CISCO

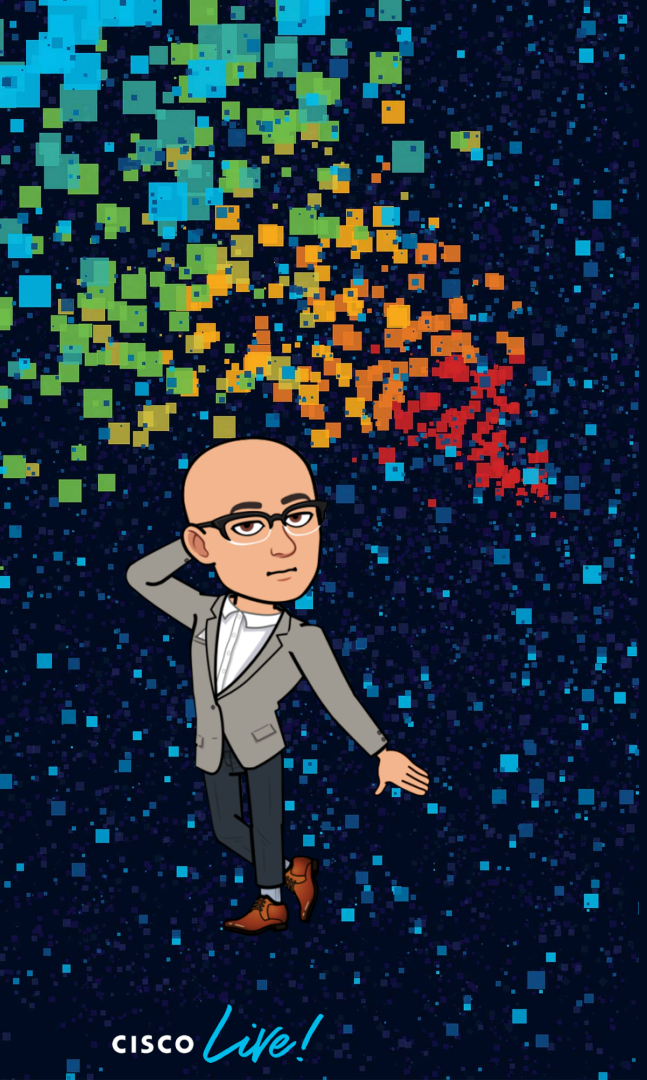
Possibilities

#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

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Hello My Name Is TK Keanini

(Pronounced Kay-Ah-Nee-Nee)

Broderbund

Cisco Systems

Morgan Stanley

nCircle Proactive Network Security

Lancope

cisco

Brief History in a Nutshell

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Fundamentals

Security Analytics versus Other Analytics

- Outcomes
- Synthesis/ Analytics
- Telemetry

Security Analytics focus on augmenting or automating these functions

- Incident Responder
- Security Analyst
- Security Operations
- Threat Hunter
- Compliance and Policy
- Business Continuity
- Cybercrime fighting

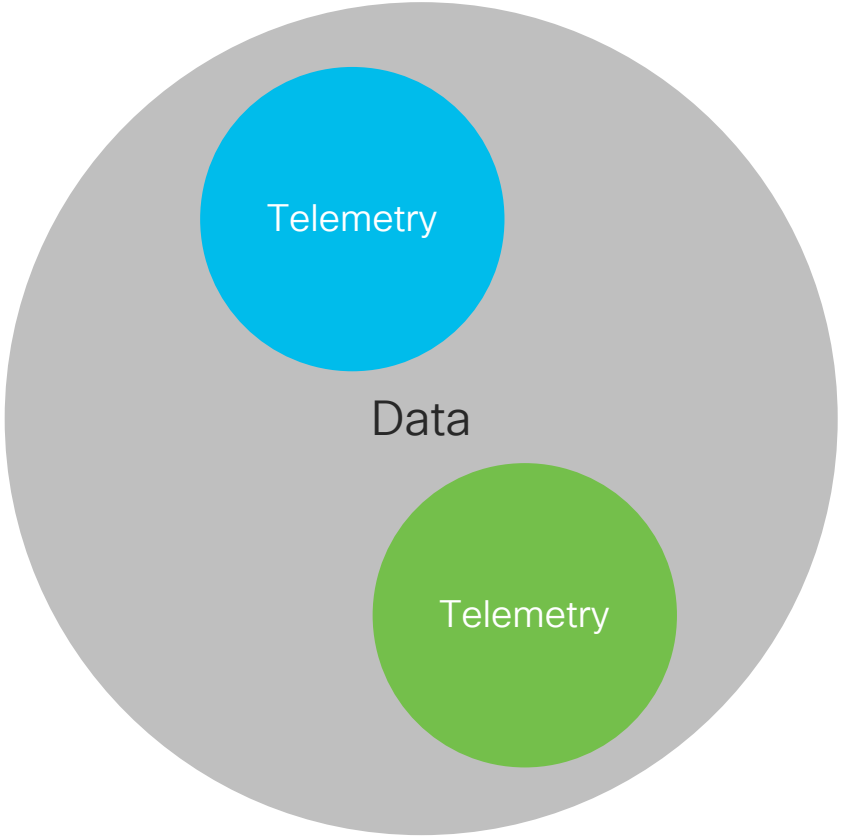
© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

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



© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Telemetry (Sensing Change)

Router	Switch	Firewall	Cloud Native
ISR | CSR | ASR | WLC	Catalyst IE ETA Enabled Catalyst	ASA | FTD | Meraki	AWS Telemetry Google GCP Telemetry Azure

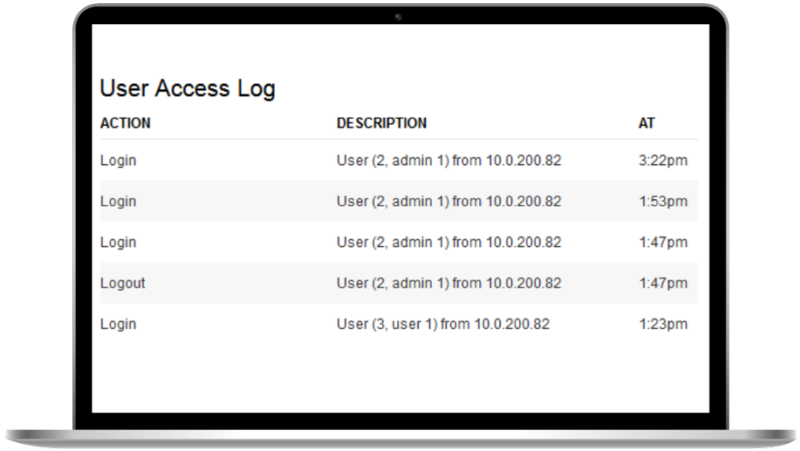
Web	Endpoint	Policy and User Info	Other
Web Security Appliance (WSA)	AnyConnect AMP	Identity Services Engine (ISE) Duo	Stealthwatch Flow Sensor Observable Network Appliance

Any Data Set That is Useful in the Analytical Outcomes

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Tracking a Users Behavior

Access logs are terse and leave gaps in the narrative



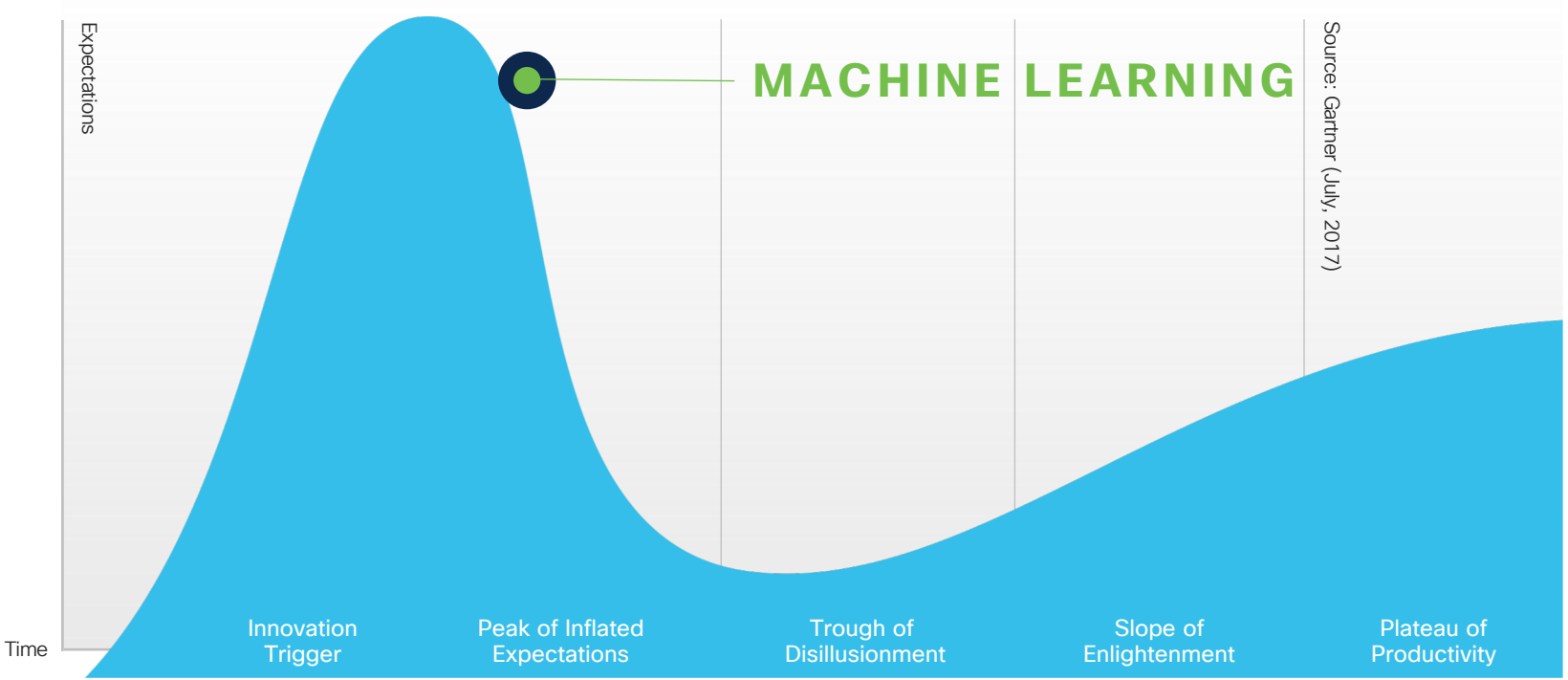
User Access Log

ACTION	DESCRIPTION	AT
Login	User (2, admin 1) from 10.0.200.82	3:22pm
Login	User (2, admin 1) from 10.0.200.82	1:53pm
Login	User (2, admin 1) from 10.0.200.82	1:47pm
Logout	User (2, admin 1) from 10.0.200.82	1:47pm
Login	User (3, user 1) from 10.0.200.82	1:23pm

- 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?

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

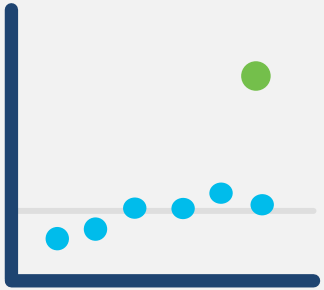
Gartner Hype Cycle for Emerging Technologies
2017



© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

What Did We Do Before Machine Learning?

Simple Pattern Matching



Statistical Methods

Rules and First Order Logic (FoL)

Use in Combination with Machine Learning

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

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

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

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!

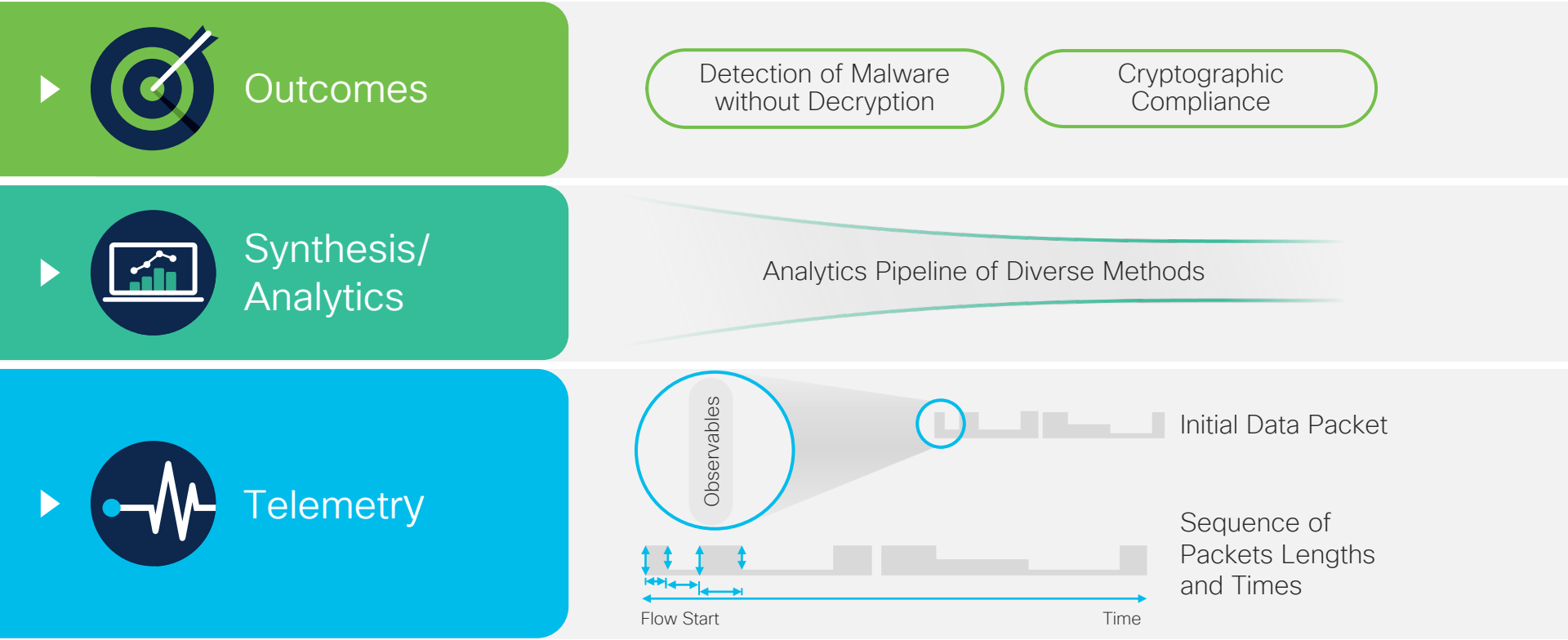
The key is to use its strengths along side other techniques in a analytics pipeline that is hard to evade and delivers the most fidelity

https://blogs.cisco.com/security/the-state-of-machine-learning-in-2019

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Example Stack: Encrypted Traffic Analytics

Example: Encrypted Traffic Analytics



© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

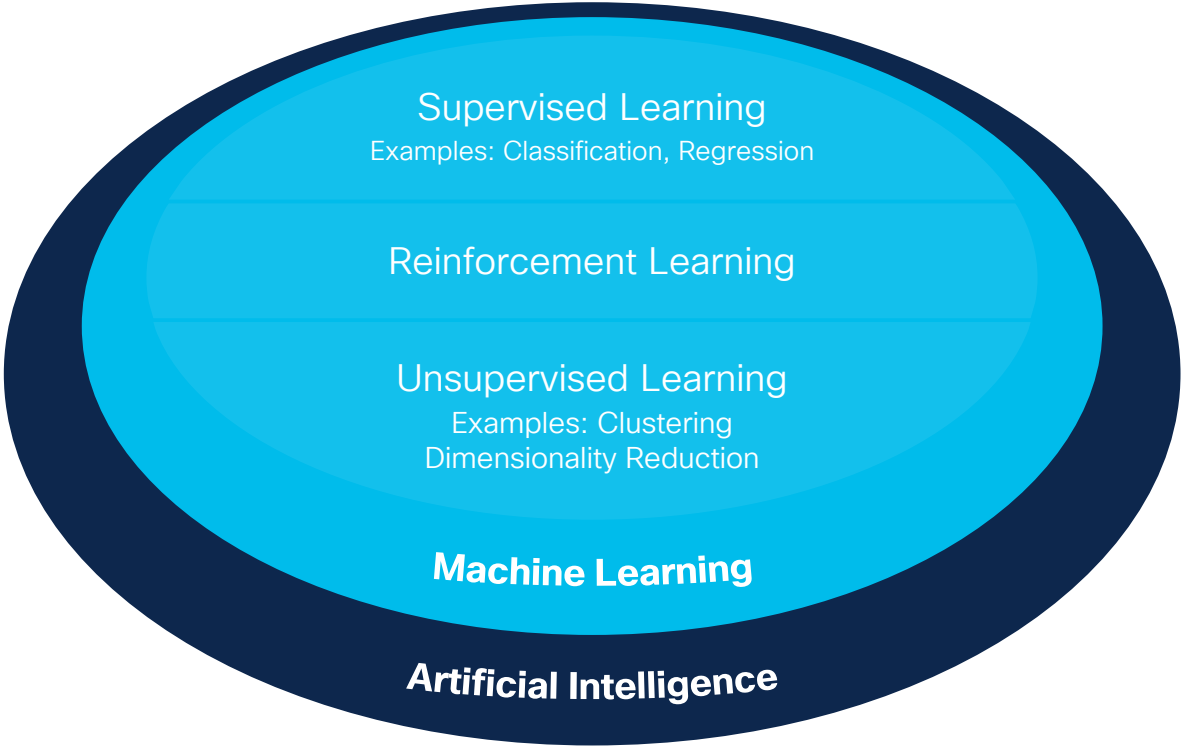
Artificial Intelligence & Machine Learning

“Field of study that gives computers the ability to learn without being explicitly programmed.”

Arthur Samuel’s definition of machine learning in 1959

 © 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Machine Learning Big Picture



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.

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

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

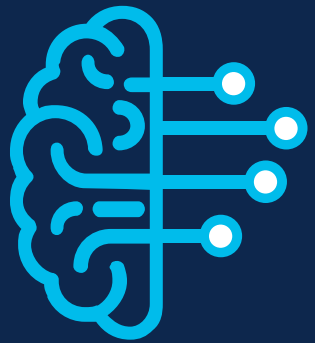
- Less structured & know little about the structure
- You don't have answers and may not fully know the questions
- Unsupervised techniques act as a tool for gaining an understanding of how elements of the set relate to each other

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

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

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**

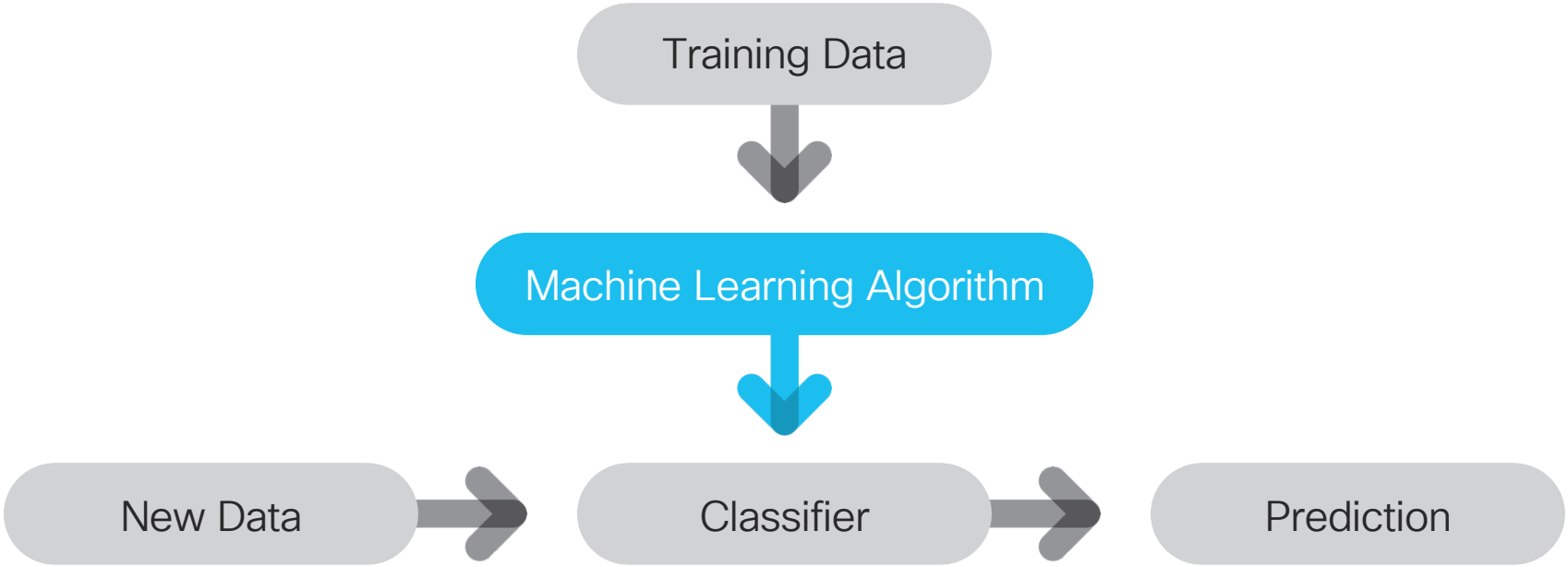
© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

"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."

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Training Classifiers

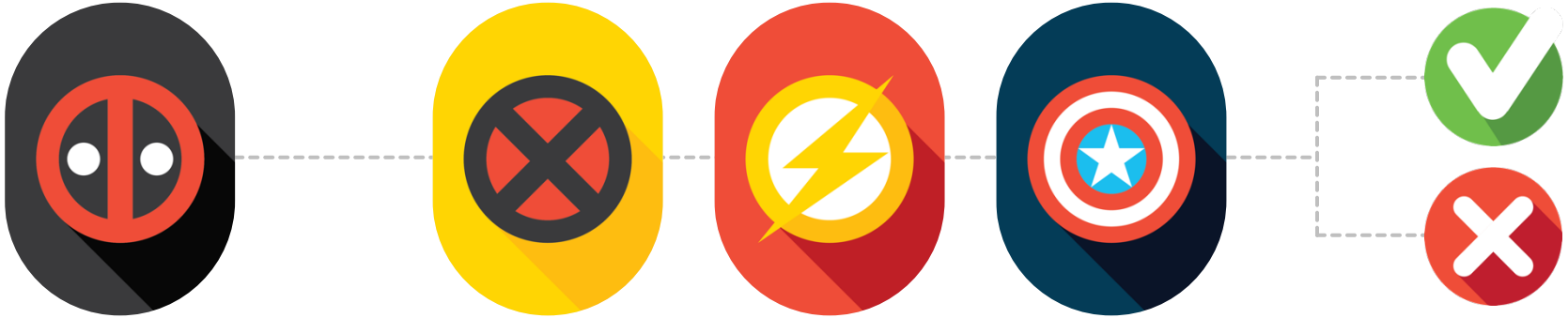


© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Efficacy and Measurement

What Is At Stake Matters

Because you watched Deadpool, you might like…



Deadpool

X-Men:
First Class

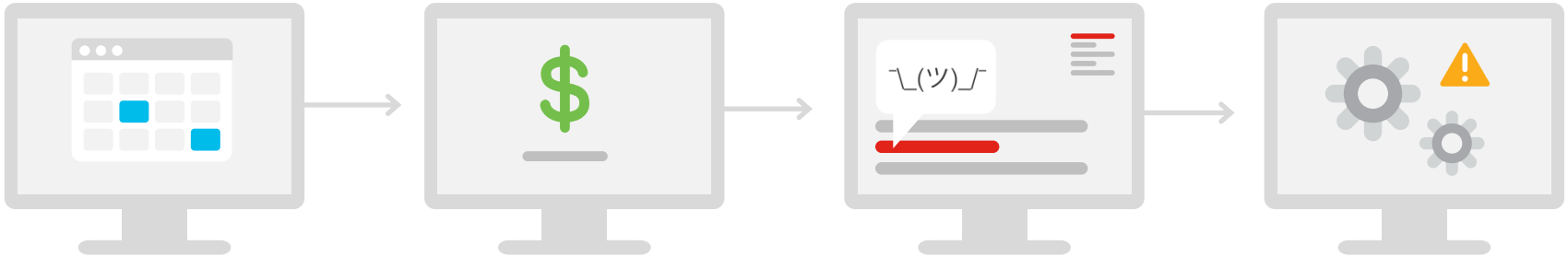
The Flash

Captain America:
The First Avenger

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

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

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

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**

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Dark Data Is Here

Direct Inspection
No Longer Possible

Inference
Becomes Our Strategy

We can accurately **infer** what we cannot observe.

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

TLS Is the New TCP

Digital
Fingerprints

All Encrypted Sessions
Begin Unencrypted

Encrypted Sessions

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

We Can Still Detect Threats Without Decryption

Unencrypted Observables

Digital Fingerprints

Attribution

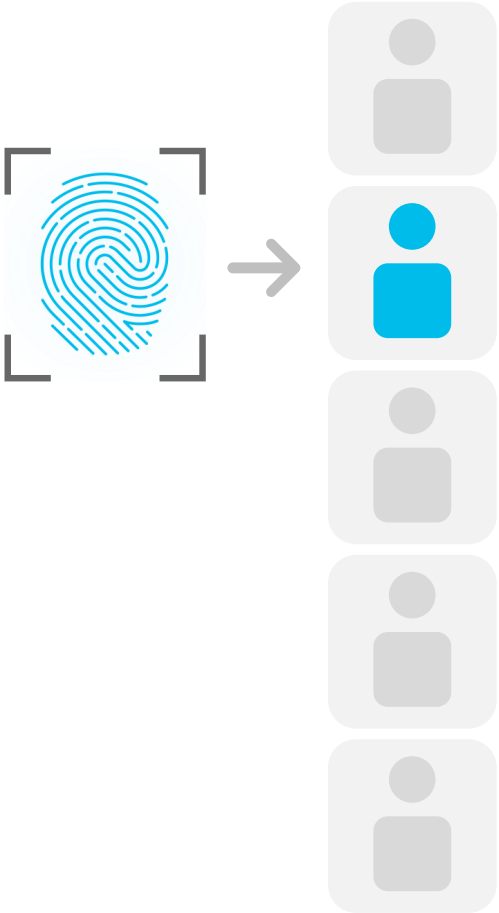
Helpful Analytics

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

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

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Fingerprinting

1:1 with Object Class

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

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Digital Network Fingerprints

Objective

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

Directly Observable (&&)	Class of Object	Instance
Has Feathers Has Webbed Feet Has Yellow Bill	Duck	Not Daffy Specifically

Duck Example *(abductive reasoning)*

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Research based on
Blake Anderson and David McGrew
Security and Trust Organization at Cisco

TLS Fingerprinting

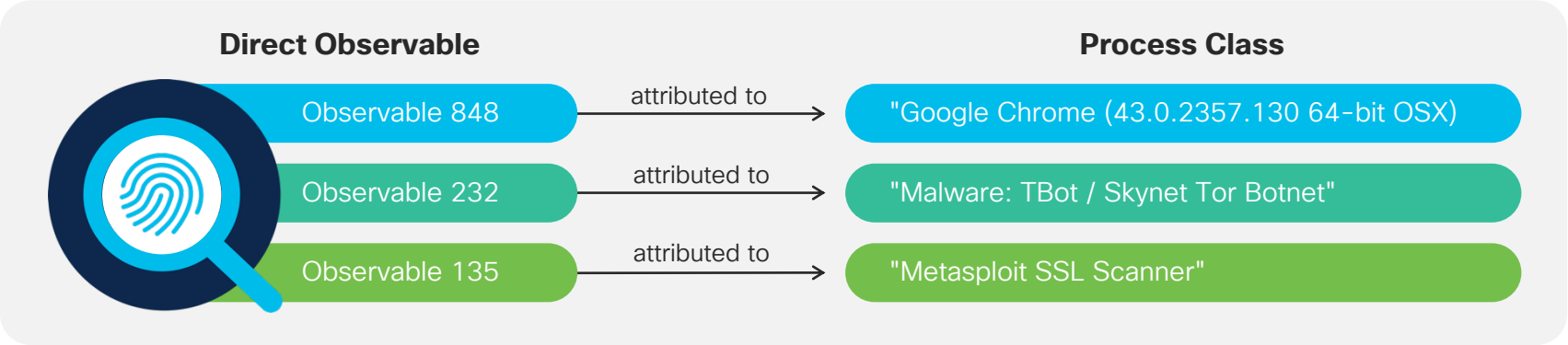
https://github.com/cisco/joy

https://github.com/cisco/joy/blob/master/fingerprinting/resources/fingerprint_db.json.gz

TLS Fingerprinting

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



© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

From Observations to Attribution

Network Visibility Module (NVM)

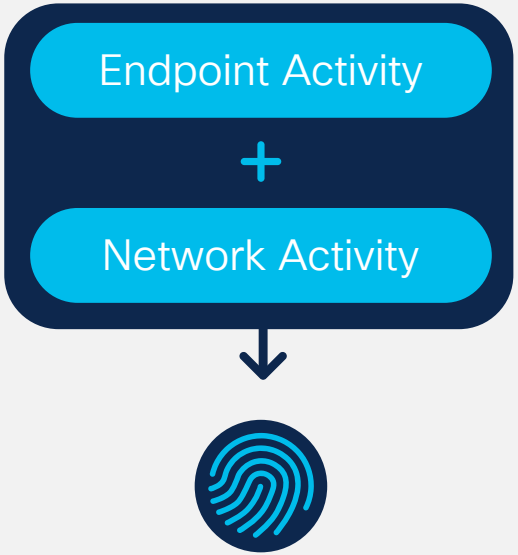
- IPFIX-Based Record (Source, Destination IP, etc)
- Unique Device ID (correlate records from same endpoint device)
- Device Name (bsmith-WIN) and OS Version (Window 7)
- Domain\User Name (Amer\bsmith)
- Account Type (win Admin/Standard/Guest mac: Standard/Root)
- Local DNS (starbucks.com) Target DNS (→ 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)

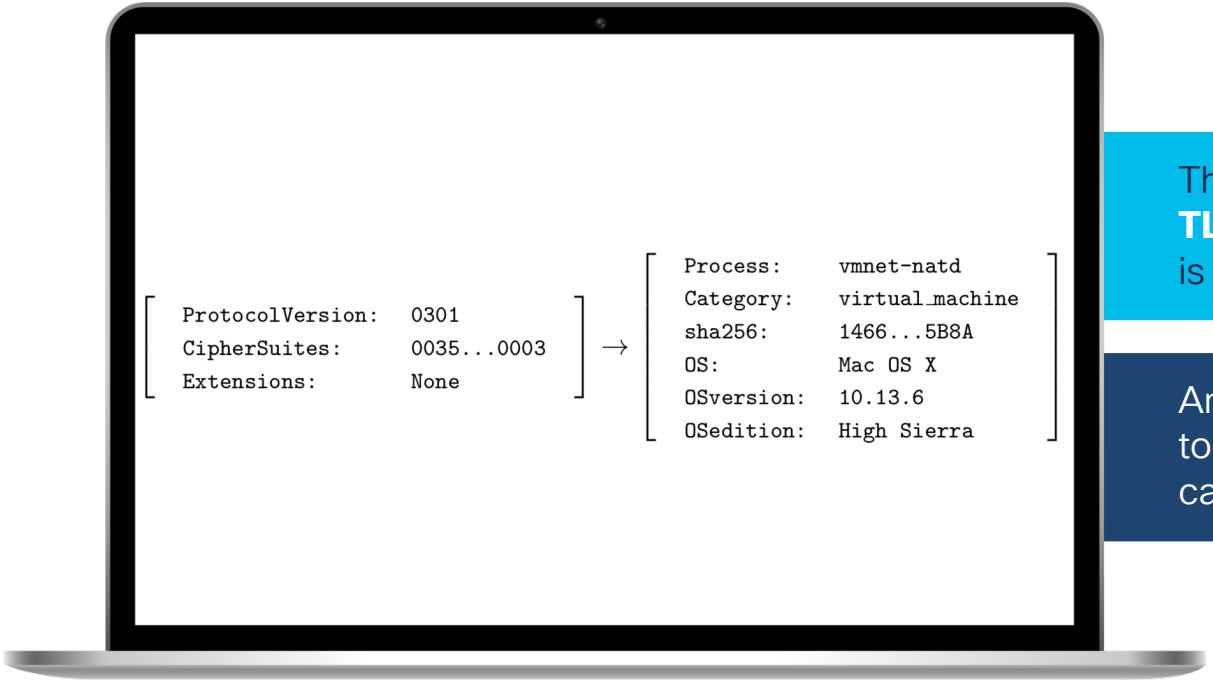
- Encrypted Traffic Analytics Telemetry (Initial Data Packet (IDP))
- Stealthwatch Flow Sensor
- Opensource Joy Package

Machines help us answer:

What deterministic patterns can be directly attributed to a process?



How does it work?



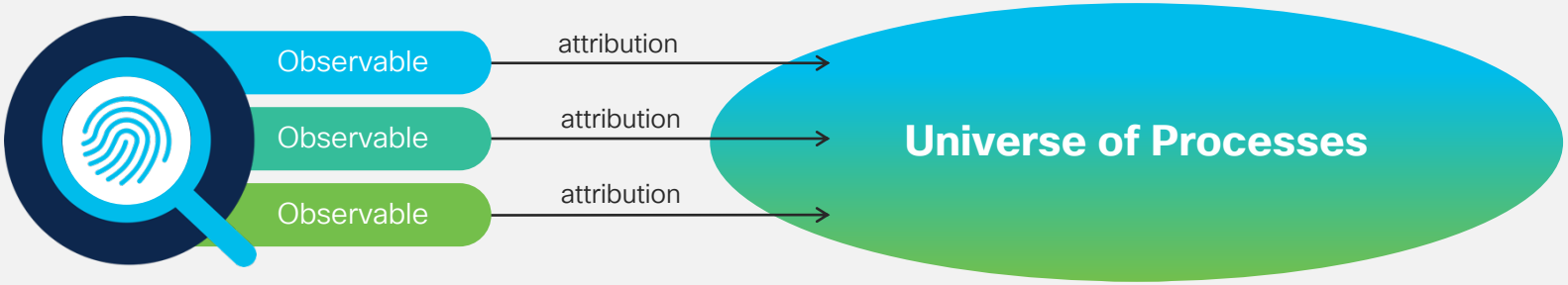
The Cisco
TLS Fingerprint Database (FPDB)
is maintained and updated daily

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

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

How To Evaluate a Fingerprint Database?

How comprehensive is the map to the territory?



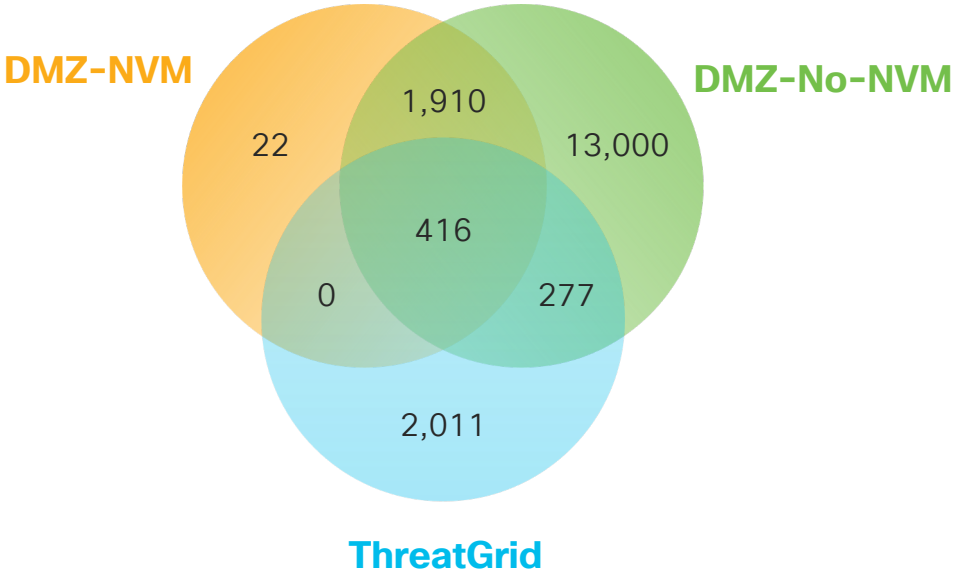
Specificity of the attributions?

Google Chrome ← less specific — more specific → **Google Chrome** (43.0.2357.130 64-bit OSX)

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Building a Better Fingerprint Database
(counts as of June 04 2019)

TLS Fingerprint Overlap (Total=17,502)



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!

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Comparing to Other TLS Fingerprinting Methods

Database	Size	Automatically Updated	GREASE Support	Static Extension Data

Cisco	4,500+	Yes	Yes	`supported_groups`
				`ec_point_formats`
				`status_request`
				`signature_algorithms`
				`application_layer_protocol_negotiation`
				`supported_versions`
				`16 others`

Kotzias et al.	~1,684	No	Discards Locality	`supported_groups` `ec_point_formats`
JA3 (Used by DarkTrace)	157	No	Discards All Data	`supported_groups` `ec_point_formats`

Fingerprint TLS	409	No	No	`supported_groups`
				`ec_point_formats`
				`signature_algorithms`

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

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

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

The Limitations of Fingerprinting

TLS Fingerprints Point to a Process

Direct Observable		Process Class	Threat Category
Observable 232	attributed to →	"Malware: TBot / Skynet Tor Botnet"	Malicious
Observable 135	attributed to →	"Metasploit SSL Scanner"	Benign
Observable 848	attributed to →	"Google Chrome (43.0.2357.130 64-bit OSX)"	Benign
		"NVIDEA GeForce Experience"	Benign
		"GitHub Desktop (tested build 216 on OSX)"	Benign
		HipChat(32bit)	Benign
		java	Unaccounted
		Firefox	Unaccounted
		Dropbox	Unaccounted

Attribution to a process class is only the first step, we still need to understand what it means to the business

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Behavioral Analytics

Signatures vs. Behavioral Detection

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

vs

A set of actions were performed, **sound the alarm**

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

The Network is Where Computers Behave

Behavioral
Telemetry

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

A Chef’s Knife

Used for food preparation in kitchens

Sharp blade of a particular shape

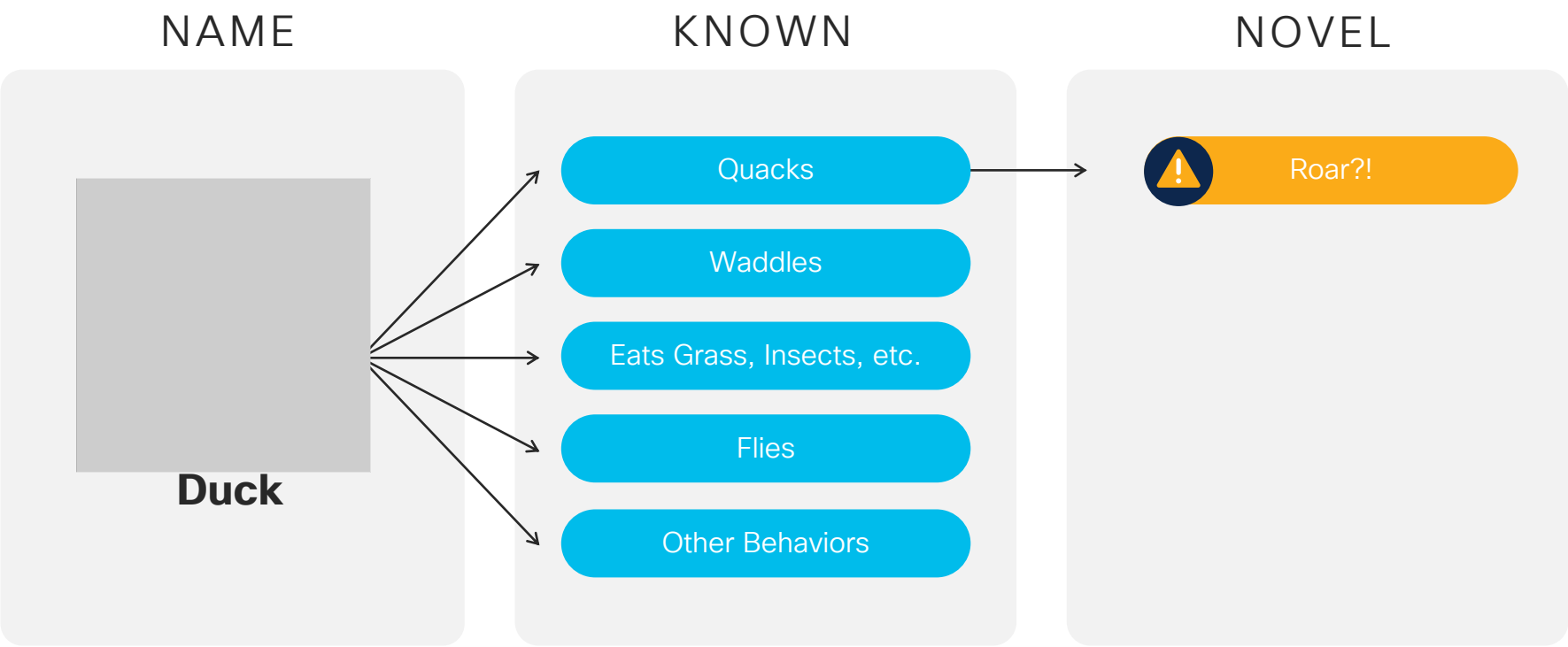
Has a handle used by role of kitchen worker

A Chef's Knife

Behaviors

- **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



 © 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Behavioral Analytics

TLS Fingerprinting

Observable 232 → attributed to → Process

Behavioral Inferences

Process → Printer → {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!

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Levels of Inference

LEVEL 1 INFERENCE

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 INFERENCE

Blacklist

CVE ####

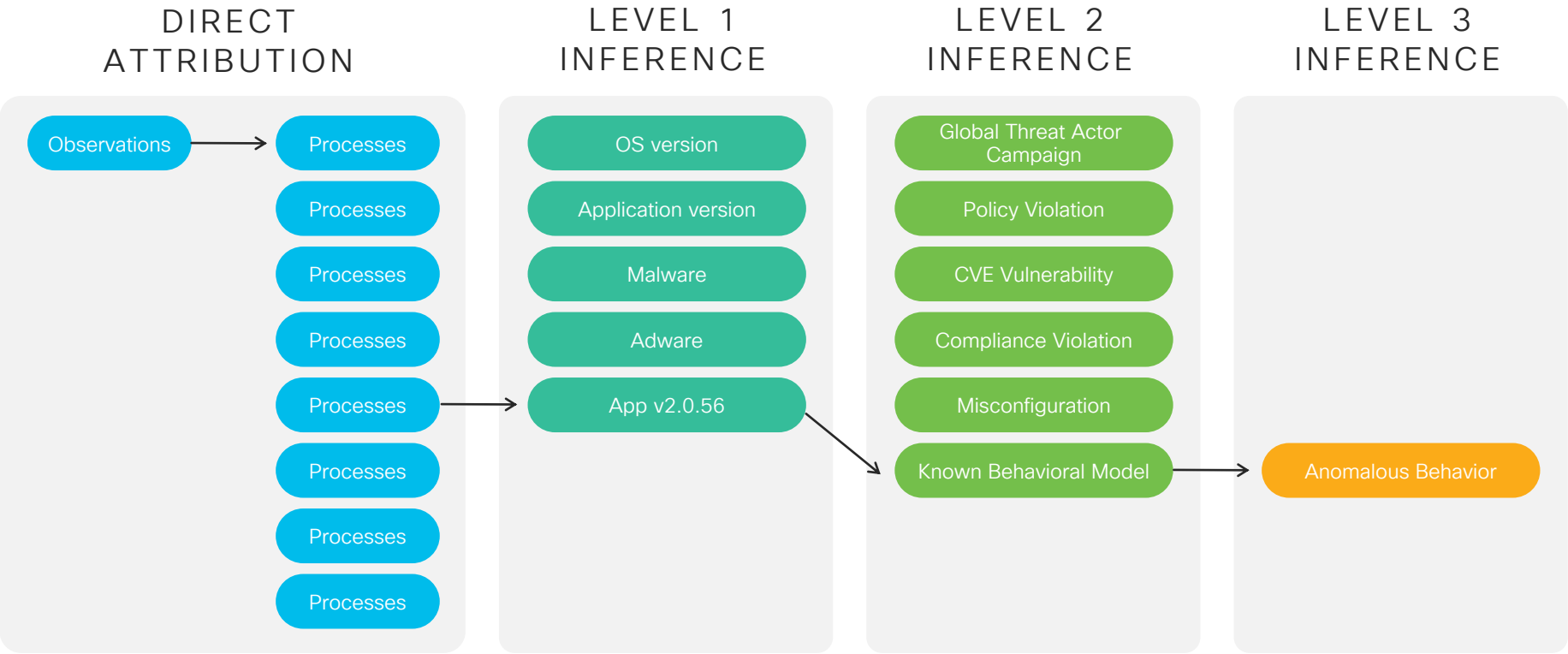
Known Bad

Whitelist

Policy/Compliance Violation

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Levels of Inference

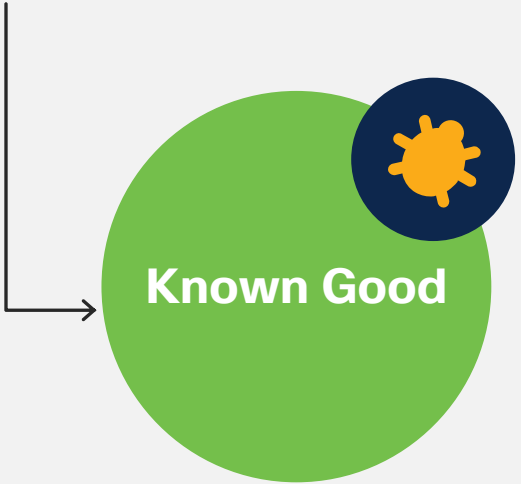


 © 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Outliers & Novelty

Objects & Behavior
Known ***a priori***

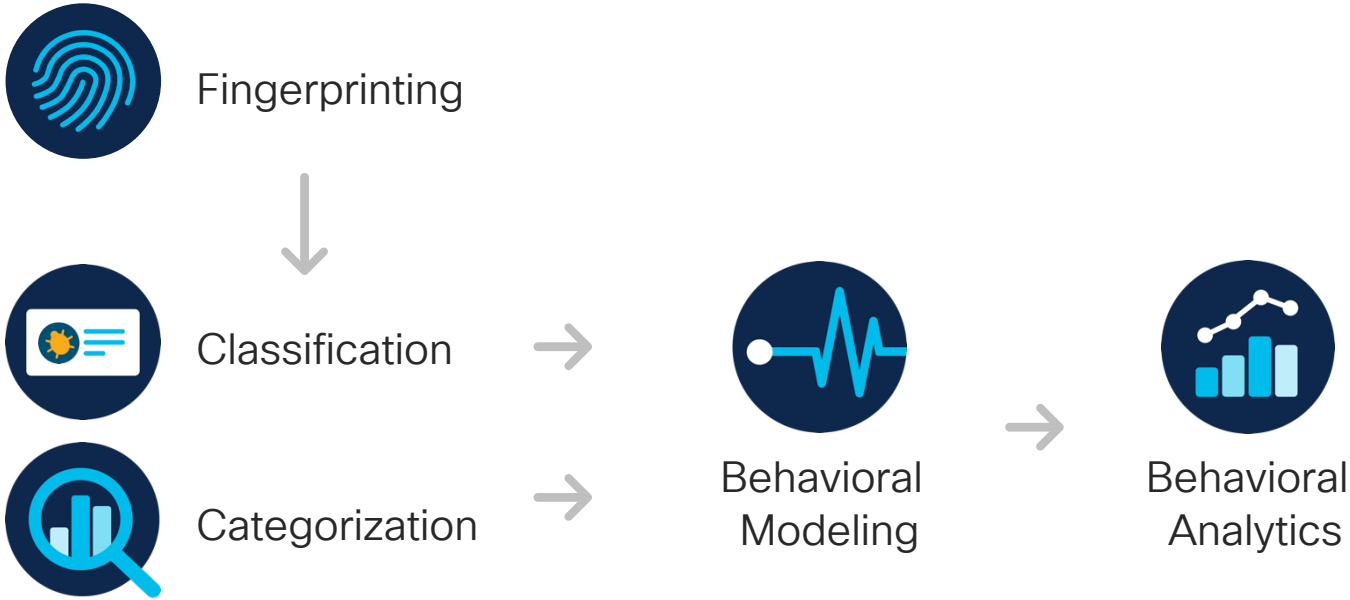
If This Is **Good**, Then What Is This?



Derived from first modeling the good

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Quick Summary of the Analytical Pipeline



© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

The Problem with Numbers

It is not your fault that you don't understand this.

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Numbers Help Us Group Things

Credit-worthiness Class

Legal to Drink/ Legally Drunk

Weight Class

Socioeconomic Class

Age Class

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

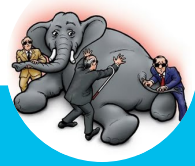
* The terms ‘Set’ and ‘Class’ are synonymous in this presentation

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

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

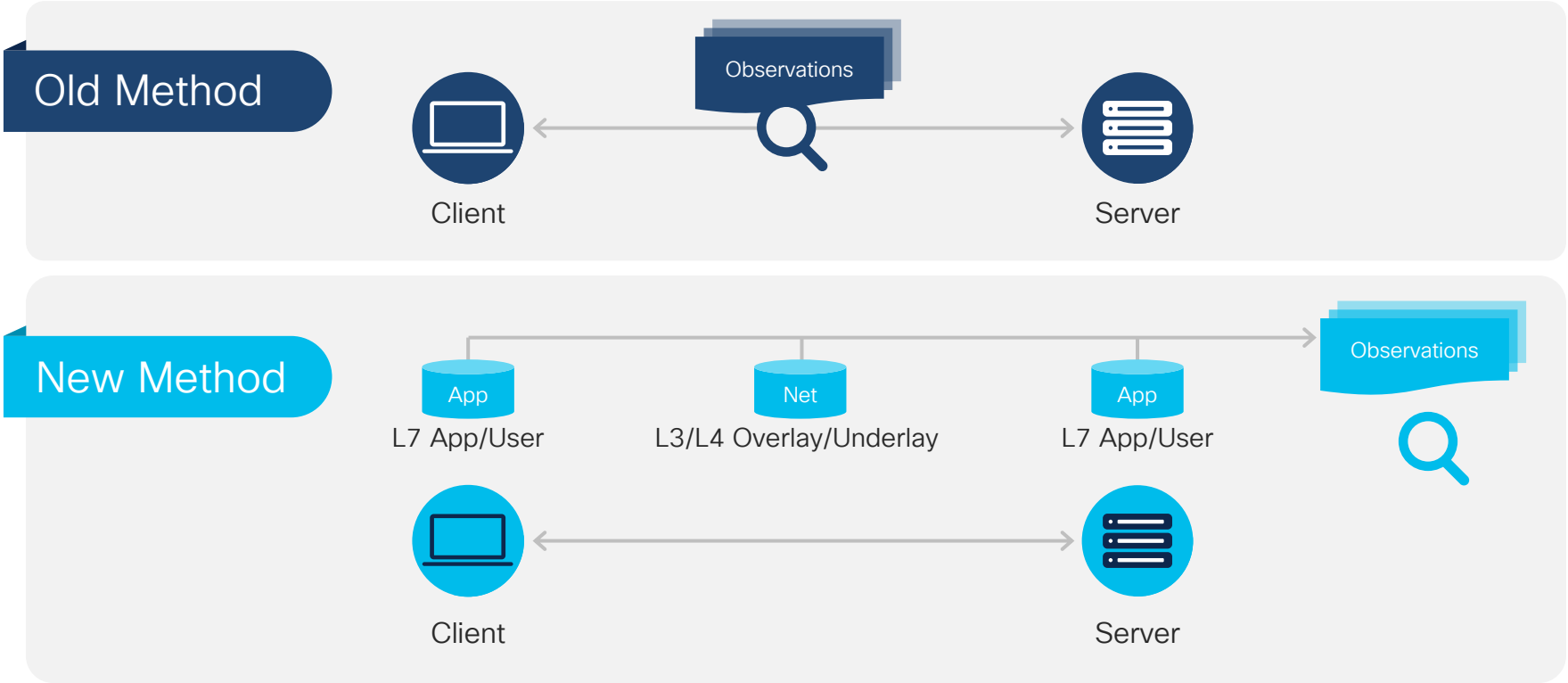
Number systems are dependent on social processes that institutionalize semantics

They often fall short when asked to support multiple perspectives and points of view

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Future Security Analytics

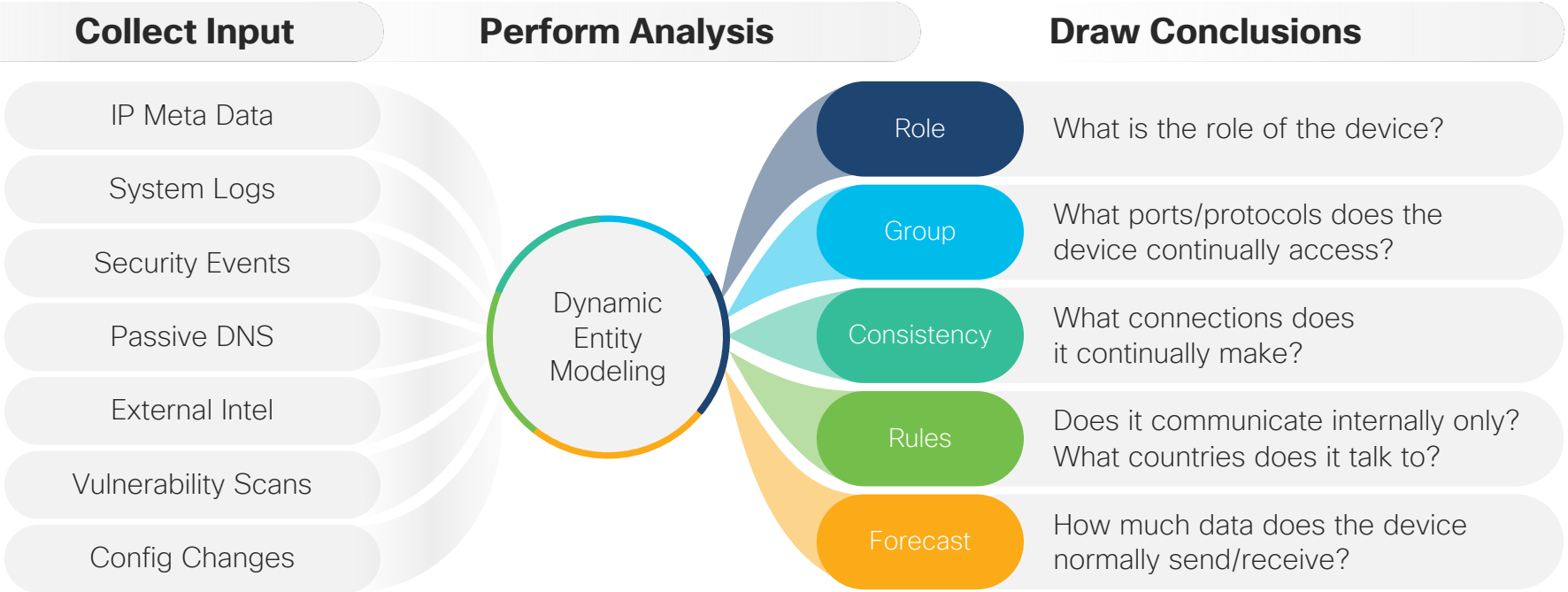
Direct Versus Indirect Observations



© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Late-Binding Modeling to Detect Security Events

Dynamic Entity Modeling



© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Classify the Observable World and Infer the Rest

Threat Actor Activity

Weird Stuff (but not threat related)

Normal Activity

 © 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

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

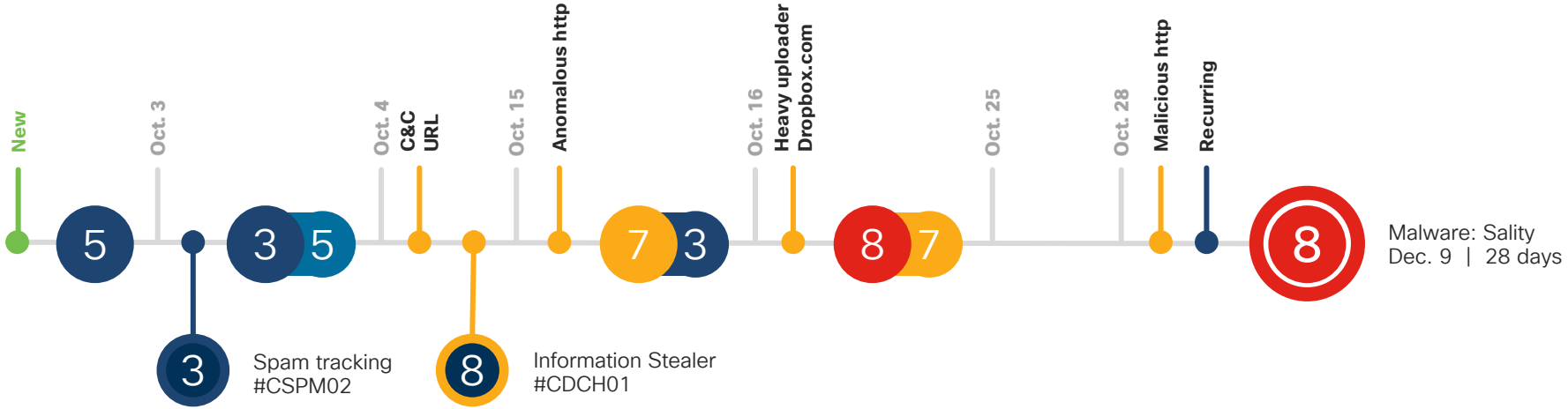
- Multiple-Instance Learning
- Neural Networks
- Rule Mining
- Random Forests
- Boosting
- ML: Supervised Learning

Relationship Modeling

- Probabilistic Threat Propagation
- Graph-Statistical Methods
- Random Graphs
- Graph Methods
- Supervised Classifier Training

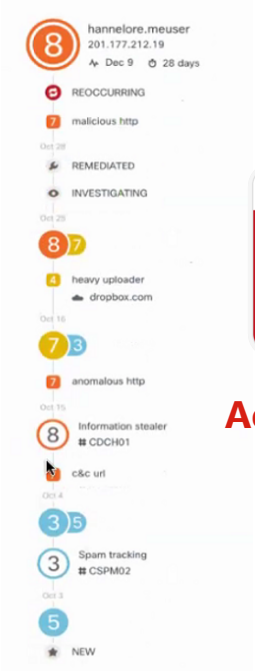
© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Security that Shows its Work



 © 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

We Track the Escalation Over the Lifetime of the Threat Actor

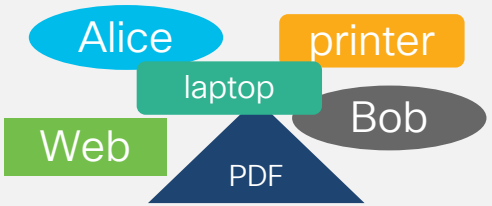


Act Now **Act Today** **Clean & Educate** **Informational**

 © 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Role-Based Behavioral Detection

Data



Labeled Data

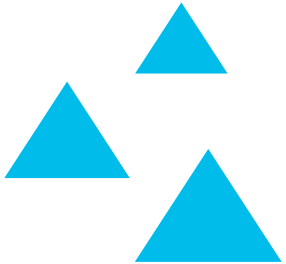
- Printers should act like printers
- Vulnerability scanners should act like scanners
- If your printer is acting like a vulnerability scanner we have a problem

Behavioral Models

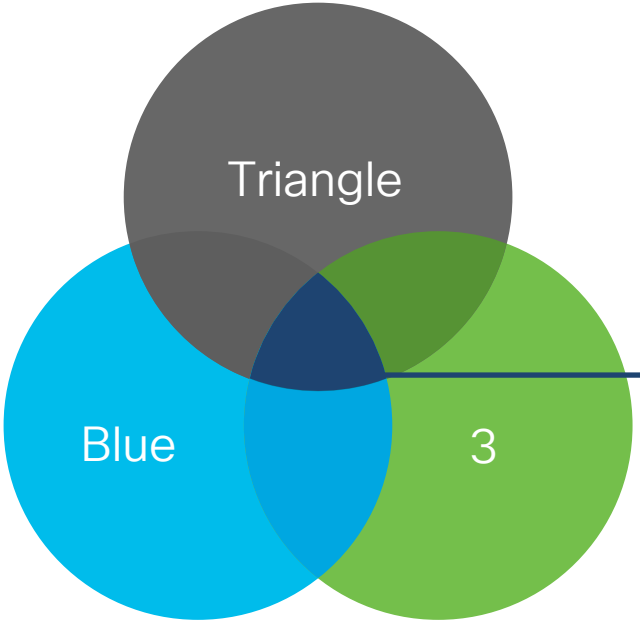
Every endpoint on the network should play a role

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Thinking in Sets/Class and Membership



There are
3 blue triangles

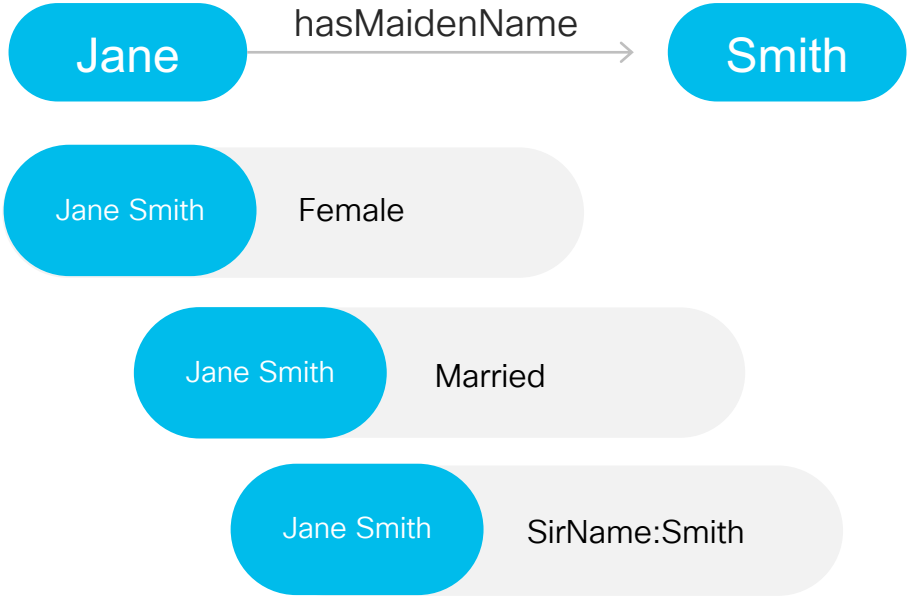


...is a member of the intersection of the set Blue, the set Triangle, and the set Three

 © 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Reasoners

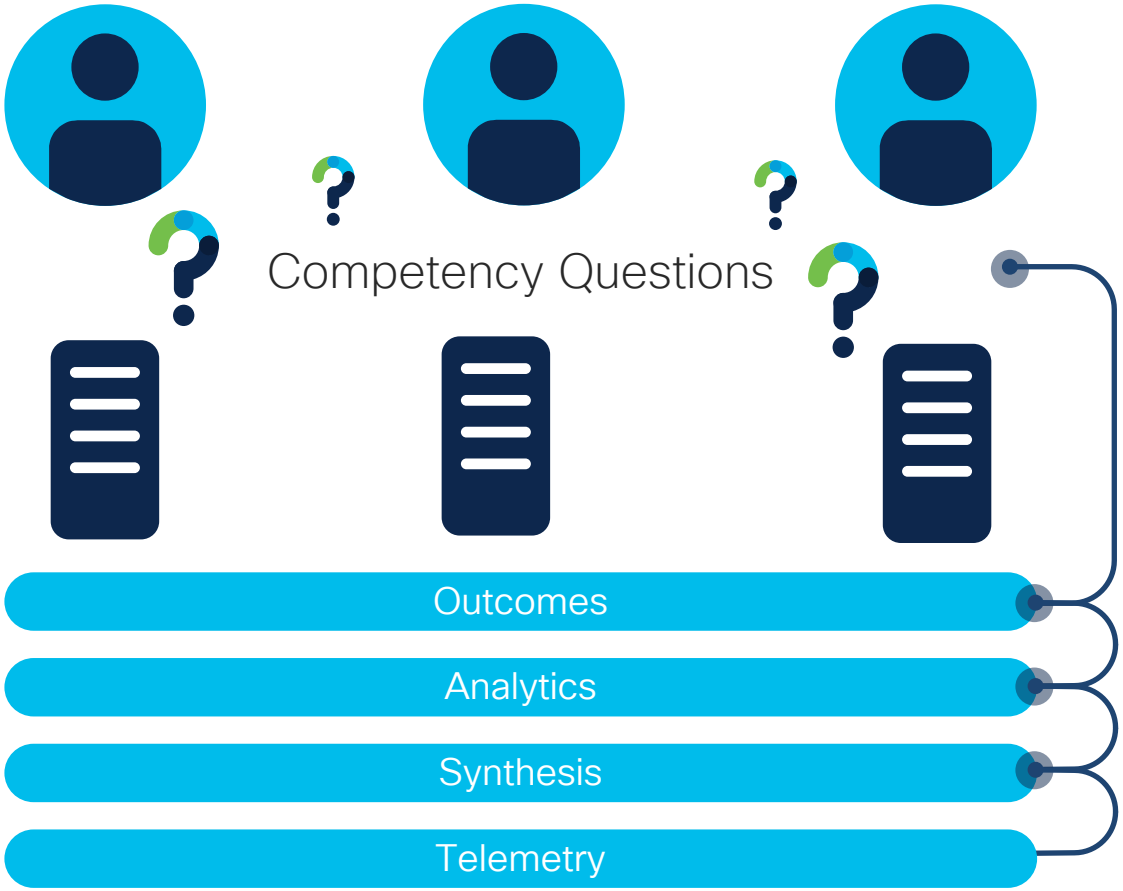
(Side Step the Numbers Problem with First Order Logic)



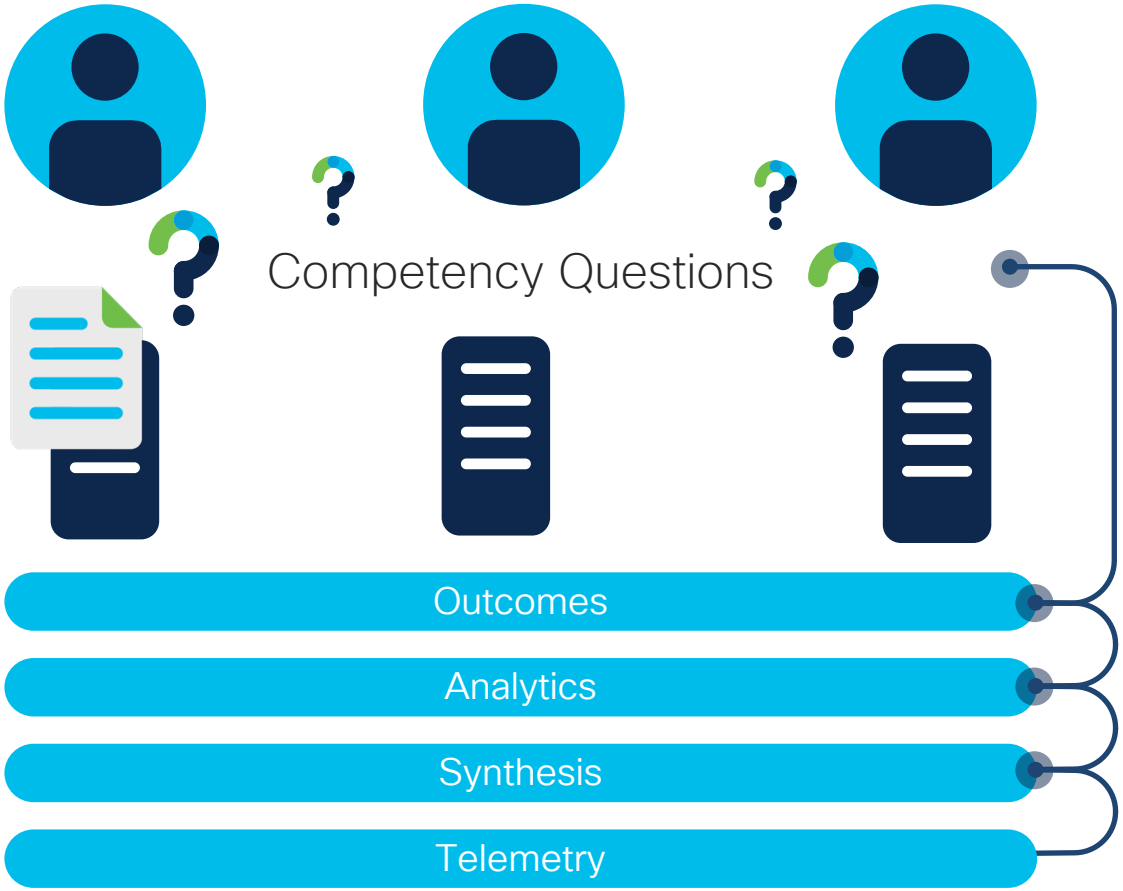
Semantic Models

DOMAIN	hasMaidenName	*RANGE*
Female	hasMaidenName	
Married	hasMaidenName	
	hasMaidenName	*SirName*

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public



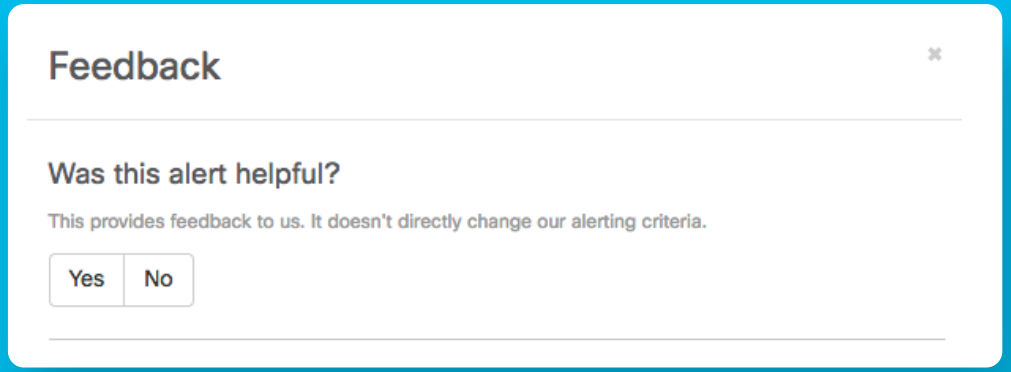
© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public



© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

While syntax can be right or wrong, analytical outcomes are helpful or not helpful to you

How Helpful Was This Alert?



2019	Stealthwatch Cloud Alerts Marked Helpful by Customers (%)
Q3 FY2018	**95%**
Q4 FY2018	**95%**
Q1 FY2019	**89%**
Q2 FY2019	**95%**
Q3 FY2019	**92%**
Rolling Average	**93%**

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

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

What to Ask Your Vendor

How are you applying Machine Learning in your product and why?

Regarding supervised learning, what are you using for ‘ground truth’?

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?

How do you measure its effectiveness?

What non-machine learning are you using and why?

For the ML based assertions, what entailments are provided?

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

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

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

References

Learn More....

- Cisco Stealthwatch Enterprise
- Cisco Stealthwatch Cloud
- Encrypted Traffic Analytics

 © 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

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

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Make Your Head Hurt Reading Material

- Identifying Encrypted Malware Traffic with Contextual Flow Data, Blake Anderson and David McGrew, AISEC '16
- Grill, M., Pevny, T., & Rehak, M. (2017). Reducing false positives of network anomaly detection by local adaptive multivariate smoothing. Journal of Computer and System Sciences, 83(1), 43-57.
- Komarek, T., & Somol, P. (2017). End-node Fingerprinting for Malware Detection on HTTPS Data. In Proceedings of the 12th International Conference on Availability, Reliability and Security (p. 77). ACM.
- Jusko, J., Rehak, M., Stiborek, J., Kohout, J., & Pevny, T. (2016). Using Behavioral Similarity for Botnet Command-and-Control Discovery. IEEE Intelligent Systems, 31(5), 16-22.
- Bartos, K., & Rehak, M. (2015). IFS: Intelligent flow sampling for network security–an adaptive approach. International Journal of Network Management, 25(5), 263-282.
- Letal, V., Pevny, T., Smidl, V. & Somol, P. (2015). Finding New Malicious Domains Using Variational Bayes on Large-Scale Computer Network Data. In NIPS 2015 Workshop: Advances in Approximate Bayesian Inference (pp. 1-10).
- Rehak, M., Pechoucek, M., Grill, M., Stiborek, J., Bartoš, K., & Celeda, P. (2009). Adaptive multiagent system for network traffic monitoring. IEEE Intelligent Systems, 24(3).

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

© 2020 Cisco and/or its affiliates. All rights reserved. Cisco Public

Thank you

Possibilities

#CiscoLive