SANS

Old Data, New Tricks

Threat Hunting with Data Science

About the Author

David Hoelzer

dhoelzer@enclaveforensics.com

COO Enclave Forensics, Inc.

SANS Fellow & Author

MGT411, MGT521, AUD507, DEV536, SEC503, AUD410, MGT512, DEV543, SEC595

30+ years in IT

20+ years in infosec

AI/ML/Data Science for Infosec Evangelist

Things Have Changed

- When internet dinosaurs roamed the earth…
  - Viruses were just a few years old
  - Publicly accessible systems/accounts had no passwords
  - We paid for time on timesharing systems

Things Have Changed

- When internet dinosaurs roamed the earth…
  - Viruses were just a few years old
  - Publicly accessible systems/accounts had no passwords
  - We paid for time on timesharing systems

- Anomaly detection required expertise and patience
  - Collect everything
  - Write filters to exclude known-good (or assumed good)
  - Evaluate/analyze everything else

Things Haven't Changed

- This might seem old, but things haven't really changed
  - We still detect malware using signatures
    - Heuristics keep trying, but often introduce pain
  - Anomaly detection has moved to the host
    - Allow lists for software / software signing for permitted signatures
  - Real-time threat defense is predominantly signature based
    - Pick your favorite firewall/NGFW/IPS vendor and check their literature
    - "Block *known* attacks"

Threat Hunting Rises

- If you're here:
  - None of this is news
  - You've bought/built/installed/use everything available for defense

Yet, you suspect there are things happening that you are completely unaware of, that none of your tools are reporting to you, and that none of your log analysis tools seem to be finding

What you know you can’t explain. But you feel it. You[feel it constantly]. That there’s something wrong with the [network].
You don’t know what it is but it’s there, like a splinter in your mind driving you mad.
It is this feeling that has brought you to me.
Do you know what I’m talking about?

What Are You Selling?

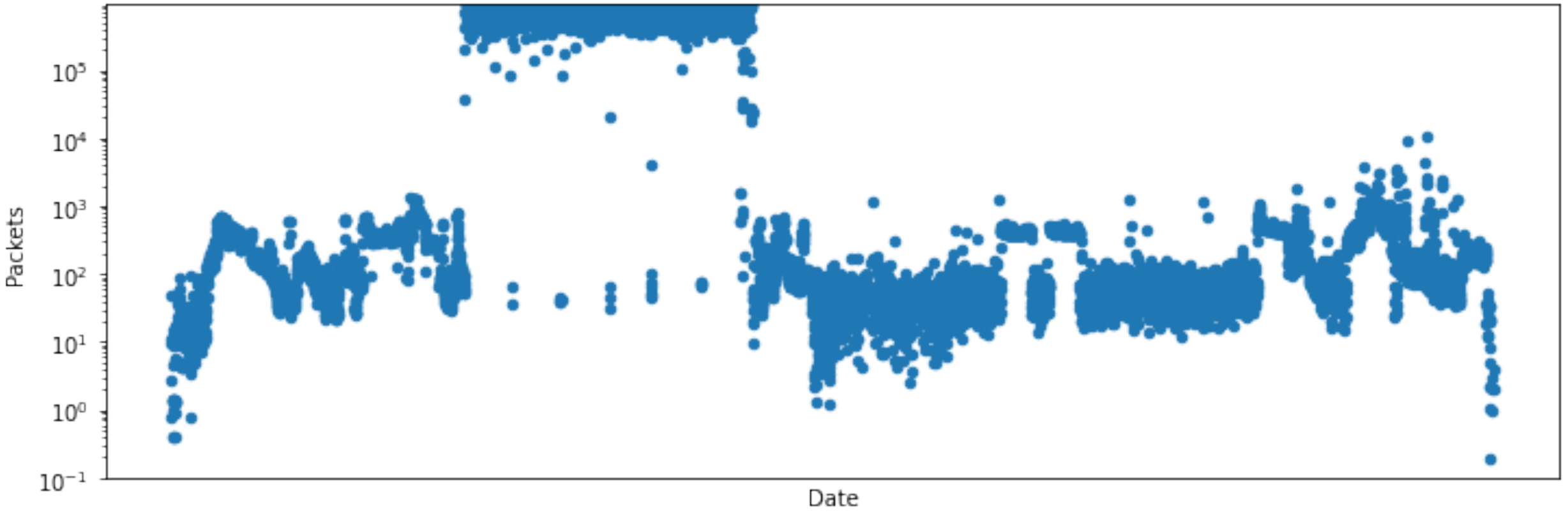
- Nothing
  - Well, not a product anyway!
- You don't need to go buy more tools
  - Everyone is marketing AI/ML tools
  - What do people in your enterprise think those are?
- You already have all of the data
  - All you need to do is *use it*

NetFlow

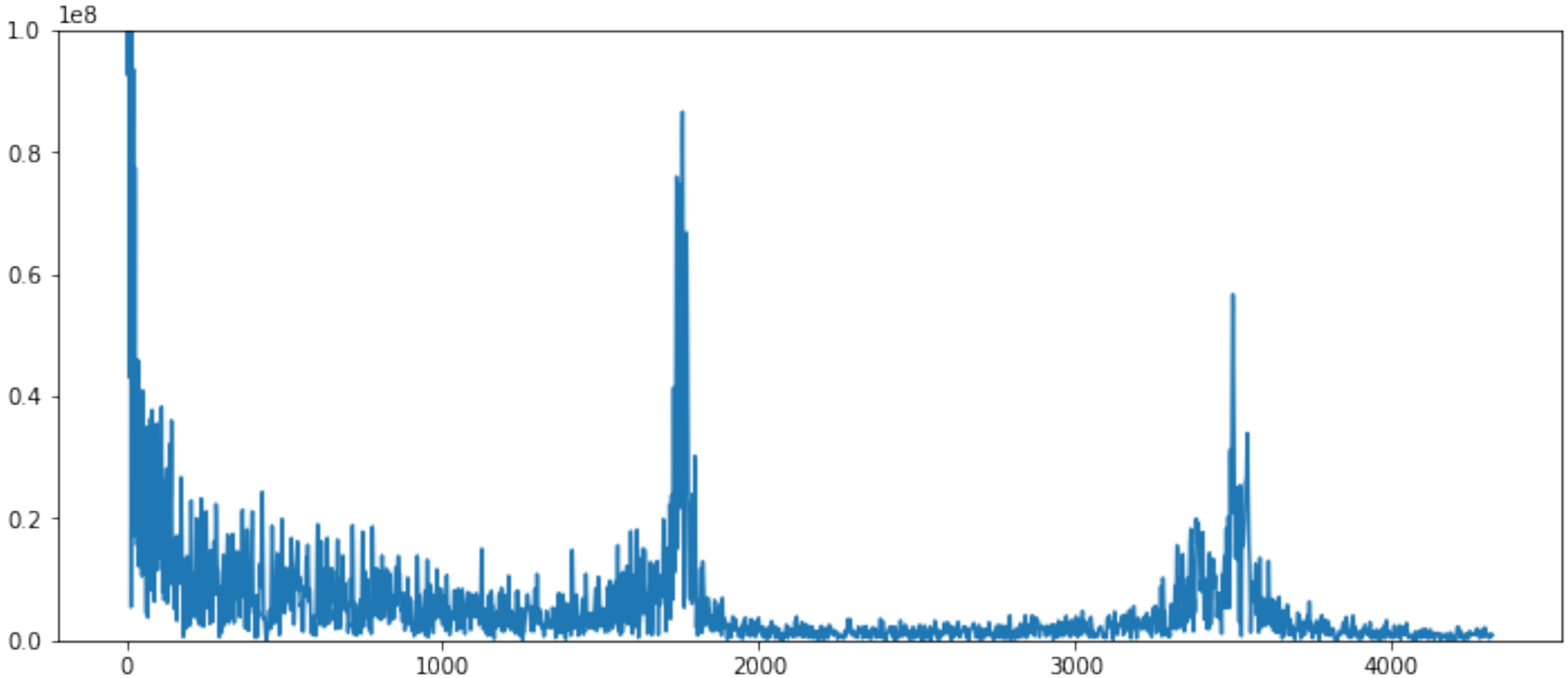
- You almost definitely already paid for this
  - If you collect it, you probably use it for periodic health reports or visuals

- If you are generating dashboards from this, what do they show?
  - Countries?
  - High bandwidth?
  - Overall usage?
  - Top talkers?

Hunting Beacons/Reverse C2



Fourier Transformation for Periodic Activity



Zeek

- Phenomenally useful tool
  - IMO, many sites are missing out!
  - Most sites I've seen just install it and point it at the SIEM
- Event based intelligence system
  - Can generate logs (obviously)
  - Can do *anything you want to do* in reaction to any event(s)
- Allows us to work with network streams at a higher level

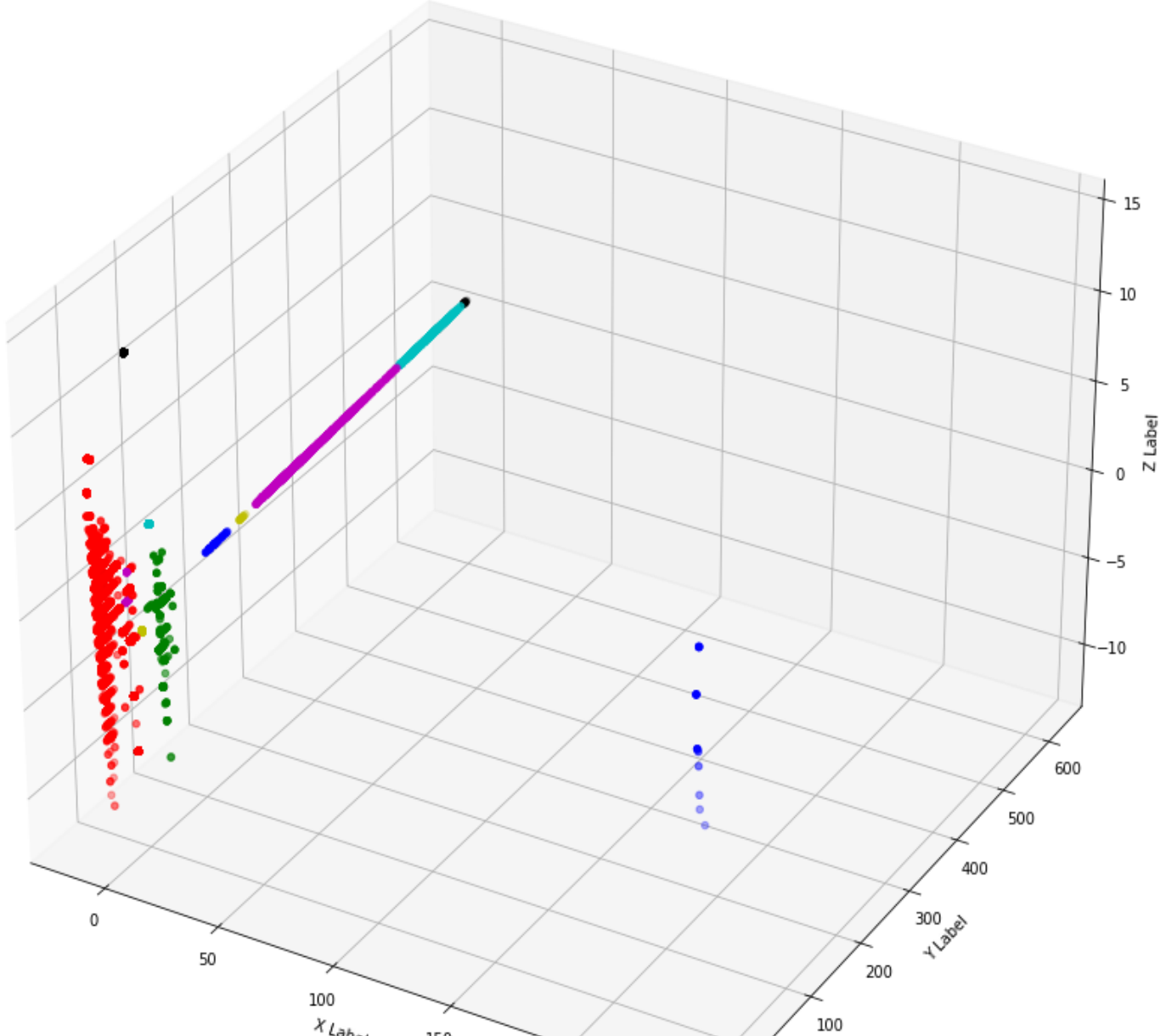
Zeek and DNS

- DNS log
  - Every query, every response
  - Imagine you have 1,000,000 unique DNS queries from a week/day/hour
    - Are there any patterns you should be aware of?
    - Are there any threats present?
- Solving this without machine learning:
  - Get a threat intel list / "known bad" list
  - See if any of those are present, generate alerts
  - See which queries are very frequent… but how often will Google appear?

Unsupervised Learning – DBSCAN Clustering

- How To:
  - Convert DNS queries to a series of numeric features
  - Transform the queries
  - Cluster with DBSCAN
  - (To plot it, apply a 3 dimension principal component analysis)
  - Enjoy analysis goodness

But what do the clusters mean?



- 9 clusters total, a few very interesting:
  - Google telemetry services
  - Ad-related telemetry services
  - Internal service lookups
  - ***Iodine C2 running in the network***

zdzqaabbccddeeffgghhiijjkkllmmnnooppqqrrssttuuvvwwxxyyzz.microsoft.com

It's Not About the Iodine

- Couldn't we have found the Iodine with the "microsoft"?
  - Sure… But couldn't Iodine have been reconfigured to use *any other domain name?*
  - It doesn't matter how the name is changed… The Iodine C2 will *always* pop out as a cluster
- But it's not about the Iodine…

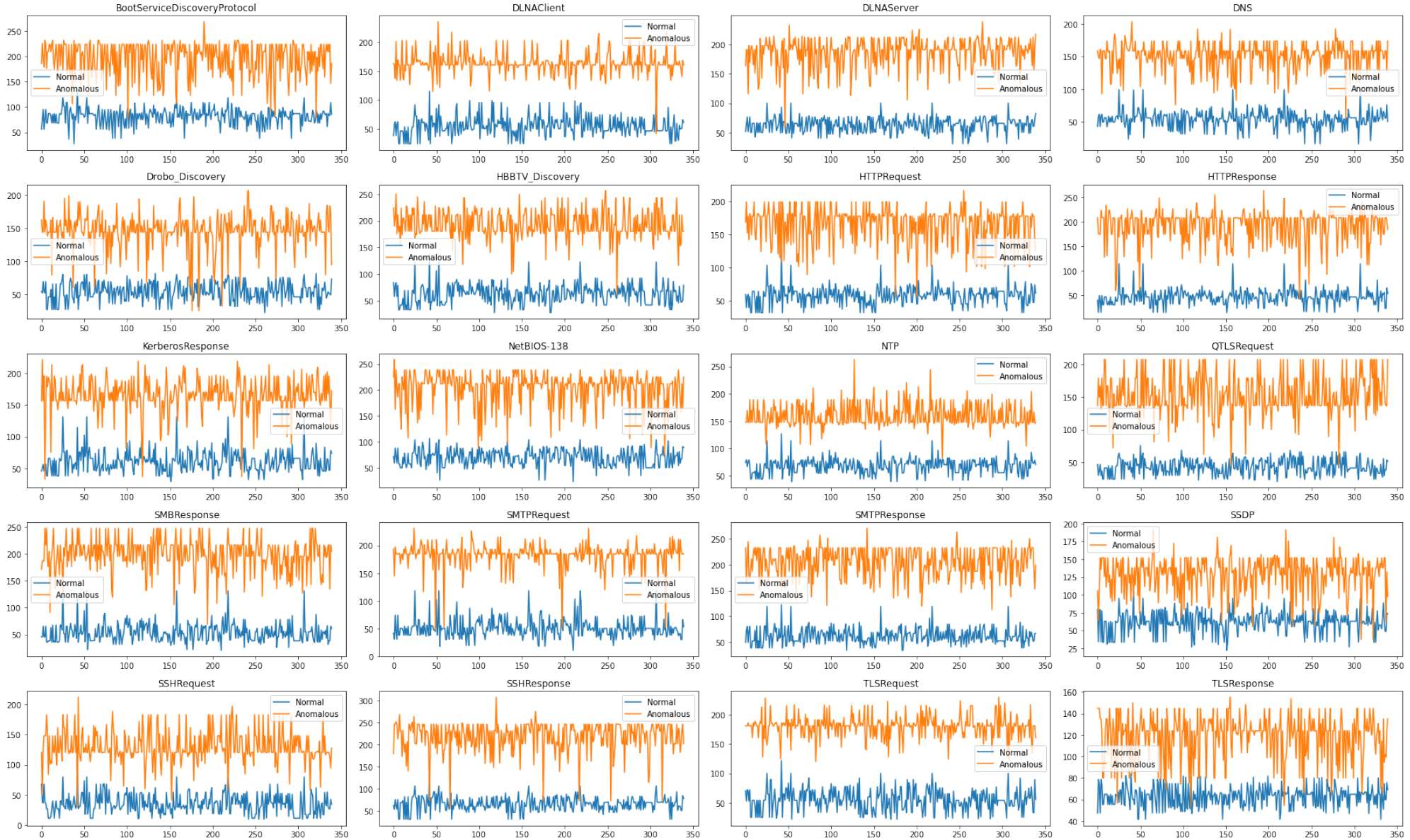
I can't look at 1,000,000 queries…
I can absolutely look at 10 clusters to see what they represent!

(and we didn't use a signature for anything!)

Finding Anomalies

- We can also tackle much smaller, more specific problems
  - Would it be useful to identify protocols running on unusual ports?
  - Would it be useful to find unexpected protocols on your network?
  - Would it be useful to find *unknown protocols* running on your network regardless of port?

Real-Time Automated Anomaly Detection



No New Data

- Think about everything we've looked at
  - Nothing here uses *any* new data!

- How can I do that?
  - Research and experiment – Statistics, Mathematics, Machine Learning
  - SANS SEC595 ("Orlando" is sold out, "Baltimore" & "London" aren't)
    - Applied Data Science and Machine Learning/AI for Cybersecurity Professionals
  - Join our biweekly livestream!
    - Tomorrow, 12 noon eastern time – Applying Machine Learning to Network Anomalies
    - https://youtu.be/PdddO1-jeQQ
    - https://www.linkedin.com/video/event/urn:li:ugcPost:6851259642048335872/

COURSE RESOURCES AND CONTACT INFORMATION

AUTHOR CONTACT

David Hoelzer
dhoelzer@enclaveforensics.com
dhoelzer@sans.edu

SANS INSTITUTE

11200 Rockville Pike
Suite 200
North Bethesda, MD 20852
301.654.SANS(7267)

BLUE TEAM OPERATIONS RESOURCES

https://sans.org/blue-team
Twitter: @SANSDefense

SANS EMAIL

GENERAL INQUIRIES: info@sans.org
REGISTRATION: registration@sans.org
TUITION: tuition@sans.org
PRESS/PR: press@sans.org