Machine Learning for Enterprise Security

Pratyusa K. Manadhata
Hewlett Packard Labs
manadhata@hpe.com

Outline

Evolution of enterprise security	Uses of machine learning	Strengths and weaknesses	Trends and research opportunities

Focus on problems, not on ML techniques

Enterprise security

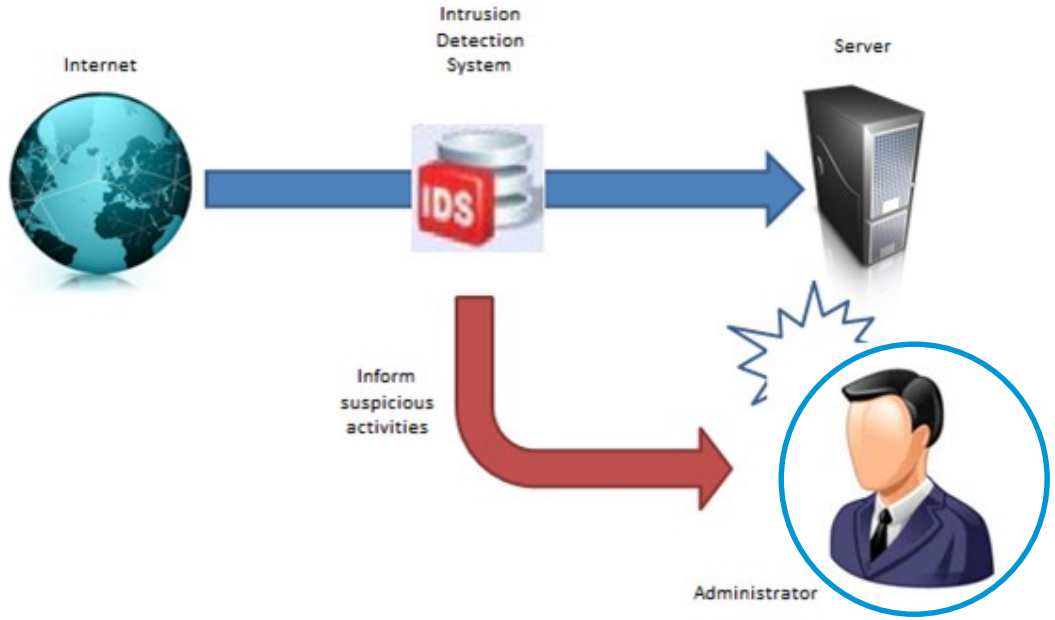
A Security framework

Prevent

Detect

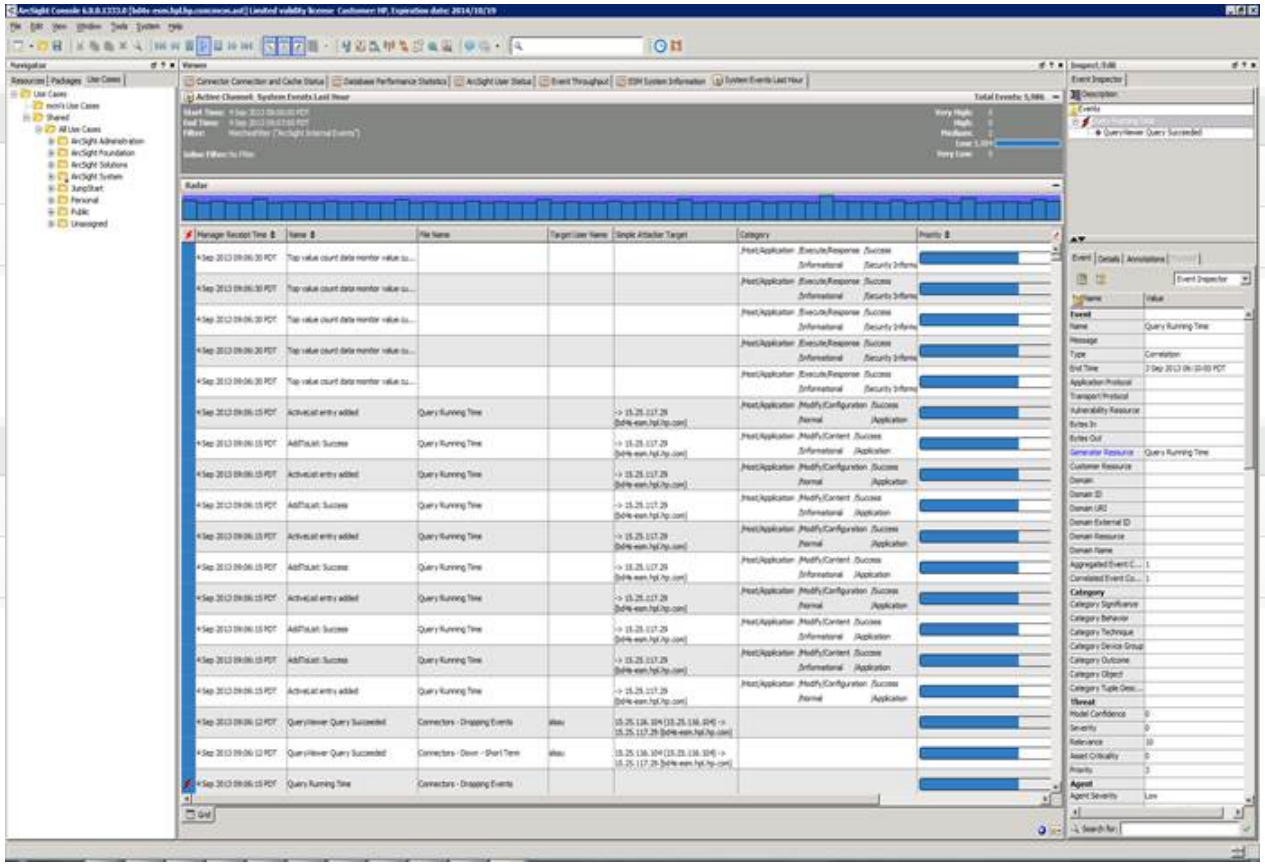
Recover

1st Generation: Point products



		Sev.	Sensor	Source IP	Destination IP	Event Signature	Timestamp
☐	☆	2	qa-eth0:eth0	172.16.116.234	217.160.51.31	ET POLICY curl User-Agent Outbound	7:41 PM
☐	☆	2	qa-eth0:eth0	217.160.51.31	172.16.116.234	GPL ATTACK_RESPONSE id check returned root	7:41 PM

2nd Generation: Security information and event management systems (SIEM)



AntiVirus

Unified UI

False positive reduction

Firewall

3rd Generation: Security Operations Centers

1st Generation Point Products

Do one thing and do it “well”

- Anti-malware products
- Data leakage prevention products
- Firewalls
- Application firewalls
- …

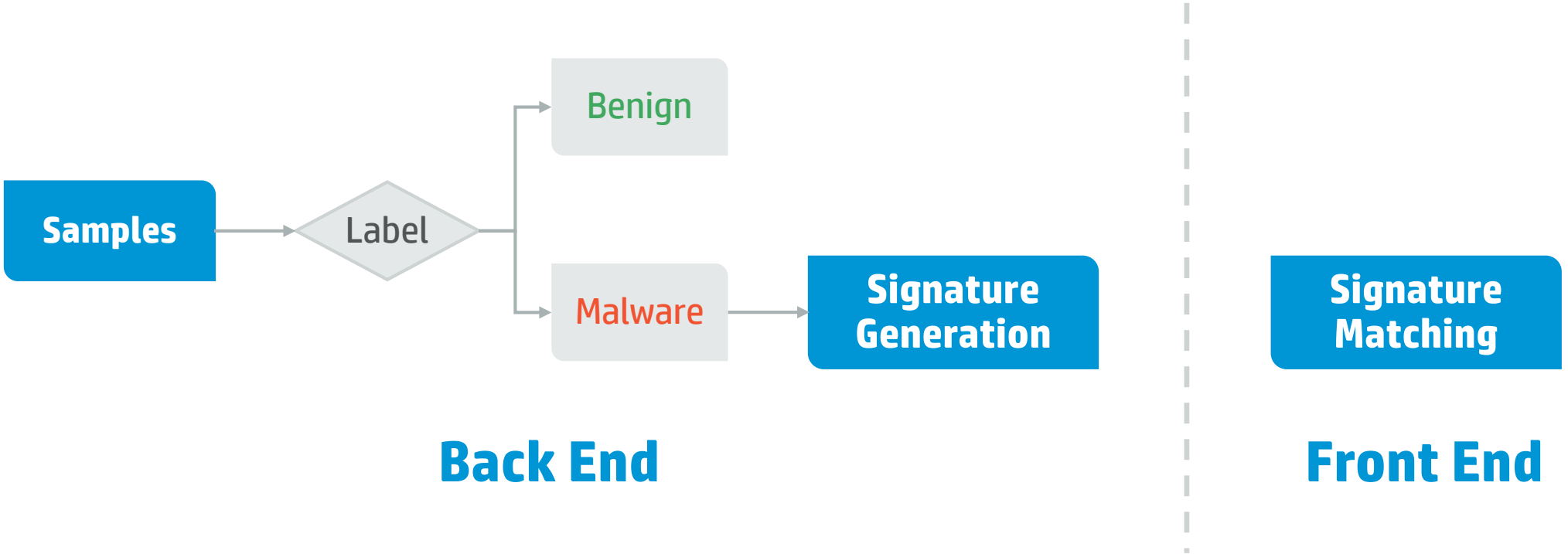
- Intrusion detection systems (IDS/IPS)
- Domain/IP Blacklists
- Sandboxes
- Web proxies
- ….

Product structure

AV
Firewall
IDS
…

AV Signatures
Firewall Rules
IDS Signatures
..

Anti-Malware products



Sample labelling

Similarity

Machine Learning: Detect similarity at scale

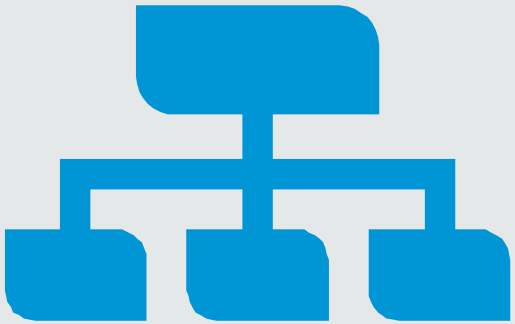
Similarity is key in other detection mechanisms too

No “true” indicators of good/bad exist

Static analysis: looks similarity

- **Features:** PE Headers, instruction sequences,..
- Classification/Clustering
- False positives
- False negatives – easy to evade

Dynamic analysis: behavior similarity



- **Features:** system call sequences, API sequences, ..
- Classification/clustering
- Computational challenges
- Hard to run samples

Reputation analysis: behavior similarity

- Program behavior data from end points, no need for samples
- **Features:** popularity, source, destinations,…
- Lightweight data collection
- Lagging detection

Front end detection: False positive-false negative trade-off

Almost no machine learning on end devices

Complex and Fragile

Malicious domain detection

Many Methods/Papers

- Classification/clustering/regression
- Graph analysis
- Pattern mining and matching
- Statistical analysis
- …

Features

- Syntactic properties of a domain name string
- HTTP/DNS protocol properties
- Access patterns
- Registration information
- Association with malware
- …

A command & control domain used in the OPM breach

opmsecurity.org

source: https://www.threatconnect.com/opm-breach-analysis/

Observations and opportunities

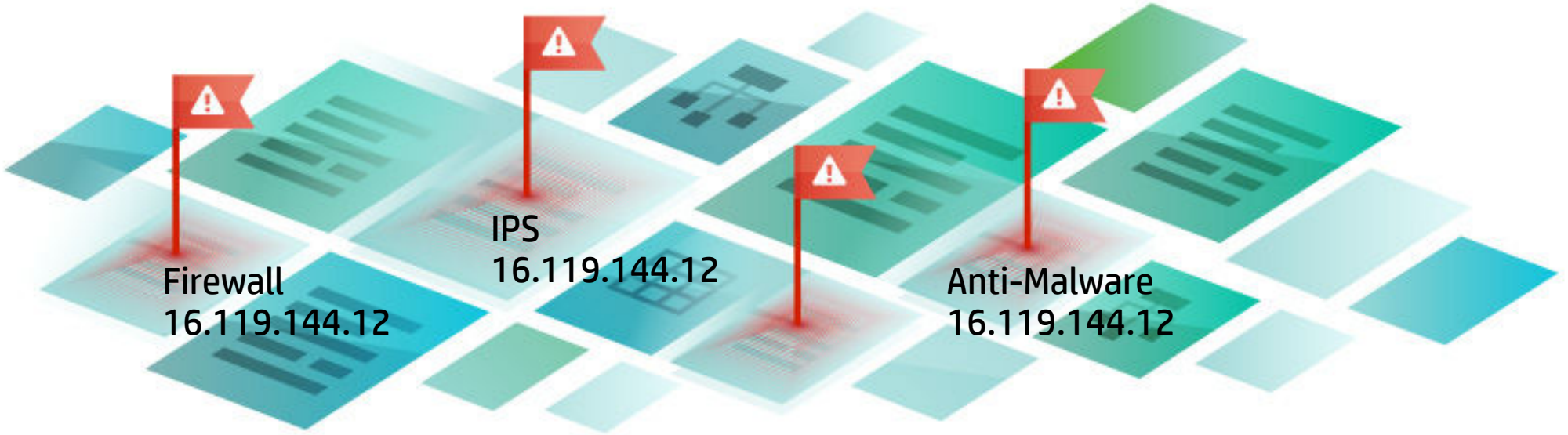
ML helps with scalability, but not much with accuracy

Complex and fragile systems

ML on end devices instead of signature matching

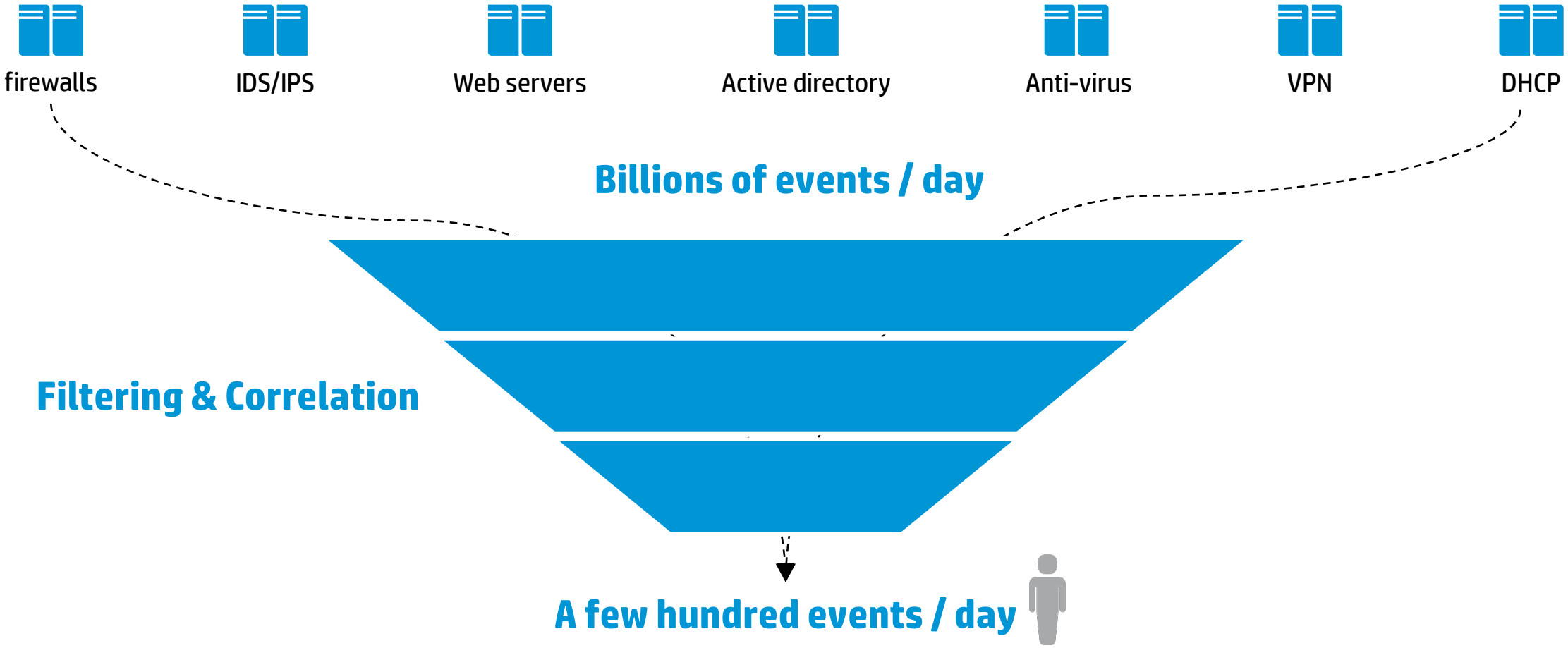
2nd Generation Security Information and Event Management

Correlation



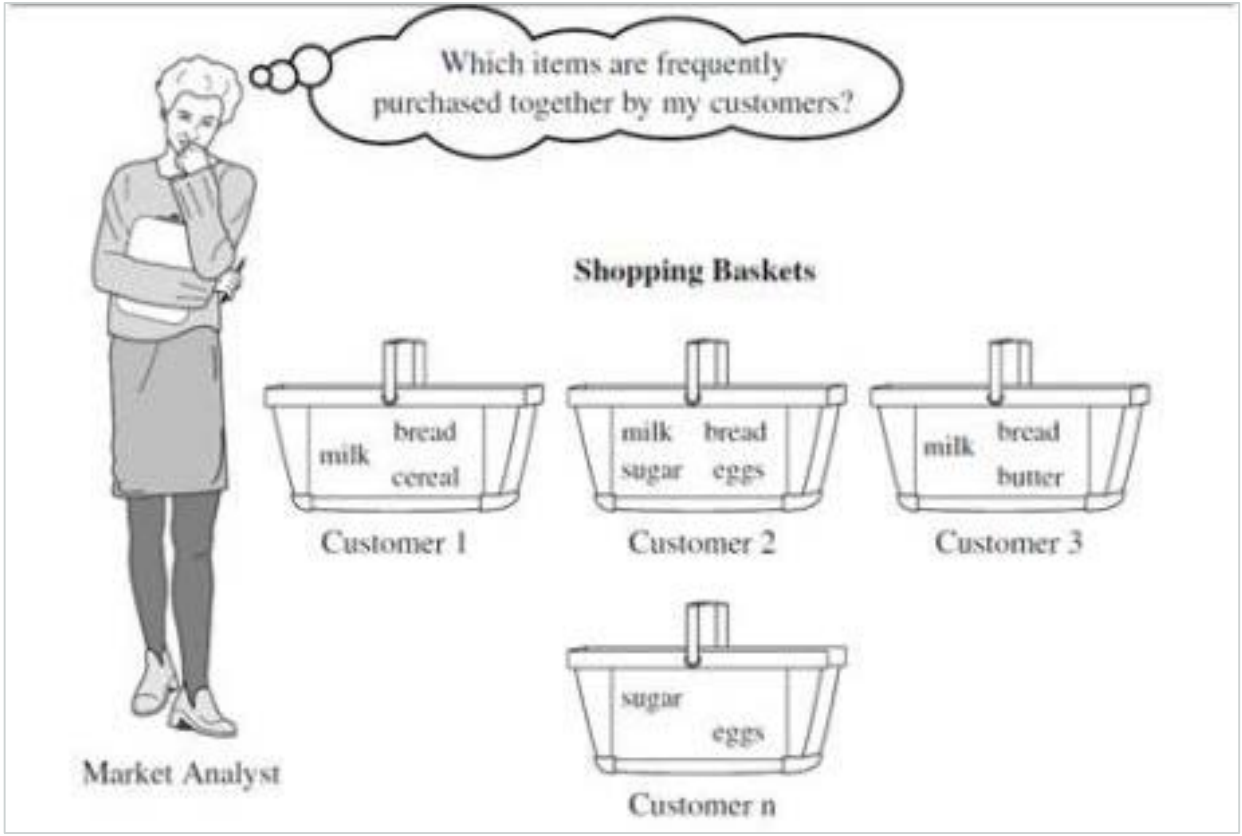
Reduce alerts by grouping
Reduce false alarms

Security Information and Event Management



How to generate **correlation rules**?

Market-basket Analysis



Observations

Rules are heuristics driven/manually generated

Correlation window order of minutes

3rd Generation Security Operations Centers

A large enterprise network

HP IT supports 6 NGDC and 86 MCS

140+ Windows Domain Controllers

Microsoft

41K+ servers owned by HP IT

450,000 end points protected with anti-virus

1.2 million connected devices

1,500+ enterprise HPN Routers

Manage **150K+ mobile devices**

15K+ HPN switches

11.5M+ Internet mails per day sent/received

2,000+ HP IT managed firewalls

2.5B security events logged per day with ArcSight

597 IPS sensors deployed

TippingPoint

39,000,000 IP Addresses including 2 contiguous Class A's

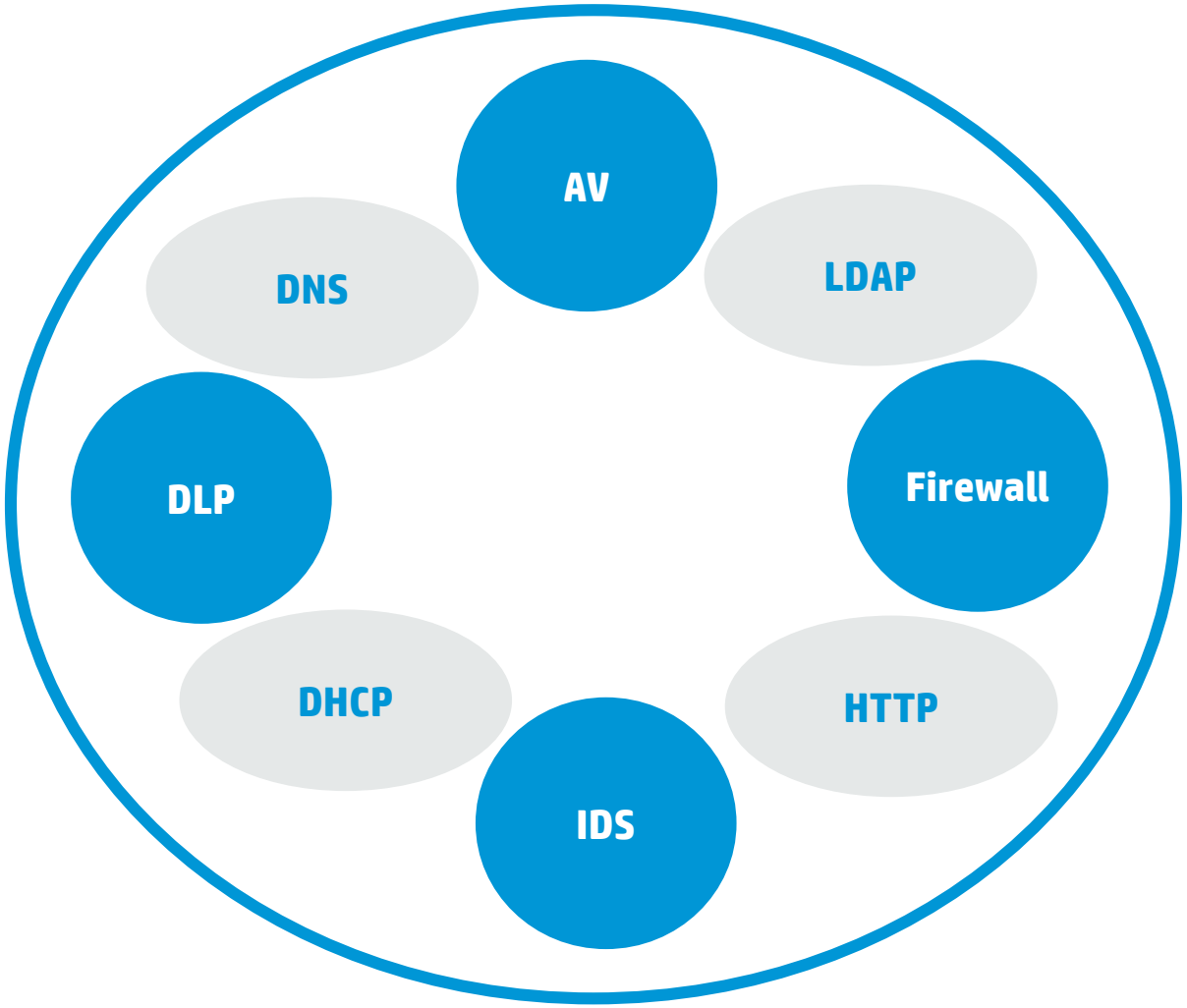
970K+ scanned devices for vulnerabilities

450,000 mailboxes managed

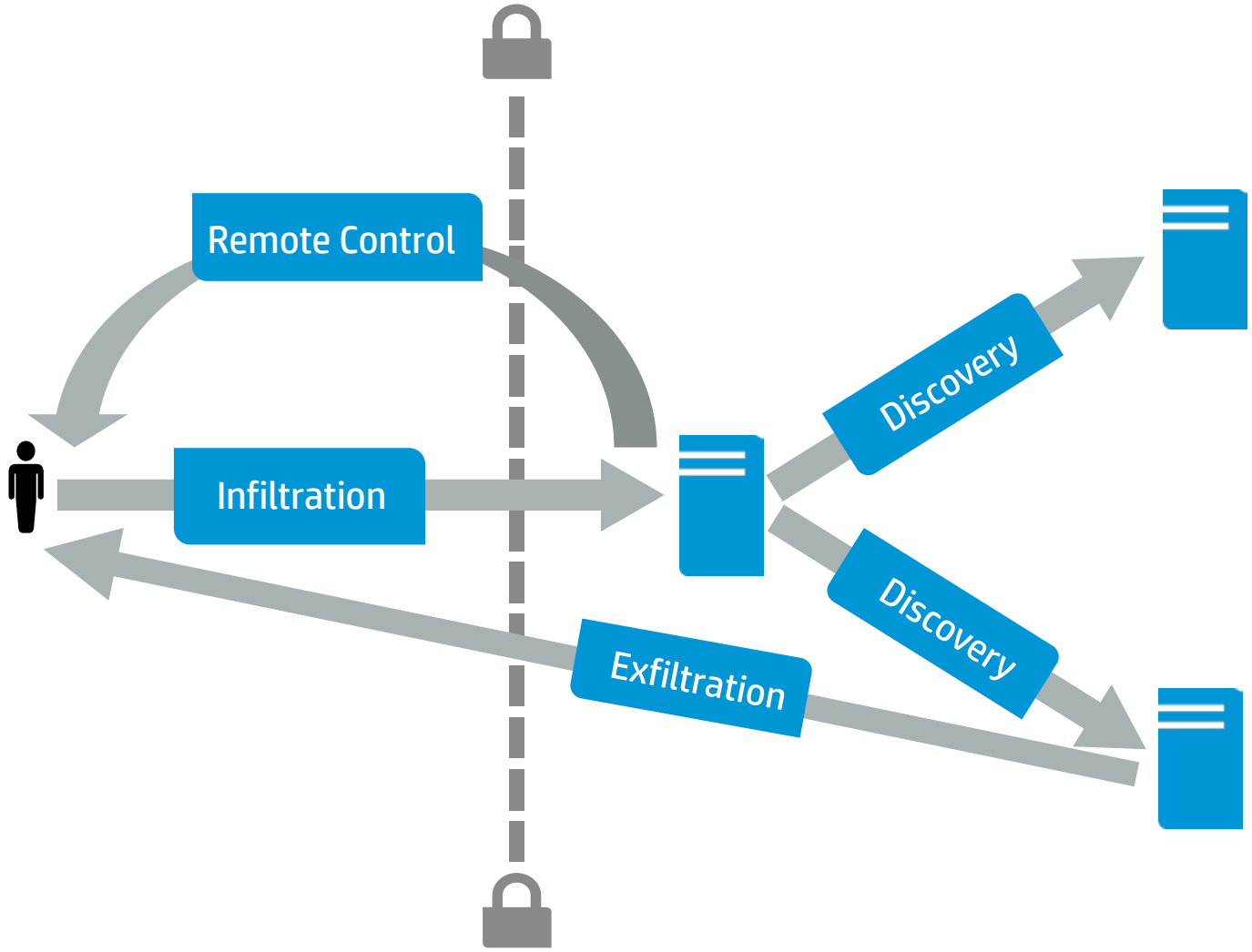
440K+ PCs deployed

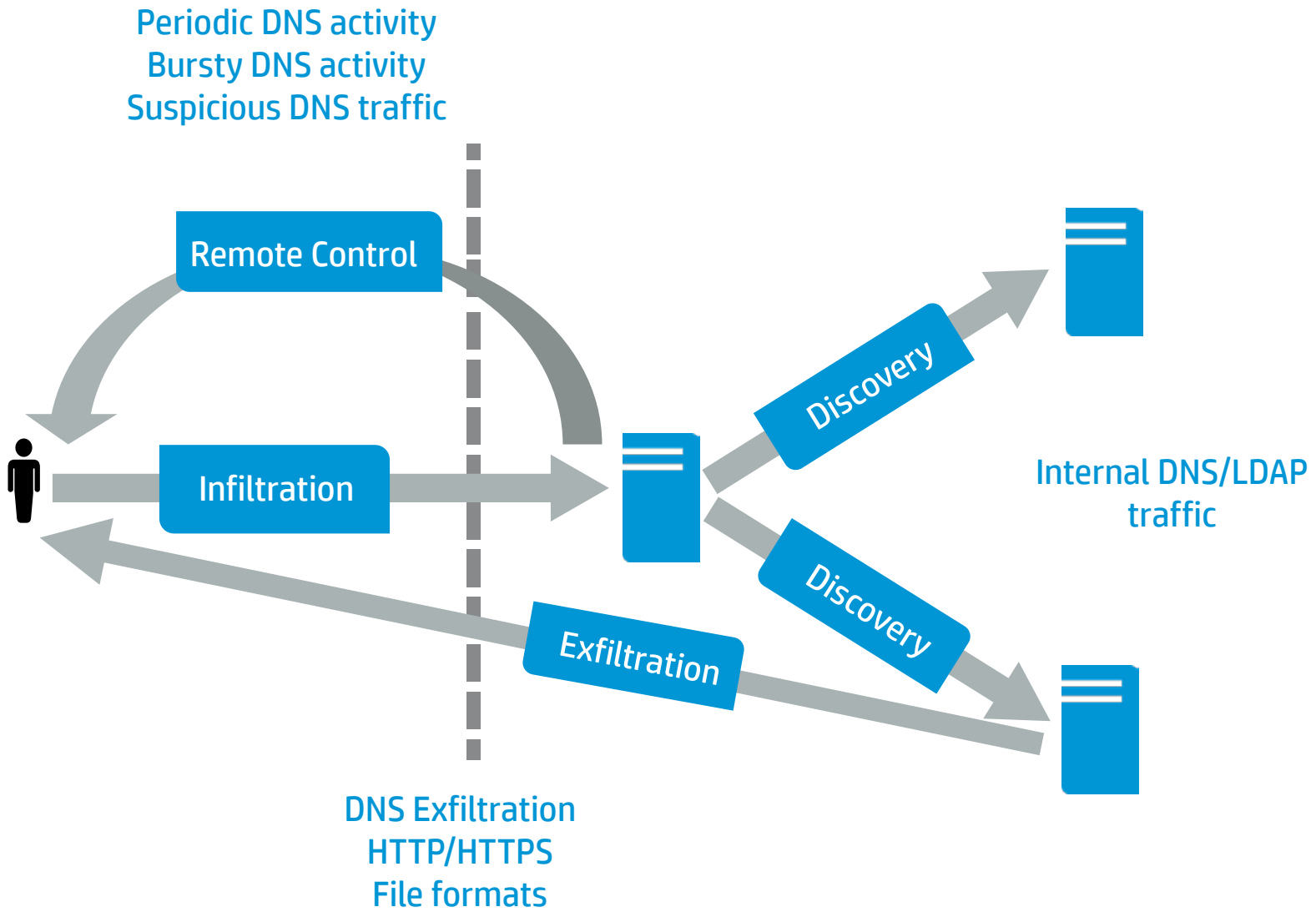
300K+ employees + contractors

Reducing false negatives



Advanced persistent threat (APT) detection





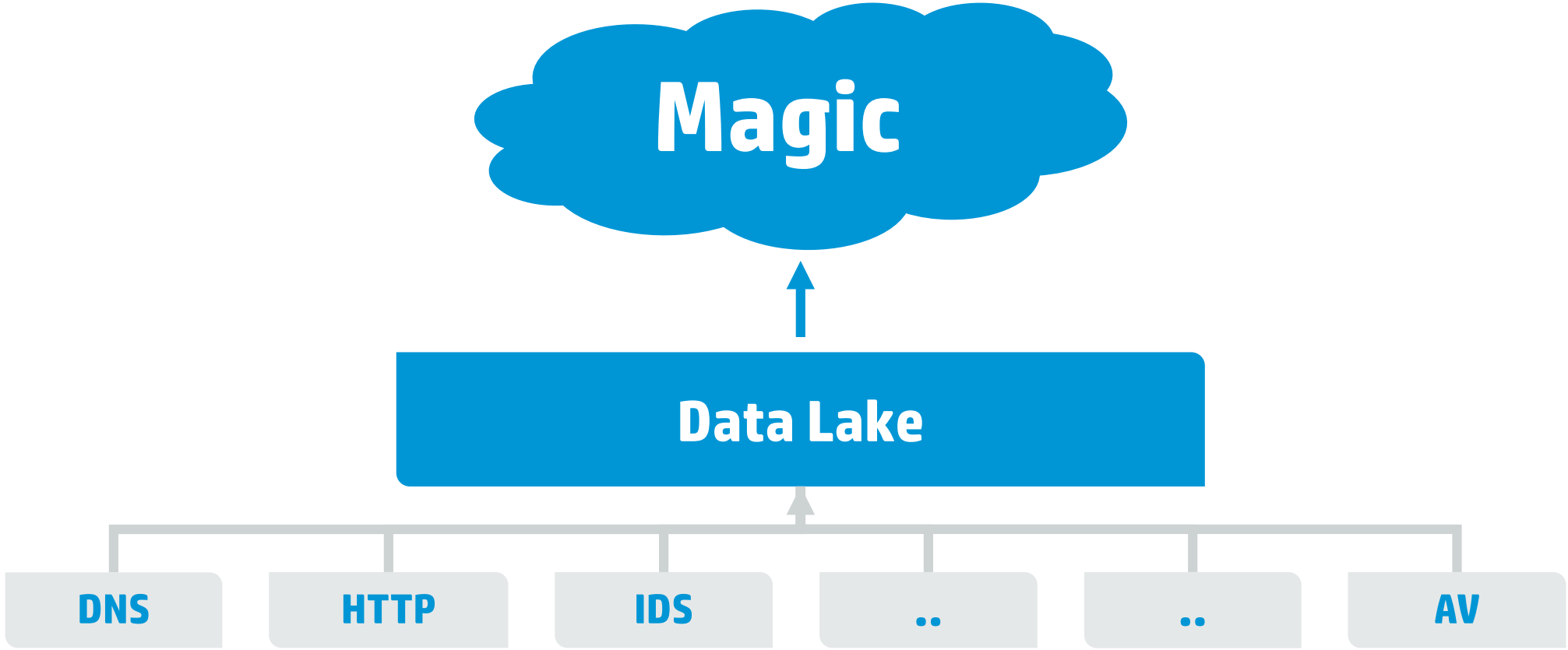
The data analysis approach holds promise

- Compromised account detection
- Lateral movement detection

- Anomalous user behavior
- Insider threat

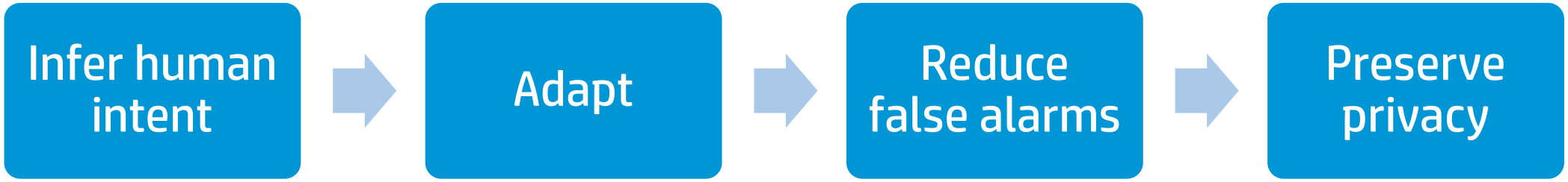
- Preemptive detection – detecting early stages of an attack

Security Analytics



Scalable, Reliable, and Timely Detection

Challenges

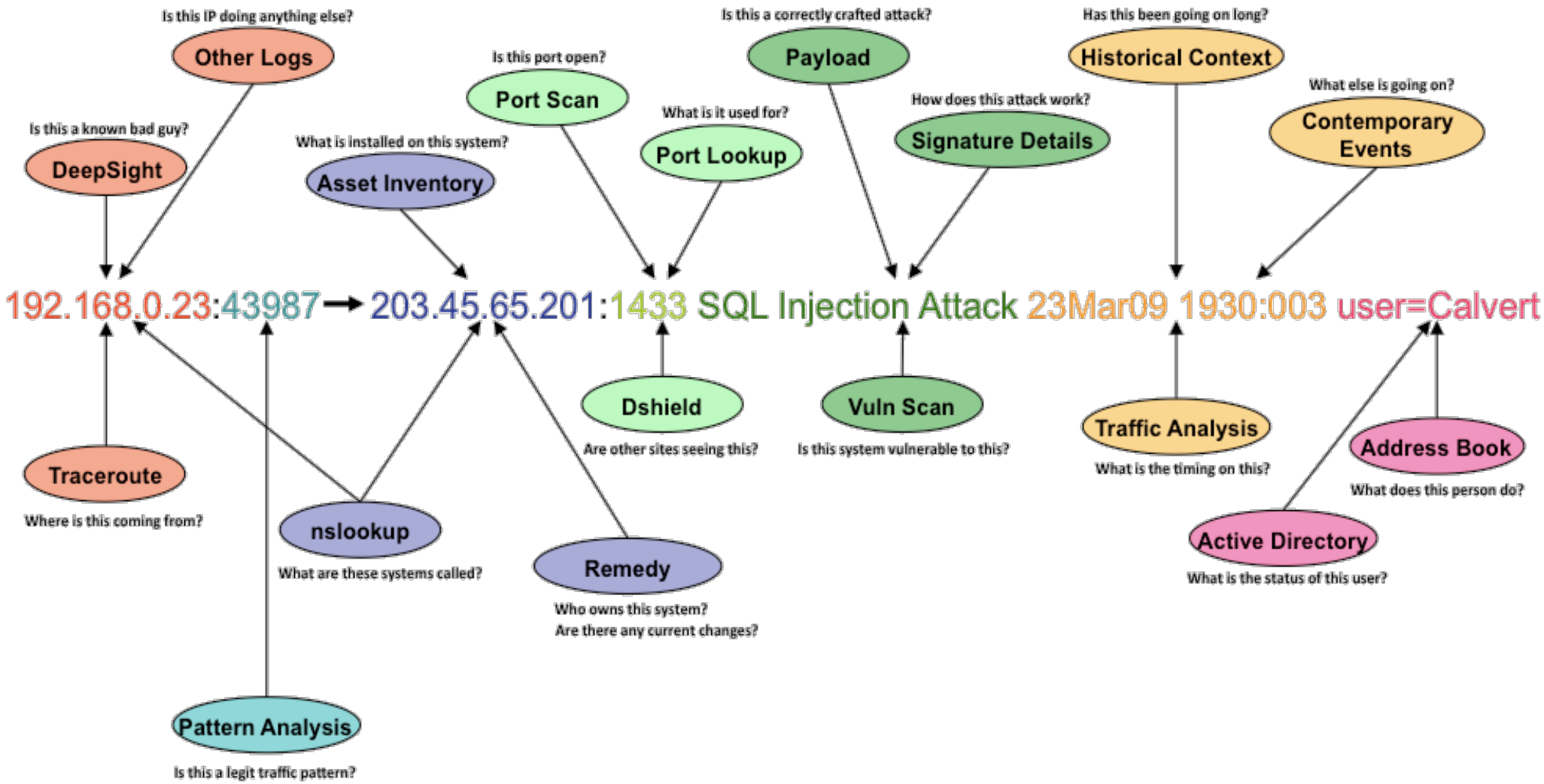


4th Generation Remediation & Recovery

Analyst sees an alert

192.168.0.23:43987→203.45.65.201:1433 SQL Injection Attack 23Mar09 1930:003 user=Calvert

Analyst builds a context



Analyst follows a remediation plan

Quarantine the infected machine

Schedule/ run clean up tools

Schedule/ run reimaging

A few minutes per alert

Target Ignored Data Breach Alarms

Target's security team reviewed -- and ignored -- urgent warnings from threat-detection tool about unknown malware spotted on the network.

http://www.darkreading.com/attacks-and-breaches/target-ignored-data-breach-alarms/d/d-id/1127712

SOCs don’t scale

Repetitive, **Manual**, and **Error Prone**

Machine learning to the rescue???

Context building as a learning problem

Predict the information needed for each alert

- alert characteristics
- source/destination
- protocol
- vulnerability
- …

Remediation plan as another learning problem

Predict remediation action for each alert

- context information
- device characteristics
- user characteristics
- location
- connectivity
- business needs

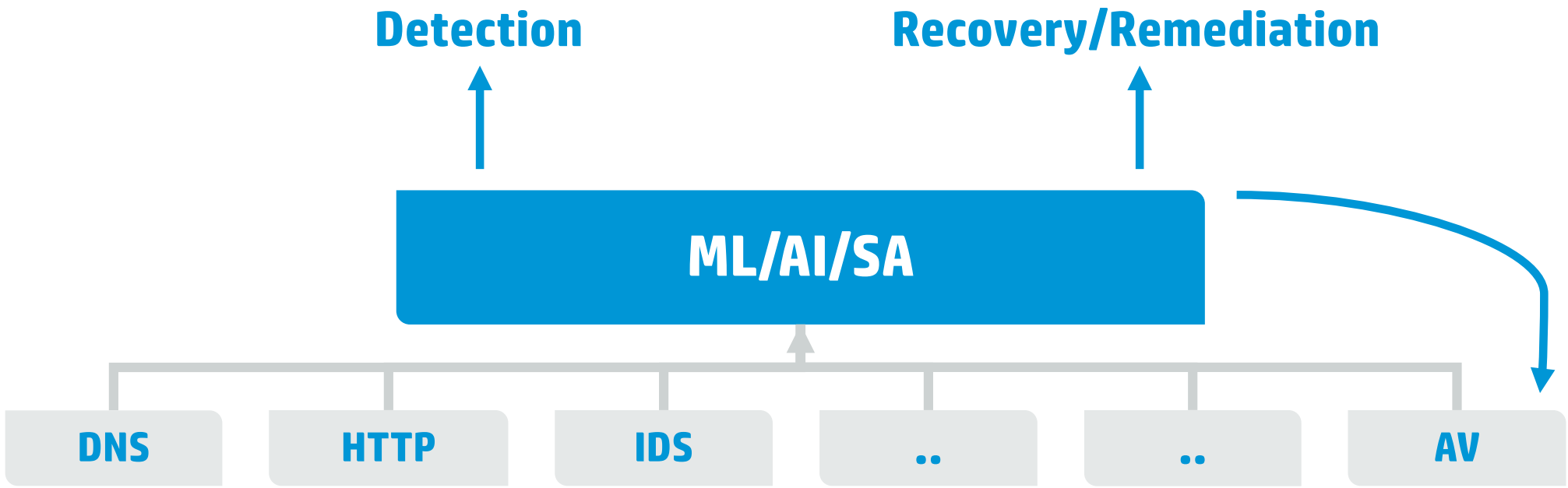
Incorporate analyst feedback

Summary

Current state

- ML has played a key role since early days of enterprise security
- ML helps achieve scale, but FPs/FNs remain
- One of the better tools at our disposal for enterprise security

Future: ML targeted toward enterprise security



Point products: Old problems require new solutions

Targeted new techniques

- False positives/False negatives
- Resource aware – power
- Advances in hardware and systems software

Security analytics and remediation: New problems

Targeted ML techniques for scalable and reliable detection

Holistic approach to detection and remediation

Incorporate user feedback

And finally, be judicious in using ML

Thank You

Pratyusa K. Manadhata

manadhata@hpe.com

