Case Studies

AI w Wykrywaniu Włamań i Reagowaniu na Incydenty

Warsaw School of Computer Science

Isolation Forest

01

3:14 w nocy w banku

4,7 GB transferu do Korei Północnej.
Anomaly score: 0.94.

CASE STUDY

Wyobraź sobie że jesteś analitykiem SOC. 15 minut do końca nocnej zmiany. Na dashboardzie odpala się alert - starszy analityk ryzyka właśnie przesłał 4,7 GB danych. O 3:14. W kierunku IP zarejestrowanego w Korei Północnej. Zanim w ogóle zobaczyłeś ten alert, Isolation Forest zdążył go oznaczyć score 0.94. Dlaczego? Normalna sesja: 08–18, max 10 MB, cel Warszawa albo Frankfurt. Ten punkt w przestrzeni cech leży sam - izoluje się w 3 cięciach drzewa zamiast 30. Pytanie: co jeśli analityk właśnie dostał awans i nowe uprawnienia oraz kto aktualizuje co jest 'normalne'?

PYTANIE DO DYSKUSJI

Jak często model powinien 'przeuczać' nowe normalny zachowania?

LSTM

02

Sekwencja, której nikt nie widział

Login OK. Email OK. PowerShell. LDAP. DC.
Score: 0.97.

CASE STUDY

Sześć zdarzeń. Każde z osobna - legalna. Ania loguje się z MFA. Otwiera email. Klika link. Uruchamia payload.exe. Z jej konta leci 3800 zapytań do AD. Potem próba połączenia z kontrolerem domeny - score 0.97. Klasyczna sieć neuronowa widzi każde zdarzenie osobno. LSTM widzi ciąg przyczynowy. Login → phishing → payload → rekon → lateral movement. Każde zdarzenie jest coraz mniej prawdopodobne biorąc pod uwagę to, co było chwilę wcześniej. To jest dokładnie ta sama logika, którą GPT używa do przewidywania słów. Tyle że zamiast słów mamy zdarzenia systemowe.

PYTANIA DO DYSKUSJI

Gdzie postawić próg automatycznej izolacji konta — score 0.7 czy 0.97? Od czego to zalezy?

Co jest gorsze zablokowanie konta Ani, czy niezablkowanie zainfekowamego konta?

Alert Fatigue

03

11 000 alertów

11 000 alertów dziennie.
44% ignorowanych.

CASE STUDY

Jeden alert co 7,8 sekundy przez całą dobę. Ośmiu analityków na zmianę. Według Ponemon Institute 44% alertów jest ignorowanych. Nie dlatego że analitycy są niedbali. Dlatego że ludzki mózg nie jest w stanie przetworzyć 11 000 sygnałów i zachować czujności wobec każdego. Tutaj wchodzi AI nie jako superinteligencja, ale jako filtr priorytetyzujący. Sentinel, Splunk AI nie eliminują alertów. One grupują je i wyciągają Top 50, które wymagają ludzkiej decyzji. Pytanie filozoficzne: czy wolisz przepuścić mniej alertów ryzykując że coś umknie — czy przepuścić wszystko ryzykując że analityk się wyłączy?

PYTANIA DO DYSKUSJI

Kto odpowiada jeśli AI odfiltruje alert, który był prawdziwym atakiem?

Jak zapobiec temu że Top 50 z czasem też stało się szumem, który analitycy zaczynają ignorować?

IR end-to-end

04

Ransomware w szpitalu

Operacje odwołane.

CASE STUDY

Szpital w Belgi 2020. Atak ransomware. Systemy padają. Operacje odwoływane. Cofnijmy się 48 godzin. Pielęgniarka pobiera 'aktualizację planowania dyżurów'. IDS nie widzi nic — email poza siecią. SIEM dostaje alert: plik exe. Priority: Low. Nikt nie patrzy. Potem PowerShell, LDAP, lateral movement. Każdy etap był wykrywalny. AI może wykryć każdy krok. Ale jeśli nie ma procesu, ludzi i budżetu na reagowanie — detekcja bez response to tylko lepszy sposób na dokumentowanie własnej porażki.

PYTANIA DO DYSKUSJI

Który etap był najłatwiejszy do zatrzymania i dlaczego nie został zatrzymany?

Jak argumentować zarządowi szpitala za budżetem na SOAR?

Harvest Now, Decrypt Later

05

Twoje maile z 2025 roku

Zbierają dziś. Odszyfrują w 2032.

CASE STUDY

Dziś wysyłasz zaszyfrowany email. Logujesz się przez VPN. Każda z tych operacji opiera się na RSA lub ECC, kryptografii, której bezpieczeństwo wynika z tego, że klasyczny komputer nie może rozłożyć dużej liczby w rozsądnym czasie. Klasyczny komputer: miliony lat. Kwantowy z algorytmem Shora: godziny do dni. Takich komputerów jeszcze nie ma — ale mogą być za 5–10 lat. Ktoś zbiera dziś zaszyfrowany ruch i czeka. Tajemnice z 2025, odszyfrowane w 2032. Nie jest to science fiction. NSA ostrzegała o tym już w 2015.

PYTANIA DO DYSKUSJI

Jakie dane z twojego życia mają shelf-life dłuższy niż 7 lat?

Kto powinien wymuszać migrację do PQC? rząd, firmy tech, czy rynek?

Mythos / Claude

Model, który hakuje i ten, który łata

271 podatności. 181 exploitów.
27-letni błąd w OpenBSD.

06

CASE STUDY

Kwiecień 2026. Anthropic publikuje raport o projekcie Mythos. Model odkrył 271 podatności w Firefoksie, 181 z nich to działające exploity. Jeden błąd miał 27 lat. Kluczowy cytat: 'The same improvements that make the model more effective at patching vulnerabilities also make it more effective at exploiting them.' Im lepszy model do znajdowania luk, tym lepszy do ich używania. Nie ma wersji 'tylko dobra'. Dostęp ograniczono do Microsoft, Google, Apple, AWS. Ale AISLE przetestowało 8 innych modeli, wszystkie 8 wykryły te same exploity. Jeden kosztuje 11 centów za milion tokenów.

PYTANIA DO DYSKUSJI

Czy ograniczenie dostępu do jednego modelu przez jedną firmę cokolwiek zmienia, skoro zdolności są dostepne gdzie indziej?

Jaką odpowiedzialność ma specjalista security, znając tę asymetrię między tempem odkrywania podatności a tempem ich łatania?

UEBA / Insider Threat

Nowy pracownik w piątek po południu

07

Konto z normalnymi uprawnieniami.

CASE STUDY

Michał pracuje w firmie od miesiąca. Piątek, 17:55. Zamiast wychodzić, zaciąga 12 000 rekordów klientów z CRM-a. Każde zapytanie było poprawne. Żadne nie przekroczyło limitów API. Ale UEBA zbudowało profil Michała przez 30 dni: zwykle 20 - 50 rekordów dziennie, głównie dla klientów z województwa mazowieckiego, zawsze między 9:00 a 17:30. Peer group Michała, inni juniorzy - nigdy nie odpytywali więcej niż 200 rekordów. Jego score wykrył odchylenie od peer group, nie od reguły. Reguły były zgodne. Zachowanie nie.

PYTANIA DO DYSKUSJI

Jak długo system powinien budować profil zanim zacznie flagować odchylenia? Tydzień? Miesiąc?

Czy monitorowanie behawioralne pracowników narusza ich prywatność? Gdzie jest granica?

Signature vs. Anomaly IDS

08

Snajper kontra śrut

Znane ataki: sygnatury wygrywają.

CASE STUDY

Wyobraź sobie dwa systemy ochrony. Pierwszy to snajper — zna cel, celuje precyzyjnie, prawie nigdy nie myli. Ale widzi tylko znane cele. Drugi to strzelec ze śrutem — na cokolwiek nieznanego reaguje alarmem, często na ptaki i cienie. Signature-based IDS to snajper: niski false positive rate dla znanych ataków, ale zero-day jest niewidoczny. Anomaly-based IDS to strzelec ze śrutem: wykrywa nowe wektory, ale generuje fałszywe alarmy. W praktyce używa się obydwu — sygnatury do znanych zagrożeń, anomaly detection jako siatka bezpieczeństwa dla reszty.

PYTANIA DO DYSKUSJI

Ile fałszywych alarmów dziennie jest akceptowalne zanim analitycy zaczną ignorować system?

Czy można zaatakować sam model anomaly detection tak, żeby nauczyć go że złośliwy ruch jest 'normalny'?

SOAR Automation

09

Automat zatrzymał nie tego

SOAR zablokował konto admina.
W środku incydentu.

CASE STUDY

Sobota, 2:00 w nocy. Atak ransomware w trakcie. SOAR playbook automatycznie izoluje host z wysokim score, logika prosta: flag → isolate. Problem: tym hostem był laptop administratora systemu, który właśnie próbował ręcznie zatrzymać propagację wirusa. Po izolacji, brak dostępu do konsoli zarządzania. Trzy kluczowe serwery zaszyfrowane w ciągu 8 minut. SOAR działał zgodnie z regułami. Reguły były złe. To jest kluczowe pytanie architektury: co powinno być w pełni automatyczne, co wymaga zatwierdzenia człowieka, a co nigdy nie może być zautomatyzowane?

PYTANIA DO DYSKUSJI

- Jak projektować playbooki żeby uniknąć takich kolizji?
- Kto odpowiada za szkodę wyrządzoną przez poprawnie działający automat?

AI-Assisted Red Teaming

10

Pentester z GPT

Okno eksploitacji skróciło się do kilku godzin.

CASE STUDY

Red teamer wynajęty przez firmę finansową. Dawniej: faza rekonesansu zajmowała tydzień. Teraz: prompt do modelu, 'znajdź publiczne informacje o tej domenie', 'zaproponuj wektor phishingowy dla dyrektora finansowego bazując na jego LinkedIn', 'napisz email w jego stylu'. Całość w 3 godziny. Model nie zastąpił pentestera, ale zwielokrotnił jego możliwości. Ten sam efekt po drugiej stronie: atakujący z dostępem do tych samych narzędzi. Okno między odkryciem podatności a jej wykorzystaniem skróciło się dramatycznie. Patch management nie nadąża.

PYTANIA DO DYSKUSJI

Jak zmienia się wartość certyfikatów takich jak OSCP, jeśli AI pomaga w każdym etapie testu penetracyjnego?

Czy firmy powinny ujawniać klientom, że używają AI w testach bezpieczeństwa?