AI in Cybersecurity

Guide to AI and Data Courses · Part VI: Systems and Applications

Intrusion detection | Malware | Imbalanced data | Adversarial attacks | Security of AI | Neighbouring courses

Sercan KÜLCÜ · Giresun University · Computer Engineering

© Sercan Külcü. All rights reserved. · Interactive HTML version of these slides: https://sercankulcu.github.io/files/ai_courses/slides/Chapter_38_AI_in_Cybersecurity.html

INTRODUCTION

What Question Does It Ask?

> **How does AI catch attacks, and how is it attacked itself?**

AI in cybersecurity asks two questions together: how can machine learning catch intrusions, malware and phishing, and how can AI systems themselves be fooled? What sets the field apart is the **adversarial setting**: on the other side is an attacker trying to evade the model.

At a glance

APPROACH security knowledge plus ML	**INPUT** network traffic, logs, files	**OUTPUT** alert, class, risk score
SUCCESS MEASURE detection rate, false alarm rate	**MATH LOAD** medium	**DATA NEEDS** imbalanced and constantly changing

A Short History

1987
Denning defines intrusion detection as a model that catches abnormal behaviour.

1999
The KDD Cup 1999 intrusion detection dataset is released; it serves as a benchmark for many years.

2002
Paul Graham popularizes the Bayesian spam filter with "A Plan for Spam".

2014
Szegedy et al. show adversarial examples: an invisible change fools a classifier.

2016
DARPA Cyber Grand Challenge: machines find and patch vulnerabilities with no human help.

2023
OWASP publishes its first Top 10 for LLM applications, led by prompt injection.

THE COURSE

Core Concepts

Intrusion detection
Signature-based and anomaly-based approaches.

Malware
Classifying files by static or behavioural features.

Phishing detection
Telling fake emails, domains and web pages apart.

Imbalanced data
Attacks are very rare next to normal traffic.

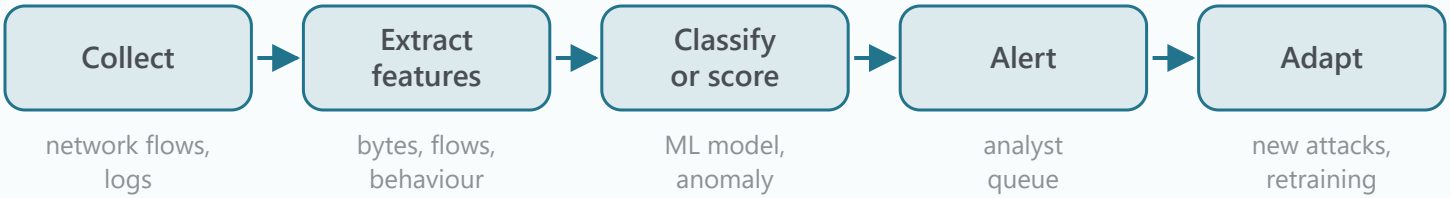
False alarms
Precision, recall; analyst fatigue.

Adversarial examples
Inputs carefully perturbed to fool a model.

Data poisoning
Slipping malicious samples into training data.

Prompt injection
Hidden instructions that change a language model's goal.

A Typical Workflow



Example: An organization's gateway collects connection logs; features such as duration, byte count and port are extracted for each connection. A model scores suspicious flows and only the highest-scoring ones go to an analyst. The analyst's decisions become new labels and the model is updated.

THE COURSE

What Does the Course Cover?

1. Threat models and the limits of ML in security
2. Feature extraction from network traffic and logs
3. Signature- and anomaly-based intrusion detection
4. Imbalanced data, false alarms and suitable metrics
5. Malware: static and behavioural analysis
6. Phishing, spam and malicious URL detection
7. Fraud and account takeover detection
8. Concept drift and attacker adaptation
9. Adversarial examples and evasion attacks
10. Data poisoning and backdoors
11. Model stealing, membership inference, privacy
12. Language model security: prompt injection

Skills you will gain

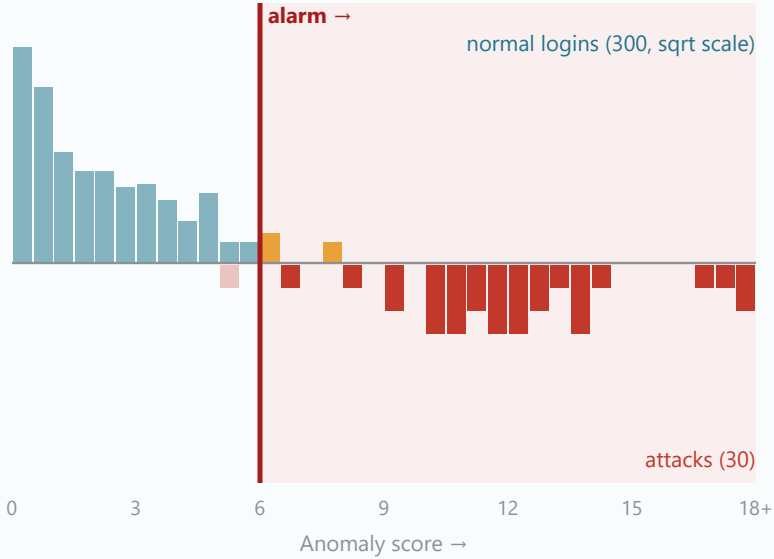
- reading traffic with Wireshark and Zeek
- an attack classifier in scikit-learn
- writing YARA rules
- choosing metrics for imbalanced data
- adversarial attacks with ART

A typical semester; the order and coverage vary between universities.

TRY IT

Catch Suspicious Logins

Password-guessing attacks are mixed in with normal logins in the company's login records. The model gives each login an anomaly score. Move the threshold or click the chart; what happens when the attacker adapts?



Alarm threshold = **6.0**

Attacker adapts

Attacks caught: **29 / 30** (97%)
False alarms: **3** normal logins (1.0%)
The threshold is well balanced.

The score is how surprising a login is compared with normal behaviour: failed attempts, hour, new device, foreign country. The model learned from normal logins only.

AI in cybersecurity **learns what is normal and catches deviations**, so it also sees attacks with no known signature. But the opponent learns too: the model and its features must be updated continuously.

CONTEXT

Prerequisites

Courses worth knowing before this one, and why they are needed.

Course	Why it is needed
Machine Learning	intrusion detection, malware classification
Operating Systems	processes, file systems and permissions: the layer attacks target
Computer Networks	TCP/IP, ports and packets: being able to read network traffic
Programming with C	memory and pointers: seeing how overflow vulnerabilities arise

CONTEXT

Are You Ready for This Course?

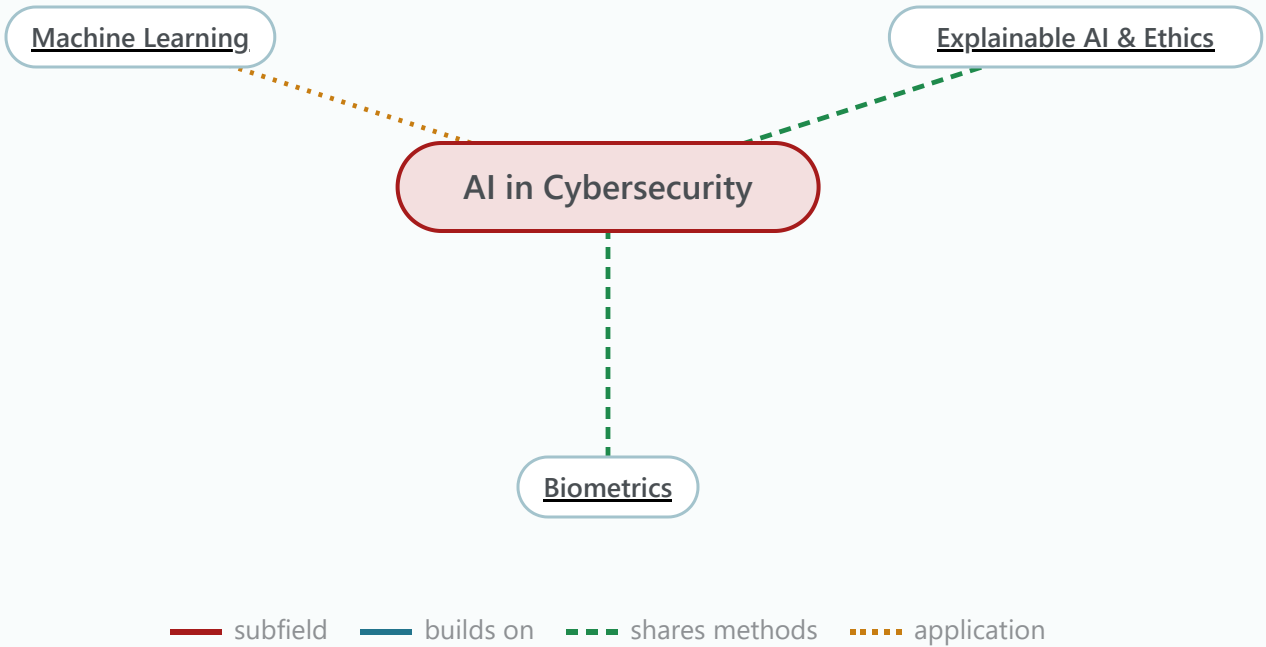
Ask yourself; for every "no", the right-hand column shows where to start.

Ask yourself	If not, start with
Can you interpret precision and recall on an imbalanced dataset?	Machine Learning
Can you explain how a TCP connection is set up and what a port is?	TCP/IP from a computer networks course
Can you inspect processes, file permissions and system logs on Linux?	An operating systems course and the Linux shell
Can you explain how a buffer overflow happens in memory?	Pointers and arrays from a C programming course

If you can say "yes" to most of these, you are ready; one or two gaps can be closed during the semester.

CONTEXT

Neighbouring Courses



Overlapping Courses

Relation	Course	Shared ground
application	This course applies the methods of Machine Learning	intrusion detection, malware classification
shares methods	Shares methods with Explainable AI and AI Ethics	robustness, adversarial examples
shares methods	Shares methods with Biometric Systems	authentication, spoof detection

DIFFERENCES

Where It Differs (1/2)

Criterion	AI in Cybersecurity	Machine Learning	Explainable AI & Ethics
Core question	How does AI catch attacks, and how is it attacked itself?	How does a program generalize from examples?	When can we trust an AI system's decision, and who is responsible for the outcome?
Approach	security knowledge plus ML	statistical learning	human-centered evaluation
Input	network traffic, logs, files	labeled or unlabeled examples	model, its data and context of use
Output	alert, class, risk score	a model that makes predictions	explanation, audit report, principles
Success measure	detection rate, false alarm rate	error on unseen data	interpretability, fairness, accountability
Math load	medium	high: linear algebra, probability, optimization	low-medium
Data needs	imbalanced and constantly changing	moderate to large	the model's own data

- **Machine Learning:** ML assumes the data distribution stays fixed; here the other side deliberately changes the input to evade the model.
- **Explainable AI & Ethics:** Explainable AI treats trustworthiness in general; this course focuses on robustness against attackers and on catching attacks.

DIFFERENCES

Where It Differs (2/2)

Criterion	AI in Cybersecurity	Biometrics	Data Mining
Core question	How does AI catch attacks, and how is it attacked itself?	How is a person recognized from their face, fingerprint or voice?	How are hidden, useful patterns found in large piles of data?
Approach	security knowledge plus ML	matching and verification	discovery-driven analysis
Input	network traffic, logs, files	face, fingerprint, iris, voice	large tables or transaction data
Output	alert, class, risk score	identity decision	rules, clusters, outliers
Success measure	detection rate, false alarm rate	FAR, FRR, EER	interestingness, support, confidence
Math load	medium	medium	medium
Data needs	imbalanced and constantly changing	few samples per person	large

- **Biometrics:** Biometrics verifies identity by recognizing a person; this course covers a wide attack surface, from networks and files to email and models.
- **Data Mining:** Data mining looks for unusual patterns to discover them; this course interprets them as attacks and turns them into alerts.

PRACTICE

Applications

Intrusion detection
Finding unusual connections in network traffic.

Malware
Judging whether a file is malicious without running it.

Phishing and spam
Filtering out fake emails and web pages.

Fraud
Detecting suspicious transactions and account takeover.

Security operations
Prioritizing and triaging thousands of alerts.

Protecting AI
Adversarial examples, poisoning, prompt injection.

PRACTICE

Risks and Ethics

Dual use

Techniques learned for detection can also produce malware that evades it, or convincing phishing emails.

Unauthorized testing

Scanning another party's network without permission is a crime; experiments belong in an isolated lab, with written consent.

Employee surveillance

Insider threat detection watches employees' email and behaviour; proportionality, GDPR and similar laws set limits.

Automated response

Cutting a connection or locking an account on the model's word alone stops legitimate users on a false alarm.

In depth: Explainable AI and AI Ethics.

PRACTICE

Tools and Datasets

Tools and libraries

scikit-learn · Zeek · Suricata · Wireshark · YARA · Elastic Stack · Adversarial Robustness Toolbox

Well-known datasets

KDD Cup 1999 / NSL-KDD · CICIDS2017 · UNSW-NB15 · EMBER (malware) · Enron-Spam · PhishTank

Old datasets (KDD Cup 1999) are still widely used but do not reflect today's traffic; results should also be tested on recent datasets.

PRACTICE

Common Misconceptions

Common belief	In fact
99% accuracy means good intrusion detection.	If attacks are very rare, a model that calls everything normal is also 99% accurate; look at precision and false alarm rate.
Once trained, a model keeps catching attacks.	Attackers change tactics; the model must be updated continuously with new samples.
AI replaces the security analyst.	AI filters and prioritizes alerts; decisions and investigation still rest with people.
Deep networks are robust to attacks.	Small, invisible changes (adversarial examples) can easily flip a model's decision.

NEXT STEPS

After This Course

Courses that build on or extend this one

This course is mostly an end point: it builds on the others.

Related careers

Security analyst (SOC) | Threat hunter | Malware analyst | Security data scientist | AI security specialist

This course opens in two directions: on the security side, network security, digital forensics and malware analysis; on the AI side, adversarial robustness, privacy-preserving learning and LLM security. Explainable AI and AI Ethics is the natural neighbour on the second path.

NEXT STEPS

Further Reading

- Chio and Freeman, *Machine Learning and Security*, O'Reilly, 2018.
- Denning, "An Intrusion-Detection Model", *IEEE Transactions on Software Engineering* 13(2), 1987.
- Sommer and Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection", *IEEE Symposium on Security and Privacy*, 2010.
- Goodfellow, Shlens and Szegedy, "Explaining and Harnessing Adversarial Examples", *ICLR*, 2015.

The End

Chapter 38: AI in Cybersecurity

← Previous chapter · Map of the field · Next chapter →