Cisco live!

January 28 – February 1, 2019 · Barcelona

cisco

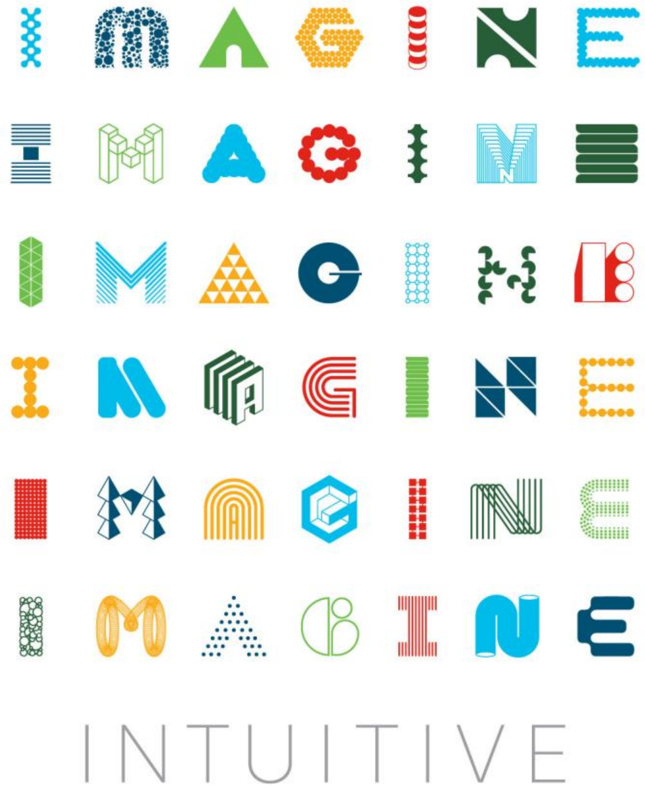
IMAGINE

INTUITIVE

BRKSEC-2068

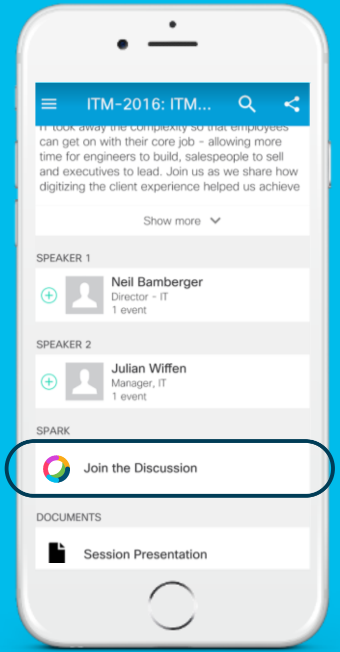
The Future of Security Analytics

TK Keanini
Distinguished Engineer, Advanced Threat Solutions



Agenda BRKSEC-2068

- Introduction
- Security Analytics Fundamentals
- Telemetry, Synthesis/Analytics, and Outcomes
- The Age of Artificial Intelligence & Machine Learning
- Trends and Changes That Shape the Future
- The Future of Security Analytics
- Conclusion & Takeaways



cs.co/ciscolivebot#BRKSEC-2068

Cisco Webex Teams

Questions?

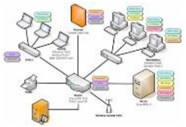
Use Cisco Webex Teams (formerly Cisco Spark) to chat with the speaker after the session

How

1. Find this session in the Cisco Events Mobile App
2. Click “Join the Discussion”
3. Install Webex Teams or go directly to the team space
4. Enter messages/questions in the team space

Hello My Name is TK Keanini

(Pronounced Kay-Ah-Nee-Nee)



Broderbund

CISCO SYSTEMS

Morgan Stanley

nCircle
Proactive Network Security

Lancope

cisco

Fundamentals

Security Analytics versus Other Analytics

Outcomes

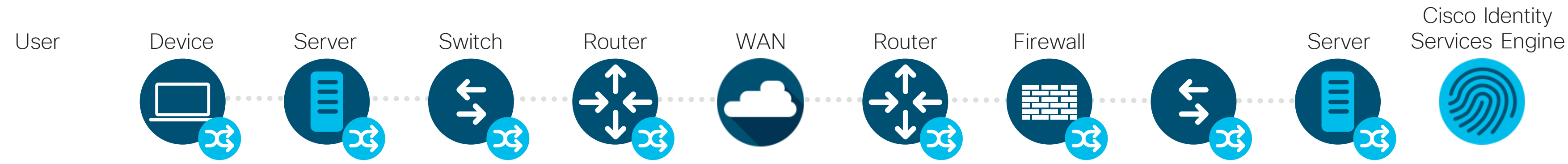
Synthesis/Analytics

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

Telemetry (changes within an observational domain)



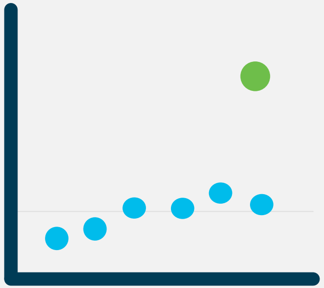
Router	Switch	Firewall	Cloud Native
ISR | CSR | ASR | WLC	• Catalyst • IE • ETA enabled Catalyst	ASA | FTD | Meraki	• AWS Telemetry • Google GCP Telemetry

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

All Telemetry is Data but not all Data is Telemetry

What Did We Do Before Machine Learning?

Simple Pattern Matching



Statistical Methods

Rules and First Order Logic (FoL)

Use in Combination with Machine Learning

When to Use Machine Learning?

If the domain is **static**, has **limited variability**, and is **well-understood**, then machine learning would not be needed.

If the domain is **evolving**, has **a large amount of variability**, or is **not well-understood**, then we can use machine learning to either **help understand the domain** or **efficiently make predictions of unseen instances.**

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. This makes it difficult to evade and delivers the highest fidelity!

Insider Threats & Behavioral Security Analytics

Attackers

They're not breaking in,
they are logging in

Detecting

Through novelty and
outliers

Events

Turn weak signals into a
strong ones

Using the Analytical Stack to explain Encrypted Traffic Analytics

Example: Encrypted Traffic Analytics

Outcomes

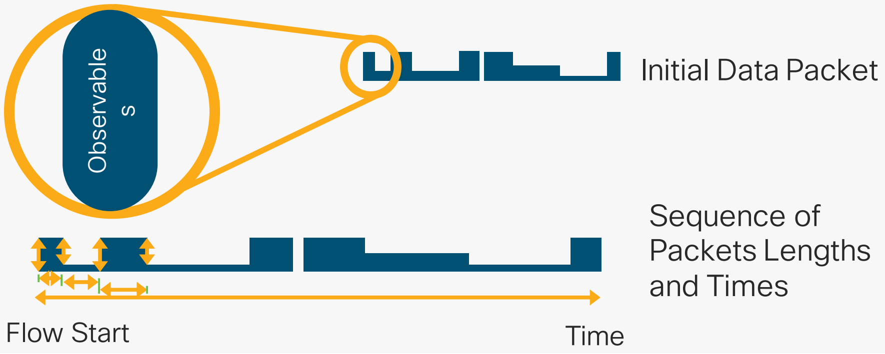
Detection of Malware without Decryption

Cryptographic Compliance

Synthesis/Analytics

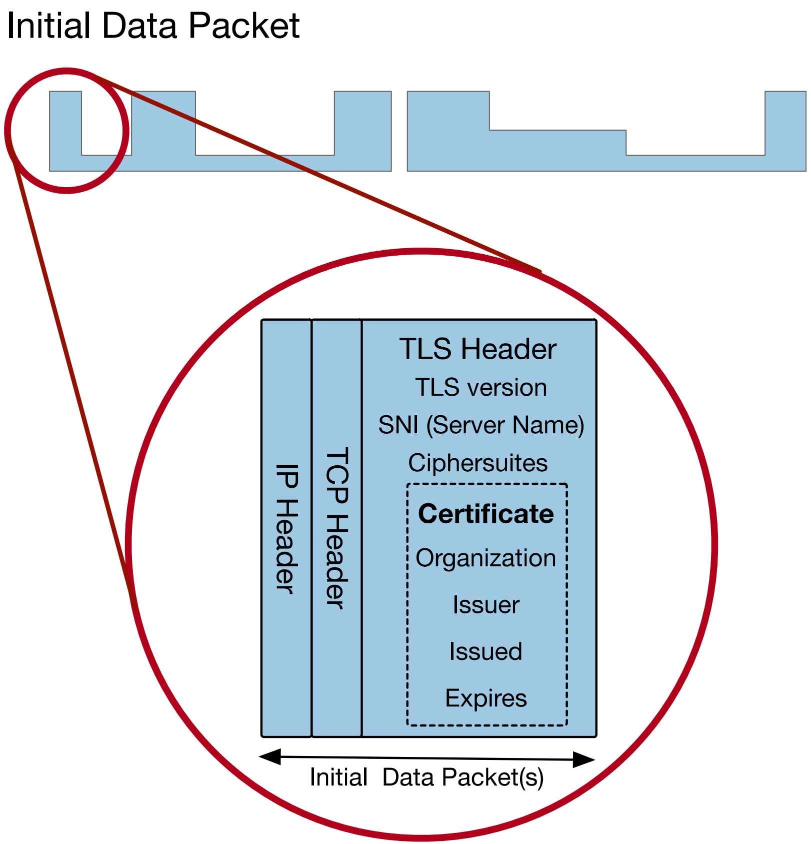
Analytics Pipeline of Diverse Methods

Telemetry

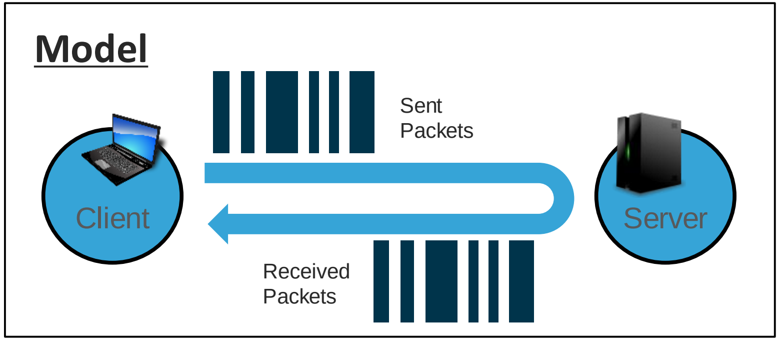


Initial Data Packet (IDP)

- HTTPS header contains several information-rich fields
- Server name provides domain information
- Crypto information educates us on client and server behavior and application identity
- Certificate information is similar to *whois* information for a domain
- And much more can be understood when we combine the information with global data



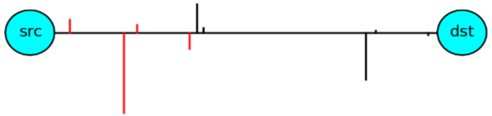
Sequence of Packet Lengths and Times (SPLT)



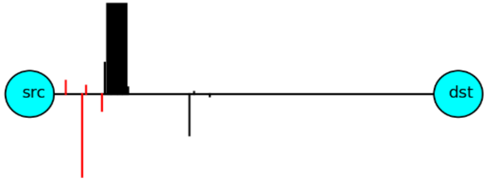
Packet lengths, arrival times and durations tend to be inherently different for malware than benign traffic.

Google search Page Download

Initiate Command & Control



Exfiltration & Keylogging

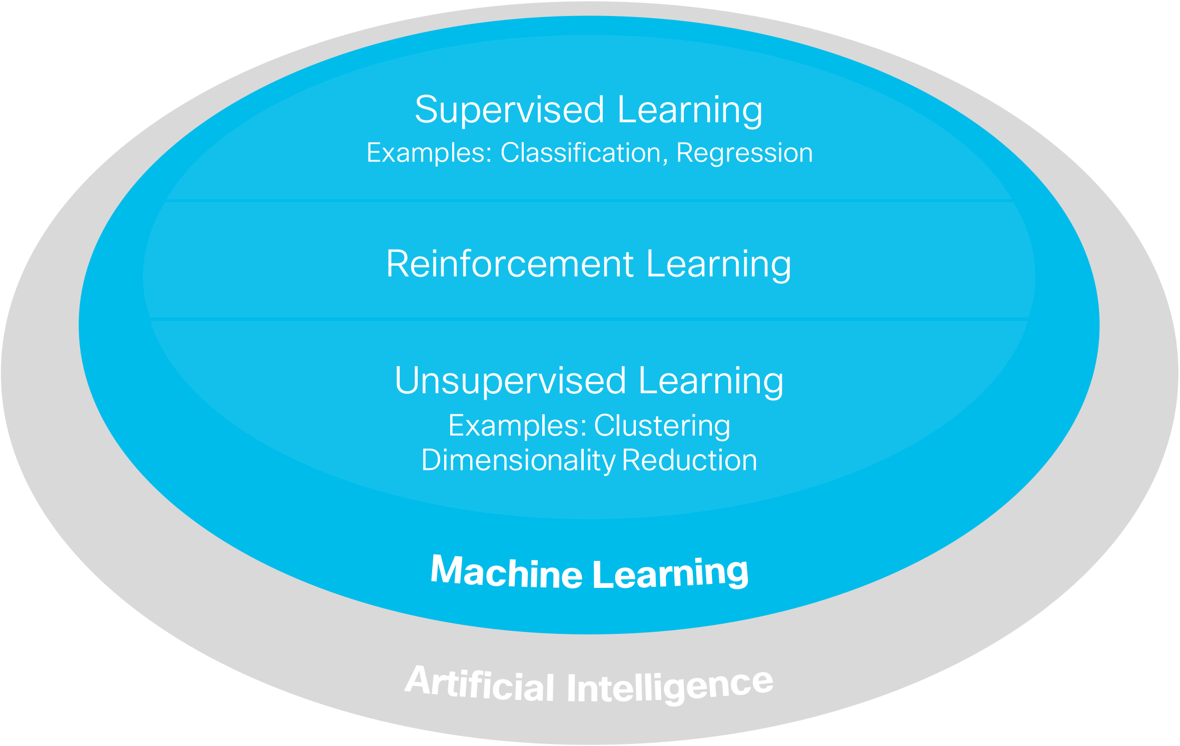


Artificial Intelligence & Machine Learning

"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 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.

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.

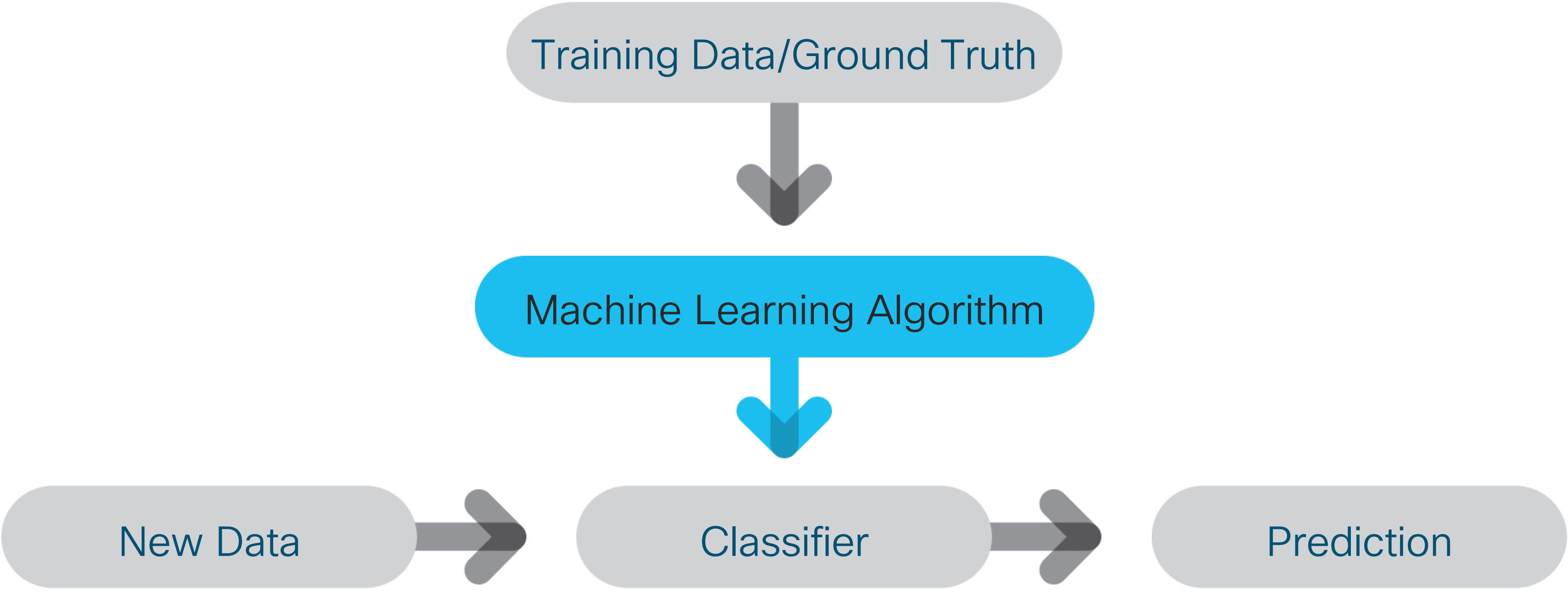
Ground Truth Used in Supervised Learning

- 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



Pitfalls to avoid with Machine Learning

How the data will explain itself?

Regression

- Answer is a real number
- Example: given the weather conditions and temperatures from the previous 10 days, attempting to predict the exact temperature of the following day

Classification

- Answer is a binary/n-ary set of labels
- Example: given the weather conditions and temperatures from the previous 10 days, attempting to predict rainy vs sunny vs windy vs snowing, etc. (labels)

The Problem with **'Just'** Machine Learning…

Hal-9000 "I have found a threat actor operating at your branch office in Los Angeles, would you like me to remove that device from the network?"

"Yes, I would like to quarantine this device but please tell me how you arrived at this conclusion?" **Dave**

Hal-9000 "I'm sorry Dave I only have the computation paths of my ML algorithms, would you still like me to remove this device from the network?"

"Don't do anything until you can share with me the logical path of your investigation!" **Dave**

Success is Domain Specific

Other ML Application ≠ Security

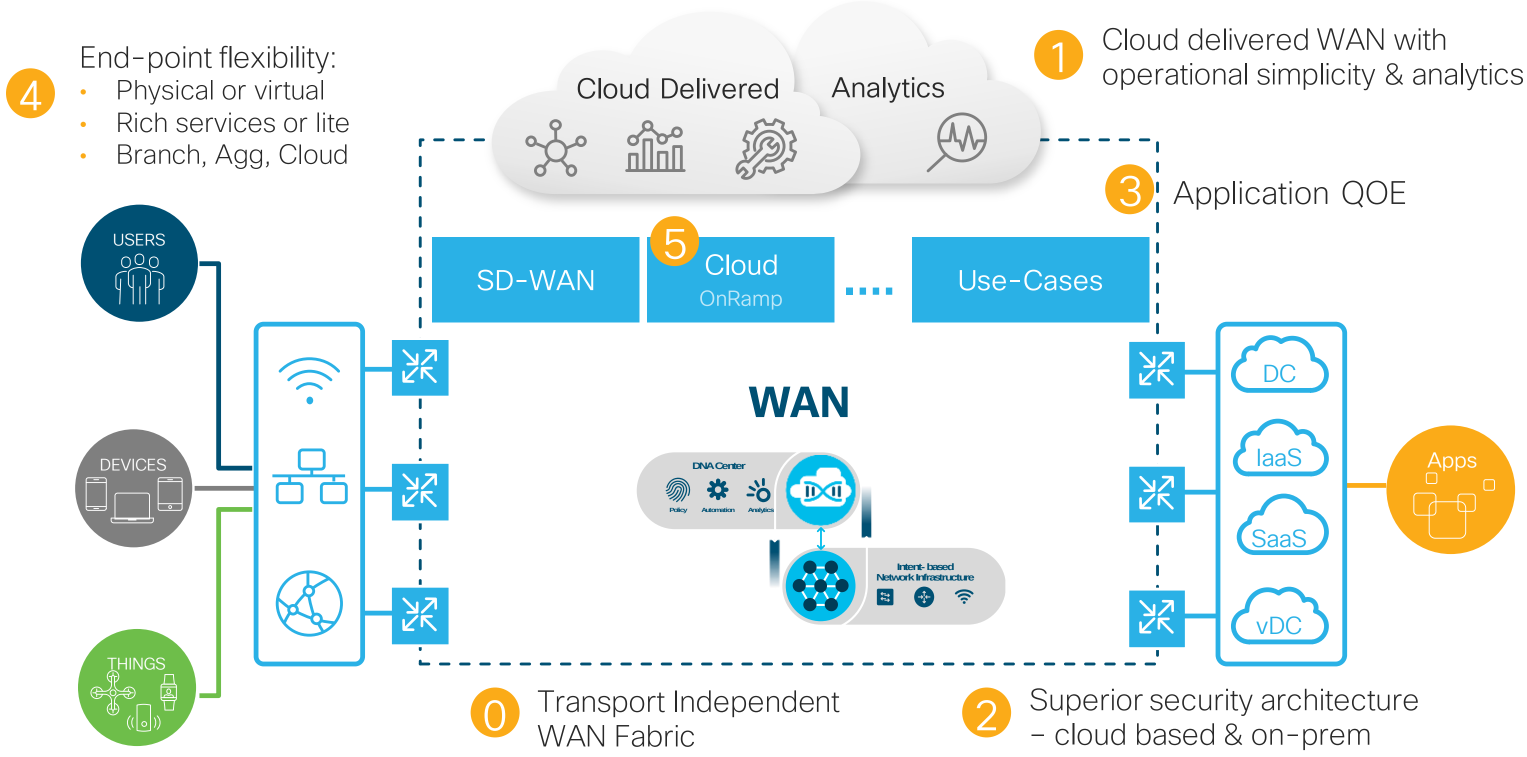
Transformational Trends

The 5 Transformative Trends

1. Overlay-based Networking and the Associated Control Planes
2. Cloud Native Architectures Like: Kubernetes Service Mesh, Lambda, etc.
3. Zero Trust Architecture
4. Transit Inspection Opacity (TLS 1.3, HTTPS by default, etc.)
5. TOR

While the security objectives have not changed, what we defend, how we defend it, & where we defend are all shifting

SD-Access/SDWAN is Networking-as-a-Service

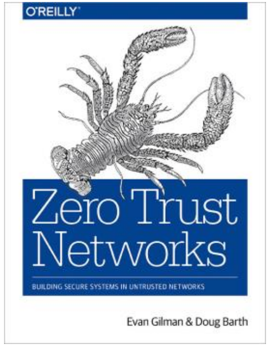


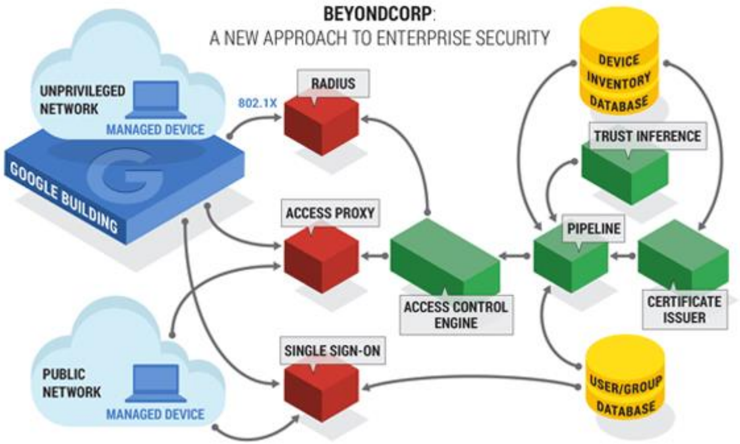
Serverless (uber for code)

Zero Trust Architectures

Zero Trust Fundamentals

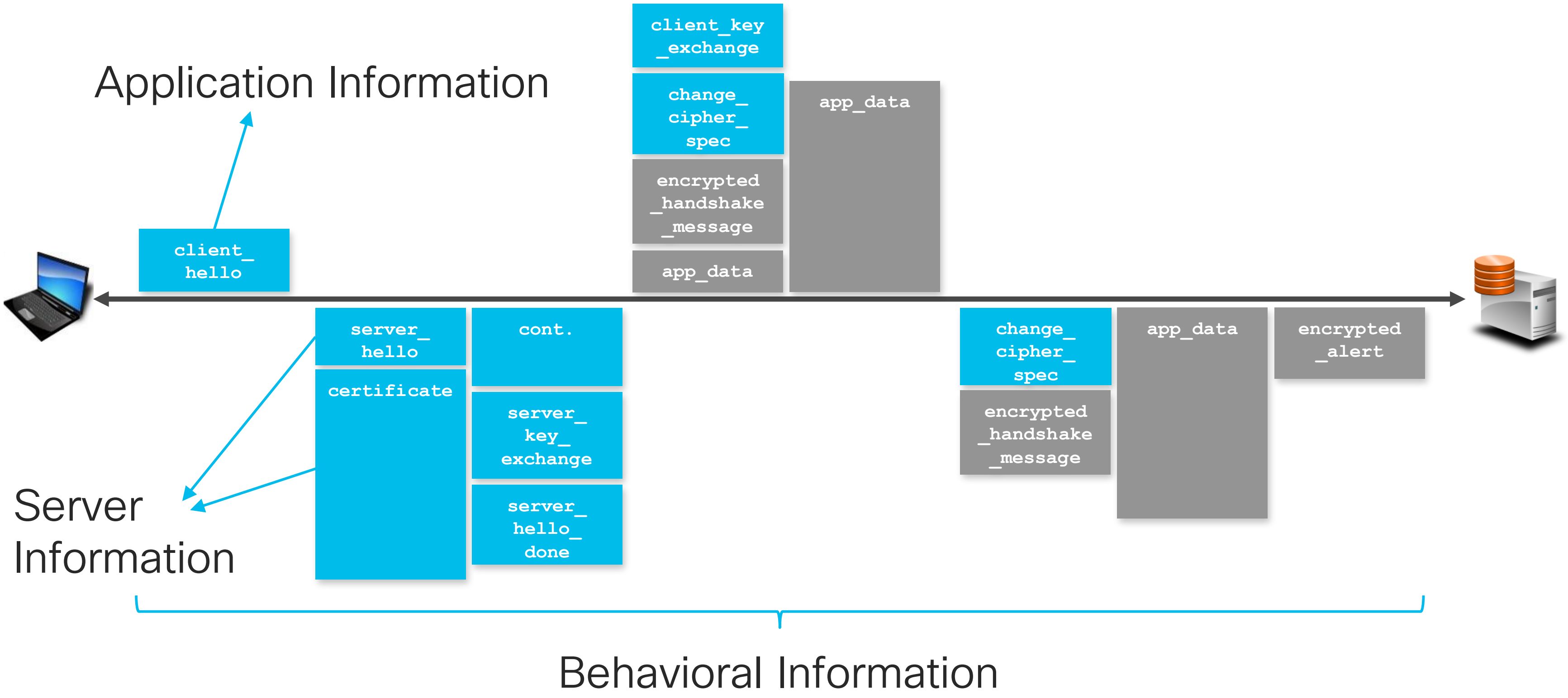
- Firewall enforced perimeters/zones are gone
- All hosts are treated as Internet facing (no firewalls or VPNs)
- Every device, user, and network flow is authenticated and authorized



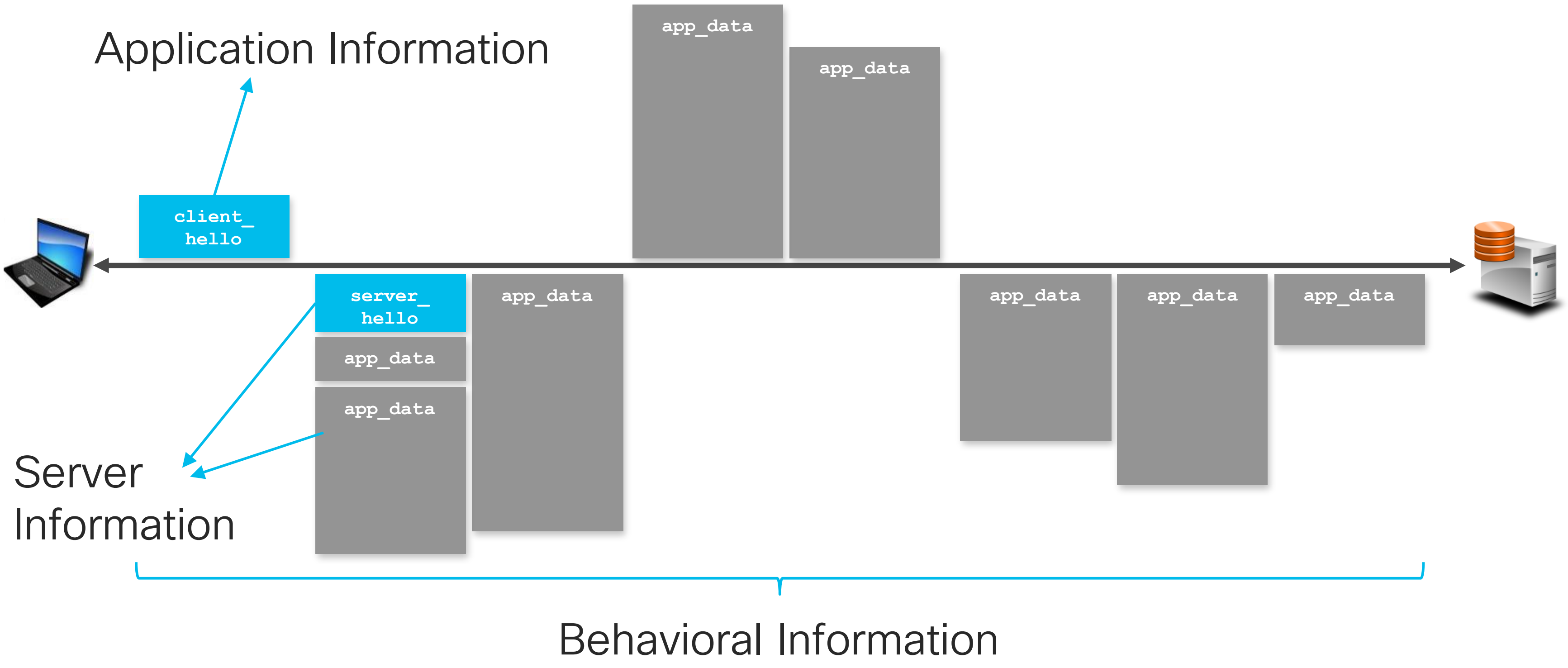


BeyondCorp components and access flow

ETA Data Features, <= TLS 1.2



ETA Data Features, TLS 1.3



The Onion Router

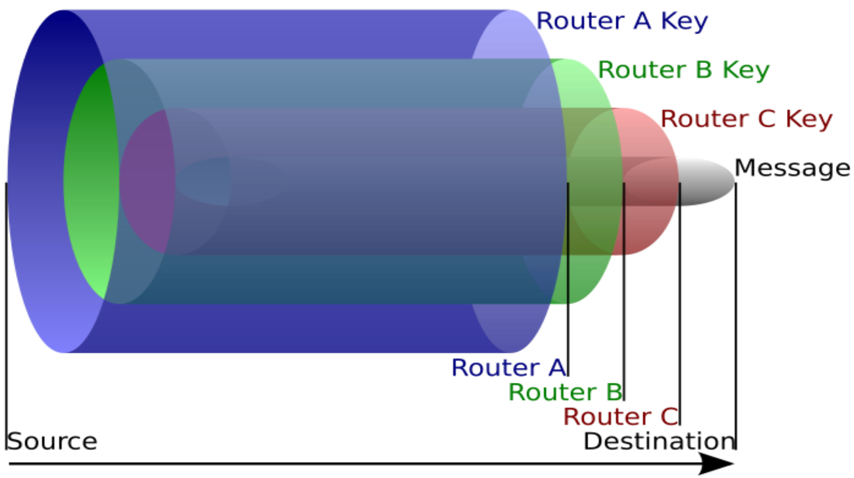
Open source SW / public design specs

Data is constantly encrypted at multiple layers

Sent through multiple routers. Each router decrypts the outer layer and finds routing instructions

Sends the data to the next router

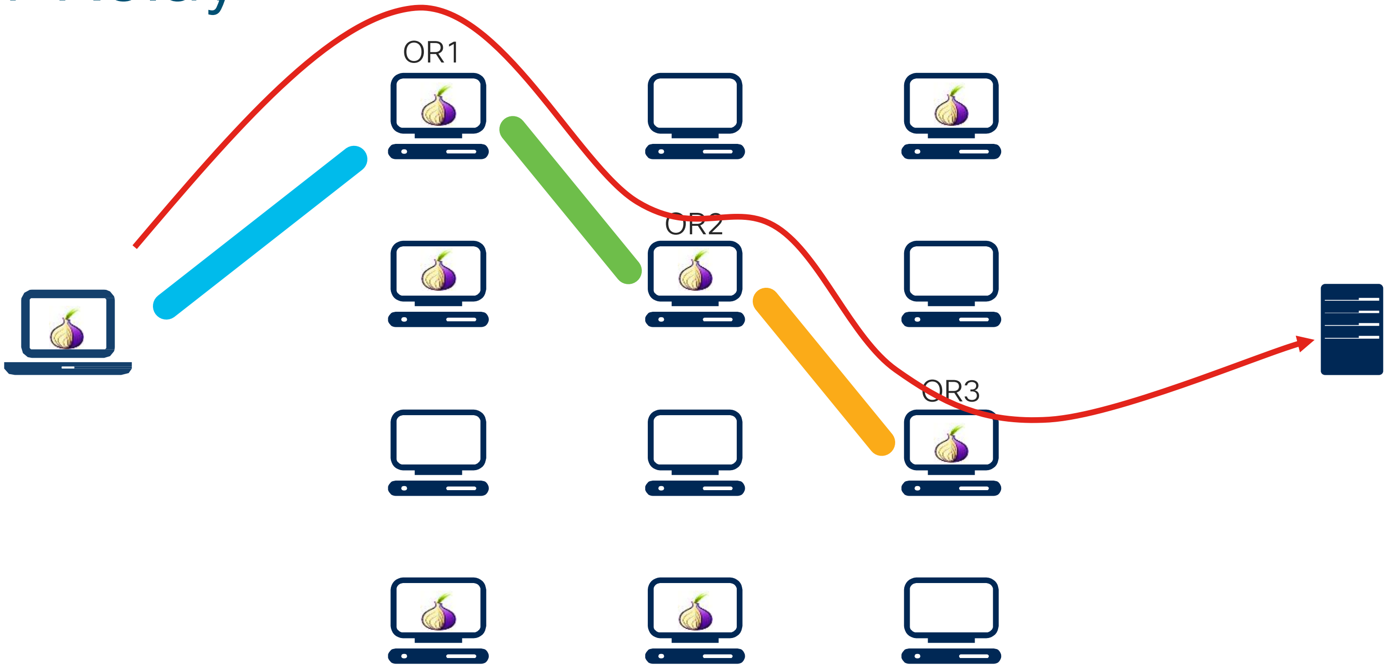
Result is a completely encrypted path using random routers



How is the Tor Network built?

- The Tor network consists of relays
- Relays are just nodes where the Tor software is installed
- They build encrypted connections to other relays, forming an overlay network
- Everyone can run a Tor relay and contribute to the network…

Tor Relay



List of all Tor Relays

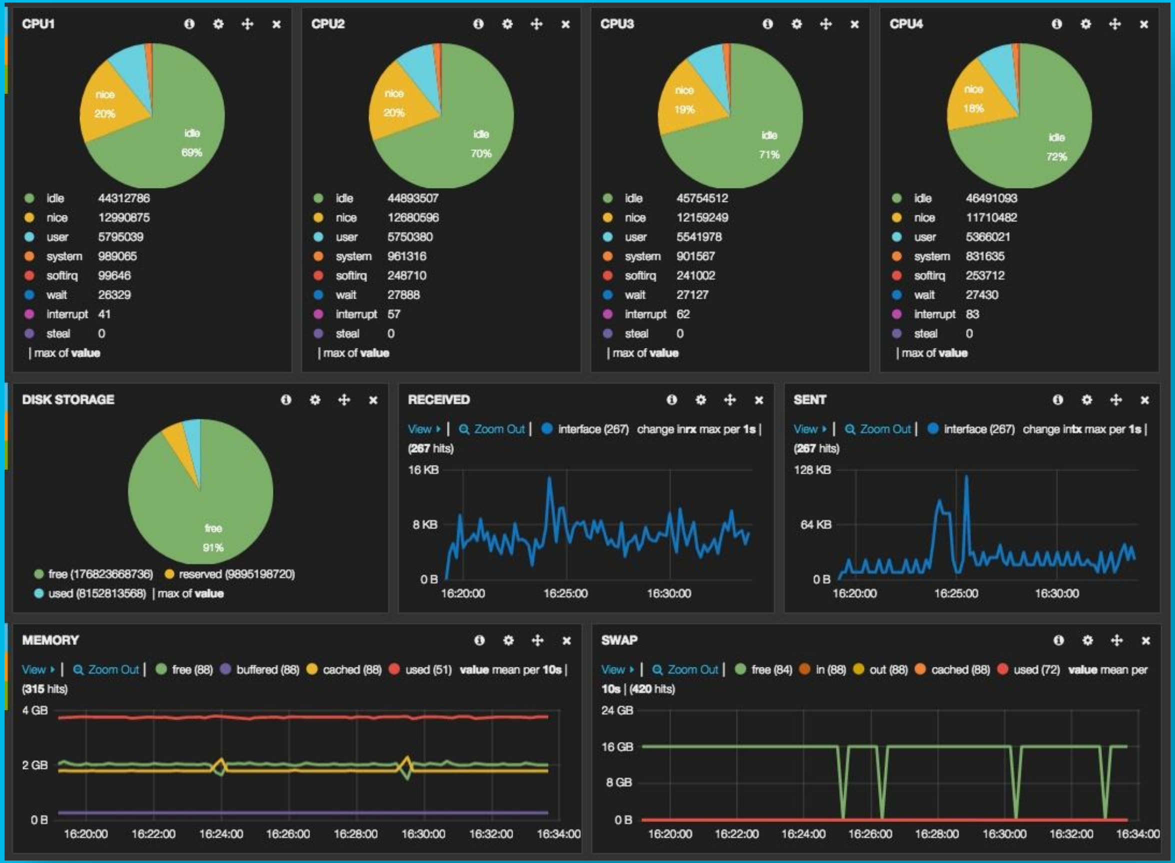
https://torstatus.blutmagie.de/

Router Name	Bandwidth (KB/s)	Uptime	Hostname	ORPort	DirPort	Bad Exit	FirstSeen	ASName
IPredator	101991	2 d	exit1.ipredator.se [197.231.221.211]	443	9030	✗	2014-04-19	CYBERDYNE,
xor	53676	12 d	public.freeflux.org [185.56.80.242]	20002	9031	✗	2018-04-04	NFORCE, NL
xor	48645	12 d	public.freeflux.org [185.56.80.242]	20000	9030	✗	2018-02-24	NFORCE, NL
LibreZone	45167	50 d	pluto.tor-exits.libre.zone [154.16.149.74]	443	None	✗	2017-09-13	OKSERVERS -
sofia	40901	12 d	chomsky.torservers.net [77.247.181.162]	443	80	✗	2017-10-03	NFORCE, NL
UOfMichigan	40608	20 h	tor-exit.eecs.umich.edu [35.0.127.52]	443	None	✗	2018-05-21	UMICH-AS-5 -
0x3d004	39983	18 d	snowden.pep-security.net [62.138.7.171]	9001	9030	✗	2016-08-24	GD-EMEA-DC-
0x3d001	38335	18 d	0x3d.lu [91.121.23.100]	8001	8030	✗	2016-05-31	OVH, FR
Onyx	36998	67 d	onyx.ip-eend.nl [192.42.115.102]	9004	80	✗	2015-04-22	SURFNET-NL
0x3d005	35594	18 d	snowden.pep-security.net [62.138.7.171]	8001	8030	✗	2016-09-23	GD-EMEA-DC-
PrivacyRepublic0001	35432	80 d	tor-exit-node.1.privacyrepublic.org [178.32.181.96]	443	80	✗	2014-11-21	OVH, FR
TotorBE1	35132	18 h	ip176.ip-5-39-33.eu [5.39.33.176]	9001	9030	✗	2016-10-22	OVH, FR
TotorBE2	33663	7 d	ip178.ip-5-39-33.eu [5.39.33.178]	9001	9030	✗	2016-12-18	OVH, FR
CriticalMass	33603	12 d	77.247.181.166 [77.247.181.166]	443	80	✗	2017-10-03	NFORCE, NL
Unnamed	33351	74 d	23.81.66.90 [23.81.66.90]	443	80	✗	2017-01-19	LEASEWEB-US
nastja	31498	5 d	envato.webcare360.com [134.19.177.109]	443	80	✗	2015-12-06	GLOBALLAYER
poiuty	31168	10 d	ns322718.ip-37-187-155.eu [37.187.155.229]	443	80	✗	2016-01-26	OVH, FR
Multivac	30347	13 d	multivac.io [163.172.53.84]	21	143	✗	2014-04-08	AS12876, FR
cry	30049	67 d	cry.ip-eend.nl [192.42.115.101]	9003	8080	✗	2015-04-22	SURFNET-NL
apx2	29898	23 d	tor-exit.r2.apx.pub [185.107.47.171]	9001	9030	✗	2015-02-02	NFORCE, NL
niftysugarglider	29357	20 d	185.220.101.34 [185.220.101.34]	20034	10034	✗	2016-08-30	FERAL Feral H
BlockHouse2	29193	4 d	static.243.91.201.195.clients.your-server.de [195.201.91.243]	9001	None	✗	2018-03-01	HETZNER-AS,
dreamatorium	29126	51 d	89.31.57.58 [89.31.57.58]	443	80	✗	2015-05-04	UNITHOST-AS,
0x3d002	29030	18 d	0x3d.lu [91.121.23.100]	9001	9030	✗	2014-04-22	OVH, FR
niftypika	28941	20 d	185.220.101.32 [185.220.101.32]	20032	10032	✗	2016-01-26	FERAL Feral H
bonjour1	28073	101 d	loft9169.serverprofi24.com [188.138.33.233]	443	80	✗	2017-10-27	GD-EMEA-DC-
GhettoColo01	27681	34 d	tor-exit01.ghettocolo.org [142.44.154.169]	9001	9030	✗	2018-02-04	OVH, FR
atlantis	26314	76 d	atlantic746.serverprofi24.com [85.25.43.31]	443	80	✗	2017-10-27	GD-EMEA-DC-
PartitoPirata	26195	23 h	pad.partito-pirata.it [188.166.16.91]	9001	None	✗	2015-03-06	DIGITALOCEAN

The public listed TOR relays (of which there are about 7000) lists about 70% of actual relays, the rest are intentionally withheld and not publicly listed

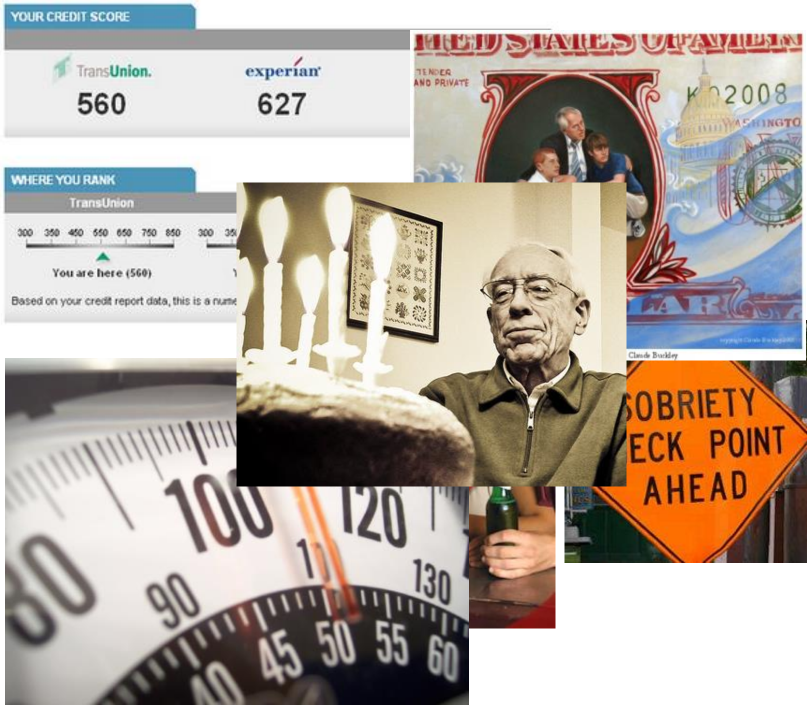
The remaining 30% must be computed using security analytics

Just one more thing…The Problem with Numbers



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

Numbers Help Us Group Things



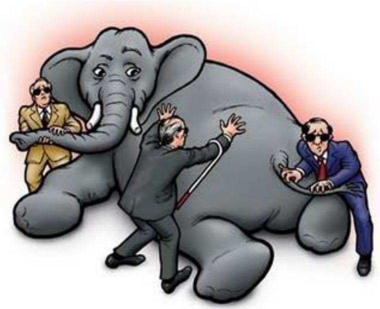
- Credit-worthiness Class
- Legal to drink / Legally drunk
- Weight Class
- Socioeconomic Class
- Age Class

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

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

Summary of the Transformational Trends

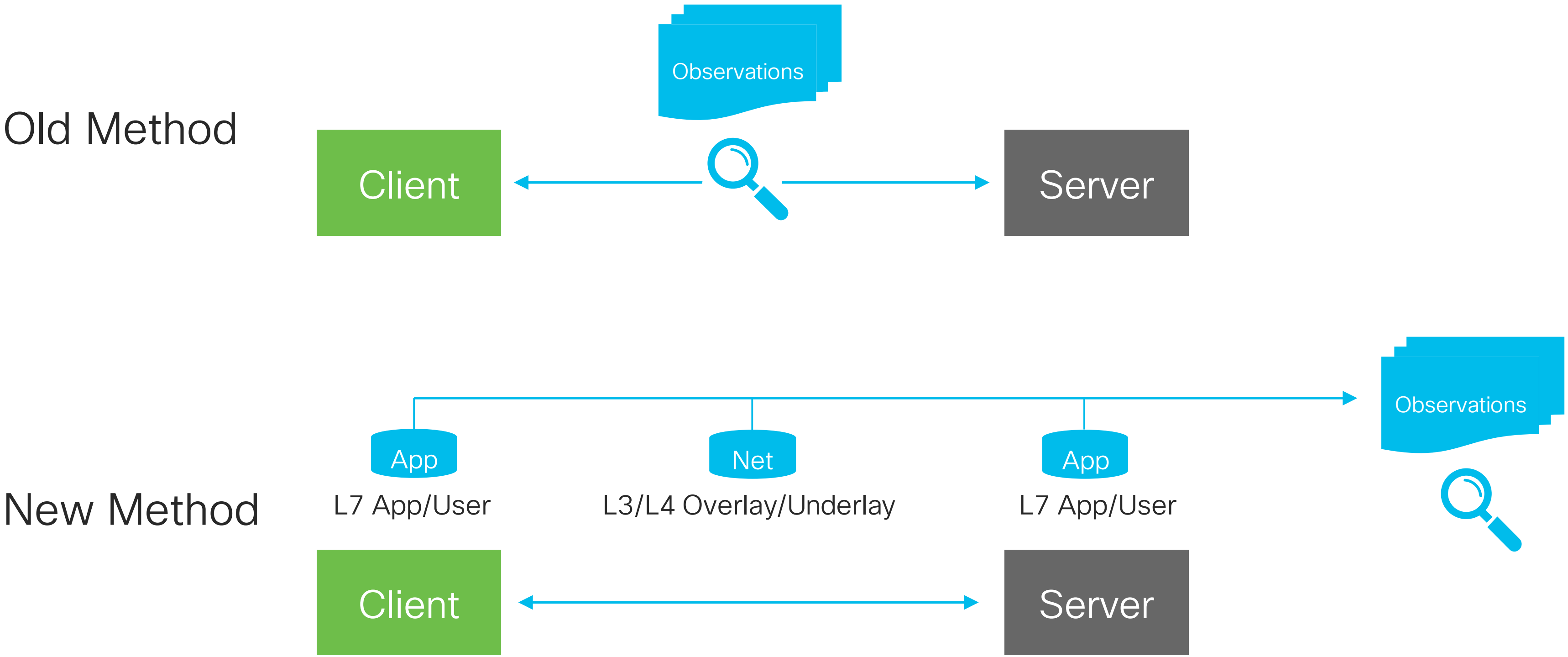
- Through the lens of these changes, will your solutions **remain effective**?
- What **new telemetry** becomes necessary and sufficient for your analytics?
- What **integrations** become deprecated or more valued?
- We no longer are able to view ‘X’ for evaluation, we must infer ‘X’!
- While your analytical outcome may remain the same, what are you are defending and the multiplicity of telemetry will change!

Challenges for Security Analytics We Must Solve

- Network Overlays and Observational Opacity
- Each observation point has less observables
- Serverless (securing a server when there is no server)
- Numbers fail us when we don’t have stable semantics on what is being counted

Future Security Analytics

Direct Versus Indirect Observations



Late-binding Modeling to Detect Security Events

Dynamic Entity Modeling

Collect Input → **Perform Analysis** → **Draw Conclusions**

Collect Input
- IP Meta Data
- System Logs
- Security Events
- Passive DNS
- External Intel
- Vulnerability Scans
- Config Changes

Dynamic Entity Modeling

Perform Analysis	Draw Conclusions
Role	What is the role of the device?
Group	What ports/protocols does the device continually access?
Consistency	What connections does it continually make?
Rules	Does it communicate internally only? What countries does it talk to?
Forecast	How much data does the device normally send/receive?

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

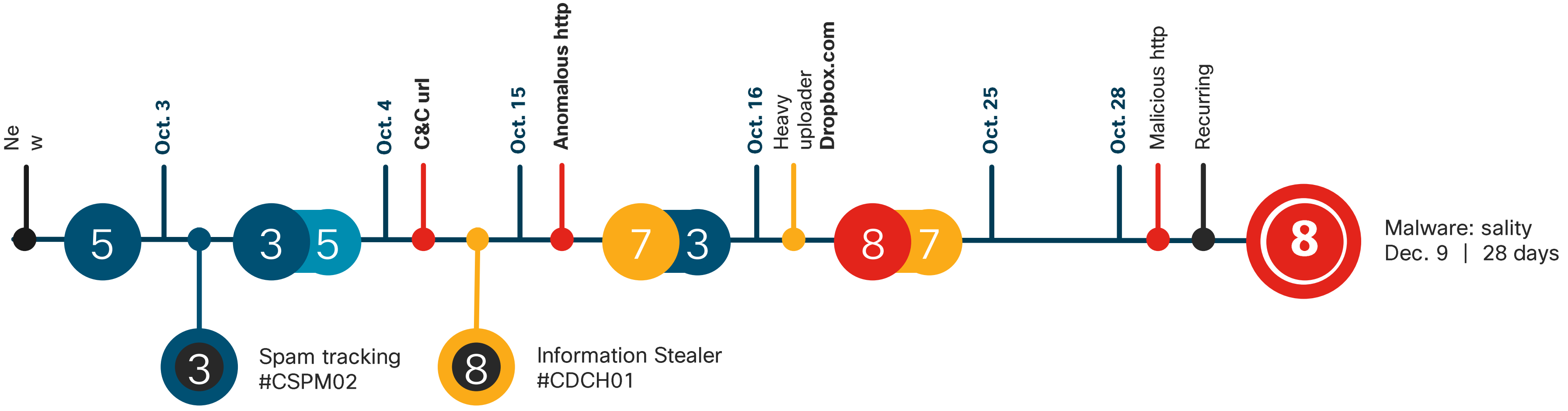
Event Classification and Entity Modeling

- 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



Serverless Security

How Can You Secure a Server When There is No Server?

Serverless Computing is a cloud computing execution model in which the cloud provider dynamically manages the allocation of machine resources (ie the servers)

Serverless Anomaly Detection

Amazon Lambda function that normally connects to two internal resources connecting to an unexpected third

Historical Outlier Observation

One of the source's metrics deviated significantly from its historical baseline.

Time	Source	Time Window	Type	Metric	Expected Value	Outlier	Probability	Sample Size	
3/13/17 12:00 AM	lambda:RDSQueryLogger	1d	device	Bytes Out	12,097,460.313	142,117,719	0.37%	42	×
3/13/17 12:00 AM	lambda:RDSQueryLogger	1d	device	Internal Bytes Out	12,097,460.313	142,117,719	0.37%	42	×

Static Connection Set Deviation Observation

Device normally talks to a static set of (internal/external) devices, but has recently started/stopped talking to new/normal devices.

			Normal Connections		New Connections		Lost Connections			
Time	Source	Type	Set	Count	Set	Count	Set	Count	History Length (Days)	
3/13/17 12:00 AM	lambda:RDSQueryLogger	internal	10.0.10.193 , 10.0.12.134	2	10.0.255.29	1	-	0	35	×

Serverless Detection of an Unusual API Call

AWS CloudTrail Event Observation

AWS CloudTrail event reported for the device.

Time	Source	Account ID	User	Source IP	Event
3/28/17 8:23 AM	Network	757972810156	awslambda_963_20170328112232282	54.91.191.63	DeleteNetworkInterface
3/26/17 12:44 PM	Network	757972810156	awslambda_346_20170326162935979	54.91.191.63	DeleteNetworkInterface

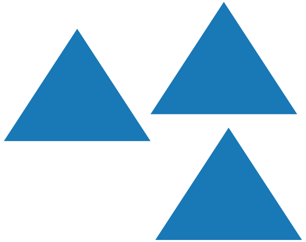
Serverless Behavioral Analytics

AWS Lambda Metric Outlier Observation

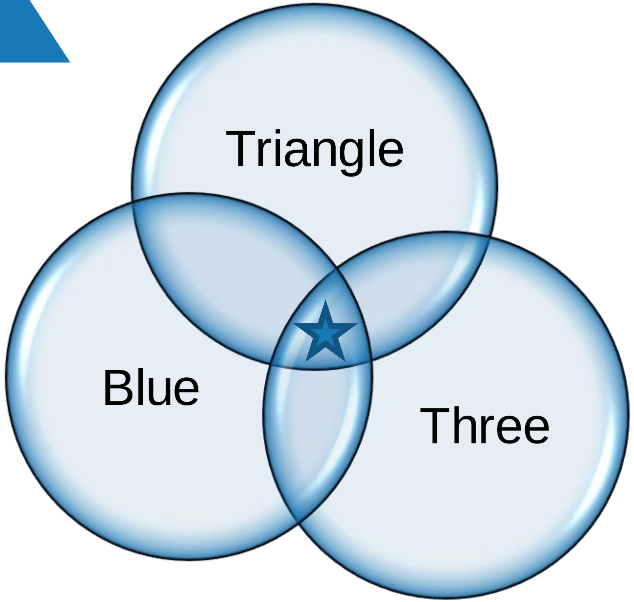
An AWS Lambda function had unusual activity on one of its metrics.

Time	Source	Account ID	Function name	Metric	Old value	New value
3/30/17 9:00 PM	192.168.43.147	23456789012	lambda:rds-poller	Invocations	21	182

Thinking in Sets/Class and Membership

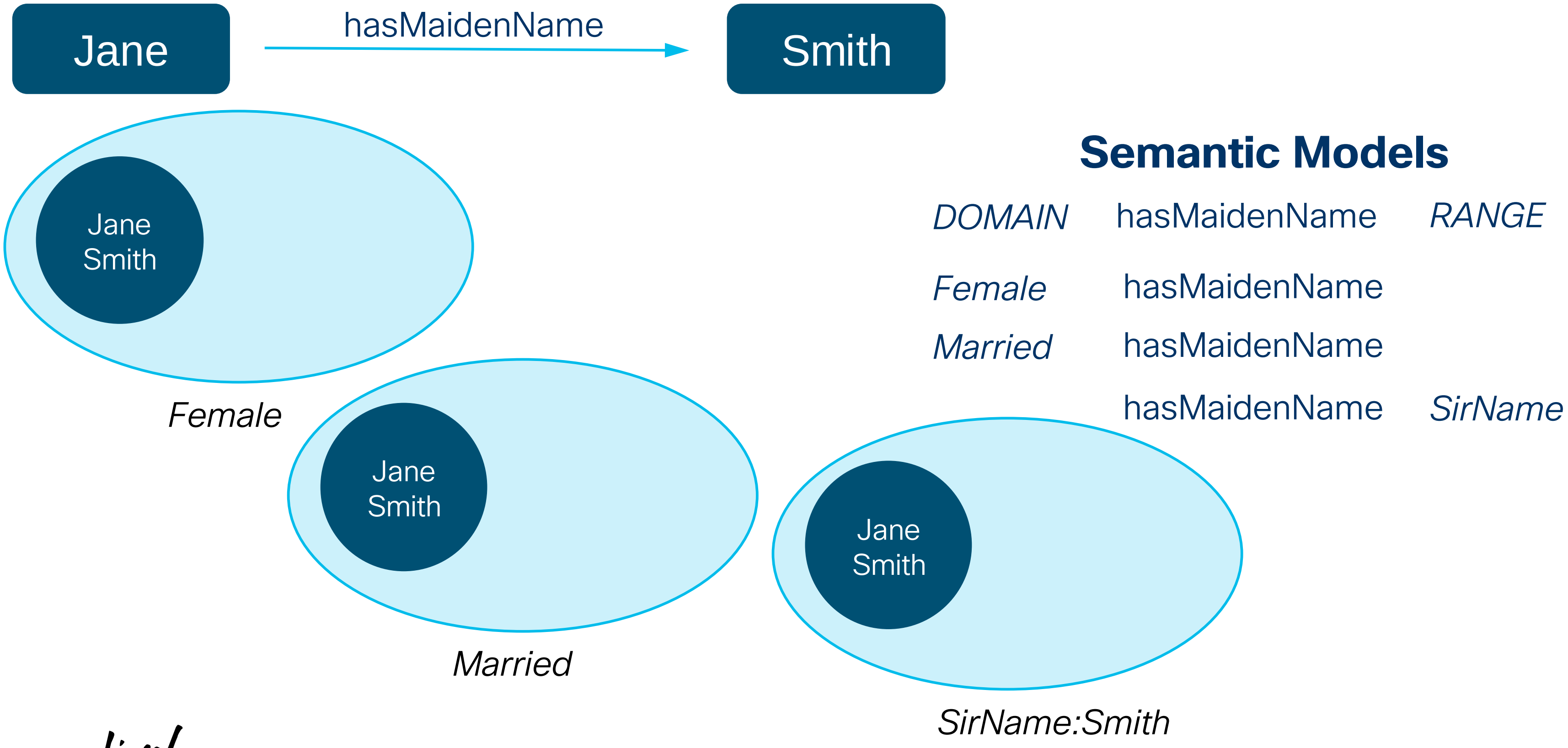


There are 3 blue triangles



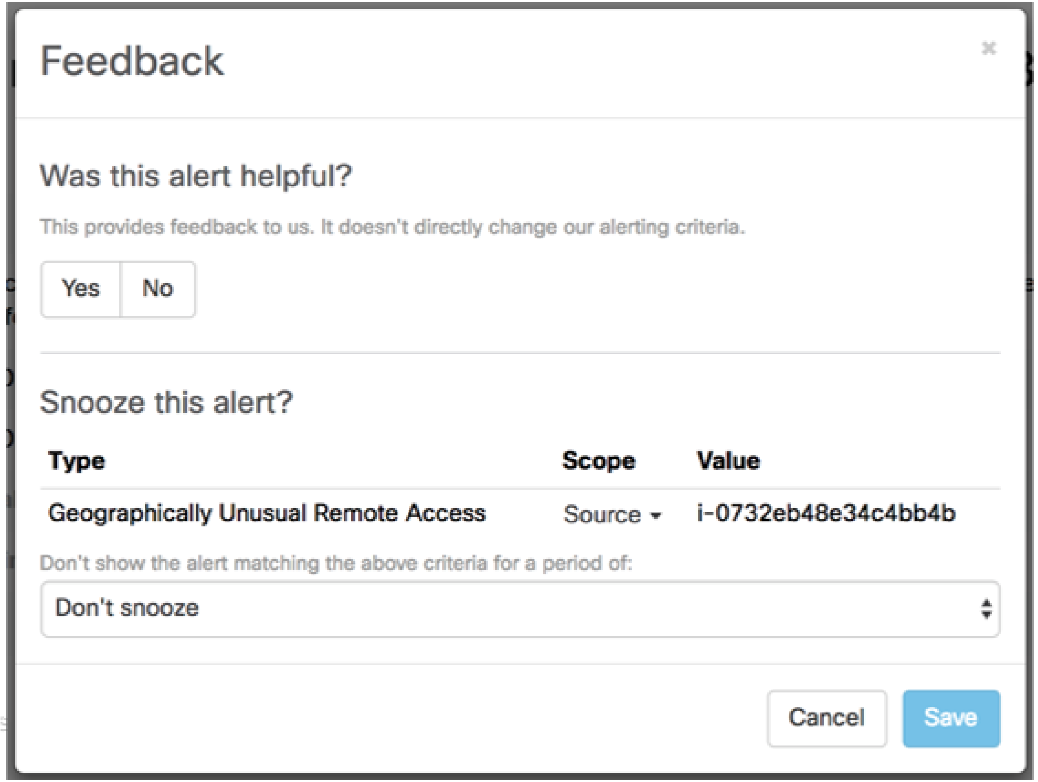
…is a member of the intersection of
the set Blue,
the set Triangle,
and the set Three

Reasoners (side step the numbers problem with first order logic)



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

How Helpful Was This Alert?



2018	Stealthwatch Cloud Alerts Marked Helpful by Customers (%)
Jan	95.91%
Feb	94.52%
Mar	94.75%
Q1 (Jan-Mar)	**94.45%**
Apr	97.23%
May	94.97%
Jun	91.70
Q2 (Apr-Jun)	**94.63%**

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’?

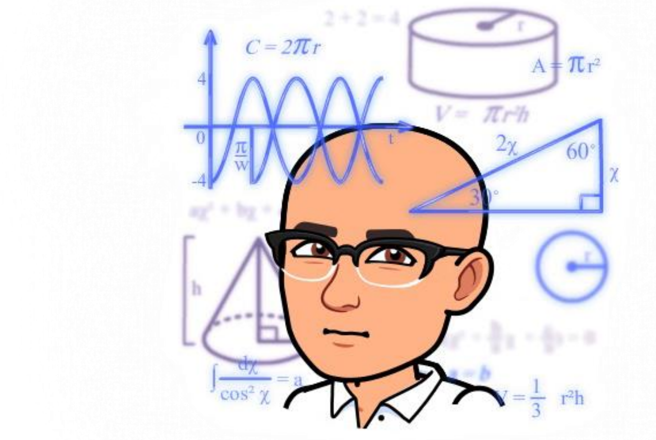
What non-machine learning are you using and why?

What papers or open-source have you published regarding your analytics?

For the ML based assertions, what entailments are provided?

Closing Thoughts

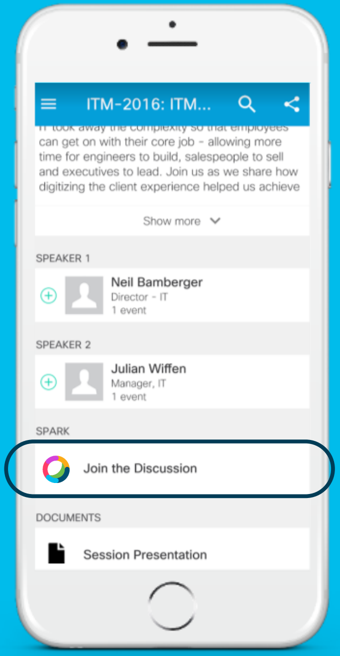
- Be Pragmatic
- Provide Entailments
- Analytical pipeline, over single technique
- Measure helpfulness, not mathematical accuracy
- Be Transparent with your science, publish papers and open source



Recommended Sessions

World of Solutions - Cloud Protect booth (Stealthwatch with Kubernetes & Serverless Security Demo)
World of Solutions – SOC and ThreatWall (Encrypted Traffic Analytics Live)

BRKSEC-3014 – Security Analytics with Stealthwatch: Operationalising Visibility and Machine Learning – Matt Robertson – Friday, Feb 1, 9:00 AM – 11:00 AM
BRKSEC-2323 – Claim Jumpers: Dealing with Illicit Bitcoin Miners – Matt Robertson – Thursday, Jan 31, 2:30 PM – 4:00 PM



cs.co/ciscolivebot#BRKSEC-2068

Cisco Webex Teams

Questions?

Use Cisco Webex Teams (formerly Cisco Spark) to chat with the speaker after the session

How

1. Find this session in the Cisco Events Mobile App
2. Click “Join the Discussion”
3. Install Webex Teams or go directly to the team space
4. Enter messages/questions in the team space

Complete your online session survey

- Please complete your Online Session Survey after each session
- Complete 4 Session Surveys & the Overall Conference Survey (available from Thursday) to receive your Cisco Live T-shirt
- All surveys can be completed via the Cisco Events Mobile App or the Communication Stations

Don’t forget: Cisco Live sessions will be available for viewing on demand after the event at ciscolive.cisco.com

Five * * * * *
5

Continue Your Education

Demos in the Cisco Showcase

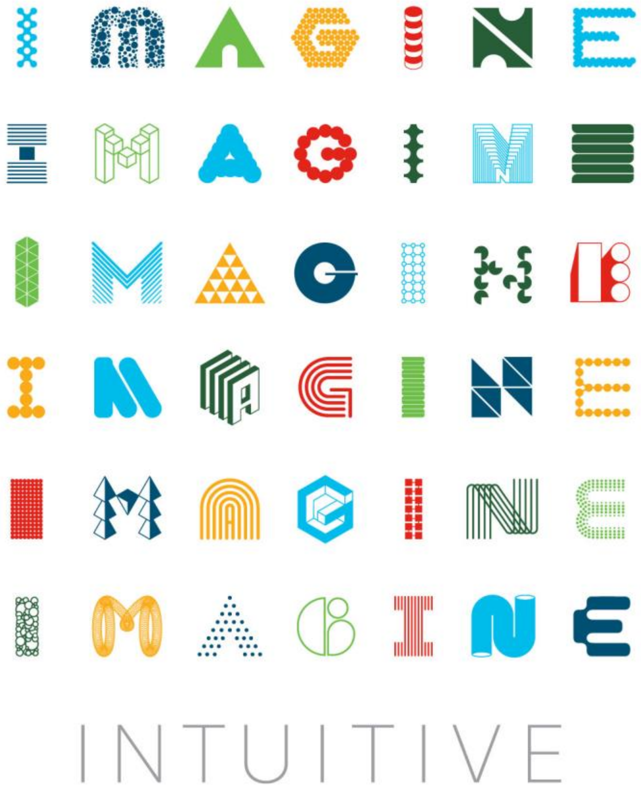
Walk-in self-paced labs

Meet the engineer 1:1 meetings

Related sessions

cisco

Thank you



IMAGINE

INTUITIVE

References

Learn More....

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

Basic References

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

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).

IMAGINE

INTUITIVE