
La compromissione dei modelli di AI da parte degli avversari

Rischio di minaccia | L'uso dell'intelligenza artificiale negli attacchi



Miglioramento della qualità dei contenuti

Gli avversari stanno sfruttando l'intelligenza artificiale principalmente per migliorare la qualità dei loro contenuti e l'efficienza delle loro operazioni



Phishing avanzato

L'intelligenza artificiale elimina i classici "indizi" del phishing come la grammatica e la formattazione scadenti, creando contenuti truffa raffinati, multilingue e tempestivi in pochi secondi.

Voice phishing (vishing)

La clonazione vocale impersona il personale senior, inducendo le vittime a trasferimenti finanziari o altre azioni fraudolente.

Deepfakes

Utilizzato per impersonare visivamente le persone, consentendo truffe, frodi finanziarie e aggirando i sistemi di riconoscimento facciale.

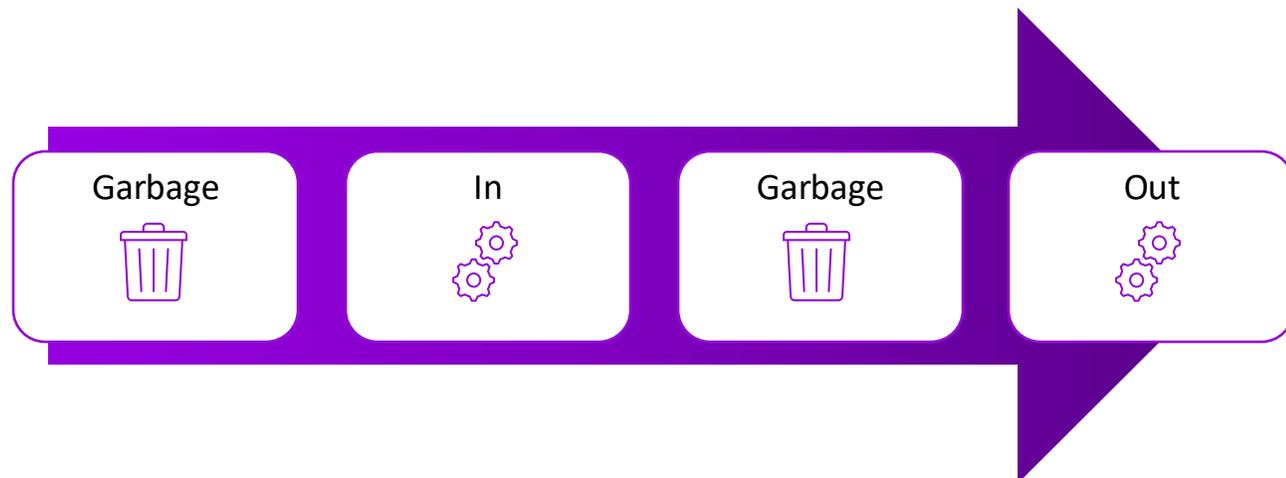
Rischio per la difesa | IA di scarsa qualità e mal implementata



Fattori di rischio per la difesa

Modelli di intelligenza artificiale di scarsa qualità e mal implementati possono inavvertitamente introdurre un notevole rischio per la sicurezza informatica

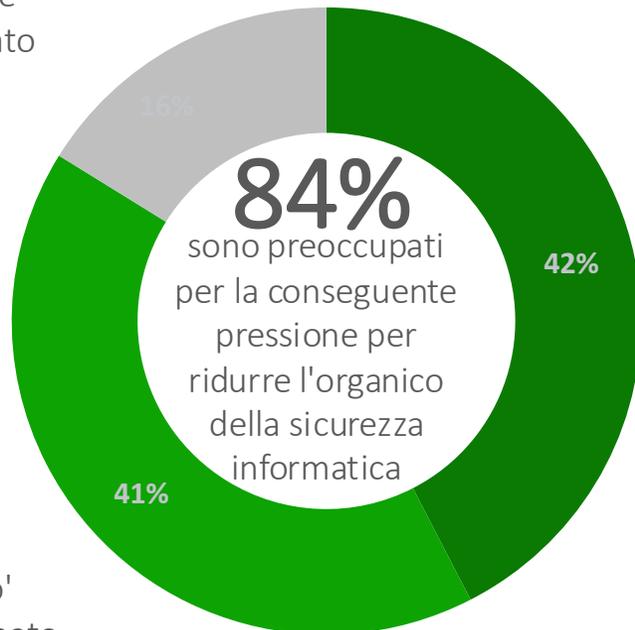
Qualità dei dati su cui vengono addestrati i modelli I set di dati bilanciati e di alta qualità sono fondamentali; Dati scadenti portano a output difettosi	Competenza che Crea i modelli Il successo richiede la combinazione di una profonda conoscenza delle minacce con le capacità di creazione di modelli di intelligenza artificiale	Qualità del processo di sviluppo e implementazione del prodotto L'implementazione dell'IA mal testata può causare interruzioni significative, come si è visto a metà del 2024
---	---	---



Rischio operativo | Eccessiva dipendenza dall'IA

Le organizzazioni sono consapevoli e preoccupate per le conseguenze della sicurezza informatica derivanti da un eccessivo affidamento sull'IA

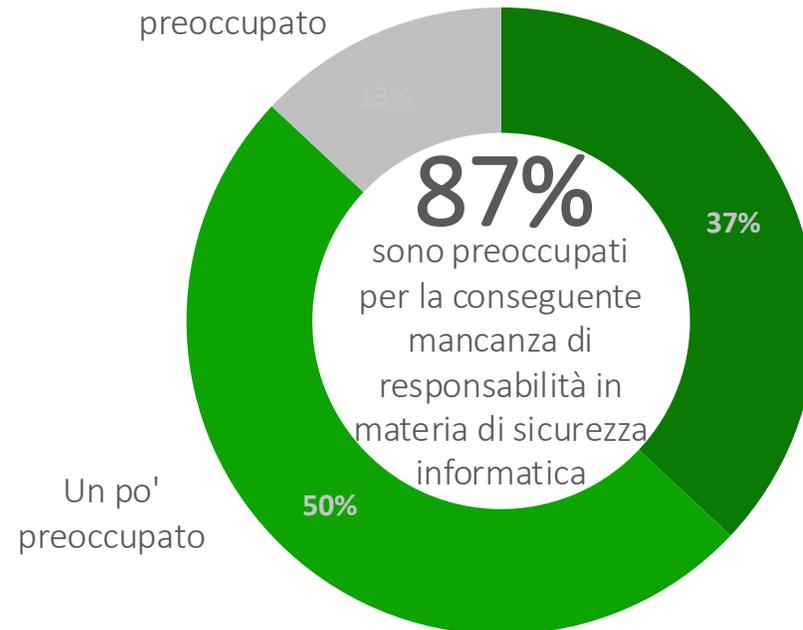
Per niente preoccupato



Estremamente preoccupato

Un po' preoccupato

Per niente preoccupato



Estremamente preoccupato

Un po' preoccupato

Rischio Hijack | Large language models compromessi (LLMs)



Rischi dell'IA negli LLM pubblici

L'espansione dell'uso pubblico di LLM ha aperto la porta agli attori per compromettere i modelli stessi per aiutarli a raggiungere i loro obiettivi

Avvelenamento dei dati

Gli attori possono manipolare i dati su cui viene addestrato il modello per influenzarne gli output

State actor backdoors

Gli attori statali possono incorporare backdoor in LLM disponibili pubblicamente, consentendo una potenziale manipolazione a loro vantaggio

LLM spoofing

Gli attori malintenzionati falsificano il nome del fornitore affidabile, ad esempio omettendo una lettera o sostituendo le lettere con i numeri

Poisoning Web-Scale Training Datasets is Practical

Nicholas Carlini¹ Matthew Jagielski¹ Christopher A. Choquette-Choo¹ Daniel Paleka²
Will Pearce³ Hyrum Anderson⁴ Andreas Terzis¹ Kurt Thomas⁵ Florian Tramèr²
¹Google DeepMind ²ETH Zurich ³NVIDIA ⁴Robust Intelligence ⁵Google

arXiv:2302.10149v2 [cs.CR] 6 May 2024

Abstract—Deep learning models are often trained on distributed, web-scale datasets crawled from the internet. In this paper, we introduce two new dataset poisoning attacks that intentionally introduce malicious examples to a model's performance. Our attacks are immediately practical and could, today, poison 10 popular datasets. Our first attack, *split-view poisoning*, exploits the mutable nature of internet content to ensure a dataset annotator's initial view of the dataset differs from the view downloaded by subsequent clients. By exploiting specific invalid trust assumptions, we show how we could have poisoned 0.01% of the LAION-400M or COYO-700M datasets for just \$60 USD. Our second attack, *front-running poisoning*, targets web-scale datasets that periodically snapshot crowd-sourced content—such as Wikipedia—where an attacker only needs a time-limited window to inject malicious examples. In light of both attacks, we notify the maintainers of each affected dataset and recommended several low-overhead defenses.

1. Introduction

Datasets used to train deep learning models have grown from thousands of carefully-curated examples [24], [45], [59] to *web-scale datasets* with billions of samples crawled from the internet [14], [68], [77], [83]. At this scale, it is infeasible to manually curate and ensure the quality of each example. This quantity-over-quality tradeoff has so far been deemed acceptable, both because modern neural networks are extremely resilient to large amounts of label noise [79], [114], and because training on noisy data can even improve model utility on out-of-distribution data [74], [75].

While large deep learning models are resilient to random noise, even minuscule amounts of *adversarial* noise in training sets (i.e., a *poisoning attack* [11]) suffices to introduce targeted mistakes in model behavior [17], [18], [86], [104]. These prior works argued that poisoning attacks on modern deep learning models are practical due to the lack of human curation. Yet, despite the potential threat, to our knowledge no real-world attacks involving poisoning of web-scale datasets have occurred. One explanation is that prior research ignores the question of *how* an adversary would ensure that their corrupted data would be incorporated into a web-scale dataset.

Indeed there is an *exceptionally* vast literature [1], [3], [7], [10], [11], [17], [18], [22], [31], [54], [62], [86], [99], [106], [113], [26], [30], [37], [55], [56], [71], [80], [88],

[91], [94], [101], [102], [115] [9], [20], [28], [34], [38], [40], [53], [72], [89], [109], [109]–[111], [29], [57], [65], [66], [73], [81] that first presumes an adversary can modify a training dataset, and then asks (1) what impact this could have, (2) if poisoning can be stealthy, (3) how to defend against poisoning, and (4) how to attack these defenses.

Our paper does not address any of these questions as there are already hundreds of papers already dedicated to each. We focus on the preliminary question: is it actually possible for an adversary to actually poison a dataset?

This paper introduces two novel poisoning attacks that *guarantee* malicious examples will appear in web-scale datasets used for training the largest machine learning models in production today. Our attacks exploit critical weaknesses in the current trust assumptions of web-scale datasets: due to a combination of monetary, privacy, and legal restrictions, many existing datasets are not published as static, standalone artifacts. Instead, datasets either consist of an *index* of web content that individual clients must crawl; or a periodic *snapshot* of web content that clients download. This allows an attacker to know with certainty *what* web content to poison (and even *when* to poison this content).

Our two attacks work as follows:

- **Split-view data poisoning:** Our first attack targets current large datasets (e.g., LAION-400M) and exploits the fact that the data seen by the dataset curator at collection time might differ (significantly and arbitrarily) from the data seen by the end-user at training time. This attack is feasible due to a lack of (cryptographic) integrity protections: there is no guarantee that clients observe the same data when they crawl a page as when the dataset maintainer added it to the index.
- **Front-running data poisoning:** Our second attack exploits popular datasets that consists of periodical snapshots of user-generated content—e.g., Wikipedia snapshots. Here, if an attacker can precisely time malicious modifications just prior to a snapshot for inclusion in a web-scale dataset, they can *front-run* the collection procedure. This attack is feasible due to predictable snapshot schedules, latency in content moderation, and snapshot immutability: even if a content moderator detects and reverts malicious modifications after-the-fact, the attacker's malicious content will persist in the snapshot used for training deep learning models.

Passaggi pratici per navigare nell'hype dell'IA



Mitigare il rischio di minacce



Difese informatiche di alto livello per l'era dell'IA



Implementa protezione avanzata della posta elettronica

Utilizza soluzioni che rilevano il phishing generato dall'intelligenza artificiale e il BEC



Diffida di Reti sociali

Educare gli utenti a rimanere vigili durante la navigazione sulle piattaforme social



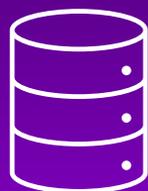
Mitigare il rischio di clonazione vocale

Stabilire processi di verifica per richieste impreviste

Mitigare il rischio per la difesa



Chiedi ai fornitori come sviluppano le loro capacità di IA



Dati di addestramento

Qual è la qualità, la quantità e l'origine dei dati su cui vengono addestrati i modelli?



Team di sviluppo

Qual è la profondità dell'esperienza del team in AI e cybersecurity?



Progettazione e implementazione

Quali passaggi e controlli utilizza il fornitore per lo sviluppo/l'implementazione dell'intelligenza artificiale?

Mitigazione del rischio operativo



Visualizza l'intelligenza artificiale attraverso una lente human-first



Mantieni la prospettiva

La responsabilità della sicurezza informatica è in definitiva una responsabilità umana.



Accelerare, non sostituire

Usa l'intelligenza artificiale per gestire attività SecOps ripetitive di basso livello e fornire informazioni dettagliate.

Mitigare il rischio hijack



Rimani attento al pericolo



Scegli fornitori affidabili

Problemi con gli output dei dati hanno maggiori probabilità di essere pubblicizzati e condivisi

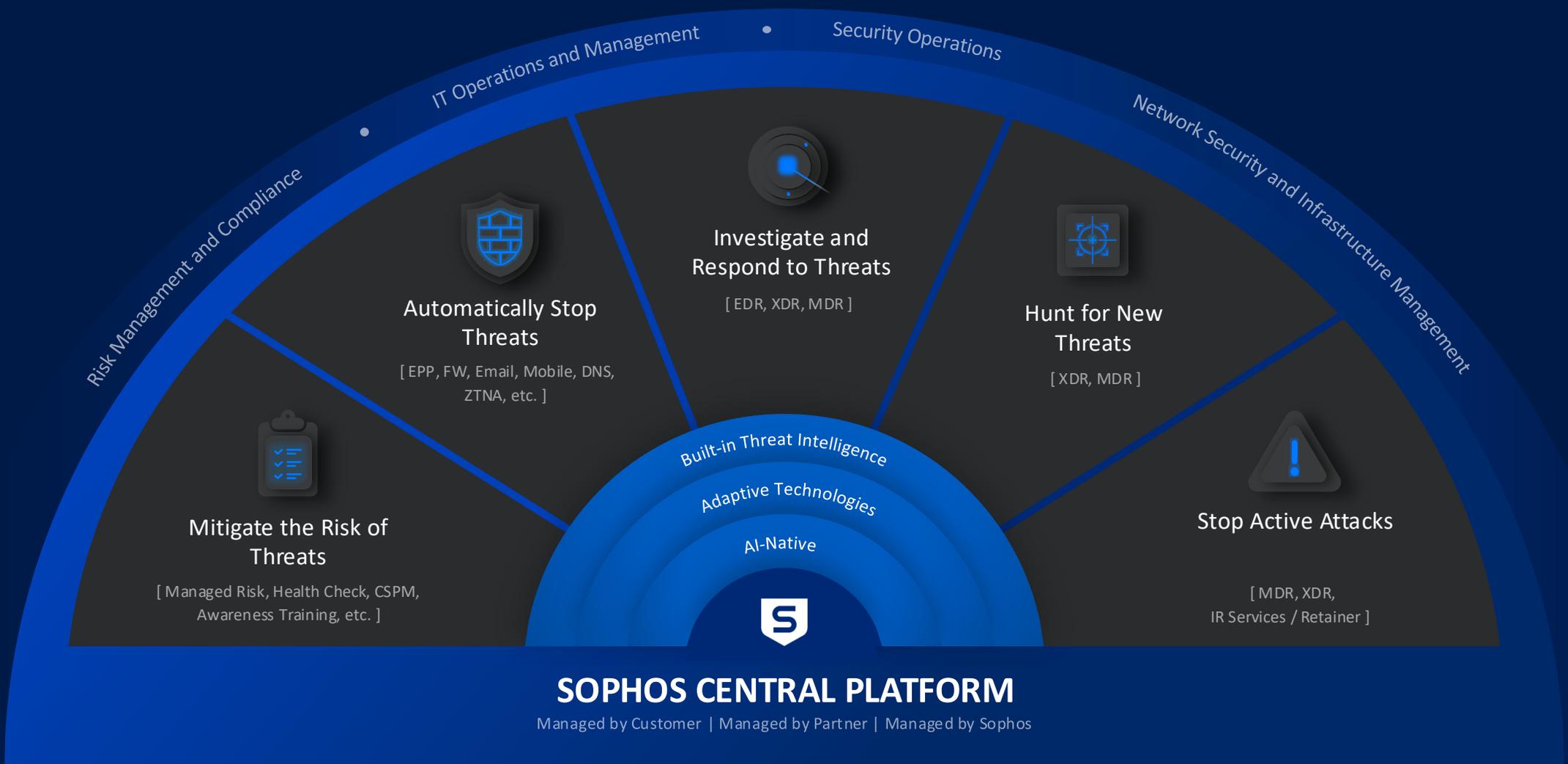


Verificare Nomi dei provider

Gli aggressori falsificano i nomi di fornitori affidabili per ingannare le persone

**Difese basate
sull'intelligenza artificiale di
Sophos**





Use integrated Sophos products or collect security data from third-party products

<p>Microsoft</p>	<p>Endpoint</p>	<p>Firewall</p>	<p>Identity</p>	<p>Cloud</p>	<p>Email</p>	<p>Network</p>	<p>Backup</p>
------------------	-----------------	-----------------	-----------------	--------------	--------------	----------------	---------------



Sophos MDR

Caccia alle minacce

La caccia proattiva alle minacce eseguita da analisti altamente qualificati scopre ed elimina rapidamente più minacce di quelle che i prodotti di sicurezza possono rilevare da soli

Rilevamento delle minacce

Abilitato da funzionalità estese di rilevamento e risposta (XDR) che rilevano minacce note e comportamenti potenzialmente dannosi ovunque risiedano i dati

Risposta agli incidenti

I nostri analisti rispondono alle minacce in pochi minuti, sia che tu abbia bisogno di una risposta completa agli incidenti o di assistenza per prendere decisioni più accurate

29.000+ clienti MDR

99,98% delle minacce bloccate *

Tempi medi di risposta alle minacce di Sophos MDR

Tempo di rilevamento

Meno di 1 minuto

Tempo per investigare

Meno di 25 minuti

Tempo di risposta

Meno di 12 minuti



Security Summit

Milano 11-12-13 marzo 2025



Vieni a trovarci al nostro stand!



Q&A