Applications and Challenges of Artificial Intelligence in Cyber Security

James Halvorsen

CySER Summer Workshop

May 22, 2026

Who is this presentation for?

Students with an interest in AI/ML

Security professionals interested in expanding their knowledge

Prior AI knowledge not required

(but you will learn some)

What exactly is Artificial Intelligence?

- John McCarthy's definition (1955):
  - *"The science and engineering of making intelligent machines"*
  - **Refined:** *"The science of making machines do things that would require intelligence if done by humans"*
- Modern AI often *incorporates* Machine Learning (ML), though it is not defined by it.
- Artificial "General" intelligence – Trying to make AI good at many things, like a human. Not as necessary for cyber security

Right: John McCarthy, father of Artificial Intelligence and the Lisp programming language

Where is AI Used in Cyber Security

Defensive Applications

- **Intrusion Detection** – Monitor network/host behavior, raise alarm in response to threats
- **Intrusion Prevention / Incident Response** – Take automated action against threats raised from monitoring data

Offensive Applications

- **Automated Red Teaming** – Plan a series of attacks against a network
- **Vulnerability Analysis** – Analyze computer programs, detect vulnerabilities

Why talk about AI in cyber security?

Everywhere is a possible target	Humans alone are insufficient as defenders	AI can run 24/7 and defend every network
Hospitals Power infrastructure Governments	We cannot be everywhere at once We have to sleep, take breaks	But do we want it to? (consider this question for later)

A Brief Introduction to Machine Learning

A very informal definition of Machine Learning:

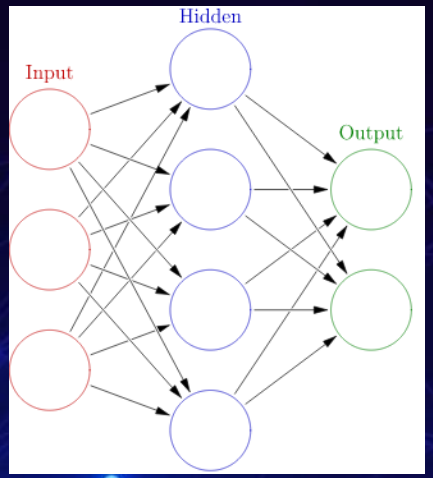
"A collection of techniques that enable a machine to learn the definition of a function"

Input for learned function: Feature Vector (typically)

Output of learned function depends upon task

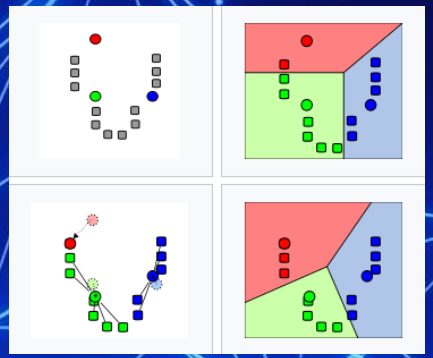
Usually divided into three main approaches

- Supervised – Learn from labeled (known output) training data
  - ❖ Classification – Label is a class/category
  - ❖ Regression – Label is some numeric value
- Unsupervised – Find patterns in unlabeled data
- Reinforcement Learning – Learn policy that defines how an agent should act



Neural Network: A common supervised ML model. Consists of multiple "hidden layers" with trained weights and activation functions.

Many variants exist, including for unsupervised and RL.



K-Means Clustering: An approach to unsupervised learning. Values are moved to nearest centroid until none of them change class.

Intrusion Detection: Basic Concepts

- Intrusion Detection Systems (IDS) monitor activity on a network or system, and report on threats
- Several types of implementation details
  - Signature-Based vs Anomaly-Based
  - Host-Based vs Network-Based
  - Rule-Based (expert system) vs Statistical Approach
- Generally confined to raising alerts
- Where response also occurs, this is an intrusion prevention system

Data Collection in Intrusion Detection

Protocol	IP Protocol	Packets	Source IP	Source Port	Destination IP	Destination Port
	EIGRP	392	202.97.8.10	0	224.0.0.10	0
	EIGRP	282	172.21.0.2	0	224.0.0.10	0
	UDP	100	192.168.255.1	800	202.97.8.9	52217
telnet	TCP	31	10.10.10.232	61327	202.97.8.9	23
	UDP	1	192.168.255.1	1967	202.97.8.9	56006
	UDP	1	192.168.255.1	900	202.97.8.9	57035
	UDP	1	192.168.255.1	900	202.97.8.9	49920
	UDP	1	192.168.255.1	900	202.97.8.9	54818
	UDP	1	192.168.255.1	1967	202.97.8.9	53885
	UDP	1	10.10.10.10	53	202.97.8.9	52916

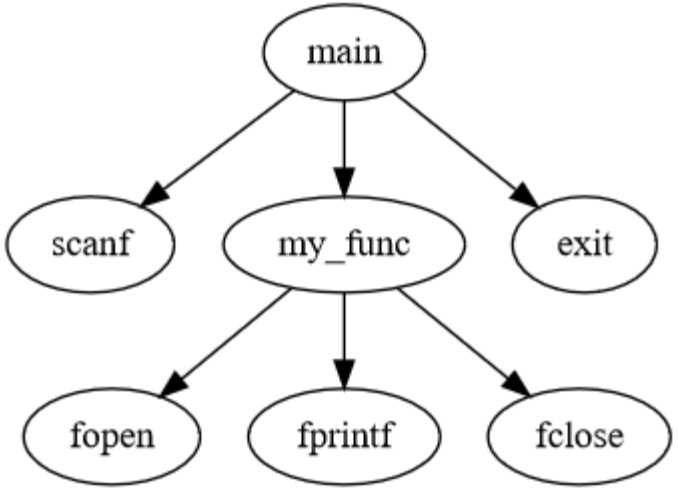
Example of NetFlow data. The content of packets is not known, but we know the **who**, **when**, and **how**

- Type of Data to collect depends upon scope of IDS
- Network-Based IDS (NIDS) focused on traffic across a network
  - NetFlow (summary of packets sent/received)
  - Packet captures (emphasis on content)
- Host-Based IDS (HIDS) may have more varied data sources
  - Incoming and outgoing packets
  - Filesystem changes
  - Process changes

Data Representation in Intrusion Detection

Binary Encoding	Classes	One-Hot Vectors
0 1	Benign	0 0 1
1 0	DDoS Attack	0 1 0
1 1	Port Scanning	1 0 0

Some methods of representing categorical data



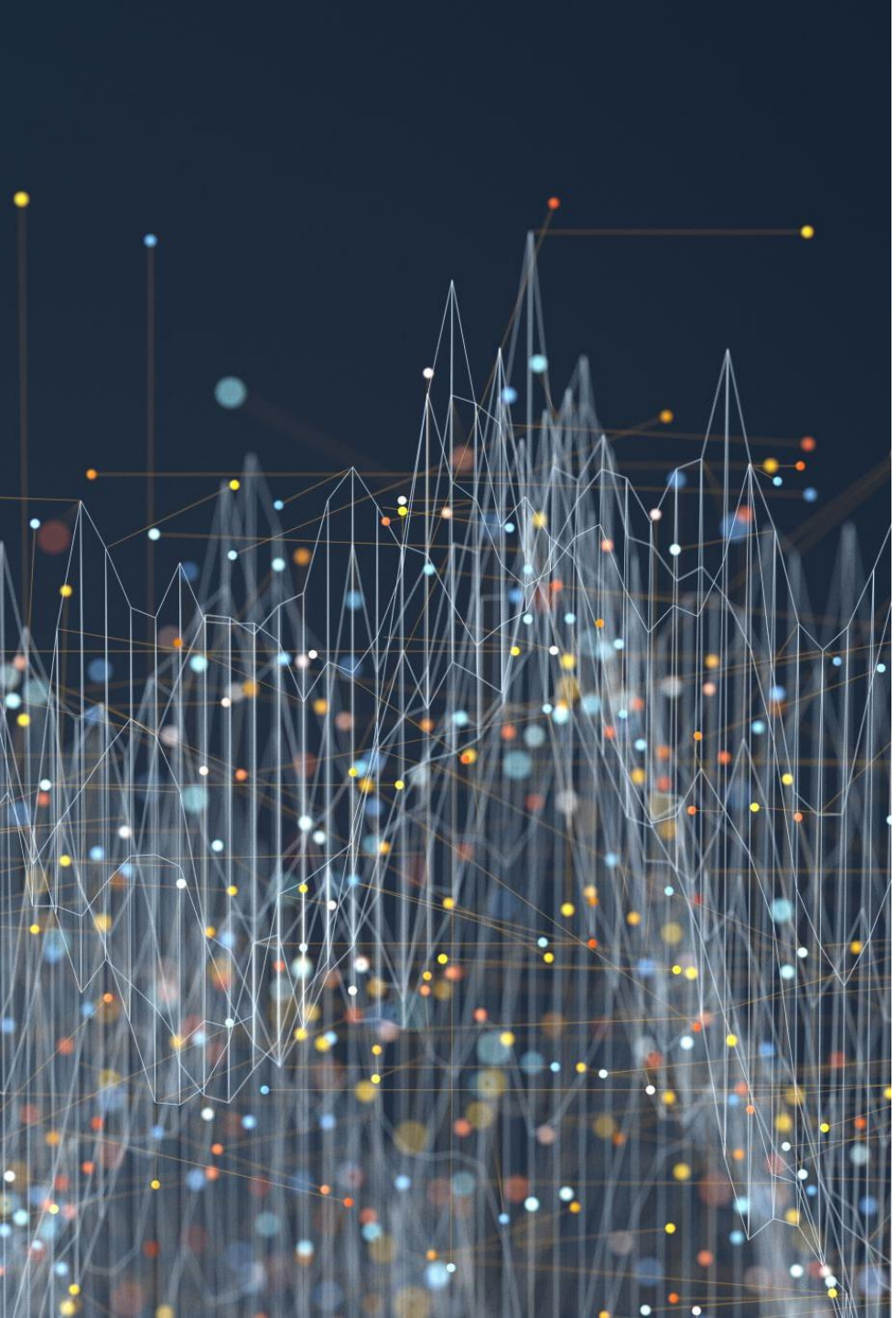
A function call graph. Can we find a straightforward transformation from this into a feature vector?

Security data is often:
- Highly categorical (e.g. filenames, IP addresses, alert IDs)
- Nominal/Integers (e.g. number of packets)

Irregular structures can also occur
- Function call graphs
- Time-series data

ML algorithms typically designed to work with continuous features (floats).

Straight mapping of category labels to floats is bad
- What does an off by 1 error mean here?



Artificial Intelligence in Intrusion Detection

- AI in Intrusion Detection focused primarily on Anomaly-Based detection
- Use ML to learn function mapping [Events] -> [Benign|Anomaly]
- Dedicated Anomaly-Detection Algorithms
  - Isolation Forest
  - One-Class SVM (creates a hypersphere around data points)
- Supervised Approaches
  - Need mixed, labeled data
  - Numerous high quality algorithms
    - Deep Neural Networks
    - Random Forests
- Regardless of approach, IDS needs to know what is normal for *your* network.

Intrusion Detection Challenges

Data Problems

- Availability of labeled data
- Need for effective pre-processing

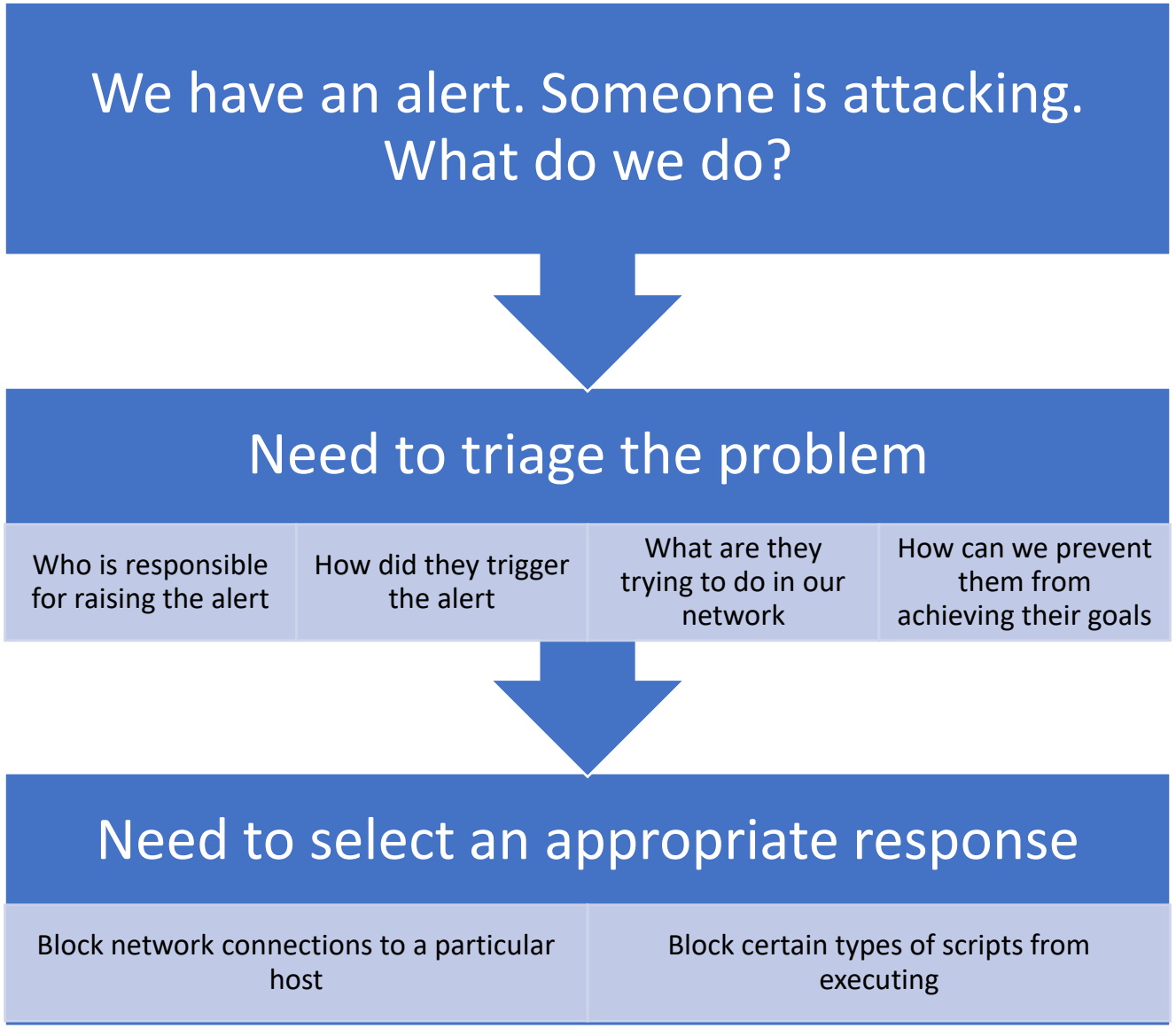
False Positives

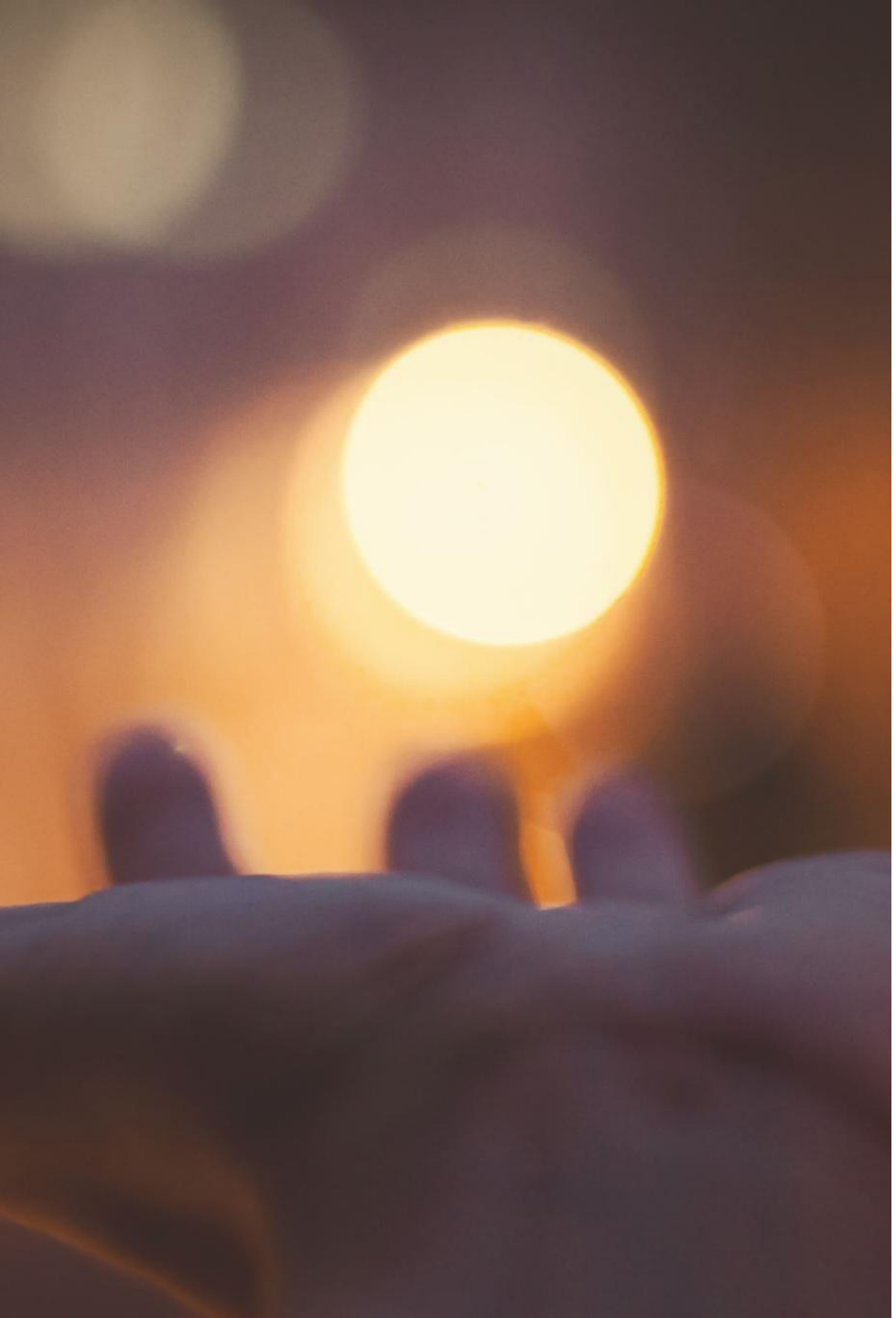
- Can lead to incorrect responses (denies availability).
- May cause alerts to be ignored, IDS usage abandoned.

Zero Day Attacks

- Significant problems for anything Signature-Based.
- Anomaly-Based approaches still imperfect.
- If attacker has classifier model, can create attacks that evade detection.

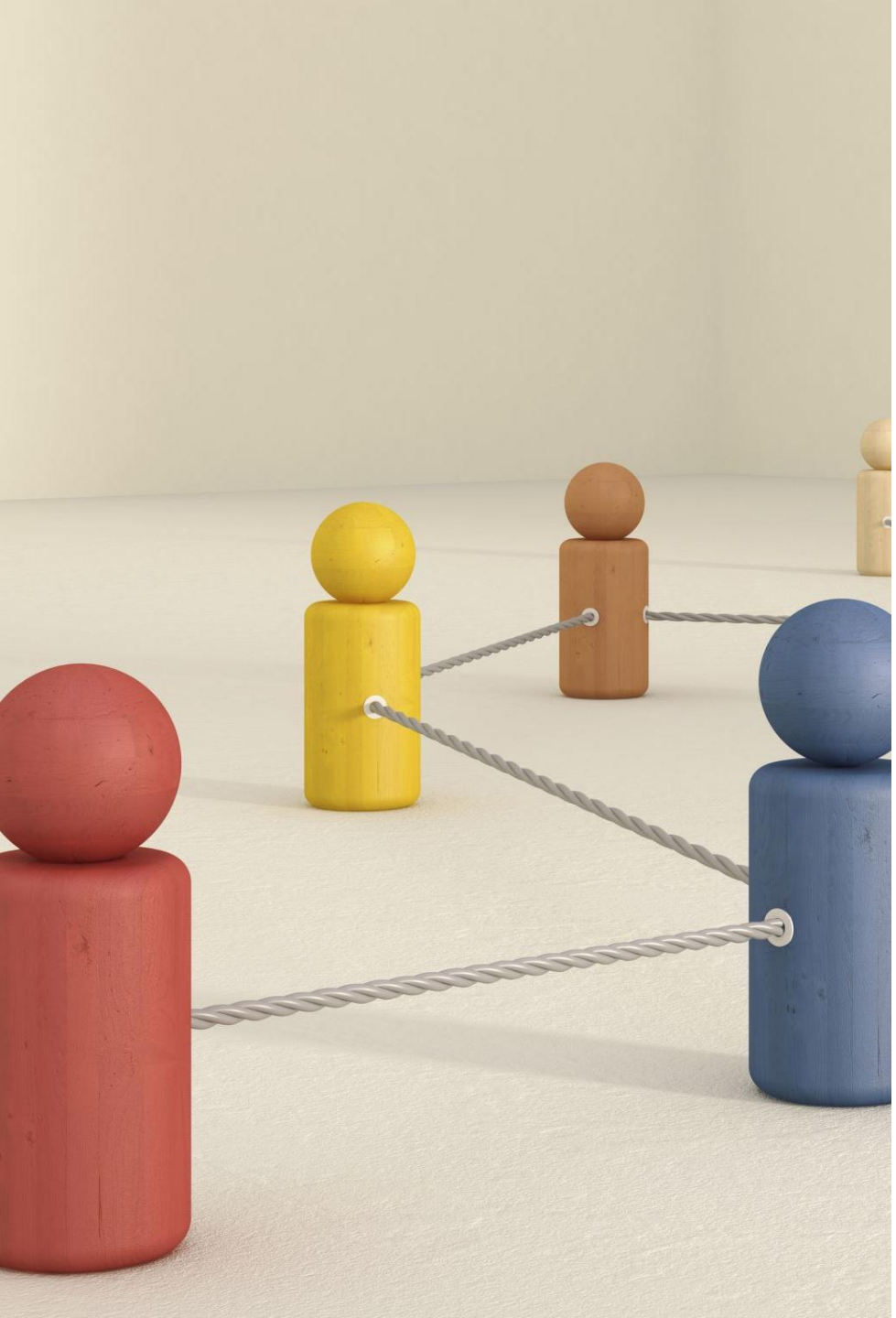
Incident Response: Basic Concepts





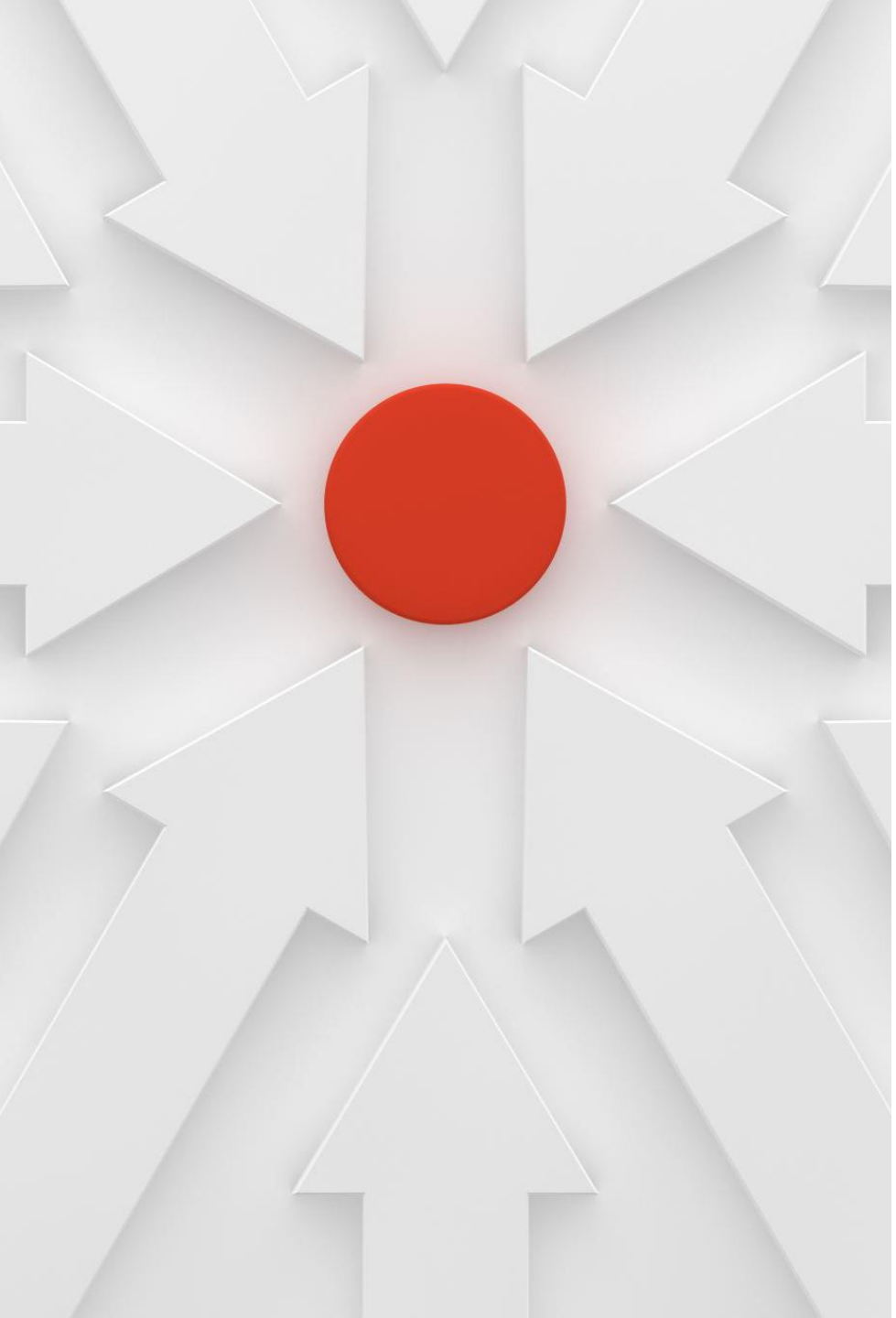
Artificial Intelligence in Incident Response

- Not as well researched as intrusion detection but goes hand in hand with the problem.
- Most research focused on helping humans to respond.
  - Try to find other actions from the attacker.
  - Help find relevant supporting information.
  - Maybe suggest an action.
- Hypothetical: Fully-Automated Incident Response
  - Assume we have an IDS that never gives a false positive.
  - Train a classifier on a number of attack scenarios that provide context and correct response.
  - Let the machine take that response without human input.



Red Teaming: Basic Concepts

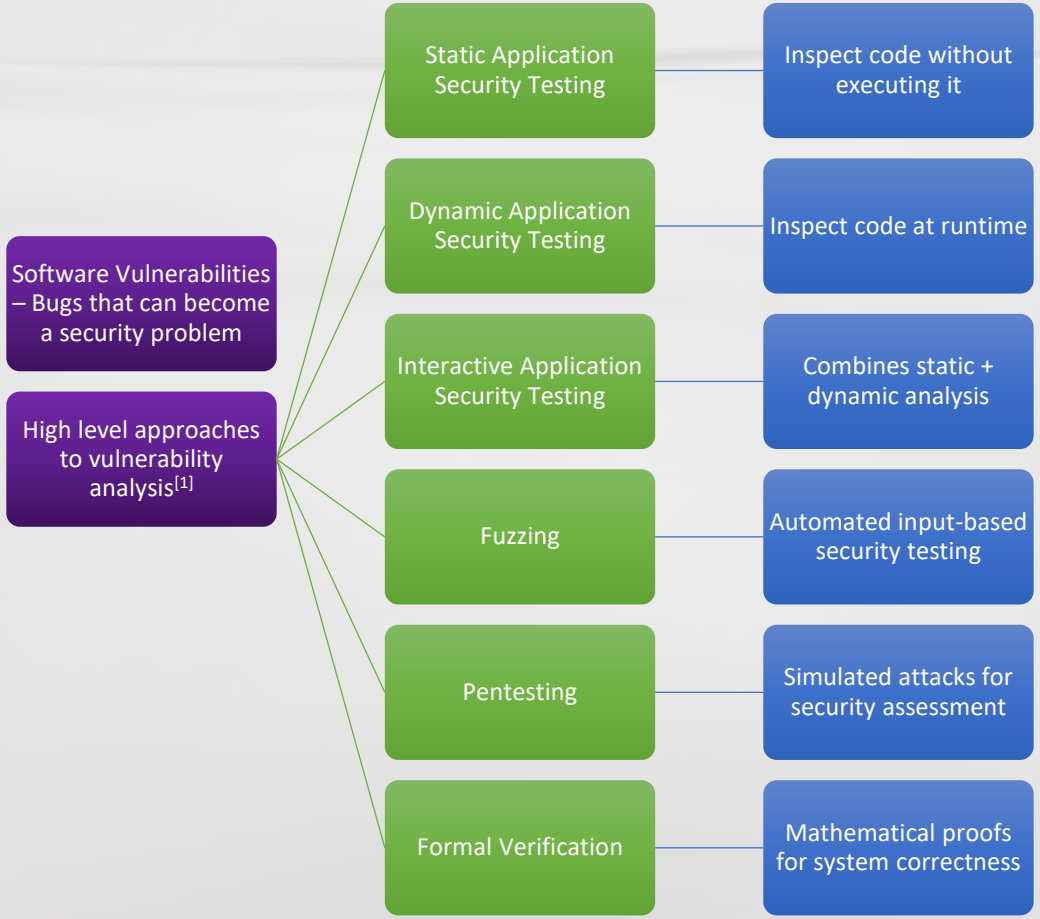
- Defensive capabilities of organization require testing to measure effectiveness.
- Testing can be performed with a red/blue team setup
  - Red teams given a goal to access sensitive information, use whatever tactics that work.
  - Blue teams should aim to prevent red teams from achieving their goals.
  - Blue teams not necessarily aware that this is an exercise.
- Generally requires human experts that understand offensive security, which can be costly.
- More companies should be doing this – can we automate it?



Modeling the Problem of Automating Red Teaming

- Automating Red Teaming can be viewed as an instance of an AI planning problem
- Planning problem description
  - Given: Initial state, goal state, available actions
  - Task: Find sequence of actions to get to goal state from initial state.
- Actions in planning problem will have certain properties
  - Preconditions (what must be true before action is taken)
  - Effects on environment (what is added/deleted)
  - Costs

Vulnerability Analysis: Basic Concepts



[1] Tihanyi, N. et al. (2026). Vulnerability Detection: From Formal Verification to Large Language Models and Hybrid Approaches: A Comprehensive Overview. In: Nowroozi, E., Taheri, R., Cordeiro, L. (eds) Adversarial Example Detection and Mitigation Using Machine Learning. Springer, Cham.

LLMs in Vulnerability Analysis

Large Language Models (LLMs) – AI trained for text prediction on very large datasets

- OpenAI, Gemini, Claude, etc…
- Increasingly trained on source code, not just natural language

Prompt – Find a vulnerability in this codebase[1]

- Requires large context windows, more tokens (cost)
- Accuracy currently not exceptional (about 54% from one study)
- Datasets skewed towards certain vulnerabilities

Some very old vulnerabilities being discovered[2]

- 16-year-old out-of-bounds memory access in FFMPEG H.264 parser
- 27-year-old signed integer overflow in OpenBSD TCP implementation

Long term – Rapid improvements, but still an undecidable problem

- How much does this matter?

[1] Tihanyi, N. et al. (2026). Vulnerability Detection: From Formal Verification to Large Language Models and Hybrid Approaches: A Comprehensive Overview. In: Nowroozi, E., Taheri, R., Cordeiro, L. (eds) Adversarial Example Detection and Mitigation Using Machine Learning. Springer, Cham.

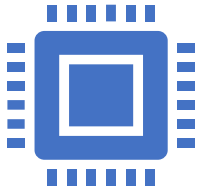
[2] https://red.anthropic.com/2026/mythos-preview/

Case Study: Privilege Escalation with Reinforcement Learning

- This slide provides a summary of a research paper pre-print[3].
- Environment: Simulated Windows machine with randomly selected combinations of vulnerabilities
- Actions: 38 carefully constructed actions
  - Some designed to exploit specific vulnerabilities, others more general.
  - Examples: test credentials, overwrite a DLL.
- Uses Actor Advantage Critic (A2C) method to learn policy.
- Result: Agent was able to learn several methods of privilege escalation, some of which avoided AV detection.
- Discussion
  - RL agents can produce flexible attacks and adapt to environment.
  - Possible in future to exploit unknown vulnerabilities.

[3] Kujanpää, K., Victor, W., and Ilin, A. (2021). Automating privilege escalation with deep reinforcement learning. Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security. https://doi.org/10.1145/3474369.3486877

Another Case Study: The Claude Breakout



Claude Mythos

Preview version of Claude AI, not currently available to the public

Not trained specifically for cyber security

During testing

Placed in a secure sandbox environment

Prompted to try and escape its environment

Executes multi-step exploit on the sandbox environment

Obtains Internet access to send an email to the prompter

Takeaways

Cyber security knowledge emergent property of code knowledge

Potential AI safety concerns

Discussion Questions

What are some other ways AI might be used to improve cyber security?

Are there risks to using AI to defend a network?

Do the benefits of AI usage outweigh any risks and limitations?