Objective For the We v4

Berlin, July 30-31 2026

macOS Malware Detection in 2020s: from Classical ML to Agentic AI

Biagio Montaruli
Sr Security Researcher @ Dynatrace Research (Vienna)
PhD on AI security from Eurecom and SAP Labs France
biagio.montaruli@dynatrace.com
biagio.montarul@gmail.com

Objective For the We v4

Berlin, July 30-31 2026

macOS Malware Detection in 2020s: from Classical ML to Agentic AI

WHY using ML for macOS malware detection? And HOW?

Biagio Montaruli
Sr Security Researcher @ Dynatrace Research (Vienna)
PhD on AI security from Eurecom and SAP Labs France
biagio.montaruli@dynatrace.com
biagio.montarul@gmail.com

Motivations

The Mac Malware of 2025 👾

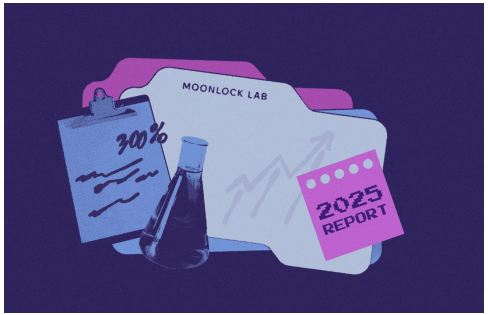
A comprehensive analysis of the year's new macOS malware

by: Patrick Wardle / January 1, 2025

Looking ahead, some predict macOS will achieve **full dominance** in the enterprise by the end of the decade:

> *"Mac will become the dominant enterprise endpoint by 2030."* — *Jamf*

Unsurprisingly, macOS malware is tracking this same growth curve, becoming more common, more capable, and more insidious with each passing year.



Here are a few key details from 2025:

- **A 67% increase** in registered macOS backdoor variants in 2025
- **17%** more macOS stealer variants seen in 2025
- **Over 80 countries affected** by major Mac stealer malware campaigns

Motivations

The Mac Malware of 2025 👾

A comprehensive analysis of the year's new macOS malware

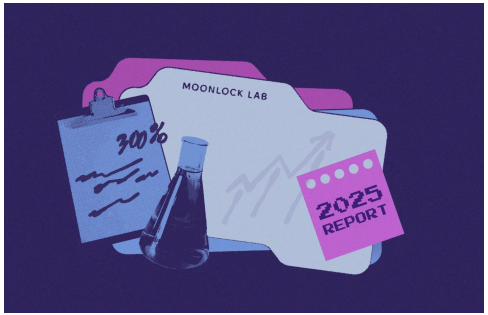
by: Patrick Wardle / January 1, 2025

Looking ahead, some predict macOS will achieve **full dominance** in the enterprise by the end of the decade:

> *"Mac will become the dominant enterprise endpoint by 2030." — Jamf*

Unsurprisingly, macOS mal[w]are is [b]ecoming more and more insidious with each passing year.

BIG GAPS IN THE CURRENT STATE-OF-THE-ART



Here are a few key details from 2025:

- **A 67% increase** in registered macOS backdoor variants in 2025
- **17%** more macOS stealer variants seen in 2025
- **Over 80 countries affected** by major Mac stealer malware campaigns

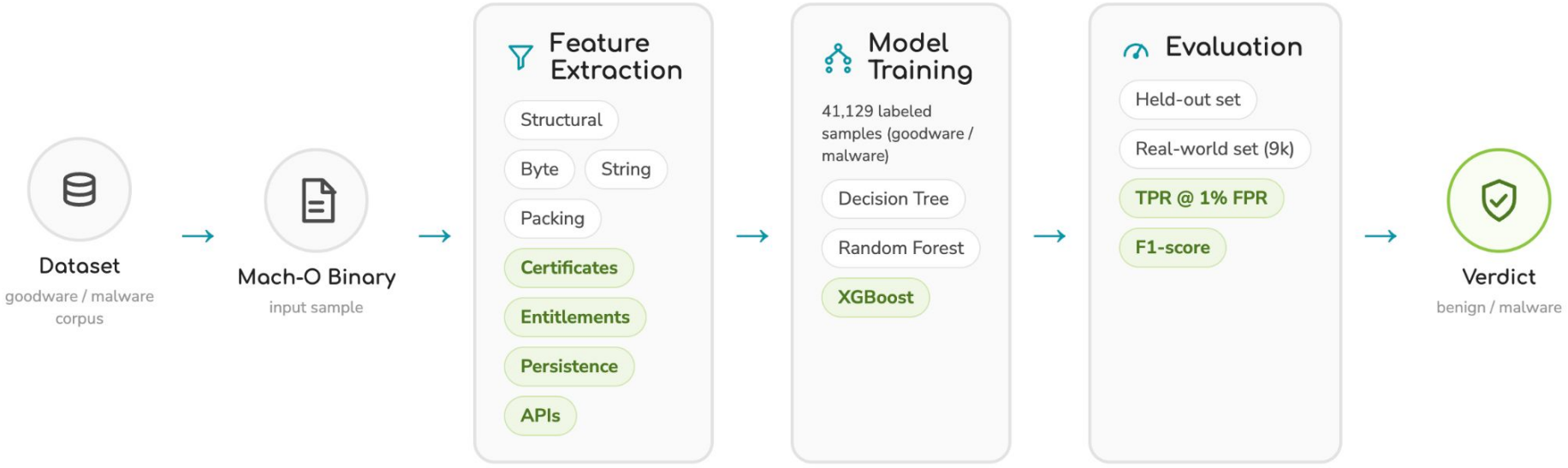
Research Gaps

- Lack of publicly available, state-of-the-art datasets of macOS binaries.
- macOS-specific features overlooked
  - Entitlements, certificates, persistence techniques, ...
- No study has investigated the effectiveness of ML-based macOS malware detectors across various dimensions:
  - Effectiveness of domain-specific features
  - Generalization capabilities in the wild

Work	Year	Dataset			Features
		Malware	Goodware	Available	
Pajouh *et al.*	2018	152	450	✗	S,T
Sahoo *et al.*	2022	152	450	✗	S,T
Chen *et al.*	2022	152	450	✗	S,T
Gharghasheh *et al.*	2022	152	450	✗	S,T
Thaeler *et al.*	2023	852	32,333	✗	S,T
This work	2025	29,716	11,413	✓	V

Table 5.14: Related work on machine learning for macOS malware detection. Features: S structural, T: string-based, V: various (also macOS-specific)

Classical ML for macOS Malware Detection



Classical ML for macOS Malware Detection

Dataset

goodware / malware corpus

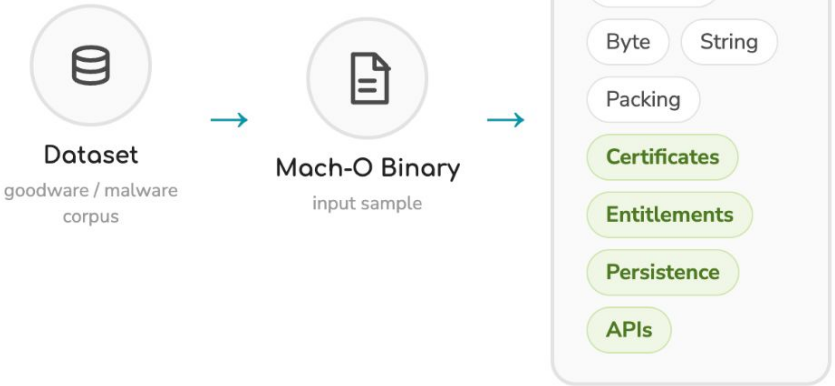
Goodware

- **Personal MacBoook (M1 Pro):** 1,239 Mach-O executables (macOS Sonoma 14.5)
- **Homebrew: 9,843 executables**
- **Open-source macOS apps (GitHub):** 1,045 binaries

Malware

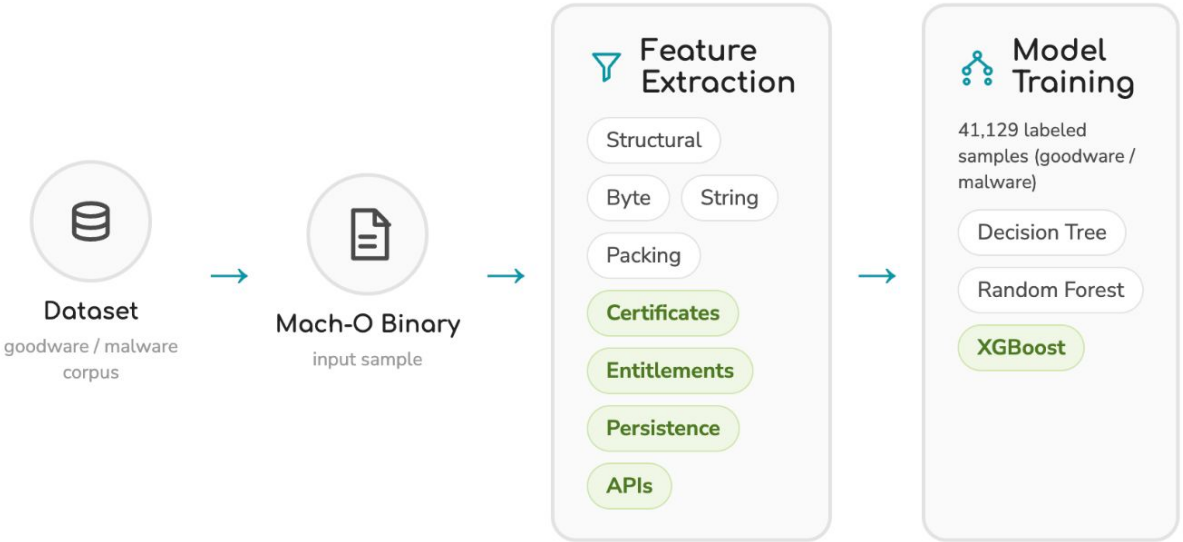
- **Objective-See dataset:** 180 samples
- **MalwareBazaar:** 90 samples (found by at least 5 AV)
- **Virus Samples Team dataset:** 103 samples
- **VirusShare:** 2,910 samples (found by at least 5 AV)
- **VirusTotal:**
  - VT feed from July 1st to November 13th 2023
  - 28,380 samples (found by at least 5 AV)

Classical ML for macOS Malware Detection

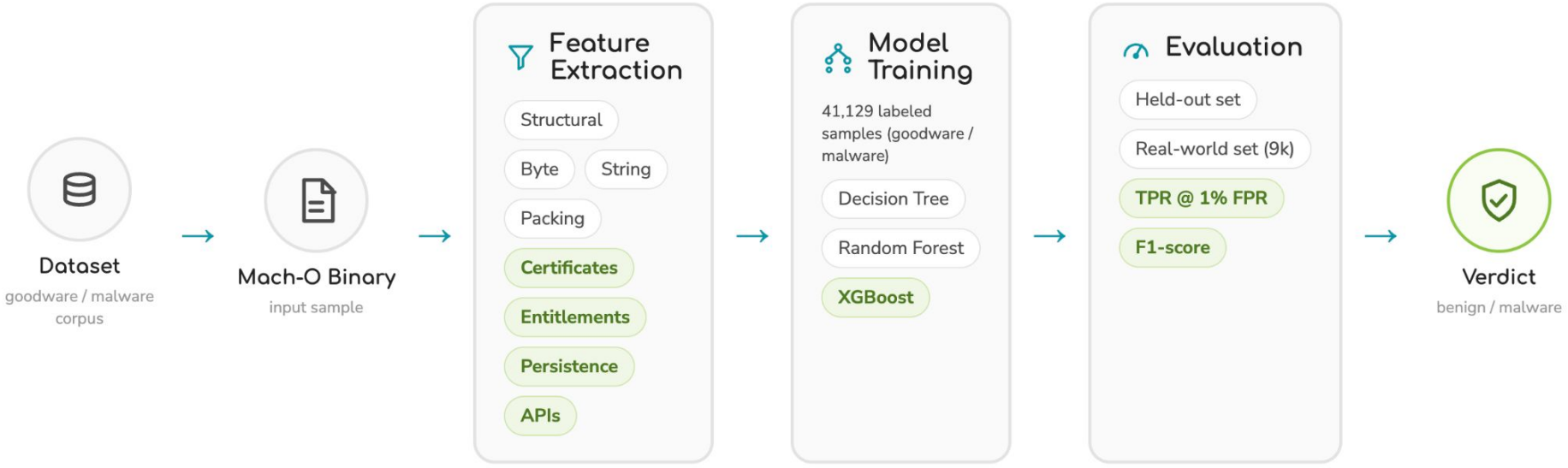


Category	Features	Type	Size
Structural	Sample Total Size	N	1
	Sample Virtual Size	N	1
	Header Flags	B	26
	Load Commands Histogram	N	50
	Min/Max/Avg Load Commands Size	N	3
	Min/Max/Avg Load Commands Entropy	N	3
	Min/Max/Avg Sections Size	N	3
	Min/Max/Avg Sections Entropy	N	3
	Min/Max/Avg Segments Size	N	3
	Min/Max/Avg Segments Virtual Size	N	3
	Min/Max/Avg Segments Entropy	N	3
	Segments Histogram	N	6
	Sections Histogram	N	11
Byte	Byte Histogram	N	256
	Bytes N-grams	N	128
String	Count of strings representing common IoC	N	5
Packing	Presence of sections with suspicious names	B	1
	Presence of segments (> 20%) with high-entropy data	B	1
API	Presence of common macOS API	B	4009
Entitlements	Presence of common and security-related entitlements	B	55
Persistence	Check for login and launch items	B	2
Certificates	Status of certificate chain	N	5

Classical ML for macOS Malware Detection



Classical ML for macOS Malware Detection



Comparison with state-of-the-art

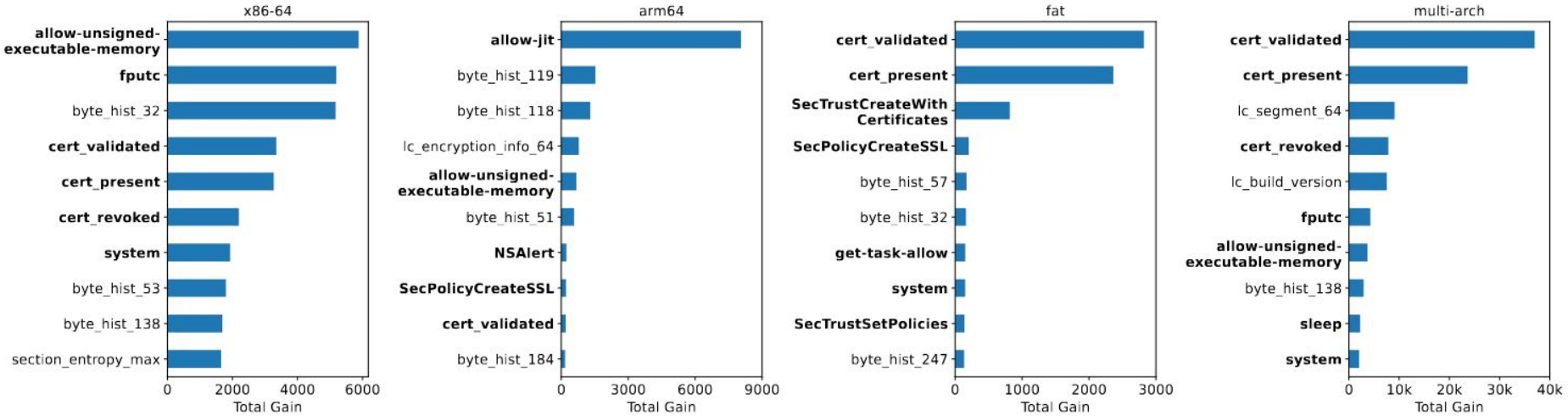
TPR @ 1% FPR

Features	Model	x86-64	arm64	fat	multi-arch
Our	DT	39.11	81.40	77.35	85.90
	RF	92.28	95.09	89.38	96.66
	XGBOOST	**96.87**	**97.07**	**90.80**	**98.50**
Pajouh-based	DT	37.38	66.43	66.01	62.81
	RF	87.68	80.91	77.18	88.62
	XGBOOST	**88.44**	**82.86**	**78.05**	**90.90**
Thaeler	DT	38.15	40.77	65.48	72.61
	RF	81.26	69.14	71.93	81.36
	XGBOOST	**86.78**	**69.25**	**72.57**	**90.23**
MalConv	-	**82.04**	**79.81**	**81.77**	**89.06**
Avg. Improv.		+13.07%	**+26.31%**	+17.49%	+9.39%

On average, across all benchmarking datasets, our detector achieves a relative improvement of **16.56%** compared to the state-of-the-art detectors

Feature Importance

Feature importance shows that the macOS-specific features play a key role



Feature Robustness

How performance changes when we remove the macOS-specific features?

Features	x86-64	arm64	fat	multi-arch
all	**96.87**	**97.07**	**90.80**	**98.50**
generic	95.16	92.60	86.49	96.68
specific	95.72	94.88	90.44	97.96
generic vs all	-1.76%	-4.60%	-4.74%	-1.85%

Removing the specific features results in only a small drop (3.23% on avg.)

- **all:** all the proposed features
- **generic:** baseline features commonly used in malware analysis such as structural, byte-based, string-based, and packing-based.
- **specific:** certificate chain status, entitlements, persistence, APIs

Feature Effectiveness for Generalization

TPR @ 1% FPR

Detector	x86-64	arm64	fat	multi-arch
our – all	**98.20**	73.47	97.42	99.50
our – generic	92.01	58.82	69.64	90.62
our – specific	97.74	**77.55**	**97.71**	**99.50**
generic vs all	-6.30%	-19.94%	-28.52%	-8.92%
specific vs all	-0.47%	+5.55%	+0.30%	+0.00%
Pajouh-based	71.98	54.43	60.86	68.38
Thaeler	70.14	53.77	57.49	67.72
MalConv	65.42	51.10	61.22	64.38
Avg. Improv.	+37.56%	+43.34%	+65.24%	+46.21%

- We built a new balanced dataset of 9,000 samples
- Collected from VT feed over a period of 3 months (September - November 2024)

The macOS-specific features are **more semantically-rich** and can maintain their efficacy even in presence of new variants.

Removing the macOS-specific features leads to **significant drop (15.92%)**

Real-world Accuracy

TPR @ 1% FPR

Detector	x86-64	arm64	fat	multi-arch
our – all	**98.20**	73.47	97.42	99.50
our – generic	92.01	58.82	69.64	90.62
our – specific	97.74	**77.55**	**97.71**	**99.50**
generic vs all	-6.30%	-19.94%	-28.52%	-8.92%
specific vs all	-0.47%	+5.55%	+0.30%	+0.00%
Pajouh-based	71.98	54.43	60.86	68.38
Thaeler	70.14	53.77	57.49	67.72
MalConv	65.42	51.10	61.22	64.38
Avg. Improv.	+37.56%	+43.34%	+65.24%	+46.21%

F1-score @ 1% FPR

Detector	x86-64	arm64	fat	multi-arch
our – all	94.92	75.91	96.41	96.74
our – generic	92.38	64.93	80.00	90.15
our – specific	**97.30**	**80.43**	**96.52**	**98.50**
generic vs all	-2.68%	-14.45%	-17.02%	-6.81%
specific vs all	+2.51%	+5.95%	+0.11%	+1.82%
Pajouh-based	81.83	60.12	64.10	75.21
Thaeler	79.52	56.75	62.84	72.85
MalConv	75.27	54.48	66.80	70.55
Avg. Improv.	+20.62%	+37.75%	+52.08%	+33.08%

Our detector maintains a very high detection rate (99.50%) and outperforms the s.o.t.a. by 50% and 37.34% in terms of TPR and F1-score

Key takeaways

1) **A novel dataset of 41,129 samples** (11,413 benign and 29,716 malware)

2) A new ML-based macOS malware detector that leverages **novel macOS-specific features**: status of certificates, entitlements, persistence techniques, and key system APIs.
   - Outstanding performance: **98.50% at 1% FPR**
   - **+16% improvement** w.r.t. state-of-the-art solutions

3) In-depth analysis of the feature importance, showing that **our detector effectively leverages the new domain-specific features**.

4) Evaluation on a new fresh dataset of 9,000 samples:
   - It confirms state-of-the-art performance: **99.50% TPR (+50% w.r.t.sota)**
   - The domain-specific features are crucial for generalizing to novel malware samples, as **their removal leads to a 15.92% performance drop**

Paper published at AsiaCCS 2026

The Role of Domain-Specific Features in Malware Detection: A macOS Case Study

Biagio Montaruli
montarul@eurecom.fr
EURECOM & SAP
France

Andrea Oliveri
oliveri@eurecom.fr
EURECOM
France

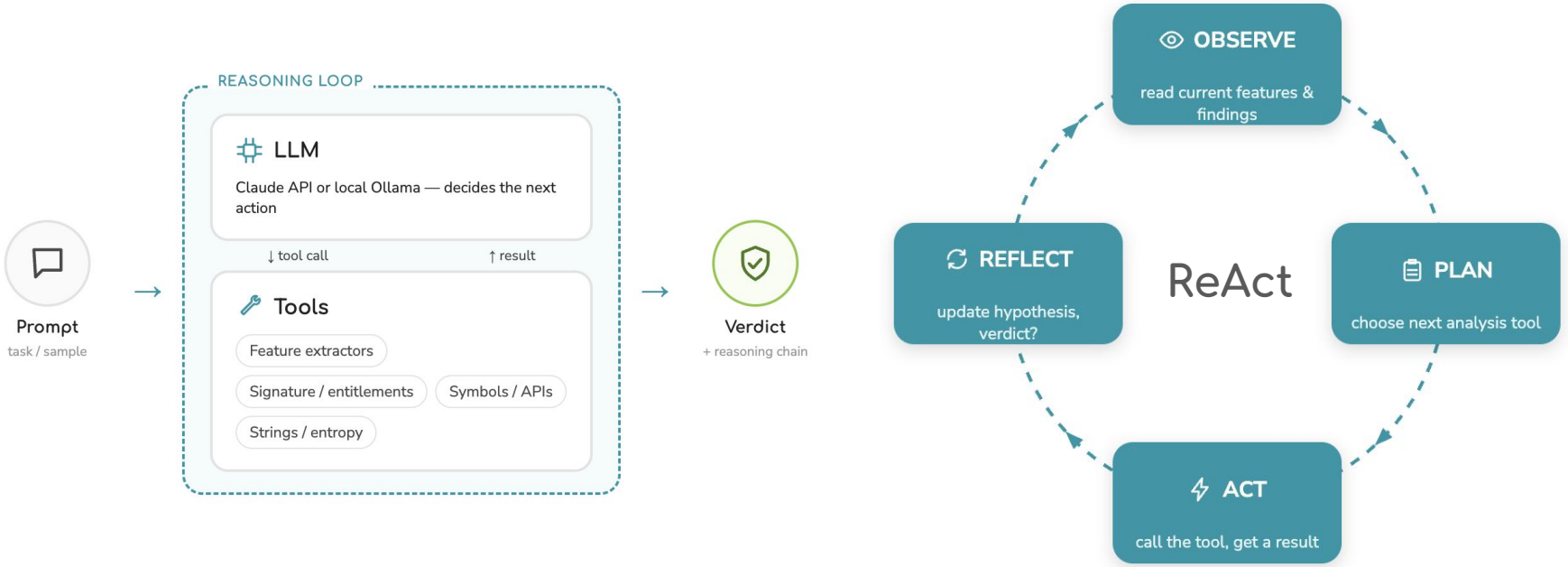
Savino Dambra
savino.dambra@gendigital.com
GenDigital
France

Davide Balzarotti
balzarot@eurecom.fr
EURECOM
France

https://dl.acm.org/doi/abs/10.1145/3779208.3785392

https://github.com/eurecom-s3/macos-malware-dataset

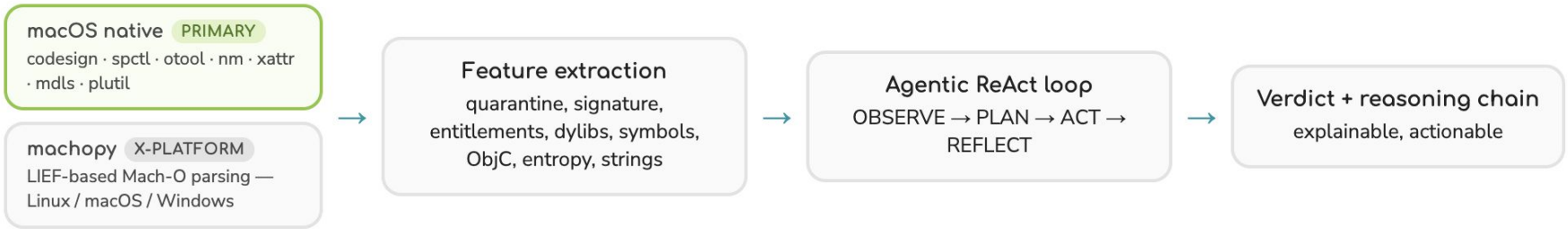
From Classical ML to Agentic AI



MacSkillet

https://github.com/biagiom/macskillet

- An **AI skill** for Agentic AI classification of macOS apps/binaries
- **Multi-frameworks:** Claude, Local LLMs (ollama), Apple Foundational Model API
- **Tightly-coupled with macOS:** use macOS cmd tools and Apple FM API
- **Analyst-friendly:**
  - risk-based score
  - full-explainable reasoning chain



Objective For the We v4

Berlin, July 30-31 2026

Thanks for the attention!

Kudos to Patrick, Andy, Objective-See and HPI for organizing such an amazing event !!!