Big Data e Machine Learning per la Cyber Threat Intelligence

Metodologie, architetture e sfide aperte

Abstract

L'industrializzazione dell'offesa cyber — campagne continue e convergenti, modelli *as-a-service*, impiego sistematico dell'intelligenza artificiale sul versante offensivo — ha reso strutturalmente insufficienti le difese fondate su firme ed euristiche. La Cyber Threat Intelligence (CTI) propone un rovesciamento di prospettiva, dalla vulnerabilità all'avversario, ma la sua praticabilità dipende oggi dalla capacità di trattare volumi di dati che eccedono ogni possibilità di analisi manuale. Questo articolo esamina in chiave critica la confluenza di Big Data e Machine Learning nel ciclo della CTI, assumendo quest'ultimo come asse ordinatore. Dopo aver fissato i fondamenti della disciplina — livelli dell'intelligence, tassonomia degli attori e delle loro *tactics, techniques and procedures*, framework di analisi strutturata (Cyber Kill Chain, Diamond Model, MITRE ATT&CK) — il lavoro propone un'architettura di riferimento per la pipeline Big Data a supporto della CTI e ne percorre gli strati: le fonti e la loro qualità; i metodi di Machine Learning per la *threat detection* su dato strutturato; le tecniche di Natural Language Processing per l'estrazione di intelligence da fonti aperte, incluso il monitoraggio di *deep* e *dark web*; gli standard, le piattaforme e i modelli organizzativi della condivisione (STIX/TAXII, MISP, ISAC, CERT/CSIRT, SOC) nel quadro normativo europeo; l'attribuzione degli attacchi e i suoi limiti epistemici; le vulnerabilità *adversarial* dei modelli e le sfide aperte. L'analisi converge su quattro tesi: il valore dell'intelligence cresce con il livello di astrazione e con esso il costo di automatizzarla; il collo di bottiglia della difesa non è algoritmico ma di dato e di organizzazione; le illusioni di valutazione che affliggono la letteratura sono sistemiche; lo strato di Machine Learning costituisce esso stesso una superficie d'attacco.

Parole chiave: Cyber Threat Intelligence, Big Data, Machine Learning, Natural Language Processing, threat detection, information sharing, MITRE ATT&CK, STIX/TAXII, adversarial machine learning, attribuzione

Indice

1. Introduzione

1.1 Contesto e motivazione

Negli ultimi anni il quadro delle minacce cyber ha assunto i tratti di un'economia industriale. L'edizione 2025 dell'*ENISA Threat Landscape* analizza 4.875 incidenti registrati tra il 1° luglio 2024 e il 30 giugno 2025 e restituisce l'immagine di un ecosistema offensivo maturo. Il phishing, nelle sue varianti di vishing, malspam e malvertising, vi costituisce il vettore primario di intrusione iniziale in circa il 60% dei casi osservati, con modelli *Phishing-as-a-Service* che distribuiscono kit preconfezionati e rendono le campagne accessibili ad attaccanti di qualunque livello di esperienza. La stessa dinamica di *commoditization* investe l'estorsione digitale. La proliferazione di programmi *Ransomware-as-a-Service* ha abbassato le barriere d'ingresso al crimine informatico, e nel solo periodo di riferimento sono state censite 82 varianti ransomware distinte impiegate contro organizzazioni degli Stati membri UE. Vi si aggiunge l'impiego sistematico dell'intelligenza artificiale sul versante offensivo: a inizio 2025 le campagne di phishing supportate da AI rappresentavano oltre l'80% dell'attività di social engineering osservata a livello globale, mentre l'emergere di sistemi AI malevoli autonomi introduce una scalabilità inedita nell'attività ostile. Il fenomeno complessivo descritto da ENISA non consiste più in singoli eventi ad alto impatto, bensì in campagne continue, diversificate e convergenti, di matrice criminale, state-nexus e hacktivista, che erodono progressivamente la resilienza dei sistemi [1].

Sul versante difensivo, la contropartita di questa industrializzazione è un problema di scala dei dati. L'istituto AV-TEST registra quotidianamente oltre 450.000 nuovi programmi malevoli e applicazioni potenzialmente indesiderate, per un archivio cumulativo che ha superato 1,56 miliardi di campioni noti [2]. A questi si sommano i volumi di telemetria interna — log applicativi e di sistema, flussi di rete, eventi endpoint, audit trail di ambienti cloud — che ogni organizzazione di media complessità produce nell'ordine dei terabyte giornalieri. La capacità analitica umana non regge il confronto. Nello studio qualitativo condotto da AlAhmadi et al. su analisti di Security Operations Center, i practitioner confermano tassi di falsi positivi talmente elevati da imporre la validazione manuale sistematica degli allarmi; il dato più significativo è che gran parte dei presunti falsi positivi risulta in realtà costituita da *benign trigger*, ossia allarmi tecnicamente corretti ma spiegabili con comportamenti legittimi del contesto organizzativo [3]. Il problema, in altri termini, non riguarda soltanto il volume degli eventi, ma l'assenza del contesto necessario a interpretarli, che è poi ciò che l'intelligence dovrebbe fornire.

In questo scenario la Cyber Threat Intelligence (CTI) si è affermata come disciplina autonoma. Le minacce di nuova generazione, evasive e resilienti, mettono in difficoltà i sistemi di sicurezza tradizionali basati su firme ed euristiche [4], concepiti per riconoscere il già noto. La CTI ribalta l'impostazione. Nella definizione ormai canonica proposta in ambito Gartner e ripresa dalla letteratura, essa è conoscenza *evidence-based* relativa a minacce esistenti o emergenti, comprensiva di contesto, meccanismi, indicatori, implicazioni e raccomandazioni azionabili, destinata a informare le decisioni di risposta del soggetto difeso [5], [6]. La distinzione concettuale tra *dato* (l'osservabile grezzo), *informazione* (il dato aggregato e contestualizzato) e *intelligence* (il prodotto analitico orientato alla decisione) attraversa

l'intero ciclo di produzione della CTI, e segna il passaggio da una postura difensiva reattiva a una proattiva e, in prospettiva, predittiva.

La tesi che questo lavoro intende articolare è che tale passaggio sia oggi tecnicamente praticabile soltanto alla confluenza di due traiettorie tecnologiche. La prima è quella del Big Data. Le fonti della CTI, dalla telemetria ai feed strutturati fino alle fonti aperte su surface, deep e dark web, presentano esattamente le caratteristiche di volume, velocità, varietà e incerta veridicità che definiscono il paradigma, e richiedono architetture di *ingestion* e *processing* progettate di conseguenza. La seconda traiettoria è quella del Machine Learning, unico strumento in grado di estrarre da quei volumi pattern, anomalie ed entità a costi computazionali sostenibili, tanto sul dato strutturato (detection di intrusioni e malware) quanto su quello non strutturato (elaborazione del linguaggio naturale su report, forum underground, social media). L'integrazione delle due traiettorie, tuttavia, è tutt'altro che compiuta: la maggior parte delle organizzazioni si ferma a casi d'uso elementari, come l'integrazione di feed di threat data nei sistemi SIEM, di prevenzione delle intrusioni e nei firewall esistenti, senza sfruttare il potenziale analitico che la nuova intelligence renderebbe disponibile [4]. A questo divario di adozione si aggiunge una fragilità strutturale, poiché gli stessi modelli di ML che abilitano la difesa costituiscono una superficie d'attacco autonoma, esposta a manipolazioni adversarial di cui si darà conto nella parte conclusiva del lavoro. La tensione tra potenziale e fragilità di questa convergenza è la motivazione di fondo del presente contributo.

1.2 Obiettivi e contributi

L'obiettivo del lavoro è sistematizzare in chiave critica l'intersezione tra Big Data, Machine Learning e Cyber Threat Intelligence, assumendo il ciclo di intelligence come asse ordinatore: dalle fonti e dalle architetture di raccolta, attraverso i metodi analitici, fino alla disseminazione e allo scambio informativo tra organizzazioni. Rispetto alle survey esistenti, tipicamente verticali su singole tecniche o singole fasi, il contributo si articola su sei piani: (i) una ricostruzione unitaria dei fondamenti della disciplina, ossia ciclo e livelli della CTI, tassonomia degli attori e delle loro tactics, techniques and procedures, framework di analisi strutturata (Cyber Kill Chain, Diamond Model, MITRE ATT&CK); (ii) un'architettura di riferimento per la pipeline Big Data a supporto della CTI; (iii) una rassegna critica dei metodi di ML per la threat detection su dato strutturato e delle tecniche NLP per l'estrazione di intelligence da fonti aperte, incluse quelle del dark web; (iv) un'analisi degli standard di rappresentazione e trasporto (STIX, TAXII), delle piattaforme di condivisione e dei modelli organizzativi della difesa cooperativa (ISAC, CERT, CSIRT, SOC), collocata nel quadro normativo europeo; (v) una trattazione dell'attribuzione degli attacchi come problema analitico dotato di limiti epistemici propri; (vi) una discussione delle vulnerabilità adversarial dei modelli e delle direttrici di ricerca aperte. Destinatari del lavoro sono tanto la comunità di ricerca quanto i practitioner, analisti SOC e CSIRT e threat intelligence team, interessati a una mappa ragionata dello stato dell'arte.

1.3 Criteri di selezione delle fonti

Il lavoro si configura come rassegna critica di tipo narrativo e non rivendica l'esaustività di una systematic literature review. La selezione delle fonti ha comunque seguito criteri espliciti. Sono state privilegiate pubblicazioni peer-reviewed del periodo 2018–2026, con preferenza

per i lavori successivi al 2020, apparse su riviste di fascia alta del settore (*IEEE Communications Surveys & Tutorials*, *ACM Computing Surveys*, *Computers & Security*, *Journal of Information Security and Applications*, *Computer Science Review*) o negli atti delle principali conferenze di sicurezza (NDSS, USENIX Security, IEEE S&P, ACM CCS). A queste si affiancano, con statuto probatorio dichiaratamente diverso, le specifiche tecniche degli enti di standardizzazione (OASIS, MITRE, NIST) e la reportistica istituzionale (ENISA), impiegata per l'inquadramento empirico del threat landscape. Il ricorso a fonti industriali è limitato ai casi in cui costituiscono l'unica base dati disponibile ed è sempre segnalato come tale.

1.4 Struttura dell'articolo

Il resto del lavoro è organizzato come segue. Il capitolo 2 fissa il lessico disciplinare: ciclo di intelligence, livelli della CTI, attori della minaccia e framework di analisi strutturata. Il capitolo 3 esamina il substrato dei dati, ne discute fonti, caratteristiche e architetture di raccolta ed elaborazione, e introduce la pipeline di riferimento richiamata nei capitoli successivi. I capitoli 4 e 5 trattano il momento analitico: il primo sul dato strutturato, con i metodi di ML per intrusion e malware detection; il secondo sul dato non strutturato, con le tecniche NLP e il monitoraggio OSINT di surface, deep e dark web. Il capitolo 6 affronta la disseminazione e lo scambio, dagli standard alle piattaforme, dai modelli organizzativi alle barriere alla condivisione. Il capitolo 7 discute l'attribuzione degli attacchi. Il capitolo 8 esamina le vulnerabilità adversarial dei modelli di ML e le sfide aperte, e il capitolo 9 conclude.

2. Fondamenti di Cyber Threat Intelligence

Questo capitolo fissa il lessico concettuale impiegato nel resto del lavoro: il ciclo di produzione dell'intelligence, i livelli in cui la CTI si articola, la tassonomia degli attori della minaccia e i tre framework di analisi strutturata che dominano la pratica e la letteratura. Si tratta di nozioni preliminari solo in apparenza. Come mostreranno i capitoli 3–5, le scelte architetturali e algoritmiche delle pipeline Big Data–ML sono in larga misura determinate da quale fase del ciclo, quale livello di intelligence e quale rappresentazione della minaccia si intende servire.

2.1 Definizioni e ciclo di intelligence

La definizione di CTI richiamata nel §1.1, conoscenza evidence-based su minacce esistenti o emergenti corredata di contesto, indicatori e raccomandazioni azionabili [5], [6], implica un processo di produzione, che la disciplina eredita dagli intelligence studies classici nella forma del ciclo di intelligence. Nella sua declinazione cyber, il ciclo è comunemente articolato in sei fasi (Figura 1). La *pianificazione e direzione* traduce le esigenze conoscitive dell'organizzazione in Priority Intelligence Requirements: quali asset proteggere, quali avversari e settori osservare, quali domande l'intelligence deve saper soddisfare. La *raccolta* acquisisce dati grezzi dalle fonti individuate, interne (telemetria di rete, log, eventi endpoint) ed esterne (feed commerciali e comunitari, fonti aperte su surface, deep e dark web). L'*elaborazione* trasforma il dato grezzo in materiale analizzabile attraverso normalizzazione dei formati, deduplicazione, arricchimento con metadati di contesto e strutturazione in rappresentazioni standard. L'*analisi* è il momento propriamente intellettivo, in cui l'informazione diventa intelligence attraverso correlazione, valutazione di ipotesi e tecniche analitiche strutturate; qui, come mostreranno i capitoli 4 e 5, il Machine Learning ha prodotto la trasformazione più profonda. La *disseminazione* consegna il prodotto analitico ai destinatari, in forma calibrata sul livello di intelligence richiesto. Il *feedback*, infine, chiude l'anello valutando la rilevanza dei prodotti erogati e revisionando i requisiti.

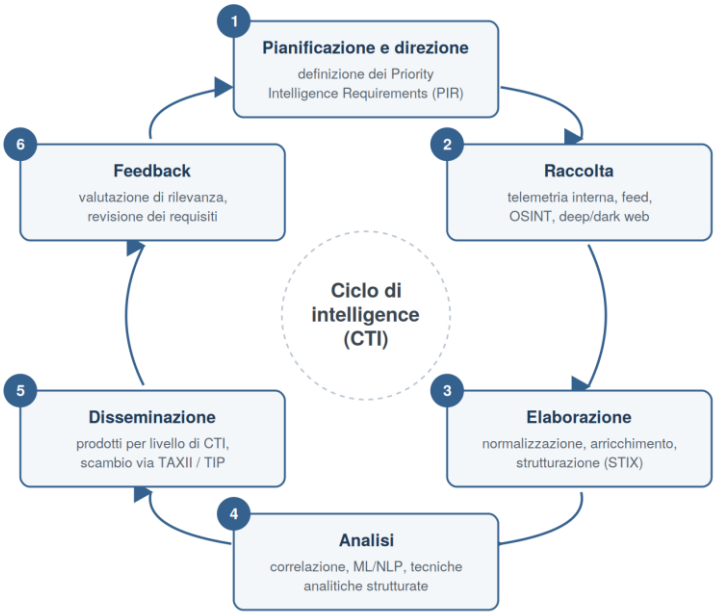

Figura 1 — Il ciclo di intelligence applicato alla Cyber Threat Intelligence. Le annotazioni indicano gli elementi caratteristici della declinazione cyber di ciascuna fase.

Il modello ciclico va assunto con la consapevolezza dei suoi limiti. Già nella letteratura degli intelligence studies Hulnick ne ha contestato l'idealizzazione, osservando che nella pratica le fasi procedono in parallelo più che in sequenza e che raccolta e analisi raramente si attendono a vicenda [7]. La critica vale a maggior ragione nel dominio cyber, dove raccolta ed elaborazione operano a velocità di macchina e comprimono l'anello fino a renderlo quasi istantaneo per la CTI di livello tecnico. Il ciclo conserva tuttavia un valore ordinatore che questo lavoro sfrutta deliberatamente: il capitolo 3 copre raccolta ed elaborazione, i capitoli 4 e 5 l'analisi, il capitolo 6 la disseminazione e lo scambio.

2.2 Livelli della CTI: strategica, tattica, operativa, tecnica

La CTI non è un prodotto omogeneo. La letteratura la articola convenzionalmente su quattro livelli, distinti per destinatari, orizzonte temporale e grado di astrazione [5], [8]. L'intelligence *strategica* si rivolge ai vertici decisionali con analisi di lungo periodo su tendenze della minaccia, motivazioni e capacità degli attori, contesto geopolitico e settoriale. L'intelligence *tattica* descrive le tactics, techniques and procedures (TTPs) degli avversari ed è destinata a chi progetta e mantiene le difese. L'intelligence *operativa* riguarda campagne specifiche, in corso o imminenti, e alimenta il lavoro di incident responder e threat hunter. L'intelligence *tecnica*, infine, è costituita dagli indicatori di compromissione (IoC) atomici come hash, indirizzi IP, domini e URL, destinati al consumo automatico da parte degli strumenti di difesa e caratterizzati da rapida obsolescenza. La Tabella 1 sintetizza la partizione. Va segnalato che la terminologia non è del tutto stabile: una parte delle fonti inverte le etichette dei livelli tattico e operativo, e la valutazione sistematica di tassonomie, standard e ontologie condotta da Mavroeidis e Bromander mostra che alla comunità manca tuttora un'ontologia capace di coprire l'intero spettro della threat intelligence [8], assenza che, come si vedrà nel capitolo 6, ha conseguenze concrete sull'interoperabilità dello scambio informativo.

Livello	Destinatari tipici	Orizzonte temporale	Contenuto caratteristico	Output
Strategica	vertici aziendali, CISO, funzioni di risk management	mesi–anni	trend della minaccia, motivazioni e capacità degli attori, contesto geopolitico e settoriale	report narrativi, threat landscape, briefing
Tattica	security architect, detection engineer	settimane–mesi	TTPs degli avversari, mappatura su framework (ATT&CK), difese aggirabili	report tecnici, regole e analytics di detection
Operativa	SOC, incident responder, threat hunter	giorni–settimane	campagne specifiche, infrastrutture e capacità impiegate dall'avversario	advisory, alert contestualizzati, hunting package
Tecnica	strumenti automatici (SIEM, EDR, TIP, firewall)	ore–giorni	IoC atomici: hash, indirizzi IP, domini, URL	feed machine-readable (STIX via TAXII)

Tabella 1 — I livelli della Cyber Threat Intelligence.

Alla partizione per livelli si affianca, come chiave di lettura economica, la Pyramid of Pain proposta da Bianco [9]. Gli indicatori vi sono ordinati per il costo che la loro neutralizzazione impone all'avversario: dagli hash, banali da rigenerare, a indirizzi IP e domini, su fino ad artefatti, tool e, al vertice, alle TTPs, la cui modifica costringe l'attaccante a riprogettare il

proprio modus operandi. Ne discende un principio che orienta l'intero lavoro. Il valore difensivo dell'intelligence cresce con il livello di astrazione, ma cresce in pari misura la difficoltà di produrla automaticamente: estrarre hash da un feed è banale, estrarre TTPs da report non strutturati richiede le tecniche NLP discusse nel capitolo 5.

2.3 Threat actor e TTPs

La nozione di TTP merita una precisazione, poiché costituisce l'unità descrittiva fondamentale dell'intelligence tattica. Nella formalizzazione di MITRE, le *tactics* rappresentano gli obiettivi tattici dell'avversario, cioè il *perché* di un'azione; le *techniques* sono i modi con cui tali obiettivi vengono conseguiti; le *procedures* le implementazioni specifiche osservate in campagne reali [10]. La gerarchia cattura livelli crescenti di specificità e, specularmente, di volatilità: le procedure cambiano di campagna in campagna, le tattiche restano stabili per anni.

Quanto agli attori, la letteratura e la reportistica istituzionale convergono su una tassonomia per motivazione. Gli attori *state-nexus* perseguono spionaggio, prepotenziamento e sabotaggio. A questa categoria è storicamente associata l'etichetta di Advanced Persistent Threat, coniata per designare avversari dotati di risorse e addestramento in grado di condurre campagne intrusive pluriennali contro informazioni economiche, proprietarie o di sicurezza nazionale [11], benché la survey di Alshamrani et al. documenti come la classe di minaccia abbia da tempo esteso la propria zona d'azione dagli apparati statali ai settori privato e corporate [12]. La *criminalità organizzata* è mossa dal profitto e opera oggi come filiera: initial access broker che commerciano credenziali, operatori di infostealer, affiliati di programmi Ransomware-as-a-Service. L'*hacktivism* persegue finalità ideologiche con azioni a forte visibilità, tipicamente DDoS e defacement. Chiudono la tassonomia gli *insider* e gli attori opportunistici a bassa capacità.

I dati empirici più recenti mostrano però quanto la tassonomia vada maneggiata come lente analitica e non come compartimentazione rigida. Nel periodo osservato dall'ETL 2025, il 79,4% degli incidenti valutati risulta a matrice ideologica, in larghissima parte campagne DDoS hacktiviste di cui appena il 2% produce un'effettiva interruzione di servizio, contro un 13,4% a movente finanziario e un 7,2% riconducibile a cyberespionage; eppure il ransomware, pur in calo dell'11%, resta la minaccia a più alto impatto [1]. L'asimmetria tra volume e impatto è un primo monito metodologico per chi addestra modelli su dati di incidenti, poiché la frequenza di una classe non ne misura la rilevanza. Un secondo monito viene dalla convergenza tra categorie: la stessa ENISA documenta gruppi hacktivisti che adottano ransomware a fini di finanziamento e visibilità, con formazioni come FunkSec che fondono messaggistica politica ed estorsione, oltre a un diffuso riuso di strumenti e tecniche tra insiemi di intrusione diversi [1]. Per la CTI, la profilazione per motivazione resta preziosa ai fini della prioritizzazione, ma i modelli, umani o automatici che siano, devono saper trattare l'ibridazione.

2.4 Framework di analisi strutturata: Cyber Kill Chain, Diamond Model, MITRE ATT&CK

Perché l'analisi della minaccia sia cumulabile e condivisibile, gli eventi ostili devono essere descritti secondo modelli comuni. Tre framework dominano la pratica; ciascuno incorpora una prospettiva diversa sull'intrusione, e la loro complementarità è riassunta nella Tabella 2.

La **Cyber Kill Chain**, formulata da Hutchins, Cloppert e Amin in ambito Lockheed Martin, muove da una diagnosi. Gli strumenti convenzionali di difesa si concentrano sulla componente di vulnerabilità del rischio, mentre la risposta agli incidenti presuppone tradizionalmente un'intrusione già riuscita, impostazione inadeguata di fronte ad avversari persistenti [11]. Il modello scompone l'intrusione in sette fasi sequenziali (reconnaissance, weaponization, delivery, exploitation, installation, command and control, actions on objectives) e fonda su questa scomposizione una difesa *intelligence-driven*: poiché l'attaccante deve completare l'intera catena per conseguire l'obiettivo, al difensore basta spezzarla in un punto qualsiasi. La courses of action matrix associa a ciascuna fase sei possibili azioni difensive (detect, deny, disrupt, degrade, deceive, destroy) mutuate dalla dottrina delle information operations del Dipartimento della Difesa statunitense [11]. Il merito storico del modello è aver spostato il baricentro dalla vulnerabilità all'avversario; i suoi limiti, ampiamente discussi, risiedono nella linearità (le intrusioni reali sono iterative), nell'impronta perimetrale e malware-centrica e nella scarsa granularità delle fasi post-compromissione.

Il **Diamond Model** di Caltagirone, Pendergast e Betz adotta una prospettiva relazionale anziché sequenziale. L'elemento atomico dell'analisi è l'*evento* di intrusione, descritto da quattro feature fondamentali (adversary, capability, infrastructure, victim) connesse dalle relazioni che danno al modello la sua forma a diamante e corredate di meta-feature, tra cui fase, risultato, direzione, metodologia e risorse, che consentono di concatenare gli eventi in *activity thread* e di coalizzarli in *activity group* [13]. L'ambizione dichiarata è portare nell'analisi delle intrusioni principi di metodo scientifico: misurabilità, testabilità, ripetibilità [13]. Il valore operativo del modello sta nel *pivoting* analitico. Nota una feature, per esempio un'infrastruttura, l'analista attraversa le relazioni del diamante per scoprire le altre. Su questo impianto concettuale poggia gran parte del lavoro di attribuzione, ripreso nel capitolo 7.

MITRE ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge), infine, è una knowledge base curata di comportamenti avversari fondata su osservazioni empiriche del mondo reale. Nata per l'emulazione degli avversari e per la misurazione della copertura difensiva, è oggi adottata trasversalmente in intrusion detection, threat hunting, security engineering, threat intelligence, red teaming e risk management [10]. Il contenuto è organizzato in matrici per dominio tecnologico (Enterprise, Mobile, ICS), dove le tecniche e sotto-tecniche sono indicizzate per tattica e collegate a gruppi avversari e software noti. Rispetto ai due modelli precedenti, ATT&CK non prescrive una struttura dell'intrusione: cataloga comportamenti, con una granularità che ne ha fatto la lingua franca de facto dell'intelligence tattica. Per l'economia di questo lavoro il punto è decisivo, perché ATT&CK fornisce lo spazio di etichette su cui operano tanto i sistemi di detection (capitolo 4) quanto le tecniche NLP di estrazione automatica di TTPs dai report (capitolo 5). I limiti sono il rovescio dei pregi: copertura sbilanciata verso ciò che è stato osservato e pubblicamente documentato, granularità disomogenea tra tecniche, natura descrittiva che lascia all'analista la ricostruzione della sequenza.

	Cyber Kill Chain	Diamond Model	MITRE ATT&CK
Origine	Lockheed Martin, 2011 [11]	CCIATR, 2013 [13]	MITRE, dal 2013; formalizzato nel 2018 [10]
Unità di analisi	intrusione come sequenza di sette fasi	evento di intrusione come relazione tra quattro feature	comportamento avversario (tattiche, tecniche, procedure)

	Cyber Kill Chain	Diamond Model	MITRE ATT&CK
Prospettiva	temporale-sequenziale	relazionale-attributiva	enciclopedico-comportamentale
Punti di forza	difesa intelligence-driven; courses of action per fase	pivoting analitico; base formale per clustering e attribution	granularità; base empirica aggiornata; lingua franca della CTI tattica
Limiti principali	linearità; impronta perimetrale; fasi post-compromissione poco granulari	scarsa prescrittività operativa; onere di modellazione	copertura sbilanciata sull'osservato; granularità disomogenea; non sequenziale
Impiego tipico nella CTI	comunicazione; gap analysis per fase d'attacco	correlazione di campagne; attribuzione	detection engineering; adversary emulation; etichettatura automatica di TTPs

Tabella 2 — Confronto tra i tre framework di analisi strutturata della minaccia.

I tre framework non sono alternativi ma componibili. La Kill Chain fornisce l'asse temporale, il Diamond Model la struttura relazionale dell'evento, ATT&CK il vocabolario comportamentale. Nelle pipeline analizzate nei capitoli successivi ricorreranno tutti e tre, con ATT&CK in posizione dominante per la sua natura machine-readable.

3. Big Data per la CTI: fonti, architetture, pipeline

Prima ancora che un problema di algoritmi, la CTI è un problema di dati: quali acquisire, da dove, con quali garanzie di qualità e attraverso quali architetture renderli analizzabili. Questo capitolo esamina il substrato su cui poggia tutto il momento analitico dei capitoli 4 e 5. Il §3.1 discute le caratteristiche che rendono il dato di sicurezza un caso paradigmatico, e per certi versi anomalo, di Big Data; il §3.2 propone una tassonomia delle fonti; il §3.3 tratta le architetture di ingestion, elaborazione e storage; il §3.4 affronta il problema, troppo spesso sottovalutato, della qualità del dato. La Figura 2 sintetizza la pipeline di riferimento, le cui macro-fasi ricalcano deliberatamente il ciclo di intelligence del §2.1.

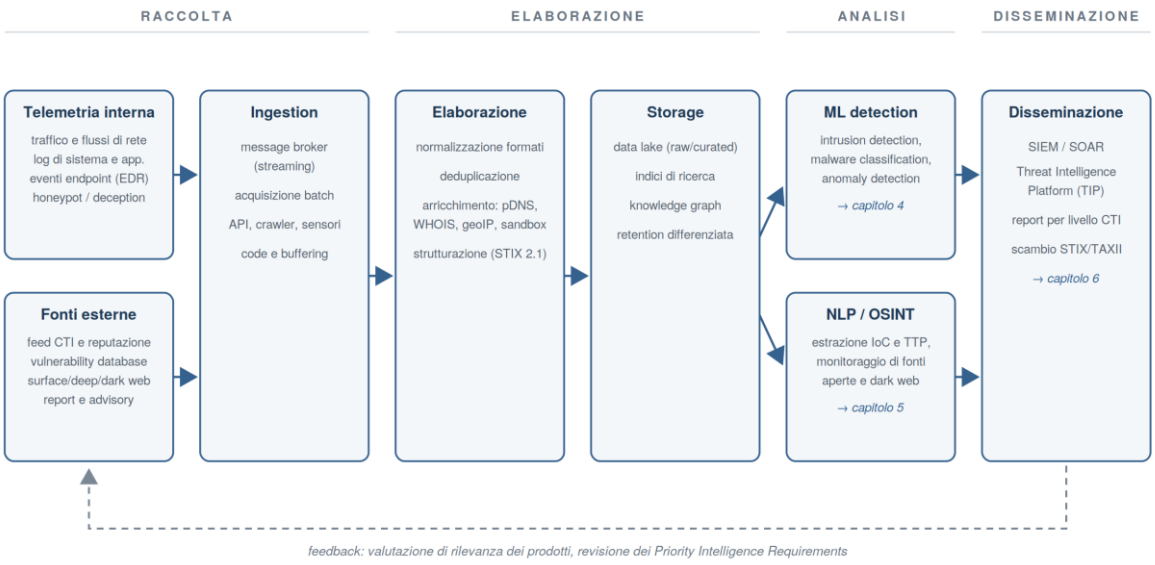

Figura 2 — Architettura di riferimento della pipeline Big Data–ML per la CTI. Le etichette superiori mappano i blocchi architetturali sulle fasi del ciclo di intelligence (Figura 1); i rimandi interni indicano i capitoli in cui ciascun blocco analitico è approfondito.

3.1 Le dimensioni del Big Data nel dominio cyber

La caratterizzazione canonica del Big Data risale alla nota con cui Laney descrisse la gestione dei dati come problema tridimensionale di volume, velocità e varietà [14], schema poi esteso nella pratica industriale con una quarta dimensione, la veridicità (*veracity*). Il dominio della sicurezza le esibisce tutte in forma estrema, con una torsione peculiare sull'ultima.

Il *volume* è la dimensione più evidente. Ai flussi di telemetria interna, che in un'organizzazione di media complessità si misurano nell'ordine dei terabyte giornalieri tra traffico di rete, log ed eventi endpoint, si sommano i volumi delle fonti esterne, a partire dagli oltre 450.000 nuovi campioni malevoli registrati ogni giorno dai repository di settore [2]. La *velocità* ha una doppia natura. Da un lato il dato arriva in streaming continuo e va processato con latenze compatibili con la finestra utile di detection e risposta; dall'altro, come mostrato in Tabella 1, il valore dell'intelligence tecnica decade in ore o giorni, sicché una pipeline lenta non produce intelligence obsoleta, produce rumore. La *varietà*, infine, copre l'intero spettro delle rappresentazioni: dato strutturato (flussi NetFlow/IPFIX, record STIX), semi-strutturato (log

eterogenei), non strutturato (report, post su forum underground, discussioni social) e binario (campioni malware).

È però la *veridicità* a distinguere il Big Data di sicurezza da quello di qualunque altro dominio applicativo. In ambito commerciale o scientifico il dato può essere rumoroso o incompleto, ma non è, di norma, prodotto da un soggetto che ha interesse a falsificarlo. Nel dominio cyber una parte del dato è generata dall'avversario stesso, che può deliberatamente inquinarlo con spoofing degli indicatori di rete, infrastrutture usa-e-getta, false flag, contenuti ingannevoli nelle fonti aperte. La veridicità non è quindi solo una proprietà statistica da misurare (§3.4), ma una superficie di attacco, tema che riemergerà nel capitolo 8 a proposito del poisoning dei dati di addestramento. Alla quaterna si aggiunge spesso una quinta dimensione, il *valore*: l'intera pipeline di Figura 2 può leggersi come una macchina per convertire dato grezzo a bassa densità di valore in intelligence azionabile, secondo la progressione dato–informazione–intelligence introdotta nel §2.1.

3.2 Le fonti: telemetria, deception, feed, vulnerability e malware intelligence, OSINT

La Tabella 3 organizza le fonti della CTI in sette categorie, distinte per natura del dato, regime di velocità e criticità caratteristiche.

Categoria	Esempi	Natura del dato	Regime	Criticità principali
Telemetria di rete	NetFlow/IPFIX, PCAP, log DNS	strutturato / semi-strutturato	volume molto alto, streaming	volume, crescente quota di traffico cifrato, privacy
Telemetria endpoint e log	EDR, syslog, audit trail, log cloud	semi-strutturato	alto, streaming	eterogeneità dei formati, rumore, copertura
Deception	honeypot, honeynet, honeytoken	strutturato + campioni	basso volume, alto segnale	rappresentatività, fingerprinting da parte dell'attaccante
Feed CTI	feed commerciali e open, blocklist, reputazione	strutturato (STIX, liste)	medio, near real-time	copertura parziale, sovrapposizione limitata, latenza, falsi positivi [16], [19]
Vulnerability intelligence	CVE/NVD, cataloghi di sfruttamento attivo, advisory	strutturato	medio	tempestività, prioritizzazione (gravità vs sfruttamento reale)
Malware intelligence	repository di campioni, report di sandbox	binario + report	medio-alto	evasione delle sandbox, qualità dell'etichettatura
OSINT non strutturata	report, blog, social media, forum, dark web	testo non strutturato	alto, eterogeneo	affidabilità, gergo e multilinguismo, accesso alle fonti (cap. 5)

Tabella 3 — Tassonomia delle fonti dati per la CTI.

Le prime tre categorie costituiscono la telemetria che l'organizzazione controlla direttamente. Quella di rete offre la copertura più ampia ma sconta la crescita del traffico cifrato, che sposta l'analisi dai contenuti ai metadati; quella di endpoint e log offre granularità comportamentale al prezzo di un'estrema eterogeneità di formati. Un discorso a parte meritano le tecnologie di *deception*. Gli honeypot rovesciano l'economia del segnale, perché ogni interazione con un sistema esca è per costruzione sospetta, e integrano le difese convenzionali come firewall e IDS attirando l'attaccante e illudendolo di aver raggiunto sistemi reali [15]. La survey di Franco et al. ne sistematizza tassonomia e fattori di progetto, dal livello di interazione alla scala al

dominio applicativo, evidenziando al contempo il limite strutturale della categoria: l'avversario evoluto esegue fingerprinting degli ambienti esca, e la rappresentatività di ciò che vi si osserva va sempre valutata criticamente [15].

Sul versante esterno, i feed CTI, commerciali, comunitari e istituzionali, recapitano indicatori pronti al consumo automatico, con i problemi di qualità documentati nel §3.4. La vulnerability intelligence (database CVE/NVD, cataloghi delle vulnerabilità attivamente sfruttate, advisory dei vendor) alimenta la prioritizzazione difensiva, il cui nodo critico è il divario tra gravità teorica e sfruttamento osservato. L'ETL 2025 documenta oltre 42.000 nuove vulnerabilità divulgate nel periodo di riferimento, con weaponization delle più critiche entro pochi giorni dalla disclosure [1], numeri che rendono insostenibile qualunque prioritizzazione manuale. La malware intelligence fornisce campioni e report di detonazione in sandbox, materia prima per i classificatori del capitolo 4. Chiude la tassonomia l'OSINT non strutturata, la fonte a più alto potenziale informativo e a più alto costo di estrazione, cui è dedicato il capitolo 5. In termini di Pyramid of Pain [9], vale una regola generale: le fonti a basso costo di acquisizione producono tipicamente indicatori dei livelli inferiori, mentre i livelli alti, tool e TTPs, richiedono o analisi umana o le tecniche di estrazione automatica discusse più avanti.

3.3 Architetture di ingestion, elaborazione e storage

Le quattro dimensioni del §3.1 si traducono in requisiti architetturali stringenti, che la letteratura di ingegneria del software ha sistematizzato. La systematic review di Ullah e Babar su 74 studi primari identifica i dodici attributi di qualità più ricorrenti nei sistemi di Big Data cybersecurity analytics e codifica diciassette tattiche architetturali per soddisfarli, rilevando al contempo che attributi pur rilevanti come interoperabilità, modificabilità, adattabilità e tutela della privacy restano privi di supporto architetturale esplicito nella letteratura [17].

Nella pratica, l'architettura di Figura 2 si articola su tre strati. L'*ingestion* disaccoppia produttori e consumatori del dato tramite message broker per i flussi in streaming, affiancati da acquisizione batch per le fonti a bassa frequenza (dump di feed, repository di campioni, crawling di fonti aperte); il buffering assorbe i picchi e garantisce la ritrasmissione. L'*elaborazione* segue il doppio regime reso canonico dalle cosiddette architetture lambda e kappa, con percorsi in streaming per ciò che alimenta la detection in tempo quasi reale e percorsi batch per arricchimento pesante, retrospettiva storica e addestramento dei modelli. Lo *storage*, infine, adotta tipicamente la forma del data lake a zone, con dato grezzo immutabile da un lato e dato curato e strutturato dall'altro, affiancato da indici di ricerca per l'interrogazione interattiva e, in misura crescente, da knowledge graph che rappresentano entità e relazioni della minaccia (attori, campagne, infrastrutture, TTPs) in forma direttamente navigabile dagli algoritmi di correlazione [4].

In questo quadro va collocata l'evoluzione del SIEM, storicamente il punto di aggregazione degli eventi di sicurezza. L'analisi di González-Granadillo et al. documenta la progressiva convergenza dei SIEM con gli strumenti di big data analytics e delinea i requisiti della generazione successiva: visibilità unificata su ambienti IT e OT, automazione della risposta, superamento dell'operatività a silos [18]. Nella pipeline di riferimento il SIEM cessa dunque di essere il contenitore universale del dato per diventare uno dei consumatori dello strato di disseminazione, accanto alle piattaforme di orchestrazione (SOAR) e alle Threat Intelligence Platform. L'anello di feedback che chiude la figura, dalla valutazione dei prodotti alla revisione

dei requisiti, è ciò che distingue una pipeline di intelligence da una semplice catena di elaborazione dati.

3.4 Data quality, normalizzazione e arricchimento

La fase di elaborazione svolge tre funzioni. La *normalizzazione* riconduce l'eterogeneità dei formati a schemi comuni, in prospettiva CTI tipicamente gli oggetti STIX 2.1 (cap. 6), e la *deduplicazione* elimina la ridondanza tra fonti sovrapposte. L'*arricchimento* aumenta la densità informativa di ciascun osservabile incrociandolo con contesto: risoluzioni storiche via passive DNS, titolarità di domini e reti via WHOIS/RDAP, geolocalizzazione e autonomous system, detonazione dei campioni in sandbox, punteggi reputazionali. È l'arricchimento a trasformare l'indicatore atomico in informazione contestualizzata, primo gradino della progressione verso l'intelligence.

Tutto questo, però, presuppone che il dato in ingresso sia affidabile, e la letteratura empirica invita alla cautela. Lo studio di Li et al. definisce formalmente un insieme di metriche per la caratterizzazione dei feed di threat intelligence (volume, contributo differenziale, sovrapposizione, copertura, accuratezza, latenza) e le applica a un ampio ventaglio di fonti pubbliche e commerciali, giungendo a una conclusione severa: i risultati delle misurazioni suggeriscono limiti e criticità significativi nell'uso dei dati di threat intelligence esistenti per gli scopi che dichiarano di servire [16]. Due fonti che coprono la stessa categoria di minaccia mostrano sovrapposizioni sorprendentemente basse, segno che nessuna osserva più di una frazione del fenomeno, e la valutazione indipendente condotta da Griffioen, Booij e Doerr sui feed aperti converge nel rilevare problemi diffusi di accuratezza e tempestività [19]. Non a caso la formalizzazione della qualità della CTI è divenuta un filone di ricerca autonomo, con proposte di dimensioni di qualità misurabili e strumenti di visualizzazione a supporto degli analisti [20].

Per l'economia di questo lavoro, le implicazioni sono due. Sul piano metodologico, e la cosa anticipa il capitolo 4, i dati che alimentano i modelli di ML ereditano tutti i difetti delle fonti, dal rumore di etichettatura al bias di copertura alla latenza, e nessuna sofisticazione algoritmica compensa un ingresso di bassa qualità. Sul piano operativo, la qualità va gestita come proprietà di primo livello della pipeline, con scoring e invecchiamento controllato degli indicatori nelle piattaforme di gestione, tracciamento della provenienza e valutazione continua delle fonti, pratiche che, come si vedrà nel §6.5, condizionano anche la disponibilità delle organizzazioni a condividere ciò che producono.

4. Machine Learning per la threat detection

Con questo capitolo il lavoro entra nel momento analitico della pipeline di Figura 2, limitatamente al dato strutturato e semi-strutturato: traffico di rete, eventi, campioni binari. I due macro-task esaminati, network intrusion detection e malware detection, sono i più maturi dell'intera intersezione tra ML e sicurezza, e proprio per questo i più istruttivi. Su di essi la letteratura ha accumulato non solo risultati, ma anche una consapevolezza metodologica dei propri errori sistematici che il resto del campo sta ancora maturando. Il capitolo si chiude infatti (§4.5) su ciò che i benchmark non dicono: metriche, falsi positivi e base-rate fallacy.

4.1 Paradigmi di apprendimento

La scelta del paradigma di apprendimento è, nel dominio della sicurezza, una scelta vincolata prima ancora che algoritmica. L'apprendimento *supervised* presuppone etichette, e il §3.4 ha mostrato quanto le etichette di sicurezza siano costose, rumorose e deperibili: un classificatore addestrato su attacchi noti eredita i confini della conoscenza che lo ha etichettato. L'apprendimento *unsupervised*, ossia clustering e soprattutto anomaly detection tramite one-class model e autoencoder, promette di svincolarsi dal noto modellando la normalità e segnalandone le deviazioni. La sua debolezza strutturale è che *anomalo* e *malevolo* non coincidono: la maggior parte delle anomalie in un ambiente reale è benigna, e la traduzione da deviazione statistica a evento di sicurezza azionabile resta a carico dell'analista. Le vie di mezzo *semi-supervised* e i paradigmi di pre-addestramento *self-supervised* rappresentano il compromesso pragmatico oggi prevalente, dato che nella pratica convivono poche etichette affidabili e volumi enormi di dato non etichettato. Il *deep learning*, trasversale ai paradigmi, ha spostato il baricentro dal feature engineering manuale al representation learning su input grezzi (sequenze di byte, pacchetti, chiamate di sistema) al prezzo di fame di dati, costi computazionali, opacità decisionale e di una fragilità adversarial che il capitolo 8 esaminerà come problema autonomo. La Tabella 4 sintetizza le famiglie algoritmiche ricorrenti e il loro impiego tipico. Va letta ricordando che nella detection operativa i vincoli di throughput, latenza e budget di falsi positivi pesano quanto l'accuratezza nominale.

Famiglia	Paradigma	Impiego tipico	Punti di forza	Limiti
Alberi ed ensemble (Random Forest, gradient boosting)	supervised	NIDS su feature di flusso; malware detection statica	efficienza, robustezza al rumore, buona interpretabilità	dipendenza dal feature engineering
SVM, k-NN	supervised	classificazione di traffico e campioni su scala media	solidità su spazi ben separabili	scalabilità limitata su volumi Big Data
Reti profonde (MLP, CNN, RNN/LSTM)	supervised / DL	malware su sequenze di byte, opcode, chiamate API; NIDS sequenziale	representation learning su input grezzi	fame di dati, opacità, fragilità adversarial
Autoencoder, one-class (OC-SVM, Isolation Forest)	unsupervised	anomaly detection su rete ed endpoint	non richiede etichette d'attacco	anomalo ≠ malevolo; calibrazione delle soglie
Clustering (k-means, DBSCAN, gerarchico)	unsupervised	raggruppamento di campioni in famiglie, triage	scoperta di struttura non nota	validazione onerosa, sensibilità ai parametri
Graph Neural Network	supervised / semi-sup.	provenance graph, call graph, correlazione di entità nei knowledge graph	modellazione esplicita delle relazioni	costo computazionale, scarsità di dataset

Famiglia	Paradigma	Impiego tipico	Punti di forza	Limiti
Transformer	supervised / self-sup.	sequenze lunghe: traffico, API, testo (cap. 5)	contesto esteso, pre-addestramento riusabile	costi, dati, opacità

Tabella 4 — Famiglie algoritmiche e impiego tipico nella threat detection.

4.2 Network intrusion detection

I sistemi di rilevamento delle intrusioni di rete si dividono classicamente in *misuse-based*, che riconoscono pattern noti con precisione alta e recall nullo sul non noto, e *anomaly-based*, dove si concentra il grosso della ricerca ML. La survey di Buczak e Guven resta la mappatura di riferimento dei metodi di data mining e machine learning applicati al problema [21].

Ogni discussione onesta del tema deve però partire dal lavoro con cui Sommer e Paxson, nel 2010, spiegarono perché l'anomaly detection basata su ML fosse così poco presente nelle reti operative a dispetto di una letteratura sterminata. La tesi centrale è che il compito di trovare attacchi sia fondamentalmente diverso dalle applicazioni in cui il ML eccelle, perché il paradigma è più adatto a riconoscere somiglianze con il già visto che a individuare novità ostili [22]. Gli argomenti a sostegno — il costo altissimo di ogni errore in entrambe le direzioni, il *semantic gap* tra anomalia statistica e report azionabile, l'estrema variabilità del traffico benigno che rende instabile ogni nozione di normalità, la difficoltà di una valutazione realistica — hanno retto così bene al tempo da valere al lavoro il Test of Time Award di IEEE S&P nel 2020 [22]. La stagione del deep learning ha reso il quadro paradossale. La letteratura recente riporta con regolarità detection rate superiori al 99% sui benchmark pubblici come UNSW-NB15 e CICIDS2017, risultati che, come mostreranno i §4.4–4.5, dicono più sui difetti dei benchmark che sulla maturità operativa dei sistemi. A complicare il quadro contribuisce la crescita del traffico cifrato, che ha spostato la detection dai contenuti alle statistiche di flusso e ai metadati, comprimendo lo spazio informativo disponibile ai modelli. Nella pipeline di riferimento, l'output di un ML-NIDS va quindi inteso come segnale di triage ad alta sensibilità da contestualizzare con l'intelligence, non come verdetto.

4.3 Malware detection e classification

Sul versante malware la tassonomia consolidata distingue per tipo di analisi. L'analisi *statica* opera sul campione senza eseguirlo, lavorando su struttura dei formati eseguibili, stringhe, import, sequenze di opcode e di byte, e copre lo spettro che va dai gradient boosting su feature ingegnerizzate fino ai modelli end-to-end che consumano il binario grezzo; è rapida e applicabile pre-esecuzione, ma cede di fronte a packing e offuscamento. L'analisi *dinamica* osserva il comportamento in ambiente controllato, registrando sequenze di chiamate API e di sistema, attività di rete e file system, ed è più robusta all'offuscamento ma costosa, lenta e vulnerabile ai campioni environment-aware che riconoscono la sandbox e sospendono il comportamento malevolo; gli approcci *ibridi* combinano i due piani. La survey di Ucci, Aniello e Baldoni organizza sistematicamente questo spazio per obiettivi, feature e algoritmi [23], mentre quella di Gibert, Mateu e Planes documenta lo spostamento del baricentro verso le tecniche deep, che apprendono rappresentazioni direttamente dal dato grezzo [24]. Accanto alla detection binaria, la *classificazione in famiglie*, supervisionata o via clustering, sostiene il triage e alimenta l'intelligence tattica, scontando però il rumore delle etichette derivate dai motori antivirus, raramente concordi tra loro.

Il problema più profondo del dominio è però temporale. Il malware è un bersaglio mobile: la distribuzione dei campioni cambia continuamente (*concept drift*), e ignorarlo in fase sperimentale produce risultati illusori. Il contributo di riferimento è TESSERACT di Pendlebury et al., che identifica due bias sperimentali sistematici, uno spaziale legato a rapporti di classe irrealistici e uno temporale legato ad addestramento con accesso a informazione "futura", e mostra come la loro eliminazione, tramite valutazioni time-aware con vincoli espliciti, ridimensioni drasticamente le prestazioni dichiarate dai classificatori di malware [25]. È una lezione che vale ben oltre il malware e che il §4.5 generalizza.

4.4 Dataset e benchmark

La Tabella 5 riassume i benchmark più usati nei due domini, con i limiti che la letteratura ha progressivamente documentato.

Dataset	Anno	Dominio	Contenuto	Limiti noti
KDD Cup 99 / NSL-KDD	1999 / 2009	rete	connessioni etichettate derivate dalla valutazione DARPA 1998	traffico ormai archeologico; ridondanze e duplicati massivi nel KDD99, solo in parte corretti da NSL-KDD [26]
UNSW-NB15	2015	rete	traffico ibrido reale/sintetico generato con appliance dedicata, nove famiglie di attacco [27]	artefatti della generazione sintetica; prevalenza d'attacco irrealisticamente alta
CICIDS2017 / CSE-CIC-IDS2018	2017–2018	rete	flussi etichettati con payload completi e traffico benigno generato da profili comportamentali [28]	errori documentati nella generazione del traffico, nella costruzione dei flussi, nell'estrazione delle feature e nell'etichettatura [29]
Drebin	2014	malware Android	applicazioni malevole e benigne con feature statiche estratte [31]	età dei campioni; ecosistema Android profondamente mutato
EMBER	2018	malware PE Windows	1,1 milioni di eseguibili con feature statiche standardizzate [30]	sola analisi statica; etichette derivate da motori antivirus; drift temporale

Tabella 5 — Dataset benchmark ricorrenti nella letteratura su ML per la threat detection.

Due osservazioni trasversali si impongono. La prima riguarda la gravità dei difetti, che non sono affatto marginali. Il caso CICIDS2017 è emblematico: il dataset più citato della sua generazione, la cui revisione da parte di Engelen, Rimmer e Joosen ha portato alla luce problemi lungo l'intero ciclo di produzione, dalla generazione del traffico all'etichettatura, con una versione corretta che modifica sensibilmente i risultati dei modelli [29]. Analisi successive hanno esteso rilievi analoghi al successore CSE-CIC-IDS2018, alimentando il dubbio che parte dei progressi dichiarati dalla letteratura misuri artefatti dei dataset più che avanzamenti metodologici. La seconda osservazione è che anche il benchmark perfetto invecchierebbe. Un dataset è la fotografia di un ecosistema di minaccia a una certa data, e il concept drift del §4.3 ne erode la rappresentatività dal giorno della pubblicazione. Per questo i risultati vanno sempre letti insieme all'anno del dato, non solo a quello del metodo.

4.5 Metriche di valutazione, falsi positivi e base-rate fallacy

L'ultima fonte di illusione è la misura stessa. In presenza dello sbilanciamento estremo che caratterizza il traffico reale, dove gli eventi malevoli sono una frazione infinitesima del totale, l'accuratezza è priva di significato e le stesse curve ROC possono lusingare. Contano precision

e recall al punto operativo, le curve precision-recall e, soprattutto, il tasso di falsi positivi assoluto rapportato ai volumi.

Il fondamento teorico ha un quarto di secolo. Axelsson mostrò che il fattore limitante di un sistema di intrusion detection è il tasso di falsi allarmi, per effetto della *base-rate fallacy*: quando la probabilità a priori di attacco è minuscola, il valore predittivo di un allarme è dominato dai falsi positivi, e perché un allarme sia credibile il false positive rate deve scendere a ordini di grandezza, intorno a 10^{-5} nei suoi scenari, fuori portata per gran parte dei modelli [32]. È la controparte teorica esatta di ciò che lo studio qualitativo di AlAhmadi et al. osserva vent'anni dopo nei SOC reali, con analisti sommersi da allarmi che richiedono validazione manuale [3]. Un modello "al 99%" su un benchmark bilanciato può essere, in produzione, una macchina generatrice di rumore.

La sistematizzazione più completa di questi e altri errori è il lavoro di Arp et al., che identifica dieci insidie ricorrenti nella progettazione, implementazione e valutazione dei sistemi di sicurezza basati su ML, dal sampling bias al data snooping, dalle correlazioni spurie alla base-rate fallacy fino alla valutazione confinata al laboratorio. Analizzando trenta pubblicazioni di venue di primo livello dell'ultimo decennio, gli autori ne documentano la pervasività, e le più frequenti risultano proprio sampling bias, data snooping e lab-only evaluation [33]. Le contromisure che ne discendono definiscono lo standard metodologico assunto nel resto di questo lavoro: partizioni temporali realistiche, rapporti di classe fedeli alla base rate operativa, valutazione cost-sensitive con il falso positivo prezzato al costo-analista, baseline non banali e dichiarazione delle prestazioni al punto operativo. In una pipeline CTI ben progettata, il modello di detection non sostituisce il giudizio: alza il rapporto segnale-rumore del triage e consegna all'analista, e agli strati NLP e di correlazione dei prossimi capitoli, un ingresso più denso di valore.

5. NLP e OSINT: dall'informazione non strutturata all'intelligence azionabile

Se il capitolo 4 ha trattato il dato che le macchine producono, questo capitolo tratta il dato che gli esseri umani scrivono: report e advisory, blog tecnici, discussioni social, post nei forum underground. È la classe di fonti a più alto potenziale informativo dell'intero panorama CTI, perché è qui, e non nei flussi di rete, che si trovano descritte le TTPs collocate al vertice della Pyramid of Pain [9]. Al tempo stesso è la più costosa da sfruttare, dato che il linguaggio naturale non si interroga con una query. Il Natural Language Processing è il ponte tra i due mondi, e la Figura 3 ne schematizza la pipeline, dalla fonte aperta all'oggetto strutturato pronto per la correlazione e la condivisione. La survey di Arazzi et al., condotta con protocollo sistematico, costituisce la mappa di riferimento delle tecniche NLP applicate all'intero ciclo, dal crawling delle fonti web all'analisi del dato alla relation extraction, e verrà richiamata trasversalmente [34].

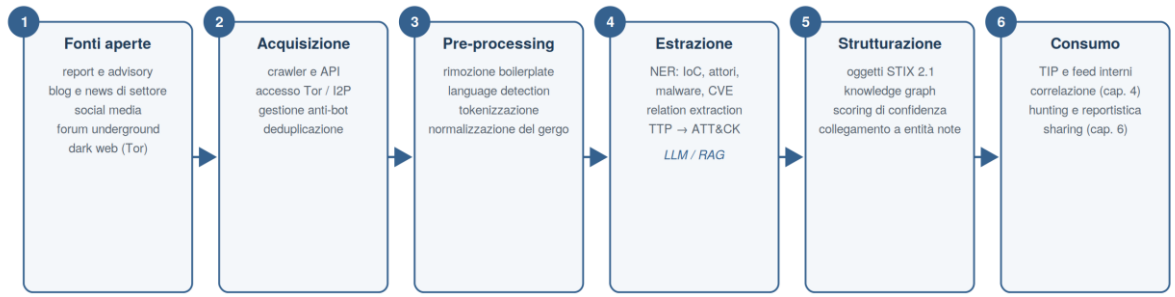

Figura 3 — Pipeline NLP per l'estrazione di intelligence da fonti aperte. L'output strutturato alimenta gli strati di correlazione (cap. 4) e di condivisione (cap. 6) dell'architettura di Figura 2.

5.1 OSINT e monitoraggio di surface, deep e dark web

La partizione convenzionale distingue il *surface web*, indicizzato dai motori di ricerca; il *deep web*, non indicizzato perché protetto da autenticazione, paywall o generato dinamicamente; e il *dark web*, raggiungibile solo attraverso overlay network anonimizzanti come Tor e I2P, dove l'anonimato è una proprietà progettuale e non un incidente. Ai fini della CTI le tre fasce offrono profili complementari di qualità e tempestività. Sul surface web risiedono le fonti curate, dai report dei vendor agli advisory istituzionali ai blog di ricerca, ad alta affidabilità ma con la latenza propria dell'analisi editoriale. I social media offrono la massima velocità al prezzo di rumore e inaffidabilità. Forum e marketplace del deep e dark web offrono qualcosa che nessun'altra fonte possiede, cioè la voce diretta dell'avversario.

Da qui il valore attribuito al dark web monitoring, sistematizzato come filone di ricerca dal lavoro del gruppo di Shakarian confluito nella monografia sul darkweb CTI mining [35]. Nei marketplace e nei forum underground si osservano in anticipo il commercio di exploit e di accessi (le inserzioni degli initial access broker), la circolazione di credenziali compromesse, il reclutamento di affiliati e, sui leak site delle operazioni ransomware, le rivendicazioni delle vittime. Il monitoraggio su scala richiede però di vincere resistenze progettate apposta: l'accesso attraverso Tor, l'autenticazione in comunità chiuse che richiedono reputazione, contromisure anti-crawling aggressive. Emblematica di questa corsa agli armamenti è la

ricerca sull'aggiramento automatico dei CAPTCHA testuali del dark web tramite generative adversarial learning, sviluppata proprio per rendere praticabile la raccolta proattiva di CTI [36]. Restano ineludibili tre cautele. Quella giuridica ed etica, perché l'interazione con mercati criminali tocca confini normativi che variano per giurisdizione e va governata da policy esplicite. Quella metodologica della rappresentatività, perché ciò che è osservabile non coincide con ciò che esiste. E quella, già discussa nel §3.1, della veridicità di fonti in cui l'inganno è parte del gioco.

5.2 Estrazione automatica di IoC e TTP: NER e relation extraction

L'estrazione di conoscenza dal testo procede per gradini di crescente astrazione, gli stessi, non a caso, della Pyramid of Pain. Il gradino inferiore è l'estrazione di IoC atomici. In apparenza è un problema di espressioni regolari, complicato dalle notazioni *defanged* (hxxp, [.]) con cui gli autori disinnescano gli indicatori; in realtà il nodo è il contesto, poiché un indirizzo IP citato in un report può essere l'infrastruttura dell'attaccante, quella della vittima o un esempio innocuo, e distinguere i casi richiede comprensione della frase. Il lavoro di riferimento è iACE di Liao et al., che combina NLP e graph mining sulle relazioni tra i token per identificare gli IoC con il loro contesto. Applicato a 71.000 articoli pubblicati da 45 blog di sicurezza in tredici anni, il sistema ha estratto circa 900.000 token IoC, documentando al passaggio la crescita del fenomeno, da una ventina a oltre mille articoli IoC-related al mese, che rende la conversione manuale in formati standard ormai impraticabile [37].

Il gradino intermedio è il riconoscimento di entità e relazioni proprie del dominio: nomi di malware e di threat actor, strumenti, CVE, vittime e settori, con relation extraction che collega le entità in triple destinate ai knowledge graph della pipeline di Figura 2 [34], [4]. Il gradino superiore, il più prezioso e il più difficile, è l'estrazione delle TTPs, ossia la mappatura di frammenti di testo sulle tecniche di ATT&CK, che funge da spazio di etichette condiviso [10]. Il capostipite è TTPDrill di Husari et al., che accoppia un'ontologia delle azioni di minaccia a tecniche di NLP e information retrieval per estrarre le azioni su base semantica, ricondurle a tecniche, tattiche e fasi della kill chain e tradurle in standard di condivisione come STIX, con precision e recall superiori all'82% nella valutazione su report reali [38]. La generazione successiva ha spostato l'obiettivo dalla singola tecnica al grafo, e sistemi come AttacKG ricostruiscono dai report interi technique knowledge graph [39]. La letteratura recente legge questa evoluzione come una successione di tre paradigmi dai limiti caratteristici. I sistemi rule-based come TTPDrill sono interpretabili ma vincolati a ontologie costruite a mano che faticano a inseguire l'evoluzione delle descrizioni d'attacco. I modelli neurali supervisionati apprendono pattern contestuali ma esigono dati annotati. Gli approcci LLM-based generalizzano meglio ma pagano in precisione (cfr. §5.3). Il problema resta comunque strutturalmente arduo: uno spazio di centinaia di etichette in regime multi-label, classi rarissime, corpora annotati scarsi e valutazioni poco confrontabili tra loro [34].

5.3 Transformer, Large Language Model e retrieval-augmented generation

L'architettura transformer ha investito il dominio in due ondate. La prima è quella dei modelli pre-addestrati adattati al linguaggio della sicurezza, varianti della famiglia BERT ri-addestrate su corpora di dominio come CySecBERT, che hanno alzato lo stato dell'arte di NER e classificazione riducendo il fabbisogno di dati annotati. La seconda è quella dei Large Language Model generalisti, che promettono di comprimere l'intera colonna di estrazione

della Figura 3: information extraction zero-shot e few-shot senza addestramento dedicato, summarization di report chilometrici, costruzione assistita di knowledge graph, interrogazione in linguaggio naturale del patrimonio informativo. Il quadro di questa produzione, in rapidissima espansione, è tracciato dalla survey di Chen et al. sui Large Language Model per la threat detection [40]. Alla promessa si accompagna il correttivo architetturale ormai standard, la retrieval-augmented generation, che ancora la generazione a corpora di dominio recuperati a runtime, riducendo le hallucination e permettendo di aggiornare la conoscenza aggiornando il corpus anziché ri-addestrando il modello.

Le ragioni di cautela, in questo dominio, sono però più severe che altrove. Anzitutto l'asimmetria del costo dell'errore: un IoC o una tecnica ATT&CK *inventati* da un modello non sono rumore, sono disinformazione iniettata direttamente nella pipeline difensiva, e in questo senso peggiori dell'assenza di informazione. Vi è poi la riservatezza, perché i report che si vorrebbero far leggere a un LLM contengono spesso dettagli di incidenti non pubblici, difficilmente conciliabili con API di terze parti. Vi è la sicurezza del modello stesso, esposto a prompt injection attraverso i contenuti che analizza, tema che il capitolo 8 riprenderà. E vi è, infine, una ragione sistemica: gli stessi strumenti armano l'avversario, come documenta l'ETL 2025 osservando che a inizio 2025 le campagne di phishing supportate da AI costituivano oltre l'80% del social engineering osservato [1]. Il bilancio ragionevole, allo stato, colloca l'LLM nel ruolo di copilota dell'analista, con output verificabili e verificati, più che di estrattore autonomo in produzione.

5.4 Analisi di forum underground e social media

Le fonti conversazionali pongono sfide linguistiche proprie: gergo criminale in perenne mutazione, code-switching e multilinguismo, ortografia deliberatamente offuscata, un rapporto segnale-rumore bassissimo. La ricerca vi ha applicato l'intero arsenale visto finora, a partire dai lavori di classificazione dei post nei forum hacker che hanno confrontato sistematicamente approcci classici e deep learning, support vector machine contro convolutional neural network, per isolare i contenuti rilevanti ai fini CTI [43].

Il salto di qualità concettuale è però l'uso predittivo di queste fonti. Il lavoro seminale di Sabottke, Suciu e Dumitraş ha mostrato che il chiacchiericcio su Twitter attorno alle vulnerabilità contiene segnale sufficiente a predire quali verranno sfruttate in the wild. Su un corpus di 1,1 miliardi di tweet a tema sicurezza, oltre 287.000 contenevano riferimenti espliciti a identificativi CVE, e un classificatore addestrato su feature estratte da quel flusso ha permesso di prioritizzare la risposta alle disclosure, con applicazioni che gli autori estendono fino al risk modeling assicurativo [41]. La stessa logica, trasposta sulle fonti underground, anima DarkEmbed, che apprende rappresentazioni neurali del linguaggio dei forum del dark web per predire lo sfruttamento delle vulnerabilità discusse [42], anello di congiunzione tra questo capitolo e la vulnerability intelligence del §3.2. Le avvertenze finali sono simmetriche a quelle del §5.1: l'avversario sa di essere osservato e può avvelenare il segnale; i social amplificano disinformazione e attività di bot; la ground truth sullo sfruttamento reale è essa stessa parziale. Il prodotto di tutta la colonna NLP, fatto di entità, relazioni, TTPs e segnali predittivi, converge comunque in un unico punto, lo strato di strutturazione in formati standard, che è la porta d'ingresso del prossimo capitolo.

6. Condivisione della CTI: standard, piattaforme, modelli organizzativi

La disseminazione chiude il ciclo di intelligence e, quando attraversa i confini organizzativi, ne moltiplica il valore: l'intelligence prodotta sull'incidente subito da un'organizzazione diventa prevenzione per tutte le altre, in un effetto rete che costituisce la promessa fondativa del threat intelligence sharing. Questo capitolo esamina l'infrastruttura di quella promessa, dagli standard di rappresentazione e trasporto (§6.1) alle piattaforme (§6.2), dai modelli organizzativi (§6.3) alle posture difensive servite (§6.4). Il §6.5 la confronta poi con la realtà empirica misurata dalla letteratura recente, che restituisce un quadro assai più sobrio. La Figura 4 rappresenta l'ecosistema complessivo dal punto di vista della singola organizzazione.

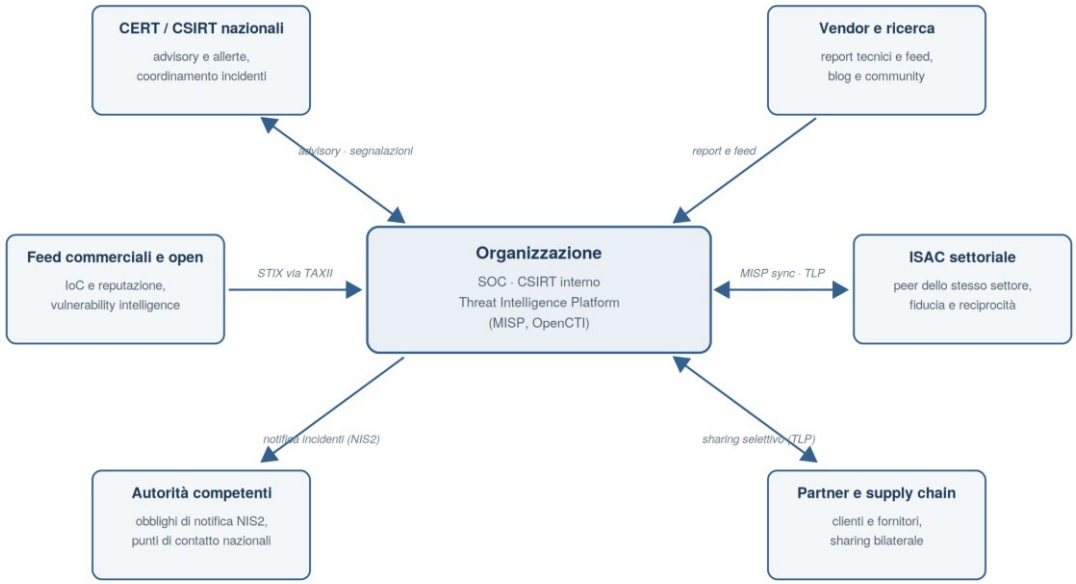

Figura 4 — L'ecosistema di threat intelligence sharing dal punto di vista dell'organizzazione: attori, direzione dei flussi e meccanismi di scambio.

6.1 Standard di rappresentazione e trasporto: STIX, TAXII e formati correlati

Lo scambio a velocità di macchina presuppone che produttore e consumatore condividano una rappresentazione della minaccia, problema tutt'altro che risolto, come mostrava già la valutazione di tassonomie e ontologie del §2.2 [8]. Lo standard che ha vinto la selezione è STIX (Structured Threat Information eXpression). Nato in ambito MITRE con il sostegno del governo statunitense e poi transitato sotto il comitato tecnico CTI di OASIS, ha abbandonato con la versione 2 la serializzazione XML delle origini in favore di un modello a grafo in JSON, dove *domain object* (indicator, malware, threat-actor, intrusion-set, attack-pattern, campaign, course-of-action, report), *observable* e *relationship* esprimono entità della minaccia e loro legami, con l'oggetto *sighting* a registrare le osservazioni sul campo [44]. La granularità del modello permette di rappresentare tutti i livelli della Tabella 1, dall'IoC atomico alla campagna, e di veicolare le TTPs nel vocabolario ATT&CK attraverso gli attack-pattern [10]. Al trasporto provvede TAXII (Trusted Automated eXchange of Intelligence Information),

protocollo applicativo su HTTPS che organizza lo scambio in collection interrogabili e canali di distribuzione [45].

La survey di Alaeifar et al., che mappa sistematicamente approcci e direzioni del CTI sharing, attribuisce la posizione dominante di STIX alla sua struttura modulare, capace di incorporare formati preesistenti come CybOX, IODEF e OpenIOC, con TAXII nel ruolo di protocollo di trasporto abilitante l'automazione [46]. Lo standard convive però con un secondo formato de facto, quello nativo di MISP, cresciuto dalla pratica della piattaforma omonima e interoperabile con STIX tramite conversione. La dualità ricorda come, in questo dominio, gli standard di successo siano nati dall'uso più che dai comitati, e come l'eterogeneità residua resti un costo strutturale dell'ecosistema [8], [46].

6.2 Threat Intelligence Platform: MISP, OpenCTI

Tra gli standard e i consumatori finali si colloca la Threat Intelligence Platform, il componente che nella Figura 2 presidia lo strato di disseminazione. Aggrega i flussi in ingresso, li normalizza e deduplica, li correla con il patrimonio interno, ne governa qualità e invecchiamento secondo le pratiche del §3.4 e li ridistribuisce verso SIEM, EDR e sistemi di risposta all'interno, e verso le comunità di sharing all'esterno.

Il riferimento open source del settore è MISP, la cui architettura è descritta nel lavoro di Wagner, Dulaunoy, Wagener e Iklody. Gli eventi aggregano attributi, cioè gli osservabili; il motore di correlazione collega automaticamente eventi che condividono attributi; tassonomie e *galaxy*, incluse quelle allineate ad ATT&CK, forniscono il vocabolario di etichettatura; la sincronizzazione tra istanze, regolata da sharing group e livelli di distribuzione, realizza comunità di scambio federate [47]. Il progetto OpenCTI rappresenta la generazione successiva, con un modello dati nativamente STIX 2.1 organizzato a knowledge graph, naturale controparte applicativa dello strato di storage discusso nel §3.3. In entrambi i casi la piattaforma è anche il punto di applicazione delle policy di disseminazione, poiché è qui che le marcature di riservatezza discusse nel §6.5 vengono imposte ai flussi in uscita.

6.3 Modelli organizzativi e flussi informativi: ISAC, CERT, CSIRT, SOC

L'infrastruttura tecnica serve una geografia organizzativa che conviene fissare con precisione terminologica, riassunta in Tabella 6. CSIRT (Computer Security Incident Response Team) è il termine generico per la squadra di gestione degli incidenti; CERT ne è il sinonimo storico, formalmente un marchio della Carnegie Mellon University concesso in uso, sopravvissuto nella denominazione di molte strutture nazionali. Il SOC presidia il monitoraggio continuo e il triage, mentre il CSIRT coordina la risposta quando l'evento diventa incidente. Gli ISAC (Information Sharing and Analysis Center) realizzano la condivisione orizzontale tra pari dello stesso settore, dalla finanza all'energia alla sanità, dove l'omogeneità della minaccia massimizza il valore dello scambio e la comunanza di interessi ne sostiene la fiducia.

Attore	Missione	Perimetro	Flussi CTI tipici
SOC	monitoraggio continuo, triage degli allarmi, risposta operativa	singola organizzazione	consuma intelligence tecnica e operativa; produce telemetria, sighting e segnalazioni
CSIRT / CERT interno	gestione e coordinamento degli incidenti	organizzazione o gruppo	consuma e produce intelligence operativa; interfaccia verso CSIRT esterni

Attore	Missione	Perimetro	Flussi CTI tipici
CERT / CSIRT nazionali e governativi	allerta precoce, coordinamento, supporto alla constituency	nazionale o settoriale	advisory e allerte in uscita; notifiche e segnalazioni in ingresso; snodo degli obblighi normativi
ISAC	condivisione e analisi tra pari di settore	settoriale	scambio peer-to-peer strutturato (STIX/TAXII, MISP), analisi e benchmark di settore

Tabella 6 — Attori organizzativi della difesa cooperativa e flussi CTI caratteristici.

Sul piano europeo questa geografia ha assunto forma istituzionale. La cornice NIS2 struttura la cooperazione attorno alla rete dei CSIRT nazionali (CSIRTs Network) e alla rete per la gestione delle crisi su larga scala (EU-CyCLONe), con ENISA in funzione di coordinamento e supporto operativo [1], [51]. Ne risultano, per la singola organizzazione, i quattro flussi della Figura 4: verticale-istituzionale (advisory in discesa, notifiche in salita), orizzontale di comunità (ISAC e sharing group), commerciale e di ricerca (feed e report), bilaterale di filiera (partner e supply chain, reso prioritario dall'esposizione documentata nel §2.3).

6.4 Difesa preventiva, proattiva e reattiva

I flussi appena descritti servono tre posture difensive, che consumano la stessa intelligence su orizzonti temporali diversi, gli stessi della Tabella 1. La postura *preventiva* agisce prima del contatto con la minaccia: l'intelligence strategica orienta investimenti e architetture, mentre la vulnerability ed exploit intelligence dei §3.2 e 5.4 trasforma il patching da rincorsa indiscriminata a prioritizzazione informata dalla probabilità di sfruttamento reale. La postura *proattiva* presume la compromissione e la va a cercare. Il threat hunting formula ipotesi a partire dall'intelligence tattica, ossia le TTPs dell'avversario rilevante mappate su ATT&CK, e le verifica sulla telemetria; l'adversary emulation, uso fondativo di ATT&CK [10], chiude il cerchio testando le difese contro comportamenti realistici. È la direzione indicata dalla stessa ENISA quando raccomanda strategie difensive intelligence-driven e sistemiche, imperniate su hunting proattivo e detection comportamentale [1], ed è il senso profondo della transizione descritta da Sun et al. dal paradigma reattivo a quello proattivo [4]. La postura *reattiva*, infine, resta ineliminabile. L'incident response consuma intelligence operativa per ricostruire la campagna in corso, e intelligence tecnica per il contenimento immediato, con la consapevolezza, maturata con la Pyramid of Pain [9], che il blocco dell'IoC atomico compra ore, non settimane.

6.5 Barriere alla condivisione: trust, privacy, qualità, quadro normativo

Se l'architettura è questa, perché lo sharing resta sotto le aspettative? La letteratura converge su una tassonomia di barriere che Alaeifar et al. riconducono a obblighi legali e regolatori, interoperabilità degli standard e affidabilità del dato, cui si sommano fiducia, qualità e vincoli di budget [46]. La fiducia è la barriera architetturale per eccellenza. La survey di Wagner et al. osserva che l'intelligence a basso rischio può circolare in forme centralizzate, mentre lo scambio decentralizzato richiede un grado di fiducia superiore, o una platea ristretta, e che la pratica gradua la confidenza delle fonti dai canali fidati e verificati fino a quelli non verificati [49]. Si aggiungono i disincentivi competitivi e reputazionali del rivelare la propria compromissione, e il free-riding di chi consuma senza contribuire. Sul piano della riservatezza, molti indicatori, in testa gli indirizzi IP, sono dati personali ai sensi del GDPR, e la tensione

tra tutela e tempestività viene gestita nella pratica con il Traffic Light Protocol, che nella versione 2.0 di FIRST gradua la ridistribuibilità dell'informazione dal livello CLEAR fino a RED [52]; la sistematizzazione istituzionale di regole, formati e cautele dello scambio resta quella della guida NIST SP 800-150 [50]. Il legislatore europeo ha infine trasformato parte dello scambio da virtù a obbligo: la direttiva NIS2 impone ai soggetti essenziali e importanti la notifica degli incidenti significativi con pre-allarme entro 24 ore, notifica entro 72 e relazione finale entro un mese, e promuove al contempo gli accordi volontari di condivisione tra imprese [51].

Alla domanda empirica — lo sharing, così com'è, aiuta davvero? — risponde lo studio di Jin et al. presentato a NDSS 2024, che con il framework CTI-Lense analizza volume, tempestività, copertura e qualità di circa sei milioni di oggetti STIX raccolti da dieci fonti pubbliche tra il 2014 e il 2023. I risultati raffreddano l'entusiasmo. Il volume condiviso cresce ma resta insufficiente a coprire il panorama delle minacce, nell'ordine di duemila oggetti unici al giorno, con una ridondanza del 37,9% persino all'interno dei singoli provider. Oltre il 90% del materiale si concentra in firme di malware e URL. La tempestività è buona per gli URL, condivisi nel 72% dei casi non più tardi della loro comparsa su VirusTotal, ma decisamente peggiore per le firme. Il 19% dei dati sui threat actor contiene informazioni errate e appena lo 0,09% degli indicator veicola regole di detection utilizzabili [48]. In termini di Pyramid of Pain, lo sharing reale è schiacciato sui livelli inferiori: proprio l'intelligence tattica, la più preziosa, è quella che circola meno, il che riporta al ruolo delle tecniche di estrazione automatica del capitolo 5 come condizione di scalabilità dell'intero ecosistema. Sul fronte delle barriere di riservatezza, infine, la ricerca esplora la via di condividere i modelli anziché i dati. Lo schema di CTI sharing basato su federated learning proposto da Sarhan et al. consente a più organizzazioni di addestrare congiuntamente un sistema di network intrusion detection senza centralizzare il traffico di ciascuna [53], e apre un ponte diretto verso i temi del capitolo 8.

7. Attribuzione degli attacchi cyber

L'attribuzione è il punto in cui la CTI smette di essere un problema tecnico e diventa un problema epistemico. Dalla risposta alla domanda «chi è stato?» dipende una scala di conseguenze che va dal blocco di un indicatore fino alla sanzione internazionale, e la si deve costruire su un tipo di evidenza, quella digitale, che è per sua natura copiabile, alterabile e deliberatamente falsificabile. È anche il capitolo in cui la CTI incontra più da vicino la digital forensics: le pratiche di raccolta, conservazione e valutazione delle evidenze qui discusse condividono con essa metodo e cautele, differenziandosene per il fine, poiché producono giudizi analitici graduati anziché prove processuali.

7.1 Principi e livelli dell'attribution

Il quadro concettuale di riferimento resta quello fissato da Rid e Buchanan. Contro la visione dell'attribuzione come problema tecnico intrattabile, risolvibile o non risolvibile in blocco, gli autori sostengono che l'attribuzione «è ciò che gli Stati ne fanno», e propongono il Q Model come guida di un processo che minimizza l'incertezza su tre piani: quello tattico, dove l'attribuzione è arte oltre che scienza; quello operativo, dove è processo sfumato e non problema binario; quello strategico, dove è funzione della posta politica in gioco [54]. Ne discendono due principi che governano l'intero capitolo. L'attribuzione è un *processo* con esiti espressi in gradi di confidenza, non un verdetto; e la sua qualità dipende da risorse, tempo e sofisticazione dell'avversario, oltre che dall'evidenza disponibile [54].

Alla gradazione della confidenza si affianca la gradazione dell'oggetto, poiché «chi» può significare tre cose diverse. Al livello più basso si attribuisce l'attività a *macchine e infrastrutture*, cioè a quali host e a quali server di command and control. Al livello intermedio a un *operatore*, l'entità umana o il gruppo che ha materialmente condotto l'operazione. Al livello più alto a uno *sponsor* o mandante, il soggetto, tipicamente statuale, nel cui interesse l'operazione è stata condotta. La pratica analitica media tra i livelli attraverso il costrutto del cluster: gli *intrusion set* e gli *activity group* del Diamond Model [13] raggruppano eventi per evidenza condivisa senza pronunciarsi sull'identità reale, ed è a questi cluster — non a persone o governi — che si riferiscono le denominazioni dei vendor (sigle APT, nomenclature faunistiche e simili), la cui proliferazione non coordinata genera un problema cronico di mappatura degli alias. La Tabella 7 sintetizza la corrispondenza tra livelli, evidenze e confidenza raggiungibile.

Livello	Domanda	Evidenze e artefatti tipici	Manipolabilità e confidenza	Chi tipicamente conclude
Macchina e infrastruttura	da quali sistemi proviene l'attività?	indirizzi IP e domini C2, log, flussi di rete, artefatti di infrastruttura	manipolabilità alta (proxy, infrastrutture usa-e-getta); confidenza elevata sul *cosa*, minima sul *chi*	SOC, incident responder
Operatore	quale gruppo ha condotto l'operazione?	riuso di codice e strumenti, stylometry, TTPs ricorrenti, orari operativi, artefatti di build e di lingua	manipolabilità media; confidenza tipicamente media, espressa su cluster (intrusion set)	vendor CTI, ricerca, law enforcement
Sponsor / mandante	per conto di chi, con quale intento?	vittimologia, allineamento geopolitico, capacità e	evidenza raramente solo tecnica; confidenza in linguaggio estimativo	Stati, comunità di intelligence

Livello	Domanda	Evidenze e artefatti tipici	Manipolabilità e confidenza	Chi tipicamente conclude
		risorse richieste, intelligence non tecnica		

Tabella 7 — Livelli di attribuzione, evidenze associate e confidenza raggiungibile.

7.2 Evidenze tecniche: code similarity, riuso di infrastruttura, stylometry

Il metodo con cui le evidenze si accumulano è il pivoting del Diamond Model: nota una feature dell'evento, l'analista ne attraversa le relazioni per scoprirne altre, concatenando eventi in thread e coalizzandoli in gruppi [13]. Le classi di artefatti seguono i vertici del diamante. Sul versante delle *capability*, l'evidenza principe è il riuso di codice: funzioni condivise tra campioni, builder e packer comuni, artefatti di compilazione (timestamp, identificatori di lingua, percorsi di debug) che sopravvivono nella toolchain dell'avversario. La frontiera scientifica di questo filone è la stylometry binaria. Caliskan et al. hanno mostrato che lo stile di programmazione sopravvive alla compilazione, al punto da consentire la de-anonimizzazione degli autori a partire dai soli eseguibili [55]. Il risultato va però letto insieme al suo contrappeso, poiché l'analisi di fattibilità di Alrabaee et al. sull'authorship attribution del malware evidenzia quanto offuscamento, riuso di codice altrui e sviluppo collettivo erodano il segnale autoriale nei campioni reali [56].

Sul versante dell'*infrastruttura*, contano le abitudini: riuso di server C2 tra campagne, pattern di registrazione dei domini, preferenze di hosting, riutilizzo di certificati. Sono evidenze che emergono precisamente dall'arricchimento con passive DNS e dati di registrazione descritto nel §3.4, a conferma che la qualità dell'attribuzione si decide a monte, nella pipeline. Sul versante della *vittima*, la vittimologia (chi viene colpito, in quale settore e geografia, con quale tempistica) parla dell'intento. Trasversale a tutto è l'evidenza comportamentale: la coerenza del tradecraft profilata sul vocabolario ATT&CK [10], gli orari operativi e i loro fusi, gli artefatti linguistici. Il principio ordinatore di questo materiale eterogeneo lo hanno formulato Skopik e Pahi, secondo cui l'affidabilità di un artefatto ai fini attributivi è inversamente proporzionale alla facilità con cui può essere contraffatto; la loro rassegna delle tecniche d'attacco lungo la kill chain valuta sistematicamente quanto ciascuna categoria di tracce sia falsificabile [60]. In cima alla scala di affidabilità stanno, ancora una volta, le TTPs: coerentemente con la Pyramid of Pain [9], ciò che all'avversario costa di più cambiare è anche ciò che lo tradisce più a lungo.

7.3 Approcci ML-assisted e limiti epistemici: false flag e deception

La domanda di automazione è naturale, poiché i volumi del capitolo 3 valgono anche qui. La survey di Rani, Saha e Shukla sistematizza il campo dell'attribuzione automatizzata delle APT proponendo una tassonomia degli artefatti utili all'attribuzione, una classificazione dei dataset disponibili e dei metodi correnti, e una discussione critica di limiti e problemi aperti [57]. I filoni principali ricalcano i capitoli precedenti. Il clustering di campioni in famiglie e il code-similarity learning estendono le tecniche del §4.3; sul versante testuale, sistemi come NO-DOUBT affrontano l'attribuzione a partire dai report di threat intelligence [58], e la ricerca sull'etichettatura automatica dell'attore principale nei report CTI tramite modelli NLP dedicati si dichiara esplicitamente primo passo verso framework di attribuzione APT completamente automatizzati [59], mettendo la colonna del capitolo 5 al servizio del «chi».

I limiti, però, sono qui più profondi che altrove. Il primo è la circolarità della ground truth: le etichette con cui si addestrano i classificatori provengono in larga parte dagli stessi report dei vendor la cui attribuzione si vorrebbe automatizzare, e la scarsità di dataset indipendenti è tra le criticità centrali censite dalla survey [57]. Il secondo è la natura mobile delle classi, poiché i gruppi si scindono, si fondono, condividono strumenti, e la diffusione di commodity malware diluisce il valore probatorio del tooling condiviso. Il terzo è adversarial in senso proprio, e riguarda le *false flag campaign*, attacchi che impiantano deliberatamente artefatti fuorvianti (stringhe in lingue altrui, riuso ostentato di codice di altri gruppi, orari operativi simulati) per deviare l'attribuzione. È il caso di scuola documentato pubblicamente da Olympic Destroyer (2018), costruito per somigliare a più attori insieme, ed è la ragione per cui Skopik e Pahi ancorano l'intero processo alla valutazione di contraffabilità degli artefatti, separando l'investigazione dell'attacco dalla profilazione dell'attore e ricongiungendole solo nella fase attributiva del loro Cyber Attribution Model [60]. La conseguenza metodologica è netta. Nell'attribuzione il ML produce *ipotesi corredate di confidenza*, mai verdetti; l'aggiudicazione resta umana, e la spiegabilità delle inferenze, su cui tornerà il capitolo 8, non è un lusso ma un requisito, perché un'attribuzione che non sa mostrare le proprie evidenze non è utilizzabile da nessuno dei fori del §7.4.

7.4 Dimensione geopolitica e standard probatori

Al livello dello sponsor, l'attribuzione cessa di essere un esercizio analitico e diventa un atto politico, il piano strategico del Q Model, dove essa è funzione della posta in gioco [54]. Gli strumenti con cui gli Stati la esercitano formano una scala: attribuzioni pubbliche congiunte, incriminazioni nominative che cristallizzano l'evidenza in atti giudiziari, regimi sanzionatori dedicati, di cui anche l'Unione europea si è dotata nel quadro della propria risposta diplomatica alle attività cyber ostili. La rilevanza pratica del tema per il contesto europeo è testimoniata dalla stessa reportistica istituzionale, che identifica negli insiemi di intrusione state-nexus gli attori più attivi contro pubblica amministrazione, difesa e telecomunicazioni dell'Unione [1].

Ogni foro applica però il proprio standard probatorio. Il processo penale richiede prova oltre ogni ragionevole dubbio, con catena di custodia e ammissibilità forense. La valutazione di intelligence si esprime in linguaggio estimativo e confidenze graduate. L'attribuzione pubblica risponde a soglie politiche prima che giuridiche. L'attribuzione privata dei vendor, infine, opera sotto incentivi commerciali e reputazionali propri, che il consumatore di CTI deve saper scontare. Per la pratica che questo lavoro ha di mira, le implicazioni si lasciano condensare in tre regole di igiene analitica: tenere separate le affermazioni sul cluster da quelle sullo sponsor, perché vivono a livelli di confidenza diversi; documentare la catena delle evidenze e propagarne la confidenza fino al prodotto finale, in forma leggibile dal destinatario; trattare la mappatura degli alias tra nomenclature come attività di manutenzione permanente del knowledge graph, non come dettaglio. Resta, sullo sfondo, una simmetria che introduce il capitolo conclusivo: gli stessi modelli che assistono l'attribuzione sono a loro volta bersagli di evasione, di avvelenamento, di inganno, e la robustezza dell'analisi dipende ormai anche dalla robustezza degli strumenti che la producono.

8. Adversarial Machine Learning e open challenges

I capitoli precedenti hanno costruito i modelli di ML come strumento della difesa; questo li tratta come suo bersaglio. È il rovescio necessario dell'intera argomentazione, poiché la stessa incorporazione di algoritmi di apprendimento che il §1.1 indicava come condizione della difesa proattiva apre, per usare le parole della survey di Sun et al., opportunità difensive inedite ma anche tipologie di rischio nuove [4]. Il capitolo sistematizza le vulnerabilità adversarial dei modelli (§8.1–8.2), discute le fragilità che ne condizionano l'uso operativo anche in assenza di un avversario attivo, ossia drift, bias e opacità (§8.3), e chiude sulle direttrici di ricerca aperte (§8.4), che riconducono a unità i fili lasciati in sospeso dai capitoli tematici.

8.1 Tassonomia degli attacchi: evasion, poisoning, extraction, inference

La consapevolezza che i classificatori possano essere sovvertiti da input costruiti ad arte non nasce con il deep learning. La ricostruzione storica di Biggio e Roli è esplicita nel collegare il filone pionieristico sulla sicurezza degli algoritmi non-deep, a partire dagli attacchi di evasione contro i filtri antispam di metà anni 2000, al lavoro più recente sulle proprietà di sicurezza delle reti profonde, mostrando connessioni tra linee di ricerca apparentemente distinte e sfatando fraintendimenti diffusi [61]. Su questa base la letteratura ha consolidato una tassonomia bidimensionale, sintetizzata in Figura 5: una dimensione per fase del ciclo di vita del modello, cioè in quale momento l'attaccante interviene, e una per conoscenza e obiettivo dell'attaccante, cioè quanto sa e cosa vuole ottenere. La sistematizzazione di riferimento nel dominio della sicurezza è quella di Rosenberg et al. su *ACM Computing Surveys*, che organizza attacchi e difese attraverso le applicazioni cyber-oriented del ML [62].

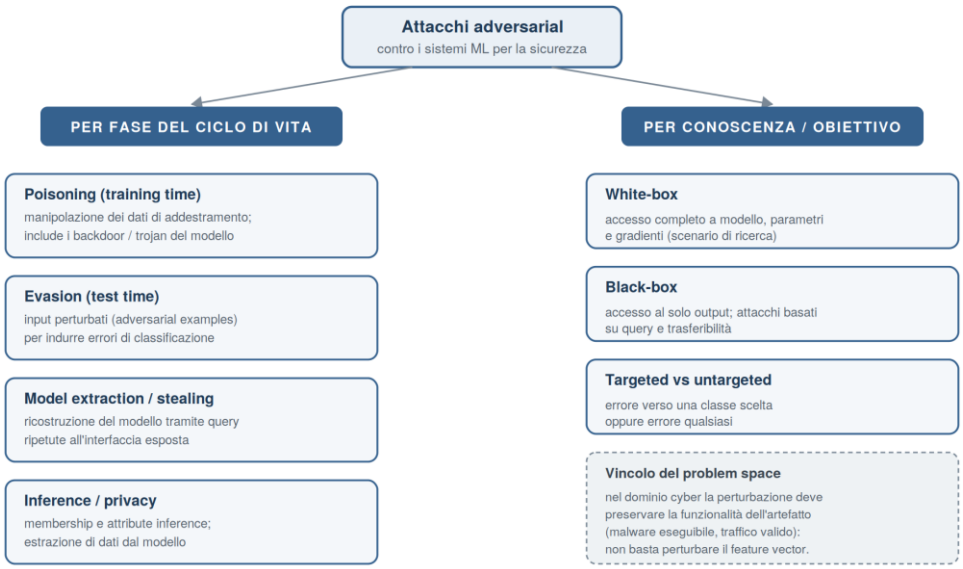

Figura 5 — Tassonomia degli attacchi adversarial contro i sistemi di ML per la sicurezza. Le due dimensioni, fase del ciclo di vita e conoscenza/obiettivo dell'attaccante, sono ortogonali; il riquadro tratteggiato evidenzia il vincolo del problem space, specifico del dominio cyber.

Lungo l'asse della fase si distinguono quattro classi. Gli attacchi di **evasion** operano a test time, con input perturbati, gli *adversarial example* o *wild pattern* [61], che inducono errori

di classificazione lasciando immutato il modello. Gli attacchi di **poisoning** operano a training time, contaminando i dati di addestramento; la loro variante più insidiosa è il **backdoor** (o trojan), in cui il modello apprende un comportamento corretto sul caso generale ma anomalo in presenza di un trigger scelto dall'attaccante. Gli attacchi di **model extraction** ricostruiscono un modello confidenziale interrogandone ripetutamente l'interfaccia esposta, mentre gli attacchi di **inference** violano la riservatezza di dati e modello: membership inference (stabilire se un record faceva parte del training set), attribute inference, estrazione di dati memorizzati. Lungo l'asse ortogonale, l'attaccante può essere *white-box* (accesso completo a architettura, parametri e gradienti) o *black-box* (accesso al solo output, con attacchi basati su query e sulla trasferibilità degli esempi tra modelli), e il suo obiettivo *targeted* (indurre l'errore verso una classe scelta) o *untargeted* (un errore qualsiasi). Per la CTI il poisoning ha una rilevanza speciale, perché salda questo capitolo al §3.1: là dove parte del dato è generata dall'avversario e la veridicità è essa stessa una superficie d'attacco, l'ingresso della pipeline è un vettore di poisoning permanente, e lo è a maggior ragione quando i dati di addestramento provengono da feed di qualità incerta (§3.4) o da fonti aperte manipolabili (§5.1).

8.2 Adversarial examples contro NIDS e malware detector

Il trasferimento di questa tassonomia dalla computer vision, dove è nata e dove la perturbazione impercettibile di un pixel è il caso paradigmatico, al dominio cyber incontra un ostacolo che ne cambia la natura. È il punto su cui la survey di He, Kim e Asghar su *IEEE Communications Surveys & Tutorials* fonda la propria analisi dei NIDS: gli studi sugli attacchi adversarial contro i sistemi di intrusion detection implementano spesso, in modo diretto, attacchi concepiti per la computer vision, ignorando le differenze fondamentali di pipeline di detection e spazi delle feature tra i due domini, sicché lanciare, e rilevare, attacchi adversarial contro i NIDS resta una sfida di ricerca aperta [63].

La differenza fondamentale è il vincolo del *problem space*, evidenziato nel riquadro di Figura 5. Un adversarial example nella computer vision vive nel *feature space*, dove qualunque perturbazione dei valori dei pixel produce ancora un'immagine valida. Nel dominio cyber le cose stanno diversamente: un malware perturbato deve restare un eseguibile funzionante e mantenere il proprio comportamento malevolo; un flusso di rete perturbato deve restare traffico protocollarmente valido e semanticamente coerente. Non è lecito, cioè, muovere arbitrariamente il vettore delle feature. Occorre trovare una modifica nello spazio del problema (byte del binario, pacchetti sulla rete) che si proietti nella perturbazione desiderata dello spazio delle feature, preservando i vincoli di validità e funzionalità. È la distinzione tra feature-space e problem-space attack che struttura la letteratura recente su malware e NIDS, e che rende gli attacchi realistici molto più costosi di quanto i risultati sulla computer vision lascerebbero supporre.

Qui interviene il correttivo di realismo più importante del capitolo. Il position paper di Apruzzese et al., significativamente intitolato «Real Attackers Don't Compute Gradients», osserva che, a fronte di una letteratura che dimostra attacchi algoritmici potenti contro ogni sorta di modello, l'evidenza empirica indica che gli attaccanti reali sovvertono i sistemi ML con tattiche semplici, e che di conseguenza i professionisti della sicurezza non hanno prioritizzato le difese contro l'adversarial ML [64]. Il divario tra la minaccia studiata dai ricercatori — l'avversario white-box che calcola gradienti — e quella effettivamente praticata, fatta di

attaccanti che aggirano il modello con mezzi banali o ne avvelenano l'ingresso, è esso stesso un rischio, perché disallinea gli sforzi difensivi. Ne discende una postura pragmatica, coerente con il §4.5: modellare la minaccia realistica prima di quella teorica, valutare la robustezza con avversari plausibili e budget di costo espliciti, e ricordare che nel dominio cyber la difesa più efficace resta spesso la qualità e l'integrità del dato in ingresso, non l'irrobustimento del solo modello.

8.3 Concept drift, dataset bias, explainability

Non serve un avversario attivo perché un modello di sicurezza degradi. Tre fragilità strutturali ne condizionano l'uso operativo di per sé, e attraversano tutti i capitoli precedenti. Il **concept drift** viene per primo: la distribuzione della minaccia muta di continuo, e un modello addestrato su una fotografia del passato perde aderenza col tempo. È lo stesso fenomeno che il §4.3 ha incontrato con TESSERACT [25] e che il §4.4 ha visto erodere la rappresentatività dei benchmark dal giorno della loro pubblicazione. La conseguenza gestionale è che un modello di detection non è un artefatto che si rilascia una volta, ma un sistema da monitorare per il drift, rivalutare in modo time-aware e ri-addestrare con cadenza governata.

Il **dataset bias** è la seconda fragilità, e ne abbiamo visto i due volti. C'è il bias della singola base dati, con gli artefatti sistematici documentati in CICIDS2017 [29] e le correlazioni spurie per cui un modello impara a riconoscere un intervallo di indirizzi anziché un pattern d'attacco [33]. E c'è il bias di copertura delle fonti, la sovrapposizione sorprendentemente bassa tra feed che misura quanto parziale sia la visione di ciascuno [16]. In entrambi i casi il modello apprende le idiosincrasie della provenienza più della semantica della minaccia, e in entrambi la contromisura sta a monte del modello, nella qualità del dato (§3.4) e nel rigore della valutazione (§4.5).

La terza fragilità è l'**opacità**, quella con le implicazioni più trasversali. Un modello che non sa spiegare la propria decisione produce allarmi che l'analista non può validare, il che riporta direttamente all'alert fatigue dei SOC del §1.1 [3], e attribuzioni che nessun foro del §7.4 può utilizzare, perché un'attribuzione senza evidenze esibibili non è azionabile. L'eXplainable AI non è quindi, nel dominio della sicurezza, un requisito estetico ma funzionale: la spiegabilità è la condizione perché l'output di un modello diventi intelligence, cioè conoscenza su cui un umano possa decidere. È anche, va detto, un requisito in tensione con le architetture deep più performanti, e la ricerca su come conciliare accuratezza e interpretabilità nei modelli di sicurezza resta aperta.

8.4 Direzioni di ricerca

I fili lasciati in sospeso dai capitoli tematici convergono in un piccolo numero di direttrici, sintetizzate in Tabella 8. Vi è anzitutto la difesa che preserva la riservatezza: il **federated learning** già incontrato come schema di CTI sharing senza centralizzazione del dato [53] promette di conciliare la collaborazione difensiva con i vincoli di privacy del §6.5, ma eredita a sua volta i problemi di questo capitolo, essendo esposto al poisoning dei partecipanti e dovendo quindi essere irrobustito di conseguenza. Vi è poi la **sicurezza degli LLM** applicati alla CTI. I modelli che il capitolo 5 ha collocato al centro dell'estrazione automatica sono essi stessi bersagli, e la prompt injection indiretta ne è la minaccia caratteristica: Greshake et al. hanno mostrato che le applicazioni LLM-integrated confondono il confine tra dati e istruzioni,

permettendo a un avversario di iniettare comandi in contenuti che il modello recupererà a inferenza [65]. Per una pipeline CTI che dà in pasto a un LLM report, post di forum e pagine web, cioè contenuto ostile per definizione, questa non è un'ipotesi di laboratorio ma il modello di minaccia operativo, e riporta al principio con cui il §5.3 raccomandava output verificati e ruolo di copilota. Vi è la difesa **robusta e spiegabile** by design, che affronti congiuntamente le fragilità del §8.3 anziché rincorrerle. E vi è, più prospettica, l'**automazione crescente del ciclo di intelligence**, dai sistemi di threat hunting autonomo fino all'agentic CTI, che sposterà la superficie d'attacco dai singoli modelli agli agenti che li orchestrano, rendendo i temi di questo capitolo più centrali, non meno.

Direttrice	Problema affrontato	Rischio ereditato	Rimando
Federated learning per la CTI	collaborazione difensiva senza centralizzare il dato	poisoning dei partecipanti; robustezza dell'aggregazione	§6.5, §8.1
Sicurezza degli LLM per la CTI	estrazione automatica da fonti non strutturate e ostili	prompt injection indiretta; riservatezza; hallucination	§5.3, §8.1
ML robusto e spiegabile by design	drift, bias e opacità come vincoli operativi	trade-off accuratezza/interpretabilità e accuratezza/robustezza	§4.5, §8.3
Automazione del ciclo (hunting autonomo, agentic CTI)	scala e velocità della difesa proattiva	spostamento della superficie d'attacco sugli agenti	§6.4, §8.2

Tabella 8 — Direttrici di ricerca aperte e loro collocazione nel lavoro.

Nessuna di queste direttrici è puramente algoritmica. Ciascuna intreccia modello, dato e organizzazione, e ciascuna conferma la tesi con cui il lavoro si è aperto, ossia che la difesa efficace vive alla confluenza di Big Data, Machine Learning e Cyber Threat Intelligence, e che la fragilità di quella confluenza va progettata anziché subita.

9. Conclusioni

Questo lavoro è partito da una tensione. Da un lato l'industrializzazione dell'offesa, con le sue campagne continue e convergenti e i suoi volumi ingestibili; dall'altro una difesa che dispone di più dati di quanti riesca a interpretarne, e di analisti sommersi da allarmi che non possono validare. La tesi proposta era che il superamento di questa asimmetria passi per la confluenza di Big Data e Machine Learning nel ciclo della Cyber Threat Intelligence, e che tale confluenza sia insieme necessaria e fragile. La rassegna condotta consente ora di articolare quella tesi in quattro affermazioni più precise, ciascuna delle quali attraversa più capitoli.

Il valore dell'intelligence cresce con il suo livello di astrazione, e con esso cresce il costo di produrla automaticamente. È il principio che la Pyramid of Pain enuncia in forma economica e che l'intero lavoro ha ritrovato su ogni piano. Gli IoC atomici si estraggono a costo quasi nullo e si neutralizzano con la stessa facilità, mentre le TTPs, ciò che all'avversario costa di più cambiare, richiedono le tecniche NLP del capitolo 5 per essere estratte dal linguaggio naturale, e restano le più difficili da rilevare, da condividere e da attribuire. La misura empirica di questo gradiente la fornisce il dato sullo sharing reale: oltre il 90% del materiale STIX condiviso si concentra in firme di malware e URL, mentre appena lo 0,09% degli indicator veicola regole di detection utilizzabili. L'ecosistema, cioè, scambia proprio ciò che vale meno.

Il collo di bottiglia non è algoritmico. È forse la conclusione meno prevedibile di una rassegna sul Machine Learning. Ma i limiti che i capitoli 3–6 hanno documentato, dalla sovrapposizione sorprendentemente bassa tra feed che dovrebbero coprire lo stesso fenomeno ai difetti di produzione dei benchmark su cui si misurano i progressi, dall'insufficienza di volume e dalla scarsa qualità dei dati condivisi fino alla fiducia come vincolo architetturale dello sharing, sono limiti di dato e di organizzazione, non di modello. Nessuna sofisticazione algoritmica compensa un ingresso di bassa qualità, e la difesa più efficace contro le manipolazioni discusse nel capitolo 8 resta spesso l'integrità della pipeline, non l'irrobustimento del classificatore.

Le illusioni di valutazione sono sistemiche, e vanno trattate come tali. Il filo che lega Sommer e Paxson ad Arp et al., passando per Axelsson, TESSERACT e la revisione di CICIDS2017, non è un catalogo di errori individuali. È la constatazione che un intero campo ha misurato per anni artefatti dei propri dataset scambiandoli per progresso metodologico. La base-rate fallacy, in particolare, non è un tecnicismo statistico ma la traduzione esatta dell'alert fatigue che si osserva nei SOC reali. Un modello «al 99%» su un benchmark bilanciato può essere, in produzione, una macchina generatrice di rumore, e la distanza tra le due cose si chiude solo con partizioni temporali realistiche, base rate operative e falsi positivi prezzati al costo-analista.

Lo strato di ML è esso stesso una superficie d'attacco. La fragilità non è accessoria, perché il dominio in cui i modelli operano è l'unico in cui una parte del dato è prodotta da un soggetto che ha interesse a falsificarlo. Il correttivo di realismo di Apruzzese et al. va però tenuto fermo: la minaccia praticata non è quella dell'avversario white-box che calcola gradienti, ma quella dell'attaccante che aggira il modello con mezzi semplici o ne avvelena l'ingresso, e disallineare gli sforzi difensivi verso la minaccia teorica è a sua volta un rischio.

Su tutte queste conclusioni pesa un limite del lavoro, che è onesto dichiarare. Trattandosi di una rassegna critica di tipo narrativo e non di una systematic literature review, la selezione delle fonti, per quanto vincolata ai criteri del §1.3, non è riproducibile in senso stretto, e l'ampiezza del perimetro adottato ha imposto profondità disuguali: alcuni sotto-domini avrebbero meritato una trattazione più fine di quanta ne consenta un contributo di ampiezza. Il lavoro non offre inoltre validazione sperimentale propria, e in un campo in cui, come si è visto, la validazione è precisamente il punto debole, questa è una limitazione da non minimizzare.

Resta, come indicazione di fondo, un rovesciamento di prospettiva. L'automazione non sostituisce il giudizio dell'analista, ne alza il rapporto segnale-rumore. Il modello di detection consegna un triage più denso di valore, la pipeline NLP trasforma il testo in oggetti correlabili, la condivisione moltiplica la copertura di ciascuno. Ma la trasformazione dell'informazione in *intelligence*, cioè in conoscenza su cui qualcuno possa decidere, richiede spiegabilità, provenienza tracciata e confidenze dichiarate. Per questo l'explainability, nel dominio della sicurezza, non è un requisito estetico: è la condizione perché l'output di una macchina possa entrare in un ciclo di intelligence. Le direttrici aperte del §8.4, dal federated learning alla sicurezza degli LLM applicati alla CTI, dalla robustezza e spiegabilità by design all'automazione crescente del ciclo, condividono tutte questa proprietà, poiché nessuna è puramente algoritmica e ciascuna intreccia modello, dato e organizzazione. È in quell'intreccio che vive la difesa efficace, ed è lì che la sua fragilità va progettata anziché subita.

Riferimenti bibliografici

[1] ENISA, *ENISA Threat Landscape 2025*, European Union Agency for Cybersecurity, ott. 2025.

[2] AV-TEST Institute, *Malware Statistics & Trends Report*, 2026. [Online]. Disponibile: https://www.av-test.org/en/statistics/malware/

[3] B. A. AlAhmadi, L. Axon e I. Martinovic, «99% False Positives: A Qualitative Study of SOC Analysts' Perspectives on Security Alarms», in *Proc. 31st USENIX Security Symposium (USENIX Security 22)*, Boston, MA, USA, 2022, pp. 2783–2800.

[4] N. Sun, M. Ding, J. Jiang, W. Xu, X. Mo, Y. Tai e J. Zhang, «Cyber Threat Intelligence Mining for Proactive Cybersecurity Defense: A Survey and New Perspectives», *IEEE Communications Surveys & Tutorials*, vol. 25, n. 3, pp. 1748–1774, 2023, doi: 10.1109/COMST.2023.3273282.

[5] W. Tounsi e H. Rais, «A survey on technical threat intelligence in the age of sophisticated cyber attacks», *Computers & Security*, vol. 72, pp. 212–233, 2018, doi: 10.1016/j.cose.2017.09.001.

[6] R. McMillan, *Definition: Threat Intelligence*, Gartner Research, G00249251, mag. 2013.

[7] A. S. Hulnick, «What's wrong with the Intelligence Cycle», *Intelligence and National Security*, vol. 21, n. 6, pp. 959–979, 2006, doi: 10.1080/02684520601046291.

[8] V. Mavroeidis e S. Bromander, «Cyber Threat Intelligence Model: An Evaluation of Taxonomies, Sharing Standards, and Ontologies within Cyber Threat Intelligence», in *Proc. 2017 European Intelligence and Security Informatics Conference (EISIC)*, Atene, Grecia, 2017, pp. 91–98, doi: 10.1109/EISIC.2017.20.

[9] D. J. Bianco, *The Pyramid of Pain*, 2013. [Online]. Disponibile: http://detect-respond.blogspot.com/2013/03/the-pyramid-of-pain.html

[10] B. E. Strom, A. Applebaum, D. P. Miller, K. C. Nickels, A. G. Pennington e C. B. Thomas, *MITRE ATT&CK: Design and Philosophy*, The MITRE Corporation, Technical Report MP18-0944, 2018 (rev. mar. 2020).

[11] E. M. Hutchins, M. J. Cloppert e R. M. Amin, «Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains», *Leading Issues in Information Warfare & Security Research*, vol. 1, n. 1, p. 80, 2011.

[12] A. Alshamrani, S. Myneni, A. Chowdhary e D. Huang, «A Survey on Advanced Persistent Threats: Techniques, Solutions, Challenges, and Research Opportunities», *IEEE Communications Surveys & Tutorials*, vol. 21, n. 2, pp. 1851–1877, 2019, doi: 10.1109/COMST.2019.2891891.

[13] S. Caltagirone, A. Pendergast e C. Betz, *The Diamond Model of Intrusion Analysis*, Center for Cyber Intelligence Analysis and Threat Research, Hanover, MD, Technical Report ADA586960, lug. 2013.

[14] D. Laney, *3D Data Management: Controlling Data Volume, Velocity and Variety*, META Group Research Note, feb. 2001.

[15] J. Franco, A. Aris, B. Canberk e A. S. Uluagac, «A Survey of Honeypots and Honeynets for Internet of Things, Industrial Internet of Things, and Cyber-Physical Systems», *IEEE Communications Surveys & Tutorials*, vol. 23, n. 4, pp. 2351–2383, 2021, doi: 10.1109/COMST.2021.3106669.

[16] V. G. Li, M. Dunn, P. Pearce, D. McCoy, G. M. Voelker, S. Savage e K. Levchenko, «Reading the Tea Leaves: A Comparative Analysis of Threat Intelligence», in *Proc. 28th USENIX Security Symposium (USENIX Security 19)*, Santa Clara, CA, USA, 2019, pp. 851–867.

[17] F. Ullah e M. A. Babar, «Architectural Tactics for Big Data Cybersecurity Analytics Systems: A Review», *Journal of Systems and Software*, vol. 151, pp. 81–118, 2019, doi: 10.1016/j.jss.2019.01.051.

[18] G. González-Granadillo, S. González-Zarzosa e R. Diaz, «Security Information and Event Management (SIEM): Analysis, Trends, and Usage in Critical Infrastructures», *Sensors*, vol. 21, n. 14, art. 4759, 2021, doi: 10.3390/s21144759.

[19] H. Griffioen, T. Booij e C. Doerr, «Quality Evaluation of Cyber Threat Intelligence Feeds», in *Applied Cryptography and Network Security (ACNS 2020)*, Roma, Italia, 2020, Proceedings Part II, pp. 277–296, Springer.

[20] D. Schlette, F. Böhm, M. Caselli e G. Pernul, «Measuring and visualizing cyber threat intelligence quality», *International Journal of Information Security*, vol. 20, pp. 21–38, 2021, doi: 10.1007/s10207-020-00490-y.

[21] A. L. Buczak e E. Guven, «A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection», *IEEE Communications Surveys & Tutorials*, vol. 18, n. 2, pp. 1153–1176, 2016, doi: 10.1109/COMST.2015.2494502.

[22] R. Sommer e V. Paxson, «Outside the Closed World: On Using Machine Learning for Network Intrusion Detection», in *Proc. 2010 IEEE Symposium on Security and Privacy*, Oakland, CA, USA, 2010, pp. 305–316, doi: 10.1109/SP.2010.25.

[23] D. Ucci, L. Aniello e R. Baldoni, «Survey of machine learning techniques for malware analysis», *Computers & Security*, vol. 81, pp. 123–147, 2019, doi: 10.1016/j.cose.2018.11.001.

[24] D. Gibert, C. Mateu e J. Planes, «The rise of machine learning for detection and classification of malware: Research developments, trends and challenges», *Journal of Network and Computer Applications*, vol. 153, art. 102526, 2020, doi: 10.1016/j.jnca.2019.102526.

[25] F. Pendlebury, F. Pierazzi, R. Jordaney, J. Kinder e L. Cavallaro, «TESSERACT: Eliminating Experimental Bias in Malware Classification across Space and Time», in *Proc. 28th USENIX Security Symposium (USENIX Security 19)*, Santa Clara, CA, USA, 2019, pp. 729–746.

[26] M. Tavallaee, E. Bagheri, W. Lu e A. A. Ghorbani, «A detailed analysis of the KDD CUP 99 data set», in *Proc. IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, Ottawa, Canada, 2009.

[27] N. Moustafa e J. Slay, «UNSW-NB15: a comprehensive data set for network intrusion detection systems», in *Proc. Military Communications and Information Systems Conference (MilCIS)*, Canberra, Australia, 2015, pp. 1–6, doi: 10.1109/MilCIS.2015.7348942.

[28] I. Sharafaldin, A. Habibi Lashkari e A. A. Ghorbani, «Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization», in *Proc. 4th International Conference on Information Systems Security and Privacy (ICISSP)*, Funchal, Portogallo, 2018.

[29] G. Engelen, V. Rimmer e W. Joosen, «Troubleshooting an Intrusion Detection Dataset: the CICIDS2017 Case Study», in *Proc. 2021 IEEE Security and Privacy Workshops (SPW)*, 2021, pp. 7–12, doi: 10.1109/SPW53761.2021.00009.

[30] H. S. Anderson e P. Roth, «EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models», arXiv:1804.04637, 2018.

[31] D. Arp, M. Spreitzenbarth, M. Hübner, H. Gascon e K. Rieck, «DREBIN: Effective and Explainable Detection of Android Malware in Your Pocket», in *Proc. Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, USA, 2014.

[32] S. Axelsson, «The base-rate fallacy and the difficulty of intrusion detection», *ACM Transactions on Information and System Security*, vol. 3, n. 3, pp. 186–205, 2000, doi: 10.1145/357830.357849.

[33] D. Arp, E. Quiring, F. Pendlebury, A. Warnecke, F. Pierazzi, C. Wressnegger, L. Cavallaro e K. Rieck, «Dos and Don'ts of Machine Learning in Computer Security», in *Proc. 31st USENIX Security Symposium (USENIX Security 22)*, Boston, MA, USA, 2022, pp. 3971–3988.

[34] M. Arazzi, D. R. Arikkat, S. Nicolazzo, A. Nocera, K. A. Rafidha Rehiman, P. Vinod e M. Conti, «NLP-based techniques for cyber threat intelligence», *Computer Science Review*, vol. 58, art. 100765, 2025, doi: 10.1016/j.cosrev.2025.100765.

[35] J. Robertson, A. Diab, E. Marin, E. Nunes, V. Paliath, J. Shakarian e P. Shakarian, *Darkweb Cyber Threat Intelligence Mining*, Cambridge University Press, 2017.

[36] N. Zhang, M. Ebrahimi, W. Li e H. Chen, «Counteracting Dark Web Text-Based CAPTCHA with Generative Adversarial Learning for Proactive Cyber Threat Intelligence», *ACM Transactions on Management Information Systems*, vol. 13, n. 2, pp. 1–21, 2022.

[37] X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing e R. Beyah, «Acing the IOC Game: Toward Automatic Discovery and Analysis of Open-Source Cyber Threat Intelligence», in *Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS)*, Vienna, Austria, 2016, pp. 755–766, doi: 10.1145/2976749.2978315.

[38] G. Husari, E. Al-Shaer, M. Ahmed, B. Chu e X. Niu, «TTPDrill: Automatic and Accurate Extraction of Threat Actions from Unstructured Text of CTI Sources», in *Proc. 33rd Annual Computer Security Applications Conference (ACSAC)*, Orlando, FL, USA, 2017, pp. 103–115, doi: 10.1145/3134600.3134646.

[39] Z. Li, J. Zeng, Y. Chen e Z. Liang, «AttacKG: Constructing Technique Knowledge Graph from Cyber Threat Intelligence Reports», in *Proc. European Symposium on Research in Computer Security (ESORICS)*, 2022, pp. 589–609.

[40] Y. Chen, M. Cui, D. Wang, Y. Cao, P. Yang, B. Jiang, Z. Lu e B. Liu, «A survey of large language models for cyber threat detection», *Computers & Security*, vol. 145, art. 104016, 2024, doi: 10.1016/j.cose.2024.104016.

[41] C. Sabottke, O. Suciu e T. Dumitraș, «Vulnerability Disclosure in the Age of Social Media: Exploiting Twitter for Predicting Real-World Exploits», in *Proc. 24th USENIX Security Symposium (USENIX Security 15)*, Washington, D.C., USA, 2015, pp. 1041–1056.

[42] N. Tavabi, P. Goyal, M. Almukaynizi, P. Shakarian e K. Lerman, «DarkEmbed: Exploit Prediction with Neural Language Models», in *Proc. AAAI Conference on Artificial Intelligence*, 2018.

[43] I. Deliu, C. Leichter e K. Franke, «Extracting Cyber Threat Intelligence from Hacker Forums: Support Vector Machines versus Convolutional Neural Networks», in *Proc. IEEE International Conference on Big Data*, Boston, MA, USA, 2017.

[44] OASIS, *STIX™ Version 2.1*, OASIS Standard, giu. 2021. [Online]. Disponibile: https://docs.oasis-open.org/cti/stix/v2.1/stix-v2.1.html

[45] OASIS, *TAXII™ Version 2.1*, OASIS Standard, giu. 2021. [Online]. Disponibile: https://docs.oasis-open.org/cti/taxii/v2.1/taxii-v2.1.html

[46] P. Alaeifar, S. Pal, Z. Jadidi, M. Hussain e E. Foo, «Current approaches and future directions for Cyber Threat Intelligence sharing: A survey», *Journal of Information Security and Applications*, vol. 83, art. 103786, 2024, doi: 10.1016/j.jisa.2024.103786.

[47] C. Wagner, A. Dulaunoy, G. Wagener e A. Iklody, «MISP: The Design and Implementation of a Collaborative Threat Intelligence Sharing Platform», in *Proc. ACM Workshop on Information Sharing and Collaborative Security (WISCS)*, Vienna, Austria, 2016, pp. 49–56.

[48] B. Jin, E. Kim, H. Lee, E. Bertino, D. Kim e H. Kim, «Sharing cyber threat intelligence: Does it really help?», in *Proc. Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, USA, 2024.

[49] T. D. Wagner, K. Mahbub, E. Palomar e A. E. Abdallah, «Cyber threat intelligence sharing: Survey and research directions», *Computers & Security*, vol. 87, art. 101589, 2019, doi: 10.1016/j.cose.2019.101589.

[50] C. Johnson, L. Badger, D. Waltermire, J. Snyder e C. Skorupka, *Guide to Cyber Threat Information Sharing*, NIST Special Publication 800-150, National Institute of Standards and Technology, 2016.

[51] Parlamento europeo e Consiglio, *Direttiva (UE) 2022/2555 relativa a misure per un livello comune elevato di cibersicurezza nell'Unione (NIS2)*, GU L 333, 27 dic. 2022.

[52] FIRST, *Traffic Light Protocol (TLP) — Version 2.0*, Forum of Incident Response and Security Teams, 2022.

[53] M. Sarhan, S. Layeghy, N. Moustafa e M. Portmann, «Cyber Threat Intelligence Sharing Scheme Based on Federated Learning for Network Intrusion Detection», *Journal of Network and Systems Management*, vol. 31, n. 1, art. 3, 2023, doi: 10.1007/s10922-022-09691-3.

[54] T. Rid e B. Buchanan, «Attributing Cyber Attacks», *Journal of Strategic Studies*, vol. 38, n. 1–2, pp. 4–37, 2015, doi: 10.1080/01402390.2014.977382.

[55] A. Caliskan, F. Yamaguchi, E. Dauber, R. Harang, K. Rieck, R. Greenstadt e A. Narayanan, «When Coding Style Survives Compilation: De-anonymizing Programmers from Executable Binaries», in *Proc. Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, USA, 2018.

[56] S. Alrabaee, P. Shirani, M. Debbabi e L. Wang, «On the Feasibility of Malware Authorship Attribution», in *Foundations and Practice of Security (FPS 2016)*, LNCS, Springer, 2017, pp. 256–272.

[57] N. Rani, B. Saha e S. K. Shukla, «A Comprehensive Survey of Advanced Persistent Threat Attribution: Taxonomy, Methods, Challenges and Open Research Problems», arXiv:2409.11415, 2024.

[58] L. Perry, B. Shapira e R. Puzis, «NO-DOUBT: Attack Attribution Based On Threat Intelligence Reports», in *Proc. IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2019.

[59] H. Abdi, S. R. Bagley, S. Furnell e J. Twycross, «Automatically Labeling Cyber Threat Intelligence Reports Using Natural Language Processing», in *Proc. ACM Symposium on Document Engineering (DocEng)*, 2023.

[60] F. Skopik e T. Pahi, «Under false flag: using technical artifacts for cyber attack attribution», *Cybersecurity*, vol. 3, art. 8, 2020, doi: 10.1186/s42400-020-00048-4.

[61] B. Biggio e F. Roli, «Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning», *Pattern Recognition*, vol. 84, pp. 317–331, 2018, doi: 10.1016/j.patcog.2018.07.023.

[62] I. Rosenberg, A. Shabtai, Y. Elovici e L. Rokach, «Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain», *ACM Computing Surveys*, vol. 54, n. 5, art. 108, 2021, doi: 10.1145/3453158.

[63] K. He, D. D. Kim e M. R. Asghar, «Adversarial Machine Learning for Network Intrusion Detection Systems: A Comprehensive Survey», *IEEE Communications Surveys & Tutorials*, vol. 25, n. 1, pp. 538–566, 2023, doi: 10.1109/COMST.2022.3233793.

[64] G. Apruzzese, H. S. Anderson, S. Dambra, D. Freeman, F. Pierazzi e K. A. Roundy, «"Real Attackers Don't Compute Gradients": Bridging the Gap between Adversarial ML Research and Practice», in *Proc. 1st IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, Raleigh, NC, USA, 2023, pp. 339–364, doi: 10.1109/SaTML54575.2023.00031.

[65] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz e M. Fritz, «Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection», in *Proc. 16th ACM Workshop on Artificial Intelligence and Security (AISec)*, Copenhagen, Danimarca, 2023, pp. 79–90, doi: 10.1145/3605764.3623985.