Possibilities

#CiscoLive
The Future of Security Analytics

TK Keanini, Distinguished Engineer, Advanced Threat
@tkeanini
DGTL-BRKSEC-2068
Agenda

• Introduction
• Security Analytics Fundamentals
• Telemetry, Techniques and Outcomes
• Artificial Intelligence and Machine Learning
• Dark Data, Encrypted Traffic Analytics and Behavioral Analytics
• The Future of Security Analytics
• Conclusion and Takeaways
Hello My Name Is TK Keanini
(Pronounced Kay-Ah-Nee-Nee)

Brief History in a Nutshell
Fundamentals
Security Analytics versus Other Analytics

Security Analytics focus on augmenting or automating these functions:

- Incident Responder
- Security Analyst
- Security Operations
- Threat Hunter
- Compliance and Policy
- Business Continuity
- Cybercrime fighting
te·lem·e·try

noun. The process of recording and transmitting the readings an instrument

It is data that represents change taking place within an observable domain

⇒ All Telemetry Is Data but, Not All Data Is Telemetry
Telemetry (Sensing Change)

Any Data Set That is Useful in the Analytical Outcomes
Tracking a Users Behavior

Access logs are terse and leave gaps in the narrative

- After logging in, what did they do?
- What protocols were used?
- Were they coming in front another machine via RDP?
- How much data was exchanged?
- Was the communication encrypted?
- What other sessions were active at the time of this session?
- What was the integrity of the device prior and during the session?
- How does this current activity compare to the historical baseline?
Gartner Hype Cycle for Emerging Technologies 2017

Source: Gartner (July, 2017)

- **Innovation Trigger**
- **Peak of Inflated Expectations**
- **Trough of Disillusionment**
- **Slope of Enlightenment**
- **Plateau of Productivity**

**MACHINE LEARNING**
What Did We Do Before Machine Learning?

Simple Pattern Matching

Statistical Methods

Rules and First Order Logic (FoL)

Use in Combination with Machine Learning
Why Is Machine Learning So Useful In Security?

**Static**
With limited variability or is well-understood

**Evolving Security**
The security domain is always evolving, has a large amount of variability, and is not well-understood
Why Use Machine Learning for Security Analytics

- Advanced Threat inherently is not static and evolving
- The data sets are often very large at scale (the 1% that matters)
- The most advanced threats are not well-understood and novel
- Machine Learning is not magic and still has problems!

The key is to use its strengths along side other techniques in a analytics pipeline that is hard to evade and delivers the most fidelity

Example Stack: Encrypted Traffic Analytics
Example: Encrypted Traffic Analytics

- Outcomes
  - Detection of Malware without Decryption
  - Cryptographic Compliance

- Synthesis/Analytics
  - Analytics Pipeline of Diverse Methods

- Telemetry
  - Initial Data Packet
  - Sequence of Packets Lengths and Times
  - Flow Start
  - Time
  - Observables
Artificial Intelligence & Machine Learning
“Field of study that gives computers the ability to learn without being explicitly programmed.”

Arthur Samuel’s definition of machine learning in 1959
Machine Learning is one of the fields in Artificial Intelligence, where machines learn to act autonomously, and react to new situations without being pre-programmed. It is about designing algorithms that allow computers to learn aimed at some outcome.

- Learn to identify faces, learn to drive a car, etc
- Learning to detect malware, learning to identify a threat actors, etc.

**Machine Learning Big Picture**

- **Supervised Learning**
  Examples: Classification, Regression

- **Reinforcement Learning**

- **Unsupervised Learning**
  Examples: Clustering
  Dimensionality Reduction

**Artificial Intelligence**
Supervised

- Used when you know the question you are trying to ask
- And have examples of it being asked and answered correctly
- If you can phrase a problem as 'we know this is right, learn a way to answer more questions of this type'

Unsupervised

- Less structured & know little about the structure
- You don't have answers and may not fully know the questions
- Unsupervised techniques act as a tool for gaining an understanding of how elements of the set relate to each other

Reinforced Learning

- Sometimes called RL and is really the 'other' category
- Learns the optimal solution by repeated trial and error
- If you can formalize your problem even at a level above even what supervised learning calls for then RL has some powerful tools for solving it
Ground Truth Used In Supervised Learning

- The 'Ground Truth' is the pairing of example questions and answers
- If you can phrase a problem as 'we know this is right, learn a way to answer more questions of this type'
- Success depends greatly on the dataset expressing the Question -> Answer mapping
“Field of study that gives computers the ability to learn without being explicitly programmed.”

“Field of study that gives computers the ability to be implicitly programmed.”
Training Classifiers

Training Data

Machine Learning Algorithm

New Data → Classifier → Prediction
Efficacy and Measurement
What Is At Stake Matters

Because you watched Deadpool, you might like...

Deadpool  
X-Men: First Class  
The Flash  
Captain America: The First Avenger
How Did The Machine Come To That Conclusion?

“The Explainability Problem”

Normal Workflow
CFO daily calendar

Irregular Activity
Machine detects “suspicious” activity and suggests remediation

Quarantined
However, Machine cannot articulate *why* it wants to remediate

Loss of Time and Resources
Dark Data
The Network Traffic Is Encrypted

90% of the network traffic today is encrypted

Threat actors are also using network encryption

All encrypted sessions begin unencrypted
Dark Data Is Here

Direct Inspection
No Longer Possible

Inference
Becomes Our Strategy

We can accurately infer what we cannot observe.
**TLS Is the New TCP**

All Encrypted Sessions Begin Unencrypted

Encrypted Sessions

---

Digital Fingerprints
We Can Still Detect Threats Without Decryption

- Unencrypted Observables
- Digital Fingerprints
- Attribution
- Helpful Analytics
Intro to Fingerprinting
What is a Fingerprint?

An impression left by the friction ridges of a human finger.

Valuable Qualities

- Detailed & unique to every person
- Difficult to alter
- Durable over the life of the person
Fingerprinting

Benefits
- Size (how much of the world do I represent?) \( x \) / domain
- Accuracy of the inference

Implementation Challenges
- Labor intensive process
- Not 100% Comprehensive
- Precision
Digital Network Fingerprints

Objective
To directly observe a pattern in transit that accurately identifies the class of process responsible for the exhibition

<table>
<thead>
<tr>
<th>Directly Observable (&amp;&amp;)</th>
<th>Class of Object</th>
<th>Instance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has Feathers</td>
<td>Duck</td>
<td>Not Daffy Specifically</td>
</tr>
<tr>
<td>Has Webbed Feet</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Has Yellow Bill</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Duck Example (abductive reasoning)
Research based on
Blake Anderson and David McGrew
Security and Trust Organization at Cisco

**TLS Fingerprinting**

https://github.com/cisco/joy

https://github.com/cisco/joy/blob/master/fingerprinting/resources/fingerprint_db.json.gz
Transport Layer Security (TLS):
is the protocol used to secure network traffic

The TLS parameters offered in the negotiations are observable and can provide attribution to a process.

Direct Observable

- Observable 848
- Observable 232
- Observable 135

Process Class

- "Google Chrome (43.0.2357.130 64-bit OSX)"
- "Malware: TBot / Skynet Tor Botnet"
- "Metasploit SSL Scanner"
From Observations to Attribution

Network Visibility Module (NVM)
- IPFIX-Based Record (Source, Destination IP, etc)
- Unique Device ID (correlate records from same endpoint device)
- Device Name (bsmith-WIN) and OS Version (Window 7)
- Domain\User Name (Amer\bsmith)
- Account Type (win Admin/Standard/Guest  mac: Standard/Root)
- Local DNS (starbucks.com) Target DNS (amceco.box.com)
- Mac Address (extension to Interface Records)
- Interface (Intel ® Dual Band Wireless)
- Process/Container Name (iexplorer.exe) Process ID (hash)
- Parent Process Name (foobar.exe) Parent ID (hash)

Encrypted Traffic Analytics (IDP Data)
- Encrypted Traffic Analytics Telemetry (Initial Data Packet (IDP))
- Stealthwatch Flow Sensor
- Opensource Joy Package

Machines help us answer: What deterministic patterns can be directly attributed to a process?

Endpoint Activity

Network Activity
How does it work?

The Cisco TLS Fingerprint Database (FPDB) is maintained and updated daily.

Any Cisco product witness to these fields in the packet can match against the FPDB.

ProtocolVersion: 0301
CipherSuites: 0035...0003
Extensions: None

Process: vmnet-natd
Category: virtual_machine
sha256: 1466...5B8A
OS: Mac OS X
OSversion: 10.13.6
OSedtion: High Sierra
How To Evaluate a Fingerprint Database?

How comprehensive is the map to the territory?

Universe of Processes

Observable → attribution
Observable → attribution
Observable → attribution

Specificity of the attributions?

Google Chrome

less specific

more specific

Google Chrome
(43.0.2357.130 64-bit OSX)
Building a Better Fingerprint Database
(counts as of June 04 2019)

Using machines, we have automated the process of identifying observable TLS patterns that can be attributed to processes.

Cisco has the most comprehensive data base growing every day!
## Comparing to Other TLS Fingerprinting Methods

<table>
<thead>
<tr>
<th>Database</th>
<th>Size</th>
<th>Automatically Updated</th>
<th>GREASE Support</th>
<th>Static Extension Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cisco</td>
<td>4,500+</td>
<td>Yes</td>
<td>Yes</td>
<td>supported_groups, ec_point_formats, status_request, signature_algorithms, application_layer_, protocol_negotiation, supported_versions, 16 others</td>
</tr>
<tr>
<td>Kotzias et al.</td>
<td>~1,684</td>
<td>No</td>
<td>Discards Locality</td>
<td>supported_groups, ec_point_formats</td>
</tr>
<tr>
<td>JA3 (Used by DarkTrace)</td>
<td>157</td>
<td>No</td>
<td>Discards All Data</td>
<td>supported_groups, ec_point_formats</td>
</tr>
<tr>
<td>Fingerprint TLS</td>
<td>409</td>
<td>No</td>
<td>No</td>
<td>supported_groups, ec_point_formats, signature_algorithms</td>
</tr>
</tbody>
</table>
Quick Summary of TLS Fingerprinting

- Every domain (not just TLS) can leverage fingerprinting methods
- All encrypted sessions begin unencrypted
- By analyzing endpoint and network activity together we can find deterministic patterns that fingerprint processes
- Cisco currently has the most robust TLS fingerprint database in terms of size and precision
The Limitations of Fingerprinting
## TLS Fingerprints Point to a Process

<table>
<thead>
<tr>
<th>Direct Observable</th>
<th>Process Class</th>
<th>Threat Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observable 232</td>
<td>&quot;Malware: TBot / Skynet Tor Botnet&quot;</td>
<td>Malicious</td>
</tr>
<tr>
<td>Observable 135</td>
<td>&quot;Metasploit SSL Scanner&quot;</td>
<td>Benign</td>
</tr>
<tr>
<td>Observable 848</td>
<td>&quot;Google Chrome (43.0.2357.130 64-bit OSX)&quot;</td>
<td>Unaccounted</td>
</tr>
</tbody>
</table>

Attribution to a process class is only the first step, we still need to understand what it means to the business.
Behavioral Analytics
Signatures vs. Behavioral Detection

If you see pattern X, **sound the alarm**

vs

A set of actions were performed, **sound the alarm**
The Network is Where Computers Behave

Behavioral Telemetry

PRIVATE

Hybrid

PUBLIC
A Chef’s Knife

- Used for food preparation in kitchens
- Sharp blade of a particular shape
- Has a handle used by role of kitchen worker
A Chef’s Knife

Beaviors

• **Behavior 01:** Monday night, the chef used it to prepare meals

• **Behavior 02:** Tuesday day, it was used as a murder weapon

• **Behavior 03:** Tuesday night, a passenger was removed from a flight because this object is not allowed on airplanes
Behavioral Profiles

NAME

Duck

KNOWN

- Quacks
- Waddles
- Eats Grass, Insects, etc.
- Flies
- Other Behaviors

NOVEL

Roar?!
Behavioral Analytics

TLS Fingerprinting

Observable 232 attributed to Process

Behavioral Inferences

{Behavioral Profiles}

- Clients Connect to it to print
- Traffic volumes have a distinct pattern
- Clients Connect for management
- Connection to AWS!
- Any activity not yet known!
Levels of Inference

**LEVEL 1 INERENCE**

**TLS Fingerprinting (Labeling)**

- Observable 232 attributed to "Malware: TBot / Skynet Tor Botnet"
- Observable 135 attributed to "Metasploit SSL Scanner"
- Observable 848 attributed to "Google Chrome (43.0.2357.130 64-bit OSX)"
- Observable 800 attributed to "Google Chrome (version with CVE)"

**LEVEL 2 INERENCE**

- Blacklist
- CVE ####
- Known Bad
- Whitelist

**Policy/Compliance Violation**
Levels of Inference

**DIRECT ATTRIBUTION**
- Observations
- Processes
- Processes
- Processes
- Processes
- Processes
- Processes

**LEVEL 1 INFERENCE**
- OS version
- Application version
- Malware
- Adware
- App v2.0.56

**LEVEL 2 INFERENCE**
- Global Threat Actor Campaign
- Policy Violation
- CVE Vulnerability
- Compliance Violation
- Misconfiguration
- Known Behavioral Model

**LEVEL 3 INFERENCE**
- Anomalous Behavior
Outliers & Novelty

Objects & Behavior
Known *a priori*

Known Bad

If This Is **Good**, Then What Is This?

Derived from first modeling the good

Known Good
Quick Summary of the Analytical Pipeline

Fingerprinting → Classification → Categorization → Behavioral Modeling → Behavioral Analytics
The Problem with Numbers
It is not your fault that you don't understand this.
Numbers Help Us Group Things

- **Credit-worthiness Class**: 720
- **Legal to Drink/Legally Drunk**: 21+
- **Weight Class**: 165
- **Socioeconomic Class**: $98,000
- **Age Class**: 56

Given a number, within a social context, we are able to infer membership to a set

*The terms ‘Set’ and ‘Class’ are synonymous in this presentation*
Syntax and Semantics

Numbers digitize certain aspects of an observable domain. They also help ignore what is not being counted!

Unlike the physical domain, before we can count things in the information domain, we must all agree on what is being counted.

The challenge is that we don’t share the same domain expertise and understanding across an enterprise.

Number systems are dependent on social processes that institutionalize semantics. They often fall short when asked to support multiple perspectives and points of view.
Future Security Analytics
Direct Versus Indirect Observations

Old Method

Client → Observations → Server

New Method

Client → L7 App/User → L3/L4 Overlay/Underlay → App → Observations → Server
Late-Binding Modeling to Detect Security Events

Dynamic Entity Modeling

<table>
<thead>
<tr>
<th>Collect Input</th>
<th>Perform Analysis</th>
<th>Draw Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP Meta Data</td>
<td>Dynamic Entity Modeling</td>
<td>Role</td>
</tr>
<tr>
<td>System Logs</td>
<td></td>
<td>What is the role of the device?</td>
</tr>
<tr>
<td>Security Events</td>
<td></td>
<td>Group</td>
</tr>
<tr>
<td>Passive DNS</td>
<td></td>
<td>What ports/protocols does the device continually access?</td>
</tr>
<tr>
<td>External Intel</td>
<td></td>
<td>Consistency</td>
</tr>
<tr>
<td>Vulnerability Scans</td>
<td></td>
<td>What connections does it continually make?</td>
</tr>
<tr>
<td>Config Changes</td>
<td></td>
<td>Rules</td>
</tr>
<tr>
<td></td>
<td>Dynamic Entity Modeling</td>
<td>Does it communicate internally only?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Forecast</td>
</tr>
<tr>
<td></td>
<td></td>
<td>How much data does the device normally send/receive?</td>
</tr>
</tbody>
</table>

System Logs

Security Events

Passive DNS

External Intel

Vulnerability Scans

Config Changes

Vulnerability Scans

Forecast
Classify the Observable World and Infer the Rest

- Threat Actor Activity
- Weird Stuff (but not threat related)
- Normal Activity
Multi-Layer Analytical Pipeline

Cascade of specialized layers of **Machine Learning** algorithms

- Billions of connections
- Anomaly Detection and Trust Modeling
  - Statistical Methods
  - Information-Theoretical Methods
  - 70+ Unsupervised Anomaly Detectors
  - Dynamic Adaptive Ensemble Creation
- Event Classification and Entity Modeling
  - Multiple-Instance Learning
  - Neural Networks
  - Rule Mining
  - Random Forests
  - Boosting
  - ML: Supervised Learning
- Relationship Modeling
  - Probabilistic Threat Propagation
  - Graph-Statistical Methods
  - Random Graphs
  - Graph Methods
  - Supervised Classifier Training
Security that Shows its Work

- Oct. 3: Spam tracking #CSPM02
- Oct. 4: C&C URL
- Oct. 15: Anomalous http
- Oct. 16: Heavy uploader Dropbox.com
- Oct. 25: Malicious http
- Oct. 28: Recurring

Malware: Sality
Dec. 9 | 28 days
We Track the Escalation Over the Lifetime of the Threat Actor
Role-Based Behavioral Detection

Printers should act like printers

Vulnerability scanners should act like scanners

If your printer is acting like a vulnerability scanner we have a problem

Every endpoint on the network should play a role
Thinking in Sets/Class and Membership

There are 3 blue triangles

Triangle

Blue

3

...is a member of the intersection of the set Blue, the set Triangle, and the set Three
Reasoners
(Side Step the Numbers Problem with First Order Logic)

Jane hasMaidenName Smith

Jane Smith Female

Jane Smith Married

Jane Smith SirName:Smith

Semantic Models

DOMAIN hasMaidenName RANGE

Female hasMaidenName

Married hasMaidenName

hasMaidenName SirName
Competency Questions

- Outcomes
- Analytics
- Synthesis
- Telemetry
Competency Questions

Outcomes
Analytics
Synthesis
Telemetry
While syntax can be right or wrong, analytical outcomes are helpful or not helpful to you
How Helpful Was This Alert?

In the end, it is not the math that matters, it is you the customer that matters!

<table>
<thead>
<tr>
<th></th>
<th>Stealthwatch Cloud Alerts Marked Helpful by Customers (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q3 FY2018</td>
<td>95%</td>
</tr>
<tr>
<td>Q4 FY2018</td>
<td>95%</td>
</tr>
<tr>
<td>Q1 FY2019</td>
<td>89%</td>
</tr>
<tr>
<td>Q2 FY2019</td>
<td>95%</td>
</tr>
<tr>
<td>Q3 FY2019</td>
<td>92%</td>
</tr>
<tr>
<td><strong>Rolling Average</strong></td>
<td><strong>93%</strong></td>
</tr>
</tbody>
</table>
What to Ask Your Vendor

- How are you applying Machine Learning in your product and why?
- How do you measure its effectiveness?
- Regarding supervised learning, what are you using for ‘ground truth’?
- What non-machine learning are you using and why?
- What papers or open-source have you published regarding your analytics?
- What detection in your product is based on known lists (a priori data)?
- What detection in your product is based on behavioral methods?
- For the ML based assertions, what entailments are provided?
Good Principals for Security Analytics

- Be pragmatic
- Always provide entailments
- Favor an analytical pipeline over single technique
- Measure helpfulness, not mathematical accuracy
- Be Transparent with your science, publish papers and open source
Learn More....

- Cisco Stealthwatch Enterprise
- Cisco Stealthwatch Cloud
- Encrypted Traffic Analytics
Basic References
Blogs

- Detecting Encrypted Malware Traffic (Without Decryption)
- Learning Detectors of Malicious Network Traffic
- Transparency in Advanced Threat Research
- Turn Your Proxy into Security Device
- Securing Encrypted Traffic on a Global Scale
- Closing One Learning Loop: Using Decision Forests to Detect Advanced Threats
- The State of Machine Learning in 2019
Make Your Head Hurt Reading Material

- Identifying Encrypted Malware Traffic with Contextual Flow Data, Blake Anderson and David McGrew, AISEC ‘16

In the past 12 years, we have published more than 50 papers (I can send you the PDF listing, just let me know)
Thank you