Data Science in Cybersecurity

ALLAN OGWANG
How Vectra applies data science for threat detection

Vectra uses AI to detect attackers in real time and enrich threat investigations with a conclusive chain of forensic evidence.

- **Automates threat detection**: Always-learning behavioral models detect attackers in real-time to enable quick, decisive response and a logical investigative starting point.
- **Empowers threat hunters**: Launch deeper incident investigations detected by Cognito and other security controls, and hunt retrospectively for covert attackers.
- **Exposes hidden attackers**: Network metadata, logs and cloud events are collected, analyzed and stored to reveal hidden attackers in workloads and user/IoT devices.
Attacker behaviors: unifying data science and security research

Attacker behavior models
- High-fidelity detection of things attackers must do
- No signatures: find known and unknown

Security Research
- Identify, prioritize, and characterize fundamental attacker behaviors
- Validate models

Data Science
- Determine best approach to identify behavior
- Develop and tune models
Who is Vectra AI?

- Vectra AI provides automated threat detection to expose hidden and unknown cyberattackers in a network.
- Apply artificial intelligence to seek out the fundamental threat behaviors that attackers simply can't avoid.
Cyberthreats in an enterprise: An advanced attack
Enterprise networks

Firewall creates a separation between inside and outside of the network

Organization firewall
Enterprise networks

Firewall prevents an attacker from connecting to network computers.

Organization firewall
Advanced attack

Organization firewall

Attacker needs a footprint inside the network
Advanced attack

Organization firewall

Infect with malware
Advanced attack

Organization firewall

Malware connects to server hosted by attacker
Advanced attack

Organization firewall

Command-and-control behaviors
Advanced attack
Advanced attack

Organization firewall

Reconnaissance behaviors
Advanced attack

Organization firewall

Lateral movement behaviors
Advanced attack

Exfiltration behaviors

Organization firewall
Progression of an attack

Initial infection → Command and Control → Reconnaissance

Reconnaissance → Data Exfiltration → Lateral Movement → Command and Control
Attack

Data

Machine learning
Different types of learning: Supervised vs. unsupervised

**Supervised Learning**
- Deep Neural Network
- ARTMAP
- Decision Tree
- Random Forest
- Perceptron
- SVM
- KNN
- Naive Bayes
- Logistic Regression

**Unsupervised Learning**
- Deep Neural Network
- MDN
- Deep Autoencoder
- ART
- ARTMAP
- RBE
- Isolation Forest
- One-Class SVM
- GMM
- DBSCAN
- K-Means
- PCA
- LogLinear
- HMM
- Naïve Bayes
- RBE
- Deep Neural Network
- DBN
- DBN
The “no-free-lunch” theorem

No single algorithm performs best for all problems
Choosing the right algorithm: Know your data

- **Concentric Rings**
- **Nested Arches**
- **Loose Clusters**
- **Elongated Clusters**
- **Tight Clusters**
- **No Clusters, Data Field**

### Clustering Algorithms
- K-Means
- Affinity
- Mean Shift
- Spectral
- Ward
- Agglomerative
- DBSCAN
- Birch
- GMM
Choosing the right algorithm: Know your data

Clustering Problems

- Concentric Rings
- Nested Arches
- Loose Clusters
- Elongated Clusters
- Tight Clusters
- No Clusters, Data Field

Algorithms:
- K-Means
- Affinity
- Mean Shift
- Spectral
- Ward
- Agglomerative
- DBSCAN
- Birch
- GMM
Choosing the right algorithm: Know your data

<table>
<thead>
<tr>
<th>Clustering Problems</th>
<th>K-Means</th>
<th>Affinity</th>
<th>Mean Shift</th>
<th>Spectral</th>
<th>Ward</th>
<th>Agglomerative</th>
<th>DBSCAN</th>
<th>Birch</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concentric Rings</td>
<td><img src="concentric_rings" alt="Image" /></td>
<td><img src="concentric_rings" alt="Image" /></td>
<td><img src="concentric_rings" alt="Image" /></td>
<td><img src="concentric_rings" alt="Image" /></td>
<td><img src="concentric_rings" alt="Image" /></td>
<td><img src="concentric_rings" alt="Image" /></td>
<td><img src="concentric_rings" alt="Image" /></td>
<td><img src="concentric_rings" alt="Image" /></td>
<td><img src="concentric_rings" alt="Image" /></td>
</tr>
<tr>
<td>Nested Arches</td>
<td><img src="nested_arches" alt="Image" /></td>
<td><img src="nested_arches" alt="Image" /></td>
<td><img src="nested_arches" alt="Image" /></td>
<td><img src="nested_arches" alt="Image" /></td>
<td><img src="nested_arches" alt="Image" /></td>
<td><img src="nested_arches" alt="Image" /></td>
<td><img src="nested_arches" alt="Image" /></td>
<td><img src="nested_arches" alt="Image" /></td>
<td><img src="nested_arches" alt="Image" /></td>
</tr>
<tr>
<td>Loose Clusters</td>
<td><img src="loose_clusters" alt="Image" /></td>
<td><img src="loose_clusters" alt="Image" /></td>
<td><img src="loose_clusters" alt="Image" /></td>
<td><img src="loose_clusters" alt="Image" /></td>
<td><img src="loose_clusters" alt="Image" /></td>
<td><img src="loose_clusters" alt="Image" /></td>
<td><img src="loose_clusters" alt="Image" /></td>
<td><img src="loose_clusters" alt="Image" /></td>
<td><img src="loose_clusters" alt="Image" /></td>
</tr>
<tr>
<td>Elongated Clusters</td>
<td><img src="elongated_clusters" alt="Image" /></td>
<td><img src="elongated_clusters" alt="Image" /></td>
<td><img src="elongated_clusters" alt="Image" /></td>
<td><img src="elongated_clusters" alt="Image" /></td>
<td><img src="elongated_clusters" alt="Image" /></td>
<td><img src="elongated_clusters" alt="Image" /></td>
<td><img src="elongated_clusters" alt="Image" /></td>
<td><img src="elongated_clusters" alt="Image" /></td>
<td><img src="elongated_clusters" alt="Image" /></td>
</tr>
<tr>
<td>Tight Clusters</td>
<td><img src="tight_clusters" alt="Image" /></td>
<td><img src="tight_clusters" alt="Image" /></td>
<td><img src="tight_clusters" alt="Image" /></td>
<td><img src="tight_clusters" alt="Image" /></td>
<td><img src="tight_clusters" alt="Image" /></td>
<td><img src="tight_clusters" alt="Image" /></td>
<td><img src="tight_clusters" alt="Image" /></td>
<td><img src="tight_clusters" alt="Image" /></td>
<td><img src="tight_clusters" alt="Image" /></td>
</tr>
<tr>
<td>No Clusters, Data Field</td>
<td><img src="no_clusters" alt="Image" /></td>
<td><img src="no_clusters" alt="Image" /></td>
<td><img src="no_clusters" alt="Image" /></td>
<td><img src="no_clusters" alt="Image" /></td>
<td><img src="no_clusters" alt="Image" /></td>
<td><img src="no_clusters" alt="Image" /></td>
<td><img src="no_clusters" alt="Image" /></td>
<td><img src="no_clusters" alt="Image" /></td>
<td><img src="no_clusters" alt="Image" /></td>
</tr>
</tbody>
</table>
Choosing the right algorithm: Know your data

<table>
<thead>
<tr>
<th>Clustering Problems</th>
<th>K-Means</th>
<th>Affinity</th>
<th>Mean Shift</th>
<th>Spectral</th>
<th>Ward</th>
<th>Agglomerative</th>
<th>DBSCAN</th>
<th>Birch</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concentric Rings</td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
</tr>
<tr>
<td>Nested Arches</td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
</tr>
<tr>
<td>Loose Clusters</td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
</tr>
<tr>
<td>Elongated Clusters</td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
</tr>
<tr>
<td>Tight Clusters</td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
</tr>
<tr>
<td>No Clusters, Data Field</td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
<td><img src="no-cluster.png" alt="Image" /></td>
</tr>
</tbody>
</table>
Choosing the right algorithm

No single algorithm performs best for all problems
Select the right option for your data and performance needs
Outline

• Metadata used for threat detection
• Approach to detection
  • Detecting Remote Access Trojans (RATs)
    • Signatures
    • Anomaly detection
    • Random forest
    • Deep learning
• Conclusions
Outline

• Metadata used for threat detection
• Vectra’s approach to detection
  • Detecting Remote Access Trojans (RATs)
    • Signatures
    • Anomaly detection
    • Random forest
    • Deep learning
• Conclusion
Metadata hits the sweet spot for security applications

- Vectra metadata designed with attacker behavior in mind
- All detection models are based on Vectra metadata
  - Metadata includes bytes, protocols, domains, ips
  - Other advanced models are based off enhanced metadata

Increasing data volume with increasing deployment complexity
Example of enhanced metadata: Beaconing behavior

- Beaconing behavior is a common sign of a command and control channel
- Whether a host is beaconing must be inferred based on the host behavior
- By applying machine learning to this raw Vectra metadata we can identify beaconing behavior
- HTTP/S tunnel model was developed using this data to help identify command and control channels
Outline

• Metadata used for threat detection

• Approach to threat detection
  • Detecting Remote Access Trojans (RATs)
    • Signatures
    • Anomaly detection
    • Random forest
    • Deep learning

• Conclusion
Remote Access Trojans (aka external remote access)

Attacker wants to establish manual control over asset inside the network

Firewalls block most inbound connection attempts

So compromised internal asset calls out to “meeting point” and attacker takes over

Examples
Blackshades
Poison Ivy
NOPEN (Shadow Brokers)
WebEx
TeamViewer
**Network Signatures**

- Based on known patterns flag known RATs

- Network
  - URLs, User Agents, Payloads, Domains, IP Addresses, etc

```
 trojan.rules:alert tcp $HOME_NET any -r $EXTERNAL_NET any (msg:"ET TROJAN DarkComet-RAT server join acknowledgement"; flow:to_server,established; dsize:12; content:"|39 34 41 35 41 44 30 41 45 46 36 39|"; flowbits:iset,ET.DarkCometJoin; reference:url, www.darkcometrat.com; reference:url,anubis.iseclab.org/?action=result&task_id=1a7326f61f6ef1ecb4ed4fb3de3f3b8cb&format=txt; classtype:trojan-activity; sid:2013284; rev:3; metadata:created_at 2011_07_18, updated_at 2011_07_18;)
```

- Great for known threats
  - Easily bypassed with changes to the malware
  - Lags behind new changes in malware
Anomaly detection

- Unsupervised
  - Assume a RAT
    - Uncommonly used port
    - Uncommon destination
    - Uncommon hour

- Everything is “uncommon”
- New ports everyday
- New domains everyday
- Time is not a great signal
- Will likely alert you to the event
  - But how do you find true event in this haystack?
Data is king – How Vectra sees RATs

- A RAT is not static
  - All behavior happens in time
    - Commands are issued
    - Information is received

- Incremental flow between a RAT server and client host
Machine learning first pass – Random forest

- A random forest is a collection of decision trees
- Not likely a single perfect decision tree model
  - Randomly look at features
  - Randomly look at data
  - Build several models
- Each model votes
  - Every model does not need to be right
  - But more that vote more confidence in decision
Random forest for RATs

- **Featurize the timeseries window – 20+ Features**
  - Data and packet client / server ratios
  - Consistency of the client / server data
  - Frequency where the server breaks silence
  - Total session length
  - Entropy of the session
  - etc…

- **Observe multiple windows and trigger on convergence**

- **Model provided value**
  - Alerted on large % of known RATs but not all
  - Did not trigger on all known RAT behaviors

- **Issues**
  - Did not properly represent the temporal nature
    - One sequence impacts the next
  - Human driven features missed behaviors
    - Can guess and test but can never be sure
Deep learning

Digit labels
0, 1, 2, 3, 4, 5, 6, 7, 8, 9

Phonemes
dh, aw, s, ax, n, d, ...

Mouse Movements
(right, left, up, down)
Deep learning:
Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM)

- **RNN**
  - Similar to feedforward NN
  - Recurrent connections == Memory

![Diagram of RNN and LSTM](image-url)
Deep Learning:
Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM)

- **LSTM**
  - Similar to RNNs
  - Replace simple neurons with LSTM blocks
  - Prevents “vanishing gradient” problem
  - Capable of learning long-range temporal dependencies
Deep learning:
Model training strategy

Model training
- Framework: TensorFlow
- Model: RNN (LSTM cell)
- Train on AWS w/ NVIDIA v100 GPUs
Deep learning:
Learning representations

- Map input time-series to embedded representation
- Classify the embedding as RAT / not RAT
- Observe convergence in classification
- Report behavior
Outline

• Metadata used for threat detection
• Approach to threat detection
  • Detecting Remote Access Trojans (RATs)
  • Signatures
  • Anomaly detection
  • Random forest
  • Deep learning
• Conclusion
Know your model

- In security, the problems are various and complex; data are sometimes unavailable, sometimes imbalanced

- Many approaches are available, but not all will perform equally well

- No free lunch! Understand the problem and choose the right model
  - Supervised or unsupervised?
  - Classification or regression?
  - Temporal factors are crucial

- Data science is not just about math. Attackers can only be detected through conjunction of deep knowledge of machine learning and security
Data science – first as an art, then apply the science
VECTRA
Security that thinks.
Detection lifecycle

Security researchers

- Collect advanced attack samples
- Come up with advanced attacks
- Abstract the behavior and form a theory
- Collect positive and negative samples
- Extract features out of the samples
- Work the theory on offline data
- Refine into detection model

Security Researchers + Data Scientists

- Improve and redeploy
- Deploy and test on live data
- Review results
- Design UI
- Develop UI
- Put detection into production
- Check efficacy; improve where necessary

Product Designer

Developers
Model Development Philosophy – Research to Production

1. Report an advanced attack behavior
   • Methodology and data sources are irrelevant

2. Provides the relevant context to investigate
   • Necessary information for rapid validation

3. Improvable over time
   • Trackable efficacy

4. Minimal noise and high coverage
   • Meets initial recall and precision requirements