RSAConference2018
San Francisco | April 16–20 | Moscone Center

SESSION ID: SPO2-R04

DEMYSTIFYING BIG DATA, ANALYTICS, AND MACHINE LEARNING IN CYBER SECURITY

Mary Writz

Director, Product Management
Micro Focus

History of SOC Evolution

And the move toward big data for breach detection

- **Birth of the Internet:** businesses not connected, or via slow connections
- **Security tools appear – SATAN / SAINT**, etc…
- **Military & Governments start to build SOCs and CERTs**

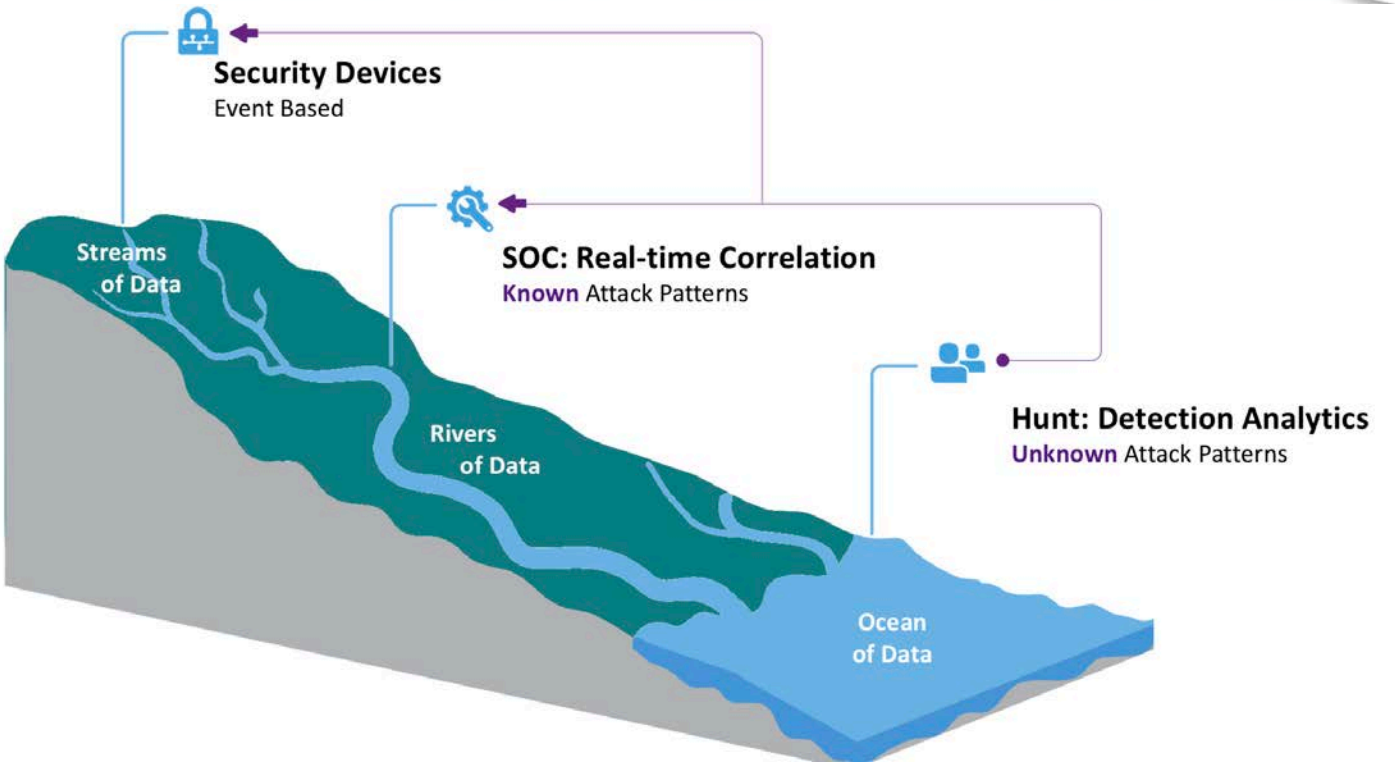
- **Malware outbreaks and Security Event Management (SEM)** concepts introduced
- **Standards & Intelligence** emerge: CVE, PacketStorm, SANS Internet Storm Center

- **Botnets, cybercrime,** IPS
- **Compliance requirements**: SOX, PCI, PII
- **Internal SOCS created Internet Meltdowns:** Slammer, Blaster, Sobig
- **Event volume problems**

- **DDoS, Hacktivism**, Intellectual Property Theft, **Advanced Persistent Threat**
- **Monetization and espionage**
- Attribution incredibly difficult

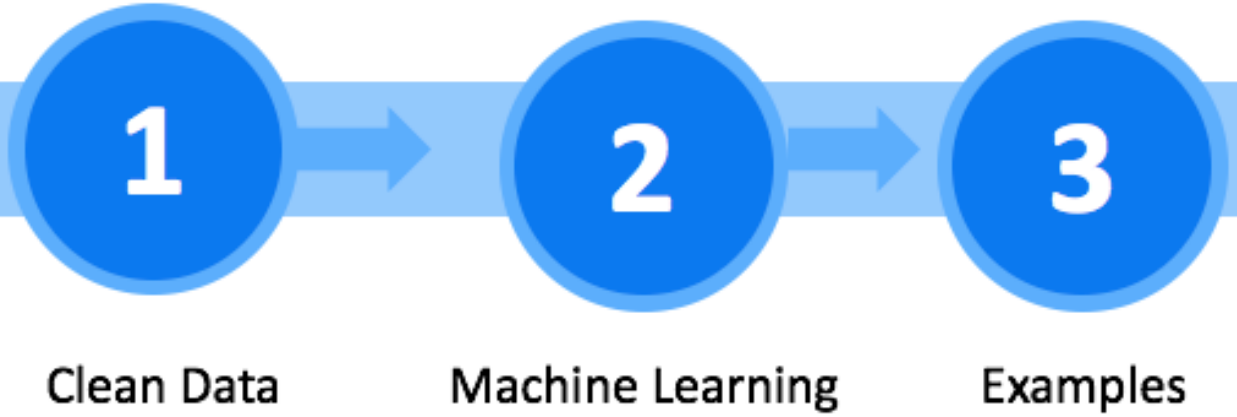
The Pivot to Big Data is Natural

Enabled by advancements in data storage, ingestion, & processing speed



Three Areas of Discussion Today

Clean data, a debrief on machine learning, practical examples



CLEAN DATA

Let's Talk About Structured Data

The hardest but most important step

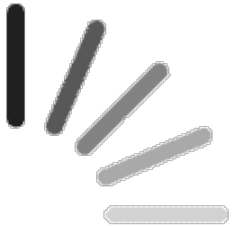
Hi, my name is Mary

I love to talk about my successes and failures building hunt teams and pioneering detection analytics

#1 Lesson:
Clean data matters

You'll Spend A Lot of Time Cleaning Data

Without this work, you will spin your wheels



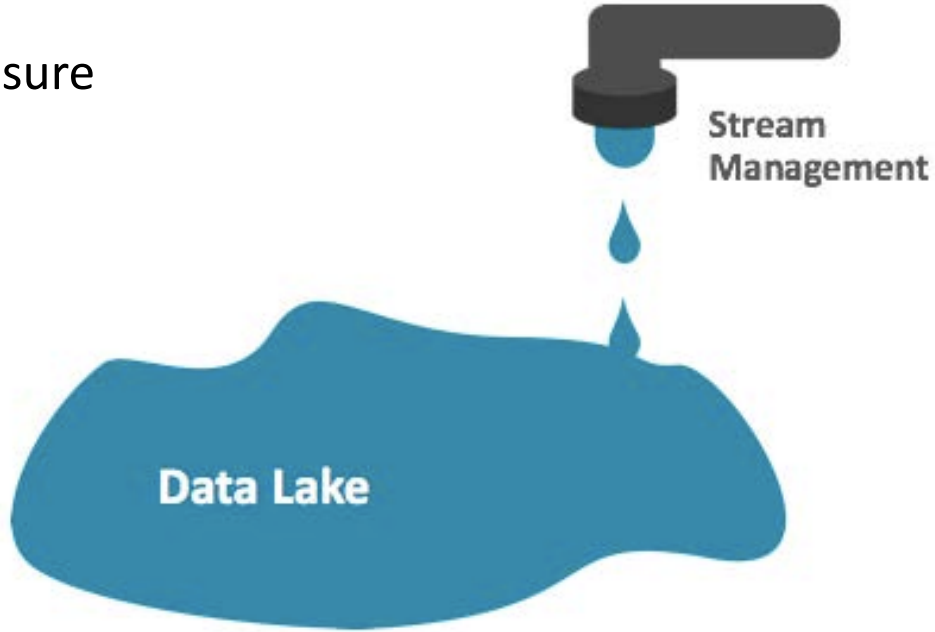
Data Integration

Making sure the data makes it to where it needs to be

Tooling and processes to ensure accurate data collection & persistence.

Challenges:

- Velocity
- Variety
- Volume



Data Exploration

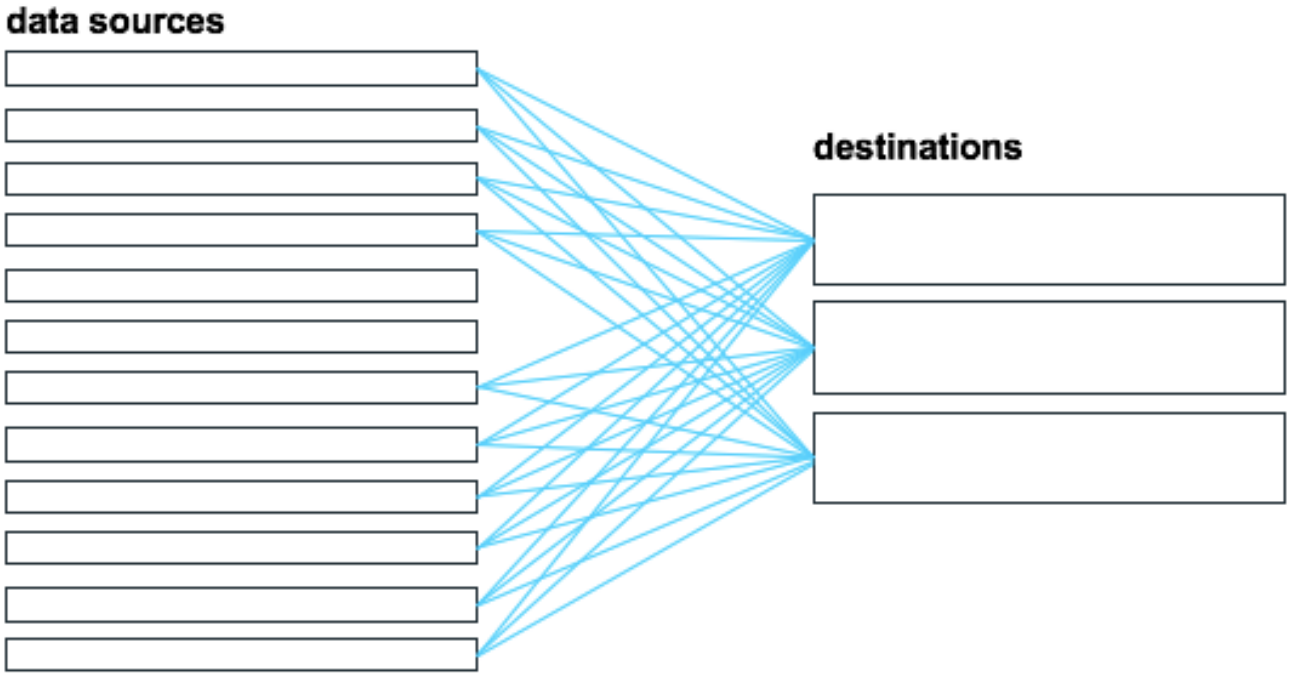
Ensuring the flexible for the many ways you may want to use it

Tooling and processes on your data which provide flexible & capable:

- access
- aggregation
- manipulation

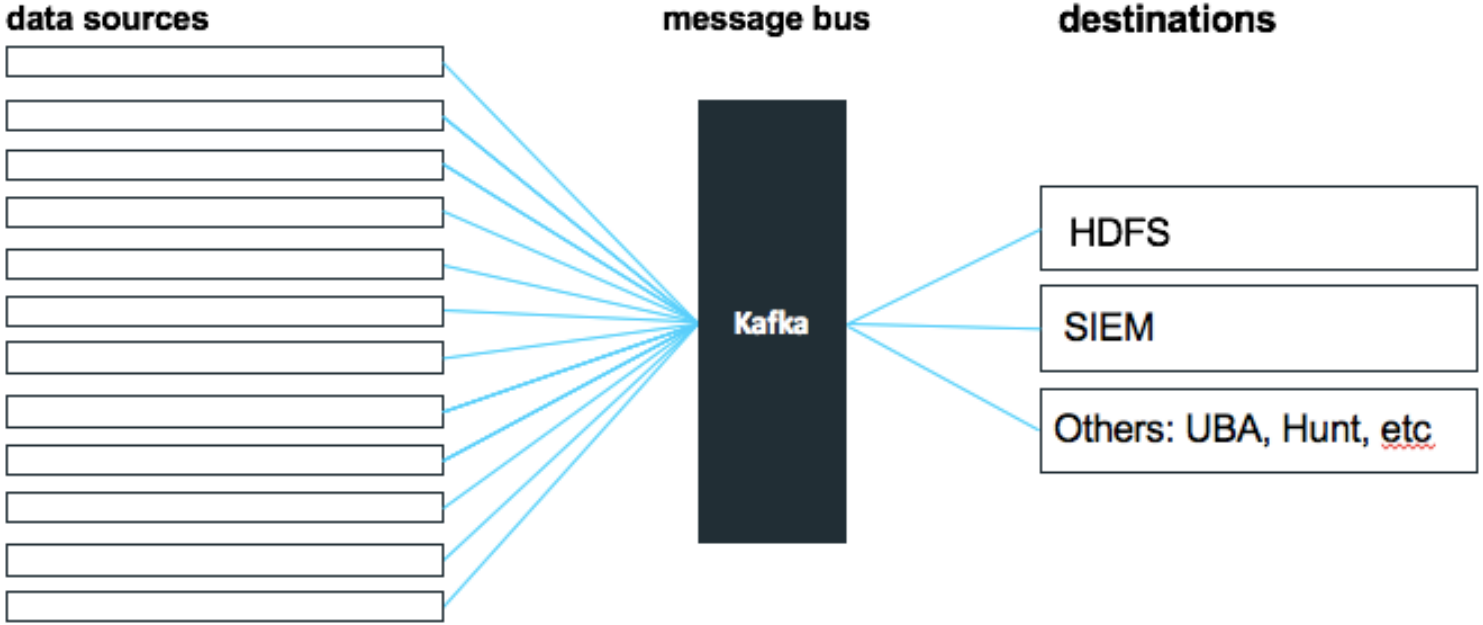
Flexibility Introduces Complexity

It starts to look a lot like this



Here's a Nice (Common) Way to Clean It Up

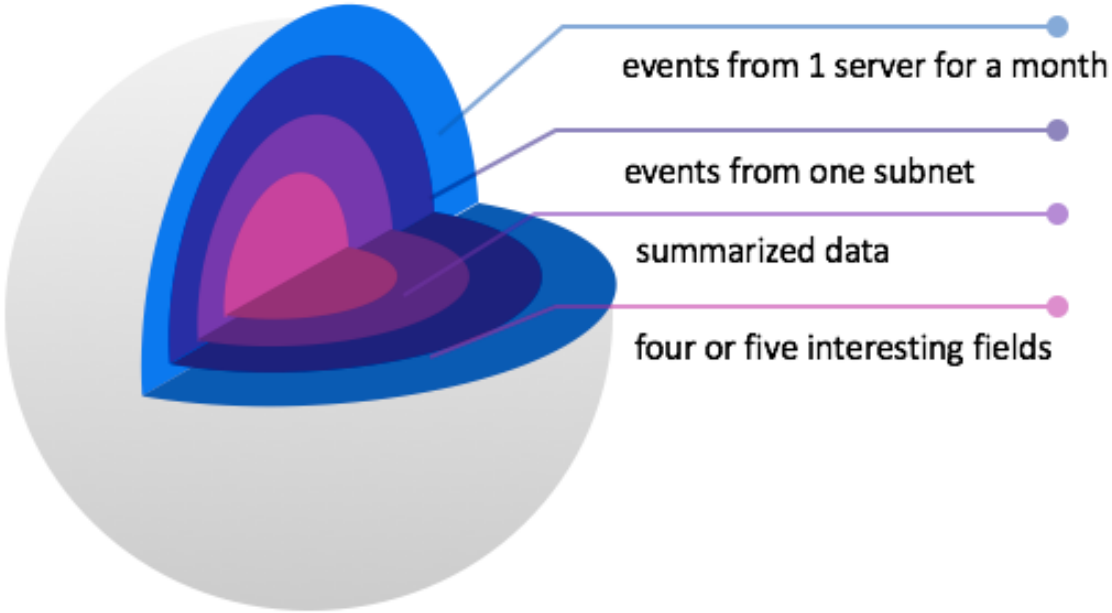
Collect once, use anywhere



P.S. A Shout Out For “Small Data”

Avoid data canyons and wasted efforts by starting small, then expand

Start small and sample several ways



DEMYSTIFYING MACHINE LEARNING

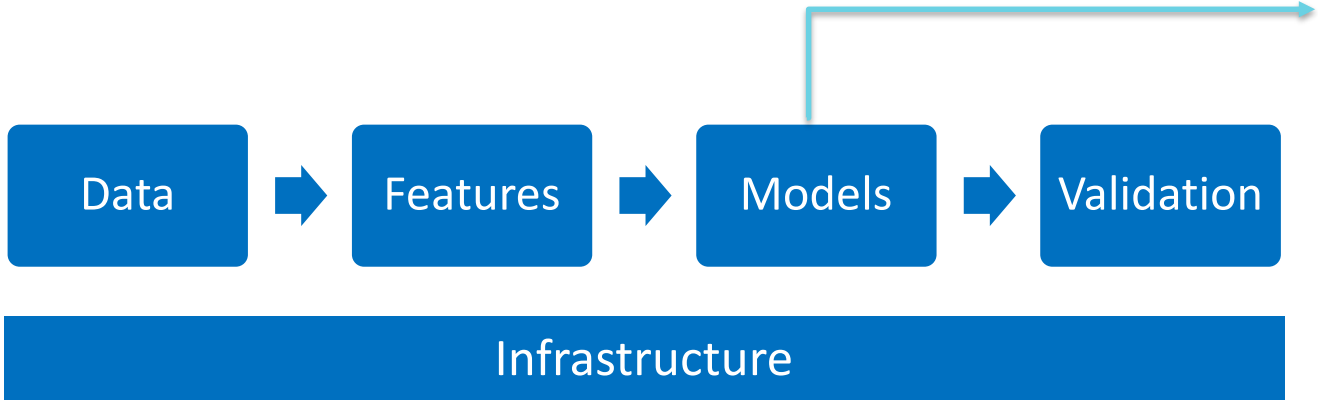
The Power & Possibility of Machine Learning

Knowns

Known Unknowns

Unknown Unknowns

A Machine Learning Approach



Common Models

Unsupervised
K means
KDE (Kernel Density)

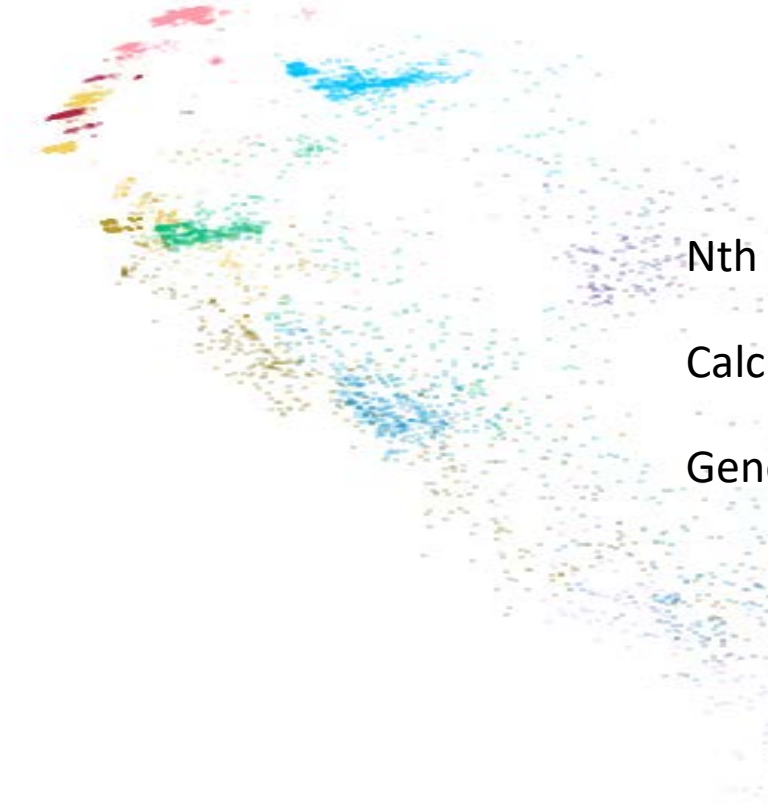
Supervised
Random Forrest
Bayesian Probability
Decision Tree
Neural Network

Models of Machine Learning

Let's talk about three models:

1. **Black box** approaches – ex. Clustering, Neural Networks
2. **Fine Tune** approaches – ex. Probability Histograms (KDE)
3. **A Special Note on Neural Networks** – Hint: cyber security not ready for this

Black Box: K Means Clustering

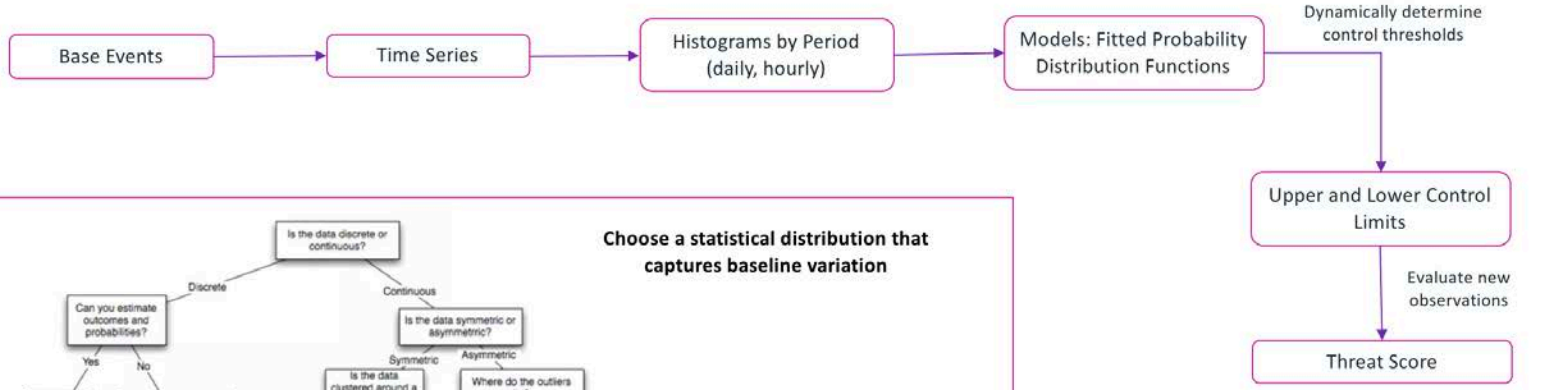


Nth degree clustering

Calculating distances from center

General Anomaly Detection

Fine Tune: Kernel Density Estimation



Source: https://www.abstractclasses.in/2017/10/08/table-probability-distribution-functions/

http://www.statisticalprocesscontrol.info/training.html

Neural Networks

You need data to learn. Cyber NOT good at sharing data.

What is it? Learning without task-specific programming

Why is it not ready for cyber security?

- Lack of training data
  - Millions of examples needed
  - Cyber not good at sharing data
- Bespoke cyber attacks
  - Threat Intel overlap of 4%
  - Detecting attack very different than detecting picture of cat
- Black box findings
  - You can't explain findings

Image from: paulvanderlaken.com

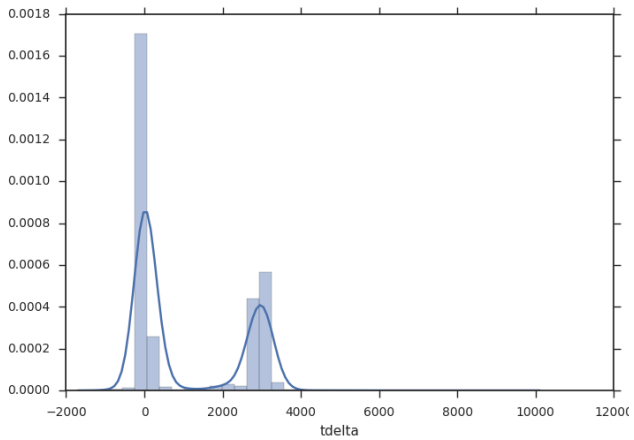
ANALYTICS EXAMPLES

i.e. Things that work!

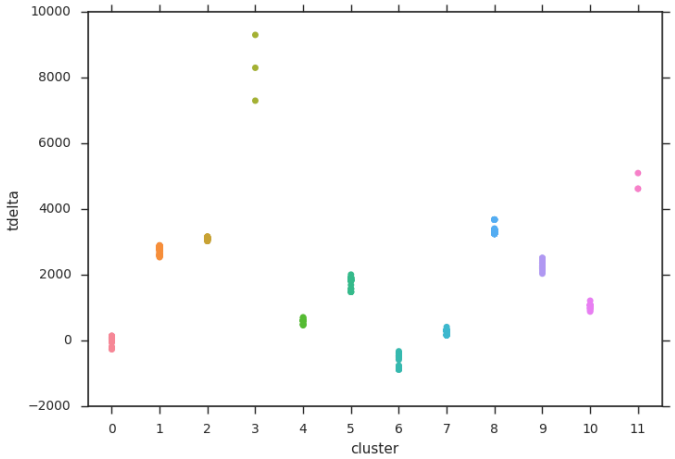
Visualization/Clusters
Anomalous file updates

TIME STAMP

What? Explore timestamps on multiple files on a file system.
Why? Typically files are updated on a schedule so all files get updated in a close time window.
How? Cluster based on time deltas and see if there are anomalies to hunt.



Visualization of clustered time deltas. 4 major times files were manipulated.



Exploding the clusters – Cluster 3 looks interesting

"With a little bit of investigation, we radically reduced our scope to be able to hone in on a few fishy ones."

Visualizations/Entropy

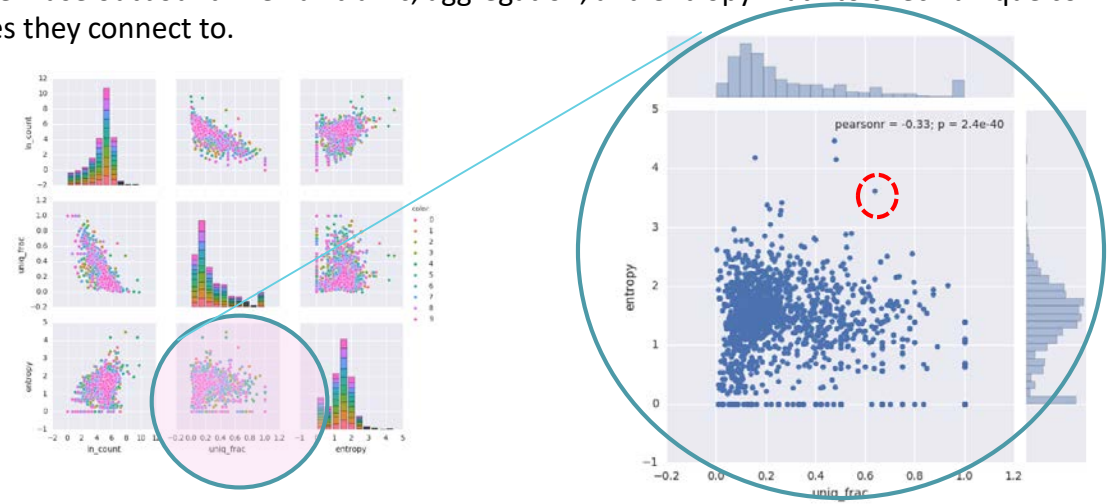
Anomalous outbound activity

OUTBOUND SCANNERS

What? Look for systems that are performing outbound scans.

Why? Compromised systems look for vulnerabilities on other systems or try to map the network – activity overlooked by IDS (little pattern and no malicious traffic sent.)

How? We'll use outbound firewall traffic, aggregation, and entropy math to check unique connection destinations and randomness of addresses they connect to.



"Using this technique, we can narrow down from thousands of IP addresses to a handful of interesting and noteworthy addresses. Sometimes, **making the haystack much smaller** is a huge win."

Scatterplots help us look for high event counts + high entropy (things floating up + right)

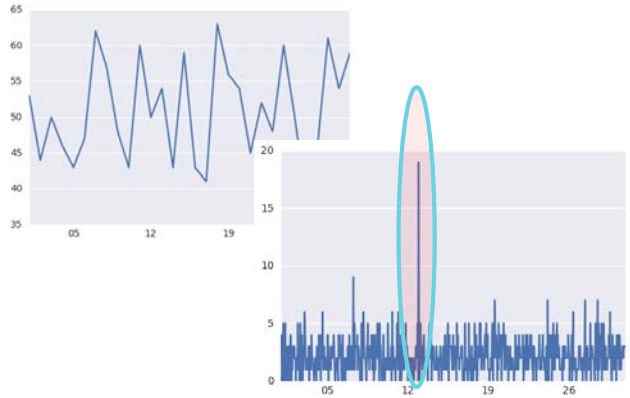
Top right corner contains hosts that we are interested in, we single out the worst offender.

Eureka! IP is malicious

The Power of Simple Averages
Anomalous audit log activity

WINDOWS 1102

What? Exploring when a users clear Windows logs, evidenced by windows creating a ”1102” event.
Why? There is no need to manually clear security event logs (in most cases).
How? We plot activity by time and pivot between time windows to see if we can unearth anomalous activity. Using avgs can be useful provided we don’t normalize an attack into our average.



	rt	name	externalId	duser	dhost
15	2015-09-12 23:44:24	The audit log was cleared.	1102	icruz	ab-ex
387	2015-09-12 23:44:25	The audit log was cleared.	1102	rbishop	unde-est
402	2015-09-12 23:44:25	The audit log was cleared.	1102	adavis	nostrum-iabore
428	2015-09-12 23:44:24	The audit log was cleared.	1102	zmorris	dolorictio-quidem
482	2015-09-12 23:44:25	The audit log was cleared.	1102	jrubio	animi-neque
577	2015-09-12 23:44:25	The audit log was cleared.	1102	mtorres	recusandae-necessitatibus
695	2015-09-12 23:44:25	The audit log was cleared.	1102	evillegas	illum-dolores
732	2015-09-12 23:23:22	The audit log was cleared.	1102	ssmith	fugiat-pariatur
762	2015-09-12 23:44:24	The audit log was cleared.	1102	kaguilar	magni-eum
770	2015-09-12 23:44:24	The audit log was cleared.	1102	jbrown	rem-saepe
807	2015-09-12 23:44:25	The audit log was cleared.	1102	jcoleman	dolor-eveniet
852	2015-09-12 23:44:24	The audit log was cleared.	1102	mrobbins	blanditiis-voluptatibus
904	2015-09-12 23:44:24	The audit log was cleared.	1102	mbautista	aliquid-consequuntur
992	2015-09-12 23:44:25	The audit log was cleared.	1102	acherry	quas-illo
1001	2015-09-12 23:44:24	The audit log was cleared.	1102	doraig	omnis-tempora
1231	2015-09-12 23:44:24	The audit log was cleared.	1102	iwilson	earum-consequatur
1262	2015-09-12 23:21:57	The audit log was cleared.	1102	lbaker	nam-quo
1375	2015-09-12 23:44:25	The audit log was cleared.	1102	dvaughan	dicta-magni
1409	2015-09-12 23:44:24	The audit log was cleared.	1102	mharper	rerum-magnam

“When we averaged by day the spike was diluted, when we averaged by hour, the spike shows prominently.”

Looking at events by day and also by hour. On the latter we see an interesting spike.

Exploding events to understand the spike near midnight on the 12th.

Logs cleared on 12 machines at the same time. Suspicious!

Visualizing Hard Coded Patterns

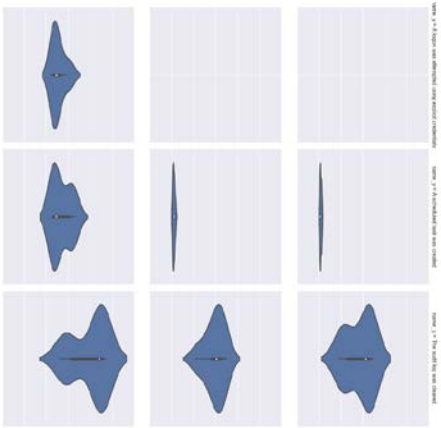
Lateral movement

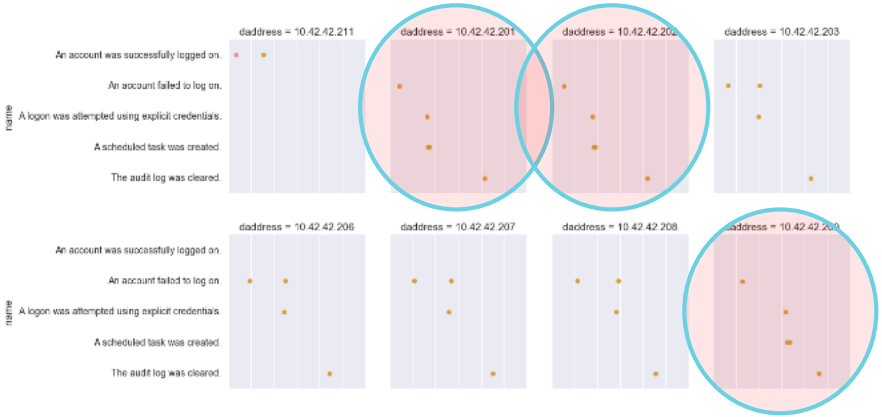
SMB ACTIVITY

What? Exploring the tracking of lateral movement through a Windows network with the use of SMB logons, privilege escalation, scheduled tasks, and possibly the subsequent clearing of audit logs.

Why? Because SMB is a trusted protocol, an attacker can use it for malicious purposes, which can be tracked and detected.

How? We look for successful logons via the SMB protocol, looking for a series of events associated with lateral movement.





When we took the previous graphs and found the common thread of the audit logs being cleared, we were able to locate the source address that was common to all destinations.

These are multiple series of pairs of events that happened sequentially over time. Empty graphs represent pairs of events that did not occur. This did not yield a positive find but helped us.

The plot graphs represent individual destination addresses and the series of events that occurred specific to that address.

We tracked back the source address to positively identify the machine that was first compromised

Distance Algorithms
Process impersonation

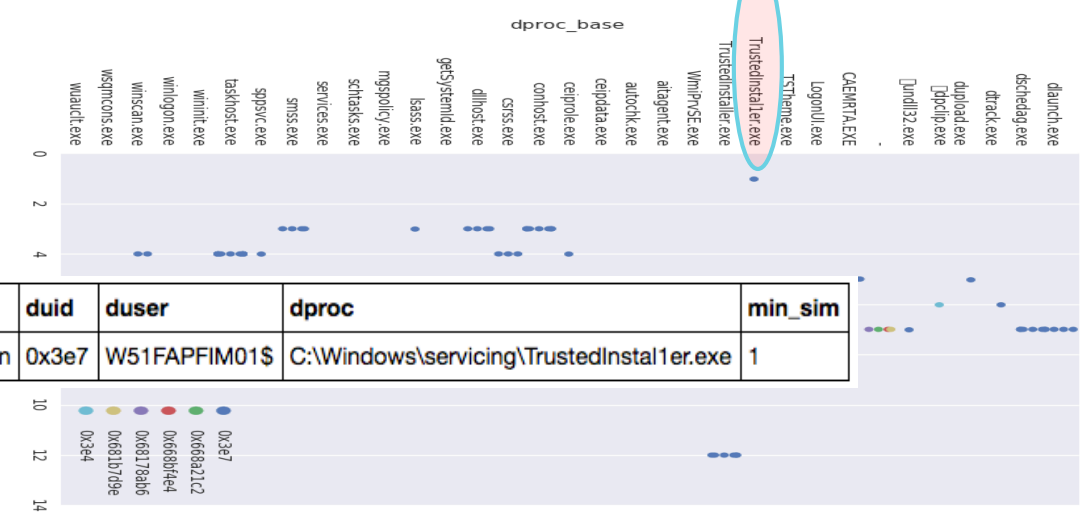
SUSPICIOUS PROCESS IMPERSONATION

What? Exploring process names on Windows systems

Why? Adversaries hide suspicious processes (malware) using similar names to legitimate windows processes.

How? Using scores from Damerau-Levenshtein distance algorithm to judge which pairs of process names are more or less similar than others.

The distance between 'svchost.exe' and 'scvhost.exe' is 1 (*transpose the 'v' and the 'c'*). With this we identify binaries potentially masquerading as critical system processes.



	rt	dhost	duid	duser	dproc	min_sim
86	1502379430300	W51FAPFIM01.rins-bank.intra.rin	0x3e7	W51FAPFIM01$	C:\Windows\servicing\TrustedInstal1er.exe	1

Using a dictionary of known good windows system32 and syswow64 binaries the algorithm calculates the distance.

The plot graphs represent individual processes with their individual distance scores.

The low score of 1 stands out – this process identified as an imposter!

Forbidden Bigrams (and Trigrams)

DGA activity

"bk", "bq", "bx", "cb", "cf", "cg", "cj", "cp", "cv", "cw", "cx",

"dx", "fk", "fq", "fv", "fx", "fz", "gq", "gv", "gx", "hk", "hv",

"hx", "hz", "iy", "jb", "jc", "jd", "jf", "jg", "jh", "jk", "jl",

"jm", "jn", "jp", "jq", "jr", "js", "jt", "jv", "jw", "jx", "jy",

"jz", "kq", "kv", "kx", "kz", "lq", "lx", "mg", "mj", "mq", "mx",

"mz", "pq", "pv", "px", "qb", "qc", "qd", "qe", "qf", "qg", "qh",

"qj", "qk", "ql", "qm", "qn", "qo", "qp", "qr", "qs", "qt", "qv",

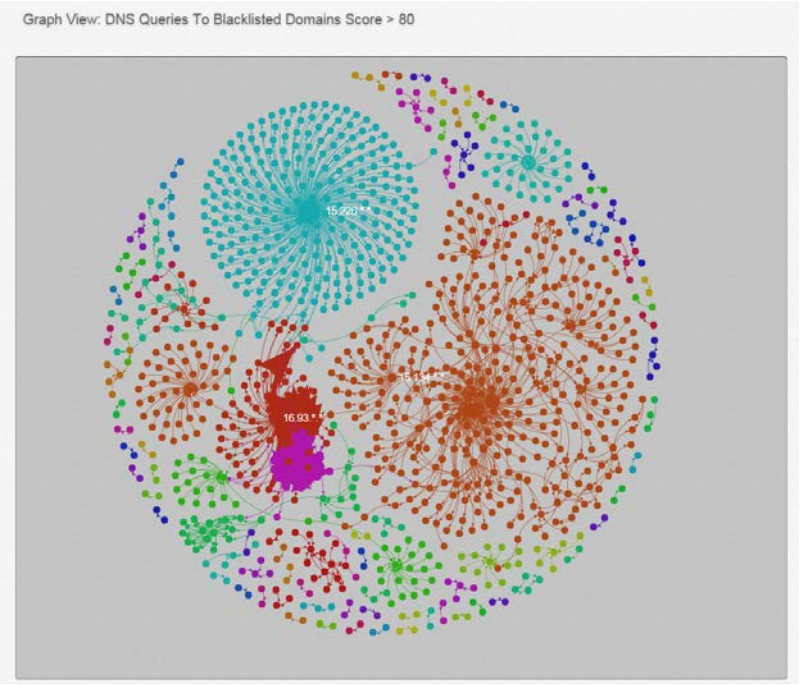
"qw", "qx", "qy", "qz", "sx", "sz", "tq", "tx", "vb", "vc", "vd",

"vf", "vg", "vh", "vj", "vk", "vm", "vn", "vp", "vq", "vt", "vw",

"vx", "vz", "wq", "wv", "wx", "wz", "xb", "xg", "xj", "xk", "xv",

"xz", "yq", "yv", "yz", "zb", "zc", "zg", "zh", "zj", "zn", "zq",

"zr", "zs", "zx"



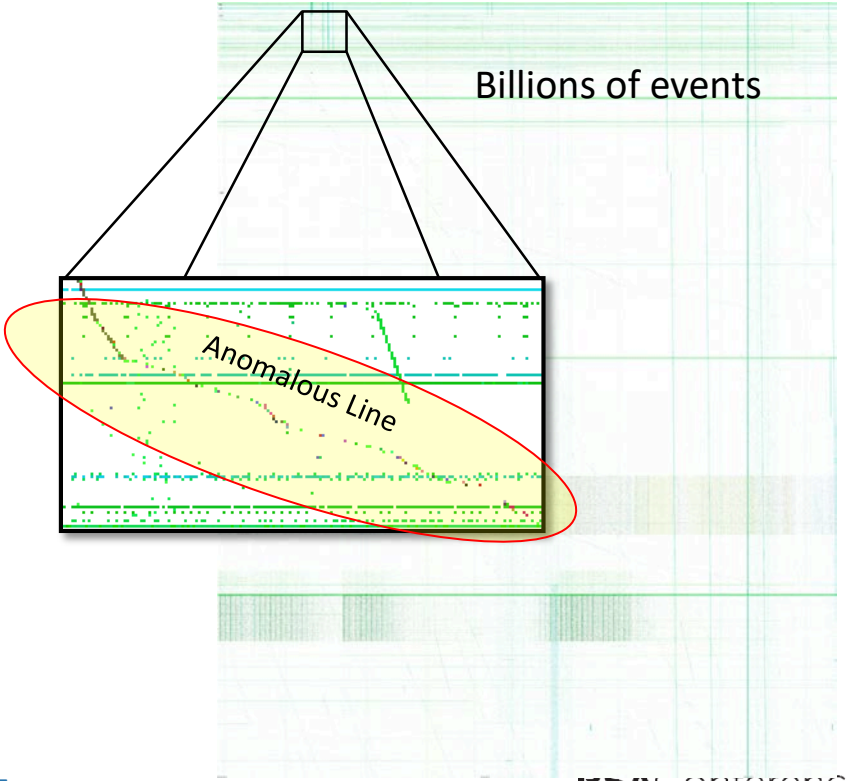
Custom Visualizations (High Cardinality Data)

Low/slow attack

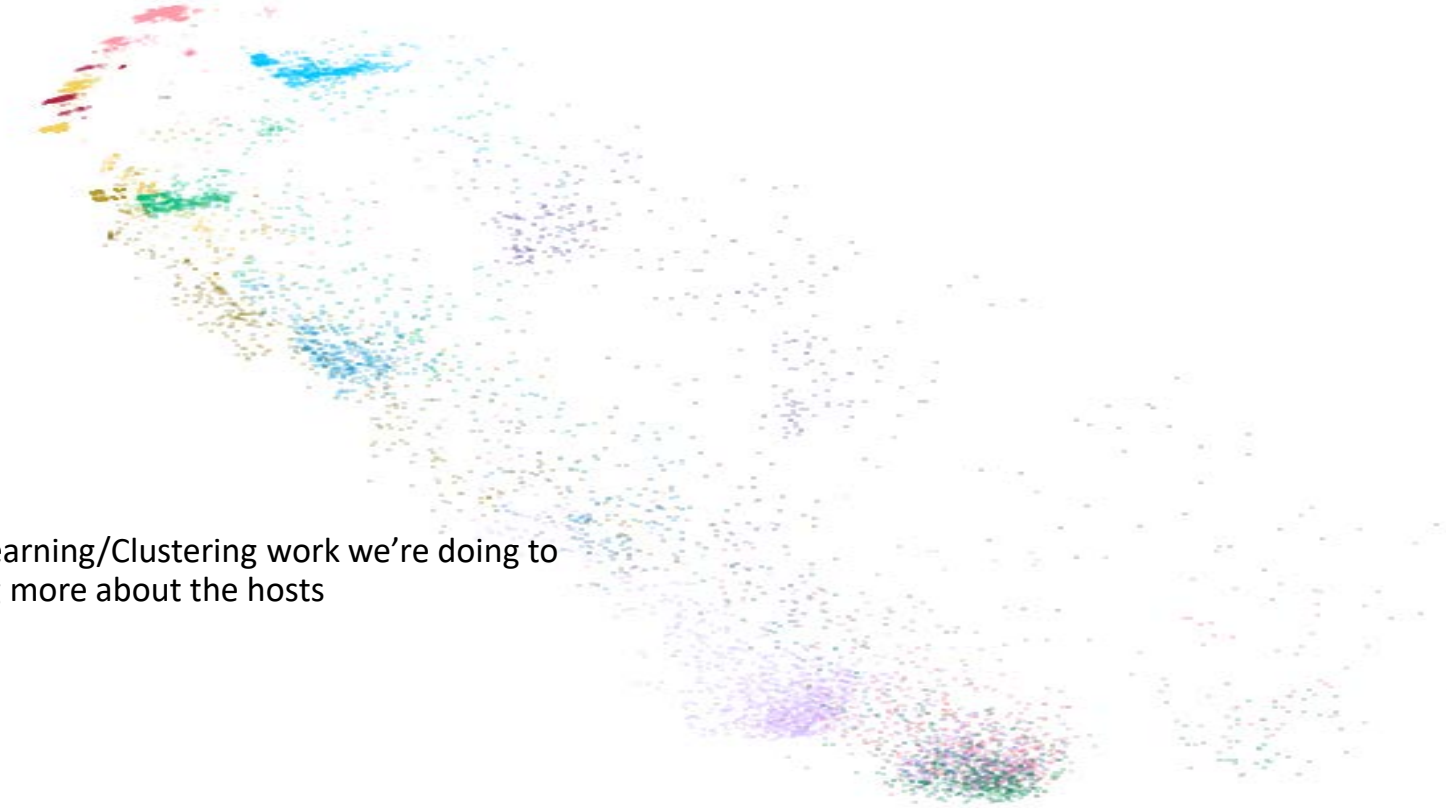
Custom D3 Scatterplot Visualization

Single multi-week scan from distributed, internal sources indicates advanced attacker performing advanced port scan activity

- Y-axis = destination port
- X-axis = Julian date
- Colored by source IP address



We Are Still Learning…



Ex: Machine learning/Clustering work we're doing to understanding more about the hosts

WRAP UP

Recap – It All Starts With Data …

Considerations for building a data store that has a long, useful runway

Data Platform Tradeoffs and considerations

- **Structuring data**: manual classification vs auto classification
- **Where compute happens**: putting compute on data store cluster scales nicely
- **Data store optimization**: ingest cost vs compute cost vs query cost
  - What's more important –**search or analytics?**
- **Combining non-similar data:** what happens when you need to fold in more data
- **Cost / compression:** big data is big…
- **Ecosystem:** open, flexible, control of data, api access, third party tool access (JDBC, ODBC)
- **Multiple data stores:** are they **harmonious**? Same data schema?
- **Separating storage and compute:** The myth of the single data lake: it's a bit of a unicorn…

Key Takeaways

Data

- The power of clean data: considerations on building data store

Analytics

- Most useful ones tend to be use case driven, designed by domain experts
- Don’t overlook simple analytics that are easier and still very powerful

Explore now!

- Small data is a good place to start, grow scale when you find a good analytics use case