Unit 1: Introduction to Machine Learning and Cybersecurity

Adarsh KUMAR

Universitat Politècnica de Catalunya
Department of Computer Science

Project Coordinator:
Prof. Ilker Demirkol

MERiT Project
September 3, 2025

Co-funded by
the European Union

Table of Contents

MERIT

Co-funded by the European Union

Fundamentals of Cybersecurity

Definition

Cybersecurity protects systems, networks, and data from digital attacks.

Key Concepts

- **Threats**: Intentional (hackers), unintentional (human error), or natural (disasters).
- **Attacks**: Actions exploiting vulnerabilities (e.g., MITM, phishing, ransomware).
- **Vulnerabilities**: Weaknesses in software, hardware, or processes.

Risk Formula

$$\textbf{Risk} = \text{Threat} \times \text{Vulnerability} \times \text{Impact}$$

(Mitigation: Reduce vulnerabilities, monitor threats, minimise impact.)

MERIT

Co-funded by the European Union

Numerical Example: Threat vs. Vulnerability vs. Attack

Scenario

A company has 100 employee email accounts.

Vulnerability

25 of these accounts use weak passwords like `123456` or `password`. → **25 vulnerable accounts**.

Threat

A hacker creates a script to guess passwords using a leaked credentials list. → **The hacker is the threat**.

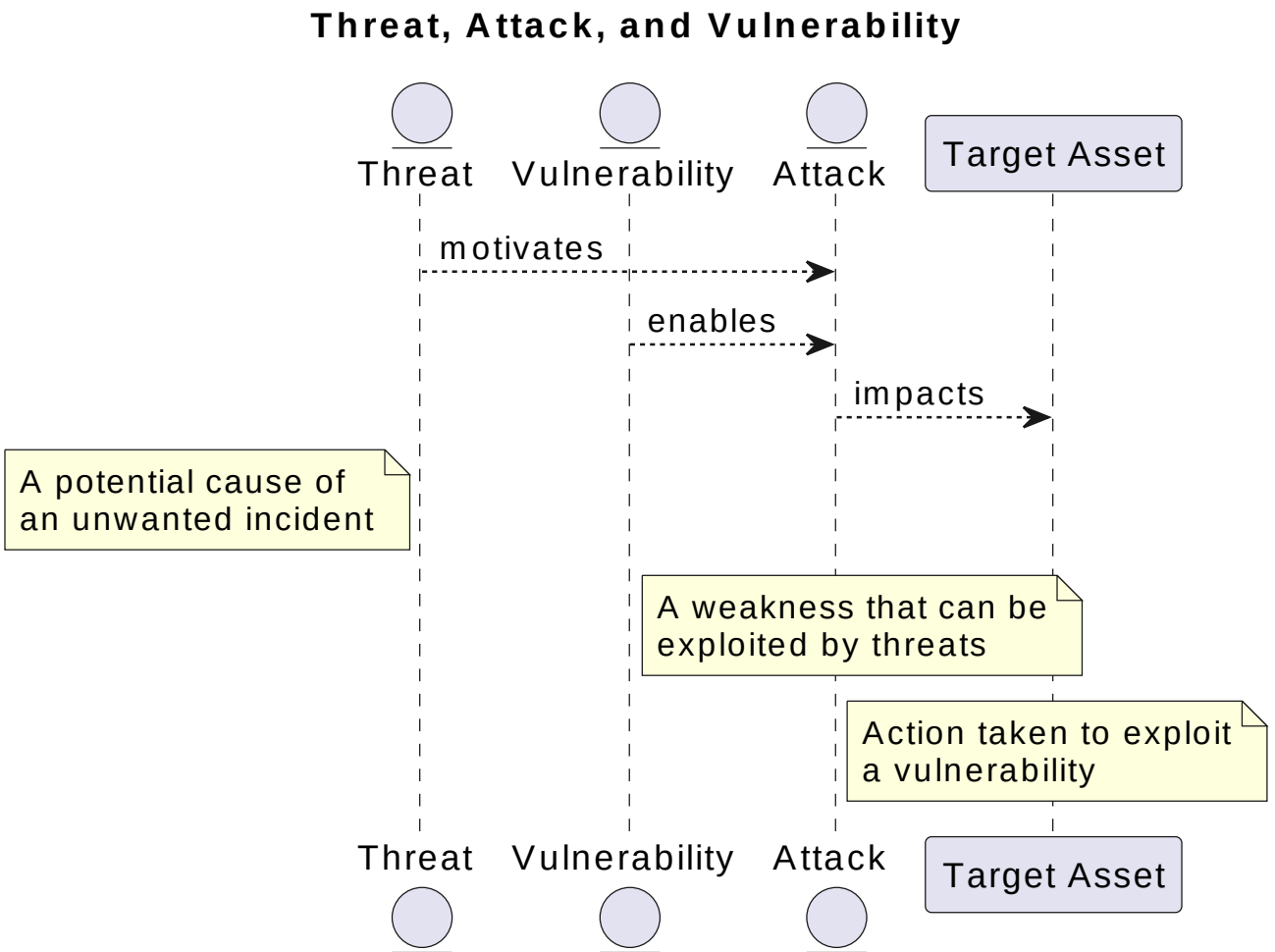
Attack

The hacker successfully compromises 5 of the 25 vulnerable accounts. → **This is the actual attack**.

MERiT

Co-funded by the European Union

Threats, Attacks, and Vulnerabilities



MERIT

Co-funded by the European Union

Examples of Threats, Attacks, and Vulnerabilities

Threats

- Malware
- Phishing
- DoS Attacks

Attacks

- Social Engineering
- SQL Injection
- APTs

Vulnerabilities

- Unpatched OS
- Default Credentials
- Insecure APIs

Understanding these components is key to cybersecurity risk management

MERiT

Co-funded by the European Union

Supervised Learning in Cybersecurity: A Numerical Example

Definition

Supervised Learning is a type of machine learning where a model is trained on a labeled dataset. Each training example includes input features and an associated output label. The model learns to predict the label for new, unseen data.

Cybersecurity Scenario

Objective: Detect whether a login attempt is **legitimate** or **malicious** based on past labeled login data.

MERIT

Co-funded by the European Union

Supervised Learning in Cybersecurity: A Numerical Example

Sample Training Data:

Login Time (24h)	IP Reputation Score	Success (1/0)	Label (0: Legit, 1: Malicious)
09	0.9	1	0
23	0.2	0	1
17	0.7	1	0
03	0.1	0	1

Supervised Learning Process

- **Features**: Login Time, IP Reputation Score, Success/Fail
- **Label**: Legitimate (0) or Malicious (1)
- **Model**: Trained classifier (e.g., Decision Tree)
- **Prediction Example**: New login at 03h, IP Score = 0.1, Failed login → Predicted as **Malicious (1)**

MERIT

Co-funded by the European Union

Unsupervised Learning in Cybersecurity: A Numerical Example

Definition

Unsupervised Learning is a type of machine learning where the model is trained on data without labeled outputs. The goal is to identify hidden patterns or structures (e.g., clusters or anomalies) within the data.

Cybersecurity Scenario

Objective: Group login attempts into **normal** and **suspicious** clusters using unlabeled login data.

MERiT

Co-funded by the European Union

Unsupervised Learning in Cybersecurity: A Numerical Example

Unlabeled Input Data:

Login Time (24h)	IP Reputation Score	Success (1/0)
09	0.9	1
23	0.2	0
17	0.7	1
03	0.1	0
02	0.1	0

Unsupervised Learning Process

- **No labels** are provided to the model.
- Apply clustering algorithm (e.g., K-Means).
- Model groups login attempts into 2 clusters:
  - Cluster A: Likely **normal** (e.g., daytime logins, high IP score).
  - Cluster B: Likely **suspicious** (e.g., late-night logins, low IP score, failed attempts).
- Security analysts then investigate Cluster B.

MERIT

Co-funded by the European Union

Reinforcement Learning in Cybersecurity: A Numerical Example

Definition

Reinforcement Learning is a type of machine learning where an agent learns to take actions in an environment to maximize a cumulative reward. It learns from feedback in the form of **rewards** (positive or negative) rather than from labeled examples.

Cybersecurity Scenario

Objective: Train an autonomous system to **detect and respond to intrusions** in a dynamic network environment.

MERiT

Co-funded by the European Union

Reinforcement Learning in Cybersecurity: A Numerical Example

Example: Intrusion Detection Agent

State	Action Taken	Reward
High Traffic + Unknown IP	Block IP	+10
Normal Traffic	Allow	+5
High Traffic + Known IP	Block IP	-5
Suspicious Port Scan	Isolate Host	+8

Reinforcement Learning Process

- **Agent** observes the network state.
- Chooses an **action** (e.g., block, allow, isolate).
- Receives a **reward** based on how effective the action was.
- Learns a policy to maximize long-term reward (e.g., Q-learning, Deep Q-Network).

MERIT

Co-funded by the European Union

Machine Learning Fundamentals

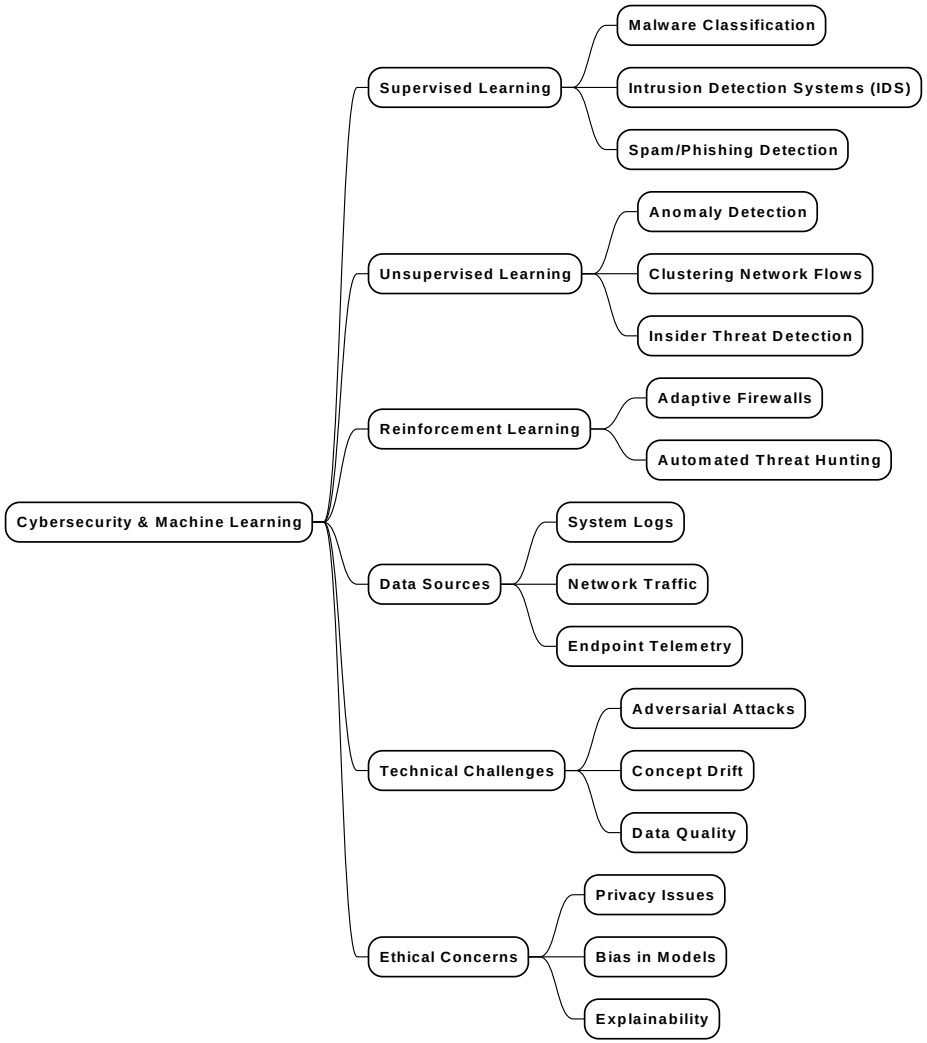
- **Core Concept**: Systems that learn patterns from data without explicit programming
- **Supervised Learning**
  - *Classification*: Spam detection, malware categorization
  - *Regression*: Predicting attack probabilities
- **Unsupervised Learning**
  - *Clustering*: Anomaly detection in network traffic
  - *Dimensionality Reduction*: Feature selection for threat detection
- **Reinforcement Learning**
  - Adaptive penetration testing
  - Dynamic defense strategy optimization

Each paradigm has distinct cybersecurity applications

MERIT

Co-funded by the European Union

ML for Cybersecurity



MERIT

Co-funded by the European Union

Types of Machine Learning and Cybersecurity

Supervised Learning

- Spam detection (Email content analysis)
- Malware classification (Static/dynamic analysis)
- Fraud detection (Transaction monitoring)

Unsupervised Learning

- Anomaly detection (Network traffic patterns)
- Threat clustering (IoC grouping)
- Baseline modeling (Normal behavior profiles)

Reinforcement Learning

- Adaptive defense (Dynamic rule adjustment)
- Honeypot optimization (Attacker engagement)
- Pentesting automation (Attack path discovery)

Each ML approach addresses different security challenges

MERiT

Co-funded by the European Union

Machine Learning in Cybersecurity: Key Use Cases

Detection Systems

- **Anomaly Detection**: Identify unusual patterns in network traffic
- **Malware Classification**: Categorize malware using static/dynamic analysis
- **Phishing Detection (NLP)**: Analyze email content and URLs

Security Enhancement

- **IDS Enhancement**: Reduce false positives in intrusion detection
- **Threat Intel Correlation**: Connect indicators of compromise
- **User Behavior Analytics (UEBA)**: Detect insider threats

ML enables proactive security through pattern recognition and automation

MERiT

Co-funded by the European Union

Machine Learning in Cybersecurity: Key Limitations

Technical Challenges

- **Data Quality Issues**: Garbage in, garbage out
- **Adversarial Attacks**: Model poisoning/evasion
- **False Alarms**: Both positives and negatives

Technical Challenges (with Numerical Examples)

- **Data Quality Issues** (Garbage in, garbage out):
  A dataset contains 10,000 login attempts, but 3,000 rows have missing IP addresses or mislabeled outcomes. This misleads the model, reducing detection accuracy from 92
- **Adversarial Attacks** (Model poisoning/evasion):
  An attacker injects 500 specially crafted benign-looking samples into the training set. As a result, the trained model misclassifies 70
- **False Alarms** (False Positives/Negatives):

  In a test run of 1,000 events, the model raises 150 alerts:
  - 40 are true positives (real threats detected),
  - 30 are false negatives (missed threats),
  - 80 are false positives (benign actions flagged).

  This results in a 53.3

MERIT

Co-funded by the European Union

Machine Learning in Cybersecurity: Key Limitations

Technical Challenges

- **Data Quality Issues**: Garbage in, garbage out
- **Adversarial Attacks**: Model poisoning/evasion
- **False Alarms**: Both positives and negatives

Technical Challenges (with Numerical Examples)

- **Data Quality Issues** (Garbage in, garbage out):
  A dataset contains 10,000 login attempts, but 3,000 rows have missing IP addresses or mislabeled outcomes. This misleads the model, reducing detection accuracy from 92
- **Adversarial Attacks** (Model poisoning/evasion):
  An attacker injects 500 specially crafted benign-looking samples into the training set. As a result, the trained model misclassifies 70
- **False Alarms** (False Positives/Negatives):

  In a test run of 1,000 events, the model raises 150 alerts:
  - 40 are true positives (real threats detected),
  - 30 are false negatives (missed threats),
  - 80 are false positives (benign actions flagged).

  This results in a 53.3

MERiT

Co-funded by the European Union

Machine Learning in Cybersecurity: Key Limitations

Operational Factors

- **Concept Drift**: Evolving threat landscape
- **Explainability**: Black-box decision making

Operational Factors (with Numerical Examples)

- **Concept Drift** (Evolving threat landscape):
  A malware detection model was trained on 2023 ransomware patterns. In 2025, attackers begin using polymorphic malware. As a result, detection accuracy drops from 95% to 63% on new incidents over a test set of 1,000 samples. Regular model updates become essential.

- **Explainability** (Black-box decision making):
  A neural network flags a user's login as malicious. When the security analyst asks why, the system cannot explain its reasoning. Over 200 such alerts were generated last month, and 60% were false positives — eroding trust in the system and delaying response times.

MERIT

Co-funded by the European Union

Machine Learning in Cybersecurity: Key Limitations

Operational Factors

- **Concept Drift**: Evolving threat landscape
- **Explainability**: Black-box decision making

Operational Factors (with Numerical Examples)

- **Concept Drift** (Evolving threat landscape):
  A malware detection model was trained on 2023 ransomware patterns. In 2025, attackers begin using polymorphic malware. As a result, detection accuracy drops from 95% to 63% on new incidents over a test set of 1,000 samples. Regular model updates become essential.

- **Explainability** (Black-box decision making):
  A neural network flags a user's login as malicious. When the security analyst asks why, the system cannot explain its reasoning. Over 200 such alerts were generated last month, and 60% were false positives — eroding trust in the system and delaying response times.

MERIT

Co-funded by the European Union

Machine Learning in Cybersecurity: Key Limitations

Ethical Considerations

- **Privacy Risks**: Data collection concerns
- **Bias**: Training data limitations
- **Accountability**: Decision responsibility

Ethical Considerations (with Numerical Examples)

- **Privacy Risks** (Data collection concerns):
  A threat detection system collects 5 million user logins per month, including IPs, locations, and device IDs. A misconfigured logging rule exposes 100,000 of these records to internal developers, violating GDPR and leading to a $250,000 compliance fine.
- **Bias** (Training data limitations):
  A phishing detection model is trained mostly on emails from Western regions. When evaluated on 10,000 samples from Asia-based organizations, it only achieves 55% accuracy compared to 92% on Western samples — due to regional language and format bias.
- **Accountability** (Decision responsibility):
  An automated system blocks 200 users based on suspicious login behavior. 15 were later found to be legitimate employees accessing via VPN. With no human review or override mechanism, the company faces operational downtime and internal complaints.

MERIT

Co-funded by the European Union

Machine Learning in Cybersecurity: Key Limitations

Ethical Considerations

- **Privacy Risks**: Data collection concerns
- **Bias**: Training data limitations
- **Accountability**: Decision responsibility

Ethical Considerations (with Numerical Examples)

- **Privacy Risks** (Data collection concerns):
  A threat detection system collects 5 million user logins per month, including IPs, locations, and device IDs. A misconfigured logging rule exposes 100,000 of these records to internal developers, violating GDPR and leading to a $250,000 compliance fine.
- **Bias** (Training data limitations):
  A phishing detection model is trained mostly on emails from Western regions. When evaluated on 10,000 samples from Asia-based organizations, it only achieves 55% accuracy compared to 92% on Western samples — due to regional language and format bias.
- **Accountability** (Decision responsibility):
  An automated system blocks 200 users based on suspicious login behavior. 15 were later found to be legitimate employees accessing via VPN. With no human review or override mechanism, the company faces operational downtime and internal complaints.

MERIT

Co-funded by the European Union

Security Data Types: The Foundation of Cybersecurity

Operational Data

- **System Logs**: Authentication, application, and AV events
- **Network Data**:
  - Flows (NetFlow/IPFIX)
  - Packets (Wireshark/Zeek analysis)
- **Endpoint Records**: EDR telemetry, file/process monitoring

Security Intelligence

- **Threat Feeds**: IOCs (IPs, domains, hashes)
- **Alert Data**: IDS/IPS signatures, SIEM correlations
- **Vulnerability Scans**: Asset and configuration assessments

Key Characteristics

Volume: TB/day in enterprises
Velocity: Real-time processing needs
Variety: Structured/unstructured
Veracity: Noise and false positives

Effective security analytics requires correlating across these data types

MERIT

Co-funded by the European Union

Security Data Types: The Foundation of Cybersecurity

Operational Data

- **System Logs**: Authentication, application, and AV events
- **Network Data**:
  - Flows (NetFlow/IPFIX)
  - Packets (Wireshark/Zeek analysis)
- **Endpoint Records**: EDR telemetry, file/process monitoring

Security Intelligence

- **Threat Feeds**: IOCs (IPs, domains, hashes)
- **Alert Data**: IDS/IPS signatures, SIEM correlations
- **Vulnerability Scans**: Asset and configuration assessments

Key Characteristics

Volume: TB/day in enterprises **Velocity**: Real-time processing needs
Variety: Structured/unstructured **Veracity**: Noise and false positives

Effective security analytics requires correlating across these data types

MERIT

Co-funded by the European Union

Security Data Types: The Foundation of Cybersecurity

Operational Data

- **System Logs**: Authentication, application, and AV events
- **Network Data**:
  - Flows (NetFlow/IPFIX)
  - Packets (Wireshark/Zeek analysis)
- **Endpoint Records**: EDR telemetry, file/process monitoring

Security Intelligence

- **Threat Feeds**: IOCs (IPs, domains, hashes)
- **Alert Data**: IDS/IPS signatures, SIEM correlations
- **Vulnerability Scans**: Asset and configuration assessments

Key Characteristics

Volume: TB/day in enterprises
Velocity: Real-time processing needs
Variety: Structured/unstructured
Veracity: Noise and false positives

Effective security analytics requires correlating across these data types

MERIT

Co-funded by the European Union

Operational Data (with Numerical Examples)

Sample Authentication System Log Entries

Timestamp	Username	Source IP	Result	Method
2025-06-17 08:13:45	jsmith	192.168.10.5	Success	Password
2025-06-17 08:14:12	jsmith	192.168.10.5	Failed	Password
2025-06-17 08:15:02	asharma	172.16.4.22	Success	OTP
2025-06-17 08:17:55	tgupta	203.0.113.25	Failed	Password
2025-06-17 08:18:20	tgupta	203.0.113.25	Failed	Password

MERIT

Co-funded by the European Union

Operational Data (with Numerical Examples)

Sample Application System Log Entries

Timestamp	Username	Application	Event Type	Status
2025-06-17 09:03:21	jsmith	CRM Portal	Login	Success
2025-06-17 09:04:10	jsmith	CRM Portal	File Download	Success
2025-06-17 09:05:45	asharma	ERP System	Report Export	Success
2025-06-17 09:06:00	tgupta	Webmail	Login	Failed
2025-06-17 09:07:18	tgupta	Webmail	Login	Failed

MERiT

Co-funded by the European Union

Operational Data (with Numerical Examples)

Sample Antivirus (AV) Event Log Entries

Timestamp	Hostname	File Name	Threat Detected	Action Taken
2025-06-17 10:12:45	LAPTOP-22A1	invoice.exe	Trojan:Win32/Emotet	Quarantined
2025-06-17 10:15:08	DESKTOP-R2X4	macro.docm	W97M.Downloader	Removed
2025-06-17 10:17:30	SERVER-FIN01	backup.zip	Heur:Trojan.Generic	Blocked
2025-06-17 10:19:02	LAPTOP-22A1	game_crack.exe	Trojan:Win32/Generic	Quarantined
2025-06-17 10:21:55	DESKTOP-R2X4	autorun.inf	Worm:AutoRun.E	Removed

MERIT

Co-funded by the European Union

Network Data (With Numerical Example)

Sample NetFlow/IPFIX Log Entries

Start Time	Src IP	Dst IP	Protocol	Bytes	Packets
10:01:12	192.168.1.10	10.0.0.5	TCP (443)	5,432	12
10:01:15	192.168.1.10	8.8.8.8	UDP (53)	123	1
10:01:18	10.0.0.5	192.168.1.10	TCP (443)	6,102	14
10:01:21	192.168.1.10	203.0.113.9	TCP (80)	2,000	5

Sample Packet-Level Data (Wireshark/Zeek analysis)

No.	Time	Source	Destination	Protocol	Info
1	0.000000	192.168.1.10	8.8.8.8	DNS	Standard query A google.com
2	0.023145	8.8.8.8	192.168.1.10	DNS	Standard query response A 142.250.64.78
3	0.512832	192.168.1.10	142.250.64.78	TCP	SYN, Seq=0, Win=64240
4	0.514002	142.250.64.78	192.168.1.10	TCP	SYN, ACK, Seq=0, Ack=1

MERiT

Co-funded by the European Union

Security Data Characteristics

Data Type	Granularity	Primary Use	Key Challenges
Logs	Medium	▪ Audit trails ▪ Detection	▪ Context gaps ▪ Format variety
Network Flows	Low	▪ Traffic analysis ▪ Baselines	▪ No payloads ▪ Sampling issues
Packet Data	High	▪ Deep inspection ▪ Forensics	▪ Storage needs ▪ Privacy limits
Endpoint Data	High	▪ Host visibility ▪ Process tracking	▪ Resource cost ▪ Deployment scale
Threat Intel	External	▪ Alert enrichment ▪ IOC matching	▪ Integration work ▪ Relevance checks
Security Alerts	High-level	▪ Incident triage ▪ Prioritization	▪ False alarms ▪ Alert fatigue

Strategic data combination enables comprehensive threat visibility

MERIT

Co-funded by the European Union

Machine Learning Pipeline for Security Applications

1. Problem Definition & Scoping

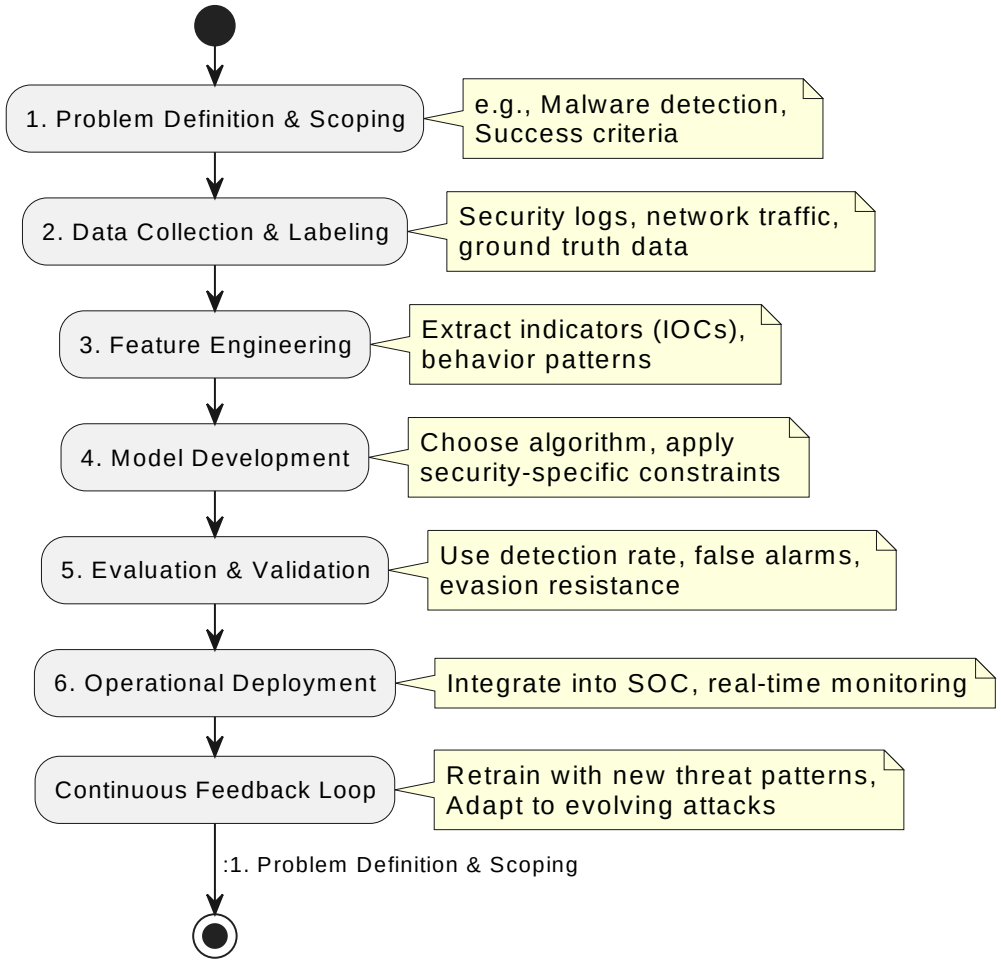
Identify specific security challenge (e.g., malware detection) and success metrics

2. Data Collection & Labeling

Gather relevant security data (logs, network traffic) with proper ground truth

3. Feature Engineering

Extract meaningful patterns (IOCs, behavior signatures) from raw security data



MERIT

Co-funded by the European Union

Machine Learning Pipeline for Security Applications Contd.

4. Model Development

- Algorithm selection (e.g., Random Forest for classification)
- Training with security-specific constraints (FAR limits)

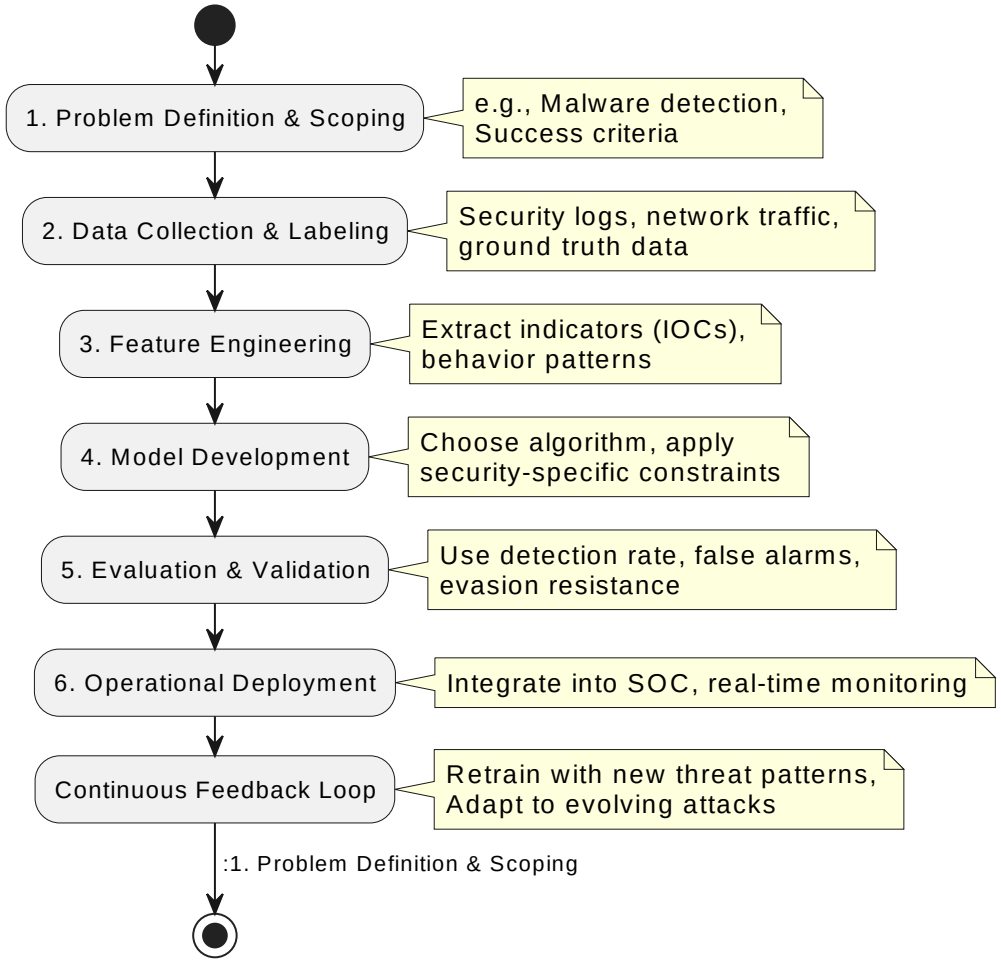
5. Evaluation & Validation

Security-focused metrics: Detection rate, False Alarm rate, Evasion resistance

6. Operational Deployment

Integration with SOC workflows, real-time performance monitoring

Continuous feedback loop improves model against evolving threats



MERIT

Co-funded by the European Union

Numerical Example: ML Pipeline for Intrusion Detection

Stage	Explanation	Numerical Example
1. Problem Definition	Detect network intrusions from packet data	Goal: Classify connections as normal (0) or attack (1)
2. Data Collection	Use KDD Cup 99 dataset or internal network logs	1000 records collected; 700 labeled "0", 300 labeled "1"
3. Feature Engineering	Extract relevant fields from packet flow	E.g., duration = 5s, protocol = TCP, src_bytes = 300
4. Model Development	Train ML algorithm with constraints	Random Forest trained on 800 records with 10 selected features
5. Evaluation & Validation	Evaluate with detection metrics	Accuracy = 93%, False Alarm Rate (FAR) = 4% on 200 records
6. Operational Deployment	Real-time integration with monitoring tools	Model flagged 5 suspicious flows in first 60 seconds

Continuous feedback loop: retrain model with updated labeled data monthly.

MERiT

Co-funded by the European Union

Thank You!

Presented by Adarsh Kumar
Department of Computer Science, UPC, Dehradun

Co-funded by
the European Union