

MODULO 08

Digital Forensics assistita dall'AI

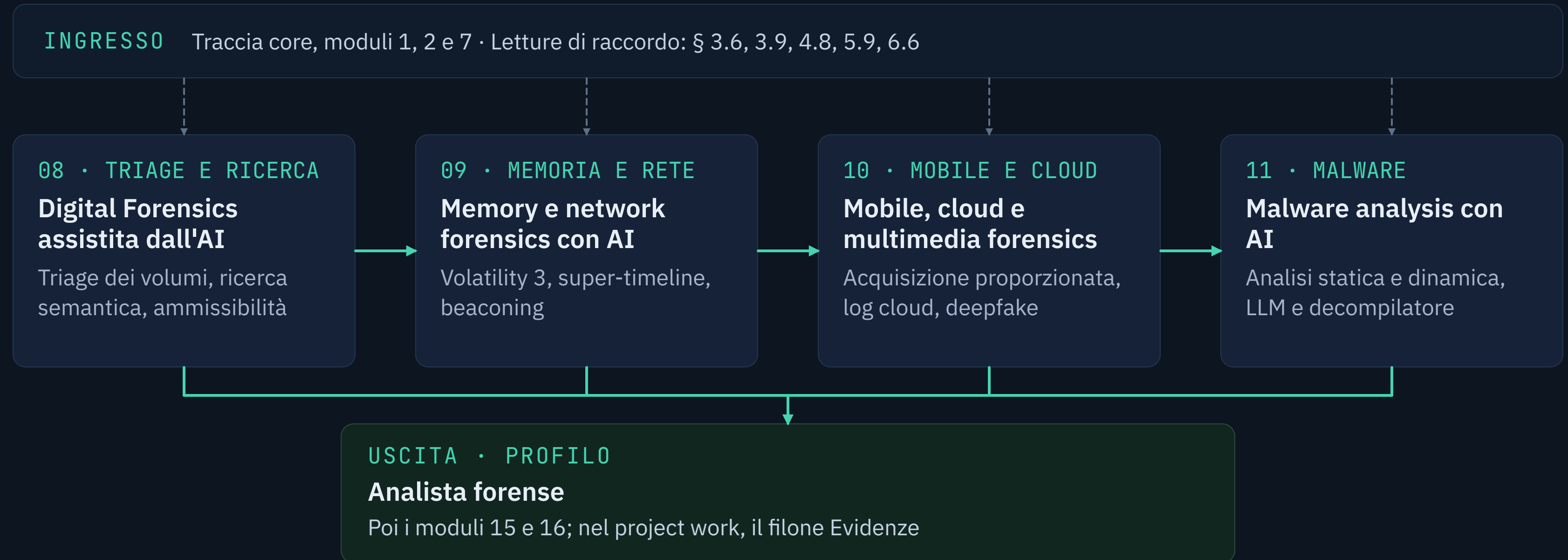
Triage automatizzato dei supporti, classificazione degli artefatti e ricerca semantica;
condizioni di ammissibilità probatoria degli esiti prodotti da un modello.

TRIAGE AI

Magnet Copilot

CRITERI DAUBERT

La specializzazione Digital Forensics



Abstract e obiettivi formativi

Abstract

Il modulo esamina l'applicazione dell'AI al processo forense: triage dei supporti su larga scala, classificazione automatica di immagini e documenti, ricerca semantica in linguaggio naturale, clustering delle comunicazioni. Una parte del modulo è dedicata a bias, riproducibilità e ammissibilità probatoria degli esiti.

Obiettivi formativi (Bloom)

- **Applicare:** impiegare strumenti AI per ridurre i tempi di triage su grandi volumi
- **Valutare:** giudicare criticamente l'affidabilità di una classificazione automatica
- **Valutare:** argomentare le condizioni di ammissibilità di un'evidenza individuata con l'AI

Traccia core

01 02 07

Detection & Response

03 04 05 06

o

Digital Forensics

08 09 10 11

o

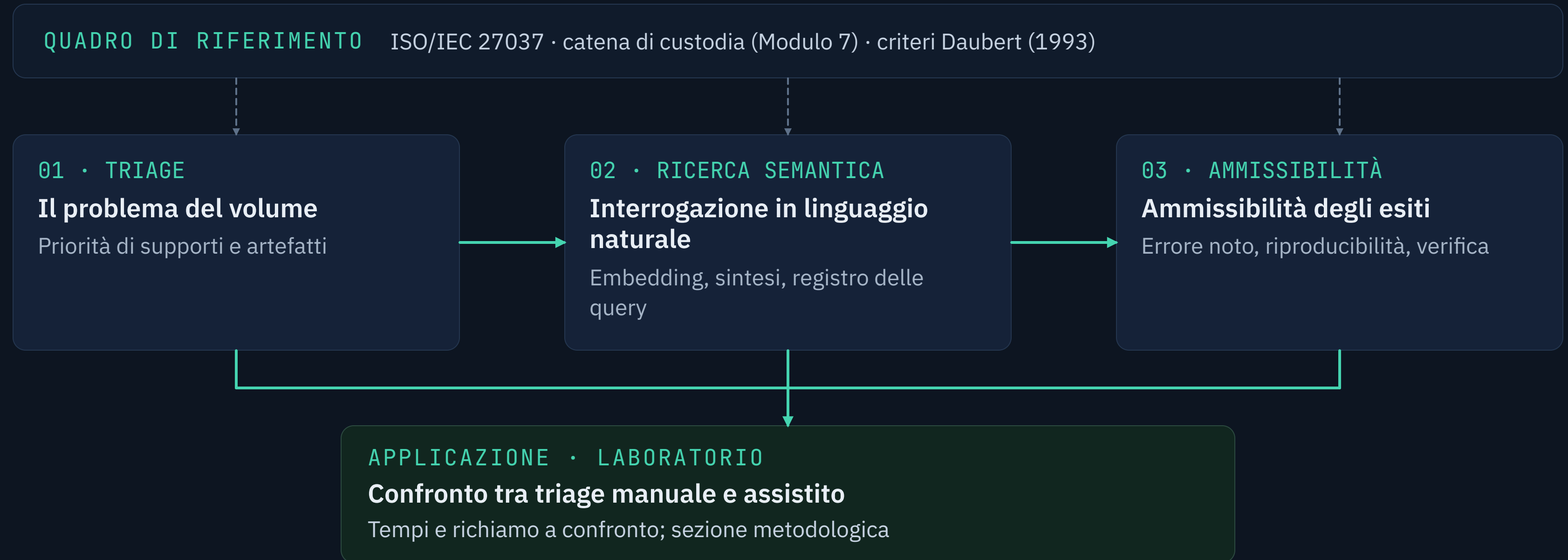
Sicurezza dell'AI

06 12 13 14

Traccia core

15 16

Selezione assistita del materiale probatorio



Il problema del volume

TERABYTE

I casi con più dispositivi raggiungono spesso terabyte di dati e milioni di file: il triage assistito ordina supporti e artefatti prima dell'analisi approfondita

Cosa automatizza

- Classificazione di immagini e documenti per categoria, incluso materiale illecito
- Prioritizzazione dei supporti per probabilità di rilevanza
- Clustering di comunicazioni e identificazione di relazioni
- Deduplicazione e filtraggio di file noti (hash set NSRL)

Nel triage probatorio si privilegia il **richiamo** rispetto alla precisione, perché un falso negativo può far perdere un'evidenza rilevante, mentre un falso positivo comporta tempo di revisione.

Interrogare l'evidenza in linguaggio naturale

Oltre la keyword search

- La ricerca semantica trova «pagamenti sospetti» anche dove la parola non compare
- Strumenti commerciali: Magnet Copilot e funzioni analoghe nelle suite forensi
- Sintesi di chat e documenti lunghi per orientare l'esaminatore

Cautele metodologiche

- Il risultato semantico orienta la ricerca; l'evidenza è il file originale, verificato dall'esaminatore
- Richiamo non garantito: la ricerca semantica si affianca a quella per parole chiave
- Ogni interrogazione e il suo esito vanno registrati nel log del caso

Ammissibilità degli esiti di un modello

Criteri Daubert (1993)

- **Verificabilità:** il metodo è controllabile ed è stato verificato empiricamente
- **Peer review** e pubblicazione del metodo
- **Tasso d'errore** noto o stimabile e standard che regolano l'applicazione del metodo
- Accettazione generale nella comunità scientifica di riferimento

Conseguenze operative

- Versione del modello, parametri e dati documentati nella relazione
- Ogni esito AI verificato dal perito, che ne risponde
- Dati di addestramento non rappresentativi producono errori sistematici, contestabili in giudizio
- La sezione metodologica della relazione dichiara l'uso dell'AI

Strumenti di scoperta e garanzie di metodo



Triage assistito e ordine della coda di lavoro

L'aumento del volume dei dati è una delle principali cause di ritardo nella digital forensics: terabyte per caso, decine di dispositivi, arretrati che nei laboratori possono durare mesi. Il triage assistito dall'AI ordina la coda di lavoro: stabilisce quali dispositivi, partizioni o file esaminare per primi e assegna a ciascun elemento un **punteggio di rilevanza** calcolato su tipo di file, hash noti, classificatori di contenuto e andamenti temporali. Le decisioni sull'esame spettano all'esaminatore, che riceve dal modello una proposta di ordine. Gli elementi collocati in fondo alla coda restano nell'insieme acquisito e possono essere esaminati in seguito, perché il punteggio non comporta l'esclusione o la distruzione di alcun dato.

Scenario didattico

Si consideri un sequestro di 12 dispositivi per 8 TB complessivi. Il triage automatico segnala un notebook con un programma di cancellazione sicura installato e migliaia di file eliminati di recente, e un telefono con un'app di messaggistica cifrata usata nelle ore del fatto. Il laboratorio inizia da questi due dispositivi e prosegue con gli altri dieci, anticipando la prima evidenza utile.

Limiti

Il punteggio riflette i dati di addestramento: un modello tarato su casi di frode può sottovalutare artefatti tipici di altri reati. Si aggiunge l'effetto di **tunnel vision** indotto dallo strumento: gli elementi in fondo alla coda rischiano di non essere mai esaminati, pur essendo formalmente disponibili.

Ricerca semantica nei corpora sequestrati

La ricerca per parole chiave restituisce solo i testi che contengono i termini cercati e fallisce con gergo, errori di battitura, lingue miste e allusioni. La ricerca semantica rappresenta ogni frammento di testo come un **embedding**, cioè un vettore numerico che ne codifica il significato, e restituisce i passaggi concettualmente vicini alla domanda anche in assenza di parole in comune. Su chat, e-mail e documenti sequestrati il corpus può così essere interrogato per temi («pagamenti in contanti non registrati», «incontri da tenere riservati») anziché per singoli termini da prevedere.

Scenario didattico

In un corpus di 200.000 messaggi la parola chiave «droga» restituisce tre risultati non pertinenti. La query semantica «consegna di sostanze dietro pagamento» individua conversazioni su «la solita roba», «quella cosa di sabato», «il pacco del giardiniere», espressioni allusive che una lista di parole chiave non avrebbe previsto.

Metodo

La ricerca semantica è uno strumento di **scoperta**: l'ordinamento dei risultati non ha valore probatorio e i falsi negativi sono possibili. Le query utilizzate vanno verbalizzate, perché sono scelte dell'esaminatore, e i risultati rilevanti vanno riscontrati sul dato originale, nel suo contesto completo.

Spiegabilità degli esiti in sede di controesame

Un risultato che l'esperto non sa spiegare è difficilmente difendibile in contraddittorio. Applicati al machine learning, i criteri Daubert (verificabilità del metodo, **tasso di errore noto**, peer review, accettazione nella comunità scientifica) portano in genere a chiedere conto dei dati di addestramento e di validazione del modello, delle metriche di errore, della riproducibilità dell'output, che richiede lo stesso risultato a parità di versione, parametri e input, e delle caratteristiche dell'input che hanno pesato sulla classificazione. A quest'ultimo scopo servono le tecniche di **explainable AI**, come SHAP, LIME e le mappe di salienza, con cui si motiva la segnalazione del modello e la si verifica.

Esempio

Domande tipiche del controesame riguardano il tasso documentato di falsi positivi del software, la popolazione su cui è stato validato e la possibilità di ottenere oggi lo stesso risultato. L'esperto risponde con il rapporto di validazione, la versione del modello indicata nel verbale e la mappa di salienza che indica le regioni dell'immagine che hanno inciso di più sulla classificazione, presentata con i limiti delle spiegazioni a posteriori.

Metodo

La relazione descrive sia la segnalazione prodotta dal modello sia la verifica condotta dall'esperto con metodo tradizionale. Se il modello è stato aggiornato o i parametri non sono stati registrati, il risultato non si può riprodurre e in contraddittorio regge solo la verifica condotta sul dato originale; per questo la conservazione della versione e della configurazione fa parte dell'acquisizione.

Hash di similarità e fuzzy hashing

L'hash crittografico (Modulo 7) dà un esito binario: due file sono identici o diversi, e un solo bit modificato cambia il valore. Le indagini richiedono spesso di misurare la somiglianza: un documento leggermente modificato, una variante di un malware, un'immagine ricompresa. A questo servono gli hash di similarità: per i file **ssdeep**, che confronta le firme di frammenti delimitati in base al contenuto, e **TLSH**, che sintetizza la distribuzione di brevi sequenze di byte, entrambi resistenti a piccole modifiche; per le immagini gli hash percettivi, che descrivono l'aspetto visivo e resistono a ridimensionamento e ricomprensione. Tecniche di questo tipo sono usate anche nel confronto con le liste di hash dei contenuti illeciti noti e nel raggruppamento dei campioni in famiglie. Data la loro natura probabilistica, in giudizio una corrispondenza di similarità vale come indizio da verificare e non come identificazione.

Scenario didattico

Il contratto «originale» prodotto in giudizio ha un hash SHA-256 diverso da quello della copia sequestrata, come atteso per un file salvato di nuovo. ssdeep, applicato al testo estratto dai due file, restituisce un punteggio di somiglianza di 96 su 100, cioè contenuti in larga parte comuni con modifiche limitate, e il confronto puntuale delle differenze individua la clausola alterata.

Limiti

Ogni schema ha punti deboli noti, come i file molto piccoli e i formati compressi, e soglie da calibrare, perché un punteggio di somiglianza non ha un significato assoluto e dipende dall'algoritmo e dal tipo di file. La relazione indica algoritmo, versione, soglia e tasso atteso di falsi riscontri, come per la spiegabilità (approfondimento 3).

Rilevamento della manipolazione di immagini

Prima di stabilire se un'immagine è generata da un modello (Modulo 10), l'analisi forense classica verifica se è stata manipolata, con tecniche consolidate. L'**Error Level Analysis** (ELA) evidenzia le regioni con una storia di compressione diversa dal resto dell'immagine, come accade quando un oggetto viene incollato. L'analisi copy-move individua porzioni duplicate all'interno della stessa immagine, mentre le incoerenze di rumore, illuminazione e prospettiva segnalano il fotomontaggio. Il **PRNU** (Photo Response Non-Uniformity), cioè l'impronta di rumore propria di ogni sensore, serve ad associare una foto a una specifica fotocamera o a rilevare porzioni non compatibili con quel sensore. I modelli di machine learning integrano questi segnali e li applicano a grandi raccolte; l'esperto deve saper spiegare in udienza i principi fisici su cui si basano.

Scenario didattico

In una controversia assicurativa la foto del danno presenta, nella zona del parafrangente, valori ELA anomali e un rumore incompatibile con il resto dell'immagine; l'analisi copy-move individua la stessa porzione di carrozzeria duplicata e ruotata. La manipolazione è sostenuta da tre segnali indipendenti, e la relazione riporta le mappe di analisi oltre alla conclusione.

Limiti

Ogni tecnica ha falsi positivi noti: l'ELA segnala anche salvataggi multipli legittimi e le piattaforme social ricomprimono le immagini. Il confronto PRNU richiede la fotocamera o un numero adeguato di immagini di riferimento. Una conclusione richiede la convergenza di più segnali e la dichiarazione dei tassi di errore.

Doppio binario di scoperta e validazione

L'integrazione dell'AI nel flusso di laboratorio si organizza su un **doppio binario**. Il binario di scoperta comprende triage, ricerca semantica, hash di similarità e classificatori: strumenti rapidi e fallibili, i cui risultati sono piste. Il binario di validazione verifica ogni pista promossa sul dato originale con metodo tradizionale, cioè con l'esame dell'artefatto, il ricalcolo dell'hash e la ricostruzione del contesto, e alimenta la relazione finale. Tra i due binari si tiene un **registro**, con cui il consulente spiega come ha selezionato il materiale esaminato: strumenti AI usati con versione e parametri, dati su cui sono stati applicati, piste promosse o scartate con la relativa motivazione.

Scenario didattico

Si supponga un caso da 6 TB: il binario di scoperta riduce il materiale a 300 file candidati in una giornata, combinando triage, ricerca semantica e similarità. Il binario di validazione li esamina in una settimana e ne porta 41 in relazione. Il registro intermedio documenta il passaggio e può essere esaminato dal consulente di parte.

Cautele

Il materiale non promosso dalla fase di scoperta resta non esaminato, e i limiti del triage (approfondimento 1) riducono il perimetro effettivamente esaminato. Le contromisure sono il campionamento casuale del materiale scartato, soglie prudenti e l'indicazione esplicita di tale perimetro in relazione.

Confronto tra triage manuale e assistito

Svolgimento su un caso didattico di migliaia di file

1. Metà dei gruppi: triage manuale per categorie e parole chiave
2. Metà dei gruppi: classificazione automatica e ricerca semantica con modello locale
3. Misura comparata di tempi e richiamo sugli elementi rilevanti predefiniti
4. Stesura della sezione metodologica della relazione, con l'uso dell'AI dichiarato

Consegna

- Tabella tempi e richiamo per approccio
- Elementi mancati e relative cause
- Sezione metodologica difendibile in contraddittorio

Lecture e autovalutazione

Lecture obbligatorie

- Casey, *Digital Evidence and Computer Crime*, capitoli metodologici
- *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993)
- ENFSI, *Best Practice Manual for the Forensic Examination of Digital Technology*, 2015

Domande di autovalutazione

1. Quali criteri Daubert sono critici per un classificatore ML?
2. Perché il richiamo conta più della precisione nel triage probatorio?
3. Come si documenta l'uso di un modello nella relazione tecnica?
4. Quali bias può introdurre un classificatore addestrato su dati non rappresentativi?
5. In quali fasi del processo forense l'AI è oggi più matura?

FINE DELLA LEZIONE • MODULO 08

Grazie

Digital Forensics assistita dall'AI • Corso «Incident Response potenziata dall'AI»

Domande e materiali: ricevimento del docente. Profilo e contatti:
vincenzocalabro.it



Vincenzo Calabrò

CYBER SECURITY • DIGITAL FORENSICS



VINCENZOCALABRO.IT