SEAS-8414

Analytical Tools for Cyber Analytics

Survey of analytical tools for analyzing cyber security data with particular attention to the use of data analytics procedures in supporting appropriate cyber security policy decisions.

Dr. M

Welcome to SEAS Online at George Washington University

SEAS-8414 class will begin shortly

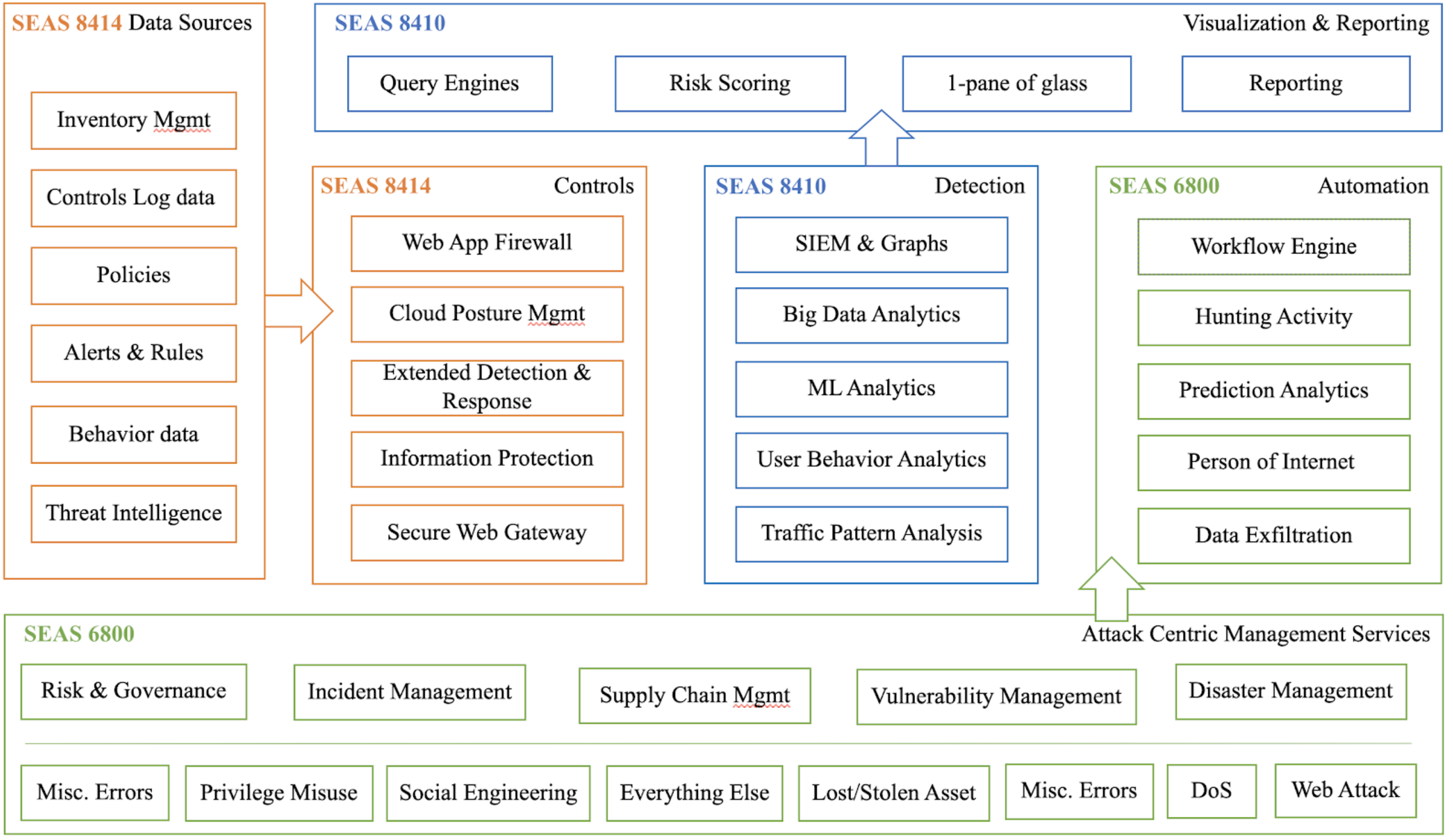
- **Audio:** To eliminate background noise, please be sure your audio is muted. To speak, please click the hand icon at the bottom of your screen (**Raise Hand**). When instructor calls on you, click microphone icon to unmute. When you've finished speaking, ***be sure to mute yourself again***.
- **Chat:** Please type your questions in Chat.
- **Recordings:** As part of the educational support for students, we provide downloadable recordings of each class session to be used exclusively by registered students in that particular class for their own private use. Releasing these recordings is strictly prohibited.

Agenda

Week-2: Learn about data-centric security tools

We will build a data-centric risk management program for gwuscc.com. We will learn about the role of assets, vulnerabilities, threats, risks, and controls. We will leverage probability to develop a risk management plan. We will discuss how hashing, fingerprinting, and signature techniques are leveraged by data-centric security controls such as Data Leak Prevention (DLP) and Content Disarm and Reconstruction (CDR) for audit, prevention, detection, and forensics.

Course Objective

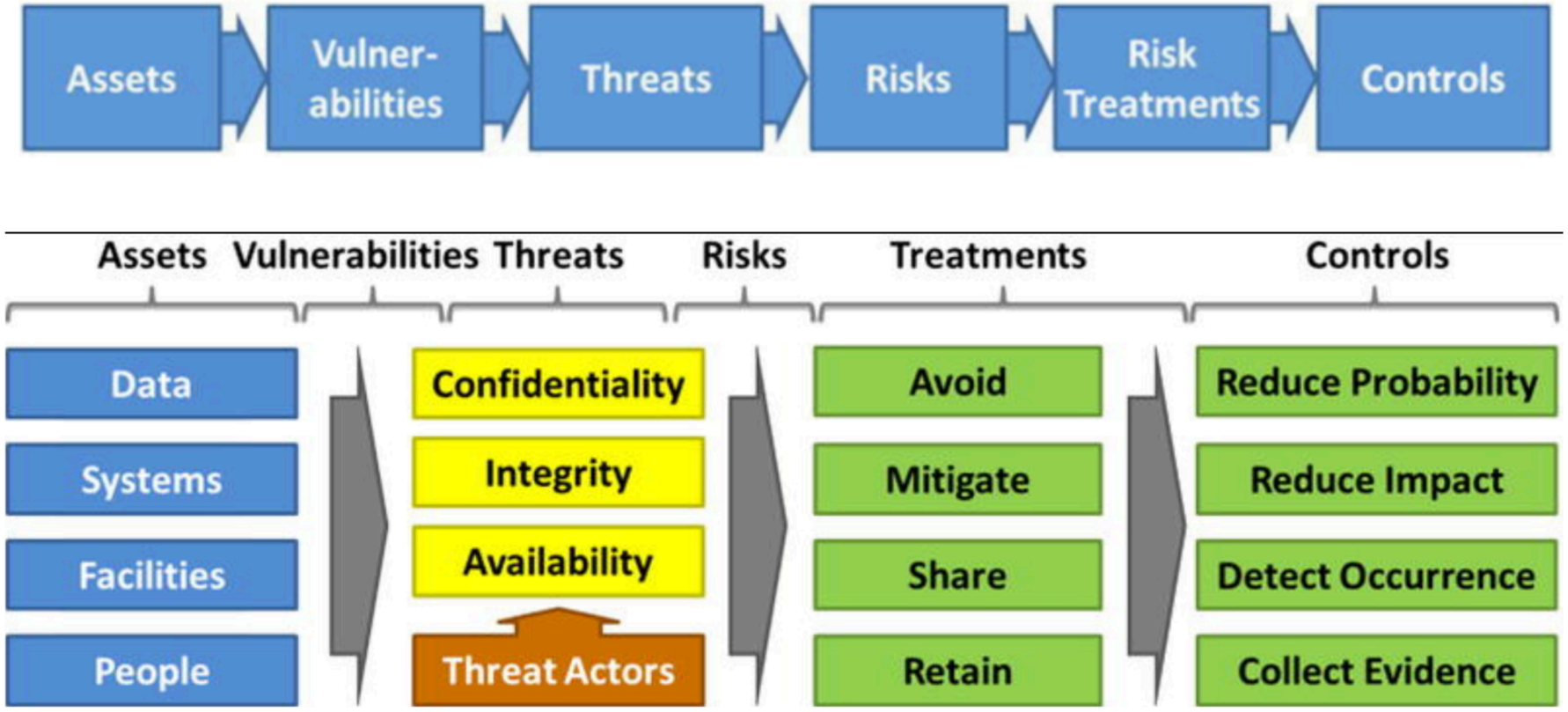

SEAS 8414 Data Sources
Inventory Mgmt
Controls Log data
Policies
Alerts & Rules
Behavior data
Threat Intelligence
SEAS 8410
Visualization & Reporting
Query Engines
Risk Scoring
1-pane of glass
Reporting
SEAS 8414
Controls
Web App Firewall
Cloud Posture Mgmt
Extended Detection & Response
Information Protection
Secure Web Gateway
SEAS 8410
Detection
SIEM & Graphs
Big Data Analytics
ML Analytics
User Behavior Analytics
Traffic Pattern Analysis
SEAS 6800
Automation
Workflow Engine
Hunting Activity
Prediction Analytics
Person of Internet
Data Exfiltration
SEAS 6800
Attack Centric Management Services
Risk & Governance
Incident Management
Supply Chain Mgmt
Vulnerability Management
Disaster Management
Misc. Errors
Privilege Misuse
Social Engineering
Everything Else
Lost/Stolen Asset
Misc. Errors
DoS
Web Attack

Class-2

Structure

1. Why do we need to protect data?
2. What is data protection?
3. How to protect data?

Quick Review



Why do we need to protect data?

Why do we need to protect data?

- **Infiltration Risk**: Company website is the primary mode of communicating with the users.
- **Exfiltration Risk**: Users or Bots sending data out to unauthorized websites.

When is data exposed?

- At rest: read the confidential data, copy, USB, steal HDD, file system copy
- In transit: remote copy, http|https
- In memory: the program is using it (MS word -> file)

Infiltration Risks

1. Man in the Middle attack
2. Website Defacement

Man in the Middle Attack

“An attack in which an attacker is positioned between two communicating parties to intercept and / or alter data traveling between them. In the context of authentication, the attacker would be positioned between claimant and verifier, between registrant and CSP during enrollment, or between subscriber and CSP during authenticator binding.”

Source: https://csrc.nist.gov/glossary/term/man_in_the_middle_attack

Man in the Middle Attack

Demo

Website Defacement

Breach of website integrity.

Source: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-44ver2.pdf

Website Defacement

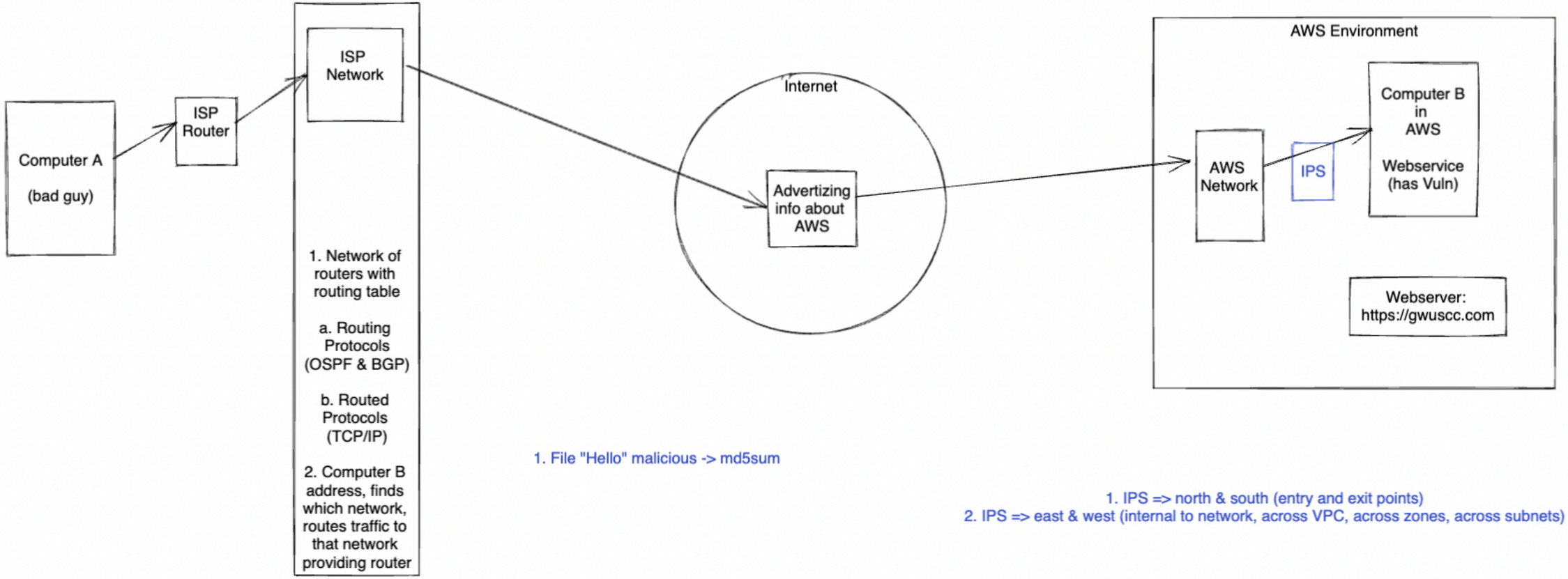
Demo

How do we protect?

Risks

- CIA: Confidential, Integrity, Availability
- At rest: encrypt -> disk, file system, file, parameters/configuration.
- Encryption: Scrambling, obfuscating, encoding some math
  - Hello -> +13 -> Eopps
  - Hello -> H * 3 + e * 3 + l *3 -> Bgwwr -> 5 (asymmetric)
  - 15 -> 3, 5
  - 35 -> 7, 5 (prime numbers)
  - Hello -> H * 3 + e *3 -> Bgwwr -> 3 (symmetric)
  - TLS -> A + S -> Both
  - AES256 (stealing encrypted ->

1. HTTPS -> network establish communication -> Encrypted TLS -> port 443



1. File "Hello" malicious -> md5sum

1. IPS => north & south (entry and exit points)
2. IPS => east & west (internal to network, across VPC, across zones, across subnets)

Pattern
A -> 1 => B
C -> 1 => B
D -> 1 => B

Pattern
A -> 2 => B
C -> 1 => B
D -> 1 => B

IPS => Known bad stuff
DLP => Known good stuff

IPS

What is IPS?

An intrusion is an unapproved and unauthorized endeavor to take possession of a corporate asset to exfiltrate data or render them unreliable. An Intrusion Prevention System's (IPS) objective is to detect and prevent intrusions by **observing network traffic** and **engaging a firewall** or **unified threat management (UTM)** system to **block network communication**. Every cyber intrusion involves network communication to deliver the **initial malicious payload**, **lateral movement**, **command & control**, and **data exfiltration** between the malicious and victim's network. IPS systems can detect **denial of service**, **network probing**, **malware**, **ransomware**, **buffer & stack overflow**, spam, phishing, DNS amplification, spoofing, and cache poisoning techniques.

UTM => IPS + Firewall + Antivirus engine + L7 Firewall

IPS => IDS + prevention

How does IPS Work?

IPS leverages two popular methods for intrusion detection (Khraisat et al., 2019):

(1) *Knowledge-based detection*, also known as **signature-based detection**, where patterns in network packets get correlated against the known pattern and rules. Knowledge-based systems leverage methodologies such as finite state machines, description languages, and expert systems.

(2) *Anomaly-based detection*, where some implementations leverage **statistics-based methods** such as univariate, multivariate, and time-series models. In contrast, others leverage machine learning-based techniques such as decision trees, genetic algorithms, or support vector machines to detect intrusions. In anomaly-based detection, suspicious activity is shortlisted based on deviation from expected behavior. It is categorized and classified using unsupervised and supervised machine learning models.

What is the difference?

Features	Signature-based IPS	Anomaly-based IPS
Accuracy of detecting a known attack	High (good)	Moderate (good)
Accuracy of detecting an unknown attack	None (bad)	Moderate (good)
Accuracy of detecting a complex attack	None (bad)	Moderate (good)
False-positive rate	Low (good)	High (bad)
Design complexity	Low (good)	High (bad)
Onboarding & training time	Low (good)	High (bad)

Why isn’t IPS enough?

1. “*Survey of intrusion detection systems: techniques, datasets and challenges*” (Khraisat et al., 2019). **IPS cannot keep up with modern networks with gigabytes of bandwidth**. Capturing network packets and processing them in real time is a significant limitation due to increasing network speed and bandwidth. It is particularly challenging for telework environments.
2. “*Evaluating intrusion prevention systems with evasions*” (Särelä, Kyöstilä, Kiravuo, & Manner, 2017). An anomaly-based IPS generates a high false-positive rate in a dynamic workplace or with a change in the network (e.g., a user moving from a cafe to a house).
3. “*Deciphering malware’s use of TLS (without decryption)*” (Anderson, Paul, & McGrew, 2018). Most IPS solutions cannot detect novel exploits in encrypted traffic with just meta-data about the session. The ability of anomaly-based IPS to detect zero-day exploits depends on the variety of data in the training dataset. It is challenging to define normal training data, and it is tough to find comprehensive labeled malicious data for training.

DLP

What is DLP?

Data leaks refer to any accidental or intentional distribution of confidential data to an unauthorized entity. Data Leak Prevention or also known as Data Loss Prevention (DLP) systems are responsible for detecting, monitoring, and protecting data in three stages: (1) Data at rest, (2) Data in motion, and (3) Data in use (Kaur, Gupta, & Singh, 2017). In practice, DLP is predominant in the space of data in motion.

Confidential => DLP =>

Secret =>

Top-secret => DLP

How does DLP work?

A typical DLP operates in three phases (Kaur et al., 2017).

1. *In the detection phase,* DLP leverages Deep Content Analysis (DCA) techniques, which are mainly rules-based -- using regular expressions, fingerprinting, file name matching, checksum, incremental hashing, and statistical analysis.
2. *In the monitoring phase*, DLP determines sensitivity based on classification (e.g., confidential, secret, top-secret), roles, and file system events to verify against access levels.
3. *In the protection phase*, DLP systems are integrated with directory services, network proxies, firewalls, and UTM systems to observe network traffic for different levels of sensitive data and allow/block traffic depending on the rules.

Bypassing DLP

Demo

Why isn’t DLP enough?

The literature review of these six DLP research papers reveals the following takeaways.

1. “*A literature survey on data leak detection and prevention methods*” (Baby & Krishnan, 2017). It is resource-intensive to detect data leaks in motion. Transport Layer Security (TLS) using encryption is generally the best way to guard the data from eavesdropping. However, DLP detection requires inspecting payload in the network communication, which requires decrypting TLS that needs an expensive computing platform to keep traffic movement near-real time.
2. “*Predicting the likelihood of legitimate data loss in email DLP*” (Faiz, Junaid, Alazab, & Shalaginov, 2019). DLP can block access to data by defining various levels of access to users by classification of the content. Defining access levels and data classification in an established organization is expensive. It takes subject matter expertise to classify every document's sensitivity in the firm. It also needs a dedicated team and software to operate Identify and Access Management (IAM) infrastructure to set the proper access levels.
3. “*Enterprise data breach: causes, challenges, prevention, and future directions*” (Cheng et al., 2017). The in-depth content inspection technique, such as DCA, requires the content of the data in network communication to inspect and make decisions. However, DCA techniques are easy to fool. They work best to detect negligent perpetrators but do not stand up against a targeted attack. DCA techniques lead to high rates of false-positive alerts in a dynamic work environment, making it difficult for security operations teams to recognize signals from noise.

Why isn’t DLP enough?

4. “*A Comparative evaluation of data leakage/loss prevention systems*” (Kaur, Gupta, & Singh, 2017). Small and slow data leaks are challenging to detect. Modern DLP systems use Partially Observable Markov Decision Processes (POMDPs) over a fixed period, called decision epochs, to detect small leaks over a large time window. The parametrization over the decision epoch window, size of leaks, and changing destination domain makes it almost impossible to distinguish the modern malware exfiltrating data from benign activity.
5. “*Data loss prevention using machine learning*” (Nagpal, & Ahmed, 2020). DLP products mislead decision-makers with the myth of anomaly detection and machine learning techniques. Some modern cloud-based DLP systems leverage anomaly detection and supervised machine learning techniques. Supervised machine learning models are only as good as their labeled data and training. The reliance on behavior analytics for anomaly detection raises more noise than filtering due to the lack of integration with employee background, psychological, and workplace behavior information.
6. “*Hypervisor-Based Sensitive Data Leakage Detector*” (Chang, Mallissery, Hsieh, & Wu, 2018). Most DLP systems cannot catch partial data leaks through images. With the advent of universal plug-and-play, DLP systems require continuous blocklisted printers and Universal Serial Bus (USB) drivers. DLP systems rely heavily on filesystem application binary calls and interrupt vector tables. It indicates that with kernel patch release, DLP systems require regression and functional testing.

EPP

What is EPP?

The Endpoint Threat Prevention Platform or Endpoint Prevention Platform (EPP) can detect, prevent, and respond to cyber threats. It observes operating system activities on an endpoint and automatically responds to known malicious activities. There are variations to the EPP solution called Endpoint Threat Detection and Response (EDR) and Extended Threat Detection and Response (XDR). While EPP can prevent known threat patterns during a session, EDR can detect complex threat patterns across multiple sessions and respond according to the policy. XDR goes a step forward to detect security incidents by correlating events from other security controls such as DLP and IDS.

EPP => DLP + IPS

EDR => DLP + IDS

How does EPP work?

EPP operates using the following methods:

1. *Signatures and heuristics-based detection* leverage regular expression type matching to detect the presence of malware in executable and binary files.

2. *Application exploits protection* guards against memory and buffer overflow attacks.

Machine learning-based predictive engine to identify malware through static file analysis. The model is trained on benign and malicious files to differentiate.

3. *Application control*, commercially known as application safe listing, enforces which binaries can be executed on the endpoint. The idea is to intercept the interrupts to examine the system calls and verify if the executed binary is listed in the approved or blocked list of vendors, checksum, hash code, or name on catching process execution calls.

How does EPP work?

4. *Application containment* is the technique that allows the shielded execution of binary on the endpoint to process conceivably malicious content. It separates the processing of application content from the rest of the applications on the operating system. Unlike application control, containment limits the impact of running software in a low-trust environment.
5. *Behavioral analysis* provides prevention by continuously monitoring for network intrusions and malware. The endpoint is monitored for known adversarial tactics, tools, and techniques. Special attention is paid to the behavior of a process deviation and file-based malware to form indicators of intrusions that may be allowed, blocked, or detected.
6. *Detection and response* monitor endpoint actions for multiple objectives, including exposure, containment, and investigation of and tracking for malicious behavior. The response actions can be automated and defined by corporate policy.

Why isn’t EPP enough?

What are the issues with EPP?

1. “*Tactical provenance analysis for endpoint detection and response systems*” (Hassan, Bates, & Marino, 2020). Most EPP and EDR products are tactics, techniques, and procedures (TTP) based that are optimized for recall, not precision. It means TTP libraries endeavor to represent all methods that are likely to be infiltration-related, even if the same techniques are extensively used for benign purposes—for example, the "File Deletion" Technique in MITRE ATT&CK (Duff, 2020). Although file deletion indicates deceptive tactics, it could also be part of benign user activities. Consequently, EPP and EDR tools are predisposed to high volumes of false alerts. EDR products are key contributors to "threat alert fatigue" for cybersecurity operations. Consequently, the true positives identified by EPP and EDR tools are in jeopardy of being succumbed the noise of false positives.
2. “*Endpoint Detection and Response: Why Use Machine Learning?*” (Sjarif et al., 2019). Most EPP functions require the Internet. For instance, EPP malware detection techniques use machine learning methodologies, which need large computation systems to run the models. Considering the teleworker’s endpoint is not the right place to run the deep learning models, most detection component of EPP solutions sends the system events to the vendor's cloud service to process. It requires the endpoint to be constantly connected to the service provider for uninterrupted protection.
3. “*Endpoint Protection: Measuring the Effectiveness of Remediation Technologies and Methodologies for Insider Threat*” (Chandel, Yu, Yitian, Zhili, & Yusheng, 2019). Matching malicious signatures on a telework endpoint is resource-intensive. EPP signature and heuristic analysis scans every execution and the run-time environment against the database of the virus, malware, and data segment hashes. This process is highly resources intensive on a busy system leading to the user experiencing slow performance.

Why isn’t EPP enough?

4. “*A Survey on Advanced Persistent Threats: Techniques, Solutions, Challenges, and Research Opportunities*” (Alshamrani, Myneni, Chowdhary, & Huang, 2019). EPP and EDR solutions cannot identify threats with varied attack patterns. EPP and EDR are trained using machine learning and programmed to detect specific patterns of vulnerability exploitations. However, whenever there is a variation in the attack pattern despite the same vulnerability, many EPP and EDRs fail to detect it. For instance, the study of exploited vulnerabilities in the APT attacks revealed the use of mostly known vulnerabilities and the occasional use of zero-day vulnerabilities. Many EPP and EDR solutions failed to detect known vulnerabilities because of the variation in exploitation patterns.
5. “*ATT&CK Evaluations: Understanding the Newly Released APT29 Results*” (Duff, 2020). EPP and EDR solutions are not silver bullets to prevent threats on the endpoint. MITRE ATT&CK evaluation finding of 20+ EDR and EPP solutions claims that there is no solution that is capable of detecting all of the APT-29 exploitation techniques. Figure 6-6 illustrates the evaluation summary of EPP and EDR products using advanced persistent threat (APT) number 29. APT29, also known as Cozy Bear, is a threat group. It is attributed to the Russian government (US-CERT, 2016). The same APT that was involved in FireEye’s SolarWinds attack in 2020. Most EDR solutions rely on collecting telemetry to introspect by the security operations team to write detection logic, triage incidents, and respond.
6. “*Application of Artificial Intelligence and Machine Learning in Producing Actionable Cyber Threat Intelligence*” (Montasari, Carroll, Macdonald, Jahankhani, Hosseinian-Far, & Daneshkhah, 2021). Threat intelligence is not a reliable source for making EPP decisions. Many EPP and EDR solutions rely on curated threat intelligence for indicators of compromise (IOC), such as domain name, IP address, and file checksum to predefined actions. However, it is prone to generate massive false-positive alerts.

Cybersecurity Strategy

Cybersecurity Team Focus

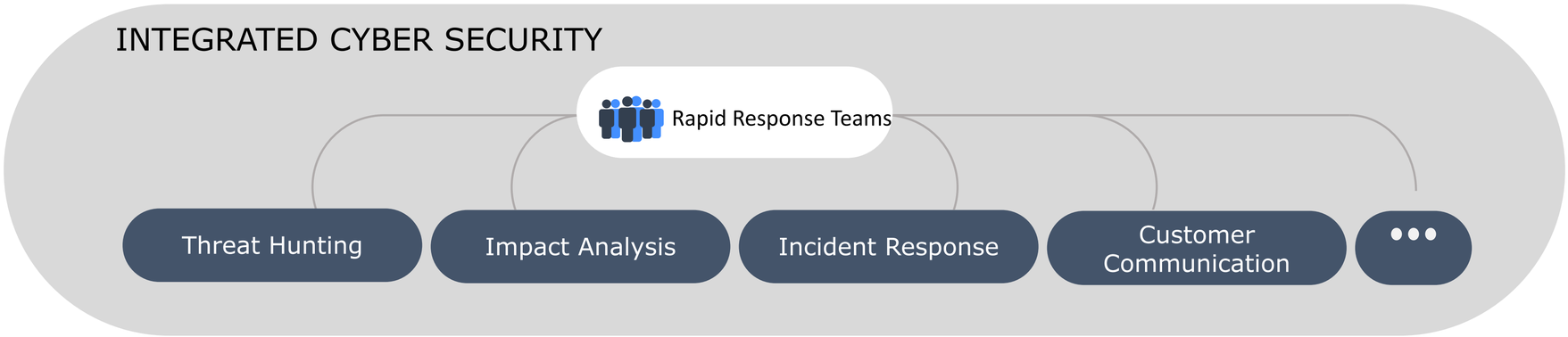
Focusing On Increasing Volume of Threats

OBJECTIVES

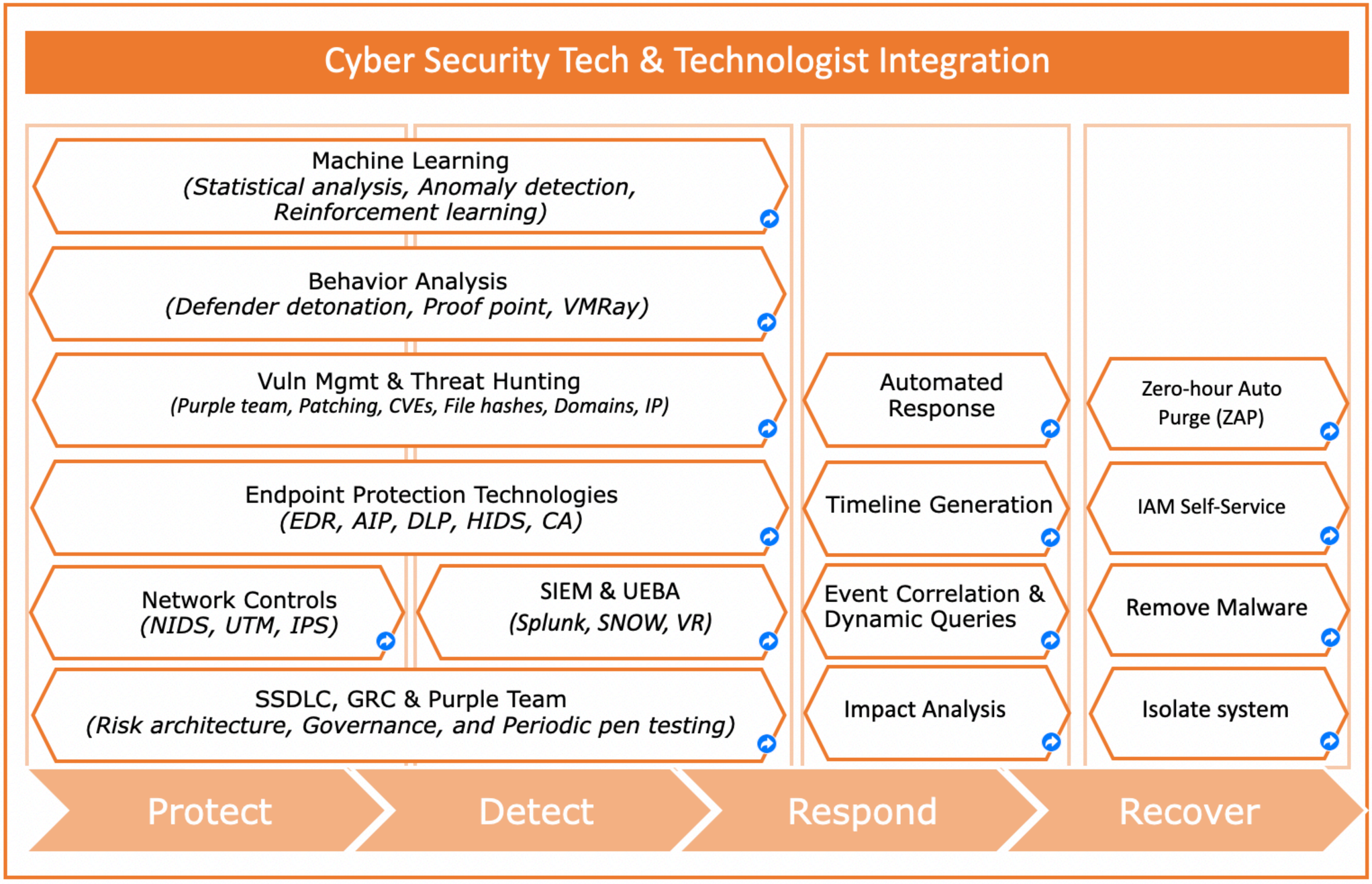
1.Prevent all known threats
2.Make it expensive for perpetrator to exploit
3.Modernize cyber detection & response
4.Integrate teams and tech for rapid response

APPROACH

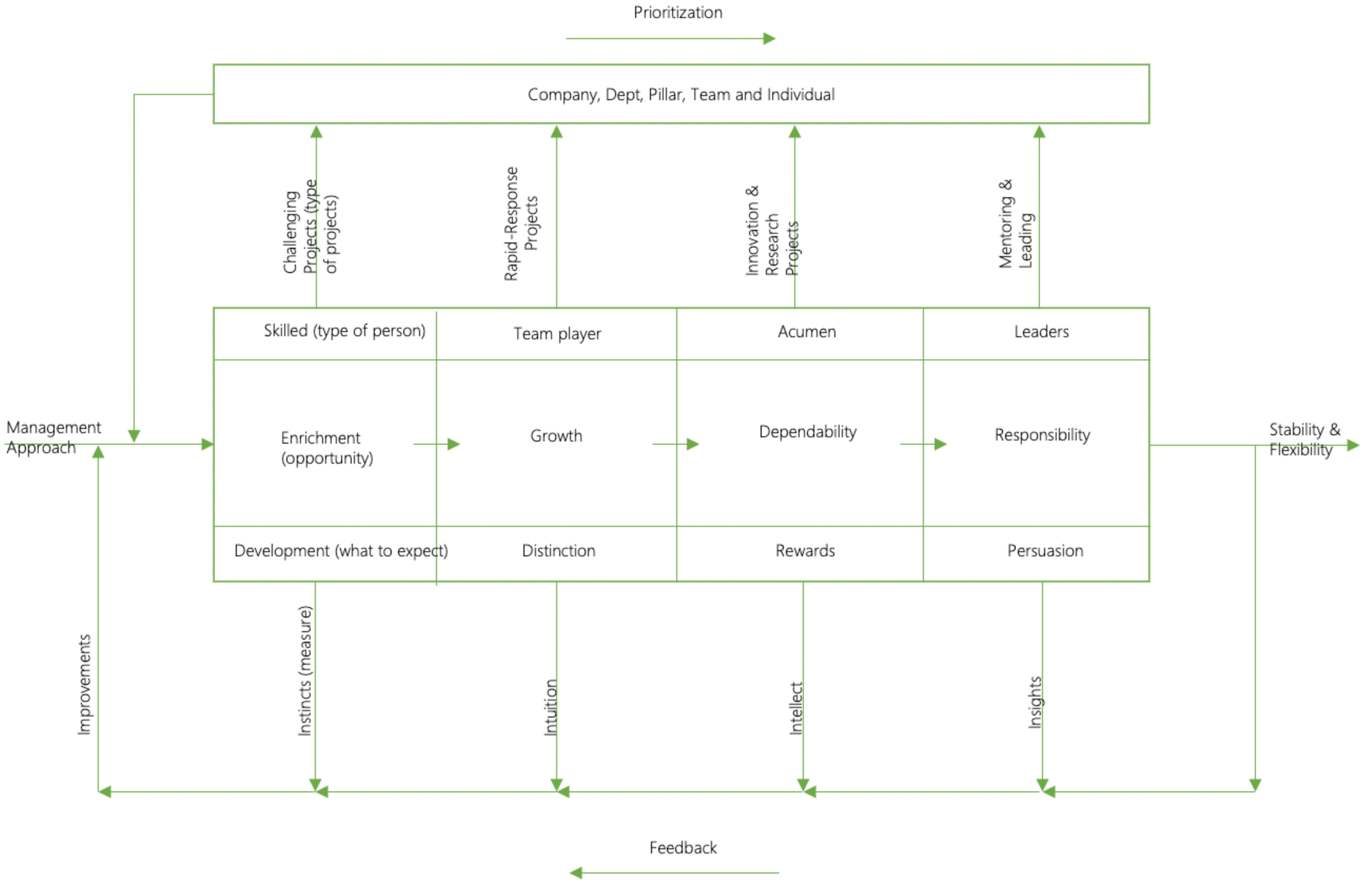
1.Maximize Visibility
   (a) high fidelity sensors
   (b) diverse intel
2.Automate manual steps – (a) triage, (b) response
3.Maximize analyst impact through continuous training
4.Regularly engage in external triangulation



Cybersecurity & Tech Integration


Cyber Security Tech & Technologist Integration
Machine Learning
(Statistical analysis, Anomaly detection, Reinforcement learning)
Behavior Analysis
(Defender detonation, Proof point, VMRay)
Vuln Mgmt & Threat Hunting
(Purple team, Patching, CVEs, File hashes, Domains, IP)
Endpoint Protection Technologies
(EDR, AIP, DLP, HIDS, CA)
Network Controls
(NIDS, UTM, IPS)
SIEM & UEBA
(Splunk, SNOW, VR)
SSDLC, GRC & Purple Team
(Risk architecture, Governance, and Periodic pen testing)
Automated Response
Timeline Generation
Event Correlation & Dynamic Queries
Impact Analysis
Zero-hour Auto Purge (ZAP)
IAM Self-Service
Remove Malware
Isolate system
Protect
Detect
Respond
Recover

Cybersecurity Team Development


Prioritization
Company, Dept, Pillar, Team and Individual
Challenging Projects (type of projects)
Rapid-Response Projects
Innovation & Research Projects
Mentoring & Leading
Skilled (type of person)
Team player
Acumen
Leaders
Management Approach
Enrichment (opportunity)
Growth
Dependability
Responsibility
Stability & Flexibility
Development (what to expect)
Distinction
Rewards
Persuasion
Improvements
Instincts (measure)
Intuition
Intellect
Insights
Feedback

Hands-on: Data Security in AWS

How to security audit the AWS Cloud Environment?

How to secure storage in AWS?

How to encrypt secrets in AWS?

How to secure object storage in AWS?

How to paper trade in newly secured AWS?

Homework

Homework - 1

Design a start-up infrastructure for running a web service with sensitive data. We will discuss various frameworks and security controls you could use during the class.

A sample reference for homework expectations:

https://blogs.lt.vt.edu/hfsecurity/