# Multi-Dimensional Network Anomaly Detection with Machine Learning

Andrew Fast, Ph.D.
*Chief Data Scientist*
*af@counterflowai.com*

Randy Caldejon
*Co-Founder*
*rc@counterflowai.com*

CounterFlow ai

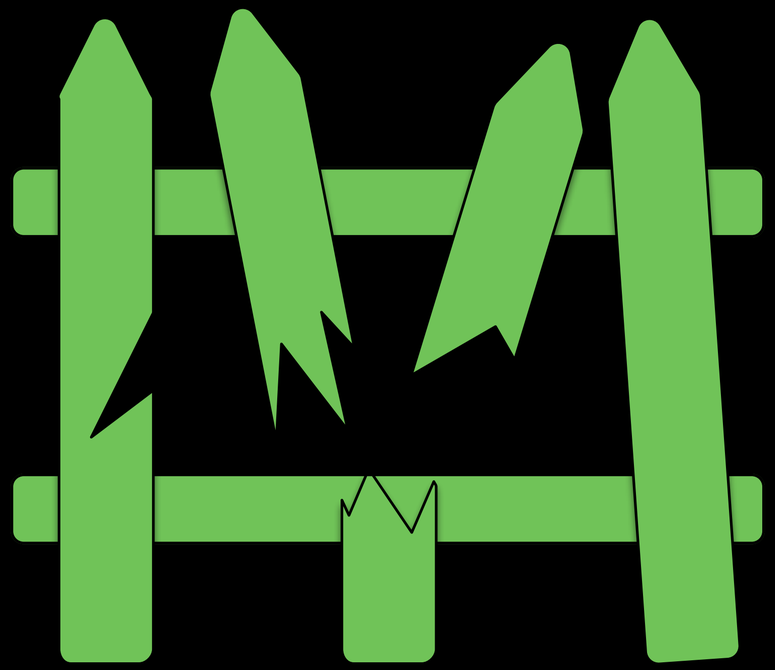
**https://counterflow.ai**

“There are two types of companies: those that have been hacked, and those who don't know they have been hacked.”

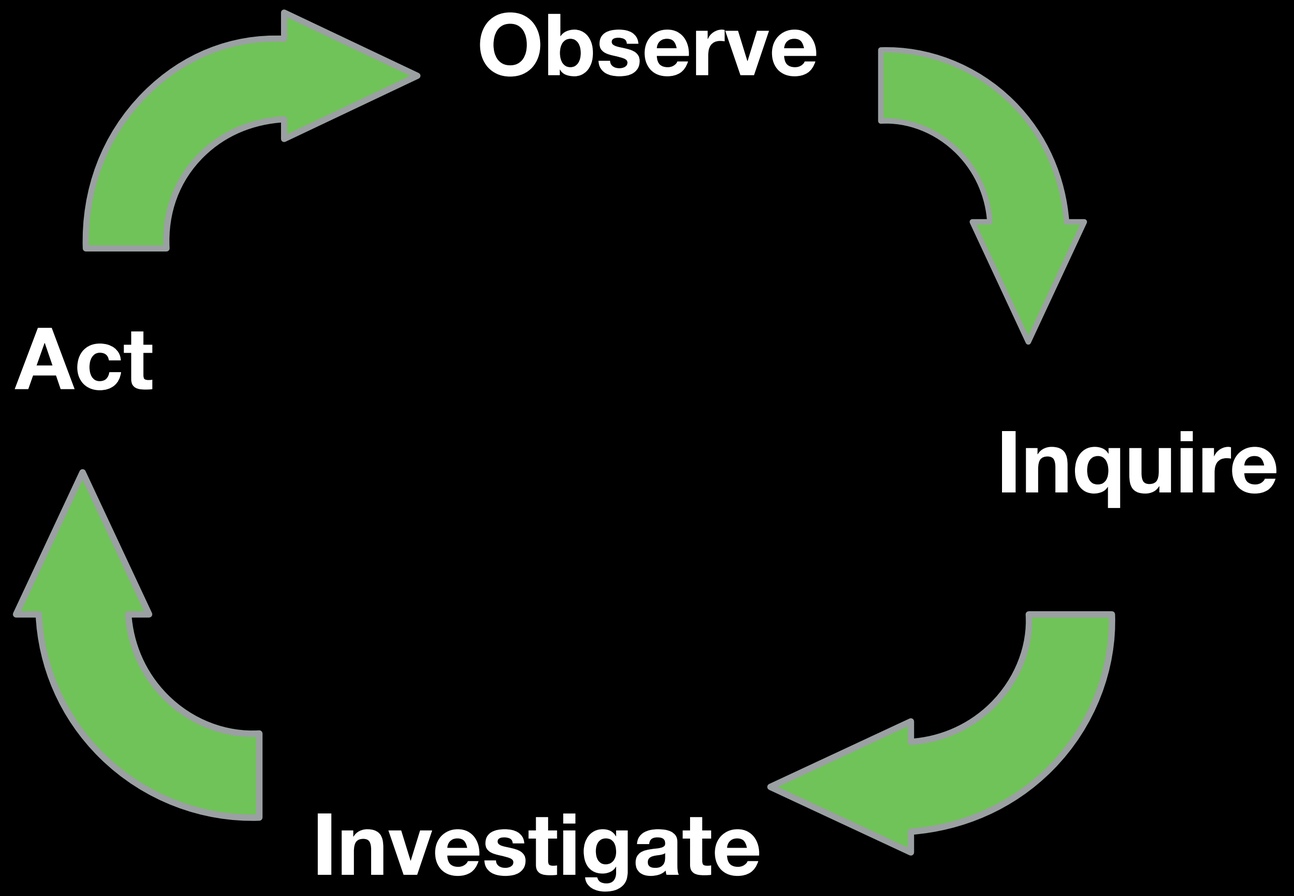
*–John Chambers,  former Cisco CEO and Chairman of the Board*

# Threat Hunting

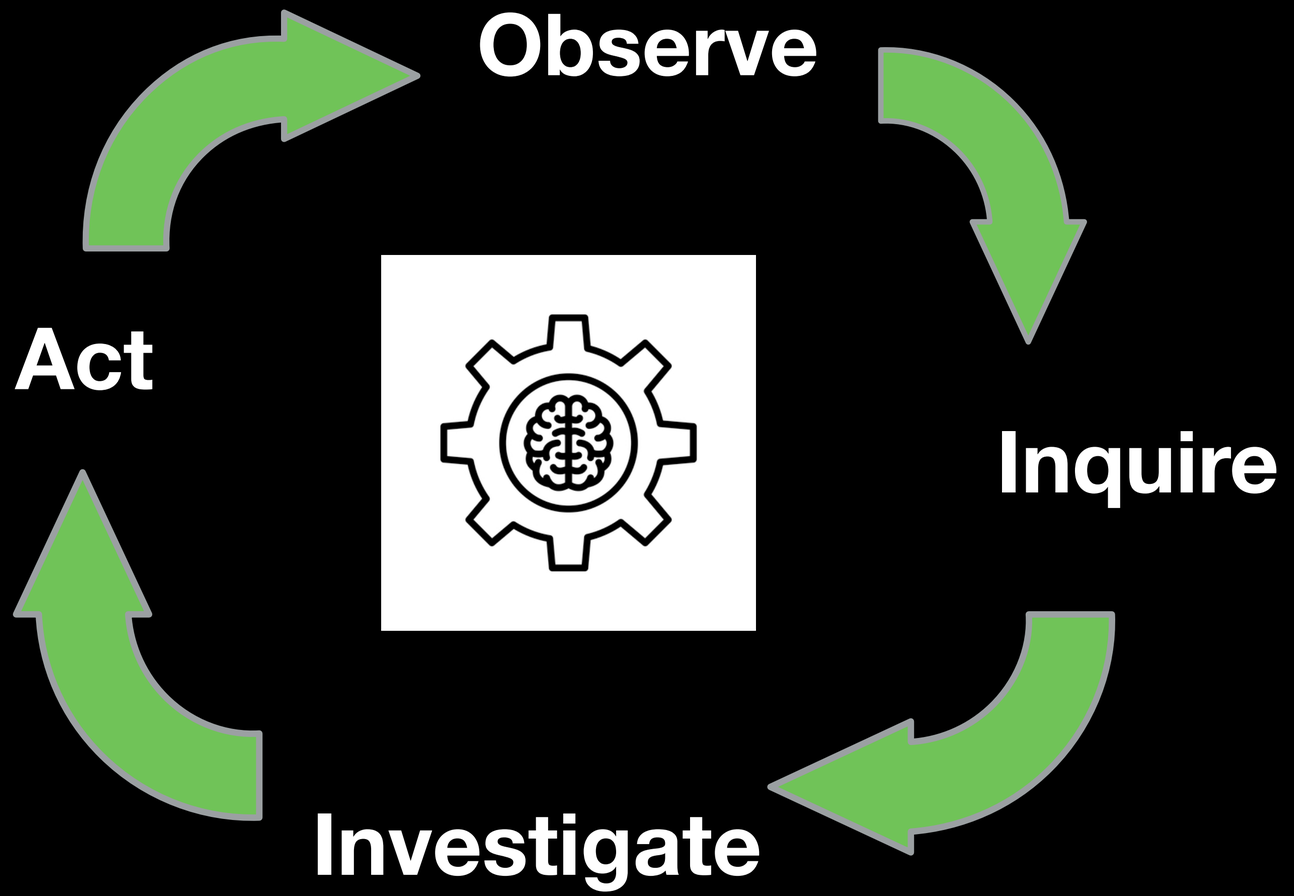
- Assume the perimeter defense is not sufficient to keep threats out
  - Zero-day Exploits
  - Advanced Persistent Threat
  - Etc.
- Need to systematically search and analyze threats *within* the network



# Threat Hunting



# Threat Hunting with ML



# Operationalizing ML for Threat Hunting

Theory

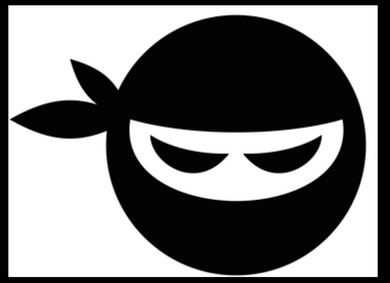
Training

Data Scientist

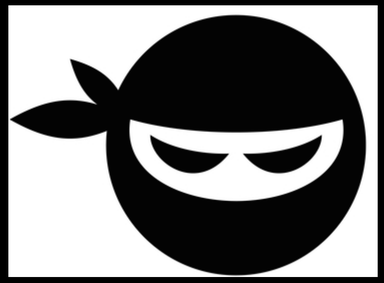
Application

Scoring

Security Analyst



# Operationalizing ML for Threat Hunting



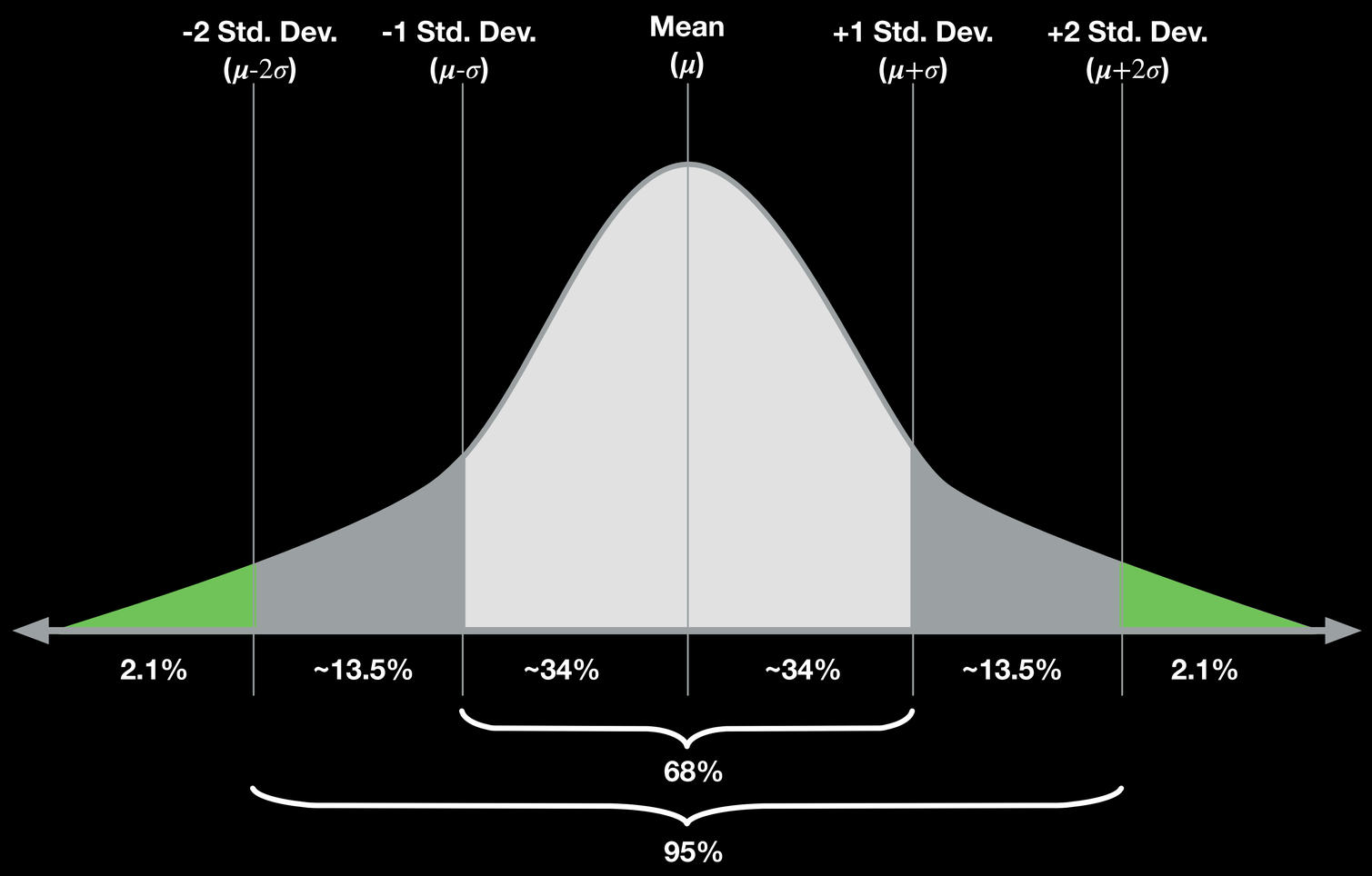
# How do we find threats when we don’t know what they look like?

# Statistical Anomaly Detection

- The Z-score is commonly used for anomaly detection in 1 dimension

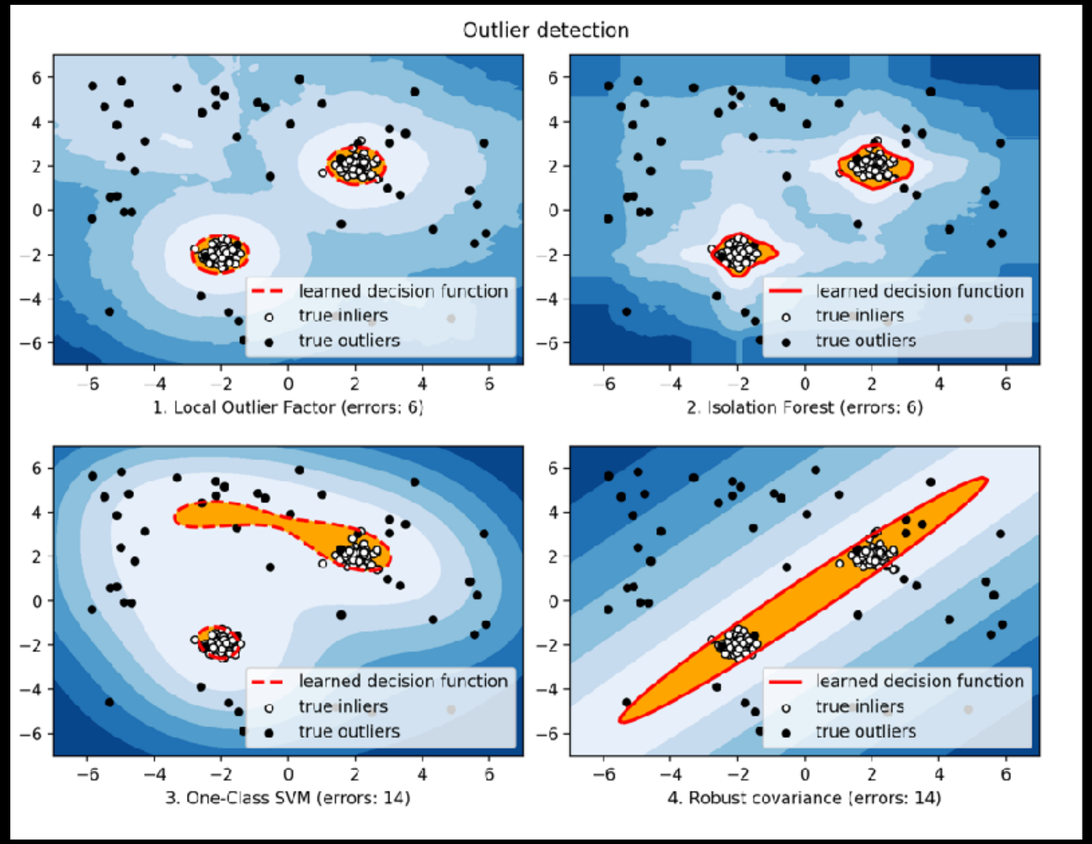
$$Z = \frac{x - \mu}{\sigma}$$

- Assumes data is normally distributed
- Curse of Dimensionality means this approach doesn't generalize well to multiple dimensions



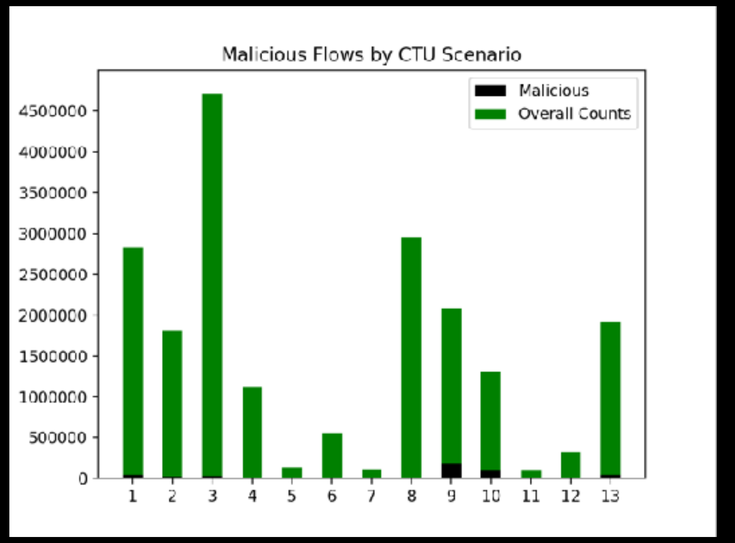
# Multi-Dimensional Anomaly Detection

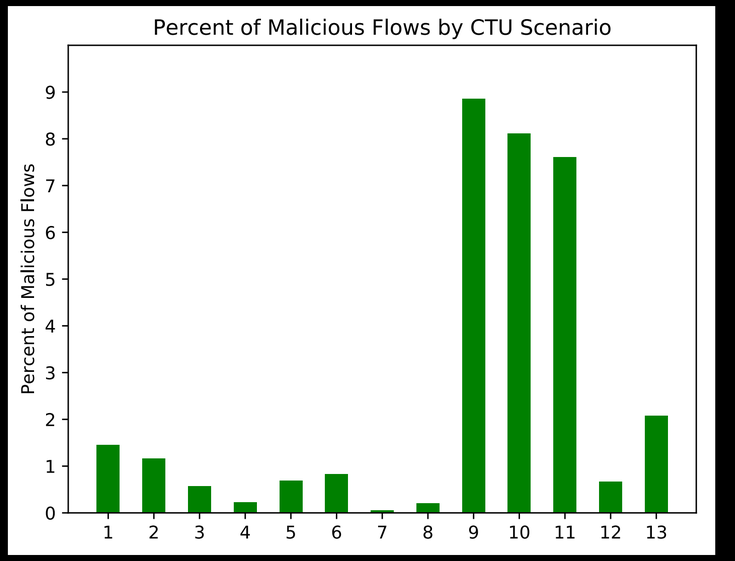
- Extends the Z-score to multiple dimensions
- Several available techniques (both parametric and non-parametric):
  - Local Outlier Factor
  - Isolation Forest
  - Robust Covariance
- Leads to improved ability to detect outliers in network traffic



# Bake-Off

- We evaluated the approaches on net flows from 13 scenarios collected in a lab environment where the ground truth was known
  - Range in size from ~100k to ~5m flows
  - Most scenarios have between ~0.5% malicious and 2% malicious.  3 scenarios are closer to 9% malicious
- "An empirical comparison of botnet detection methods" Sebastian Garcia, Martin Grill, Honza Stiborek and Alejandro Zunino. Computers and Security Journal, Elsevier. 2014. Vol 45, pp 100-123. http://dx.doi.org/10.1016/j.cose.2014.05.011



# Approaches Considered

| Algorithm | Description |
|---|---|
| Scaled Duration | 1-dimensional anomaly detection (comparable to a top-10 list) |
| Scaled Total Bytes | |
| Scaled Total Packets | |
| Apache Spot | Open-source Threat Hunting |
| Local Outlier Factor (LOF) | Multi-dimensional, data-driven anomaly detection approaches |
| Isolation Forest (ISO) | |
| Robust Covariance (RC) | |

# Apache Spot

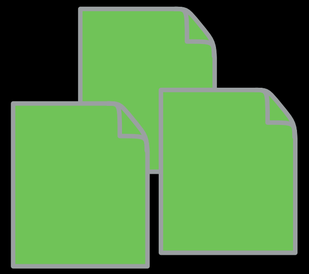

**Apache Spot combines open-source technology to identify potentially malicious threats from stored network traffic.**

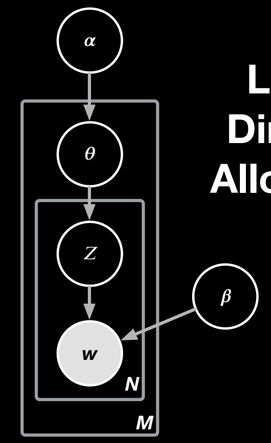
Generate Flow Metadata → Convert to "Words" → Group "Words" by IP → Apply Clustering

{"timestamp":"2013-06-26T13:58:19.249230-0400","flow_id":1068807640305038,"pcap_cnt":1875259,"event_type":"flow","src_ip":"128.169.74.229","src_port":36252,"dest_ip":"128.142.211.249","dest_port":53,"proto":"UDP","flow":{"tot_pkts":5,"tot_bytes":0,"bytes_toclient":0,"start":"2013-06-26T13:58:00.270388-0400","end":"2013-06-26T13:58:18.238342-0400","label":"flow=Background-TCP-Attempt" }}

**SRC: "80_TCP_13_3_0"**

**DEST: "-1_80_TCP_13_3_0"**






**Latent Dirichlet Allocation**





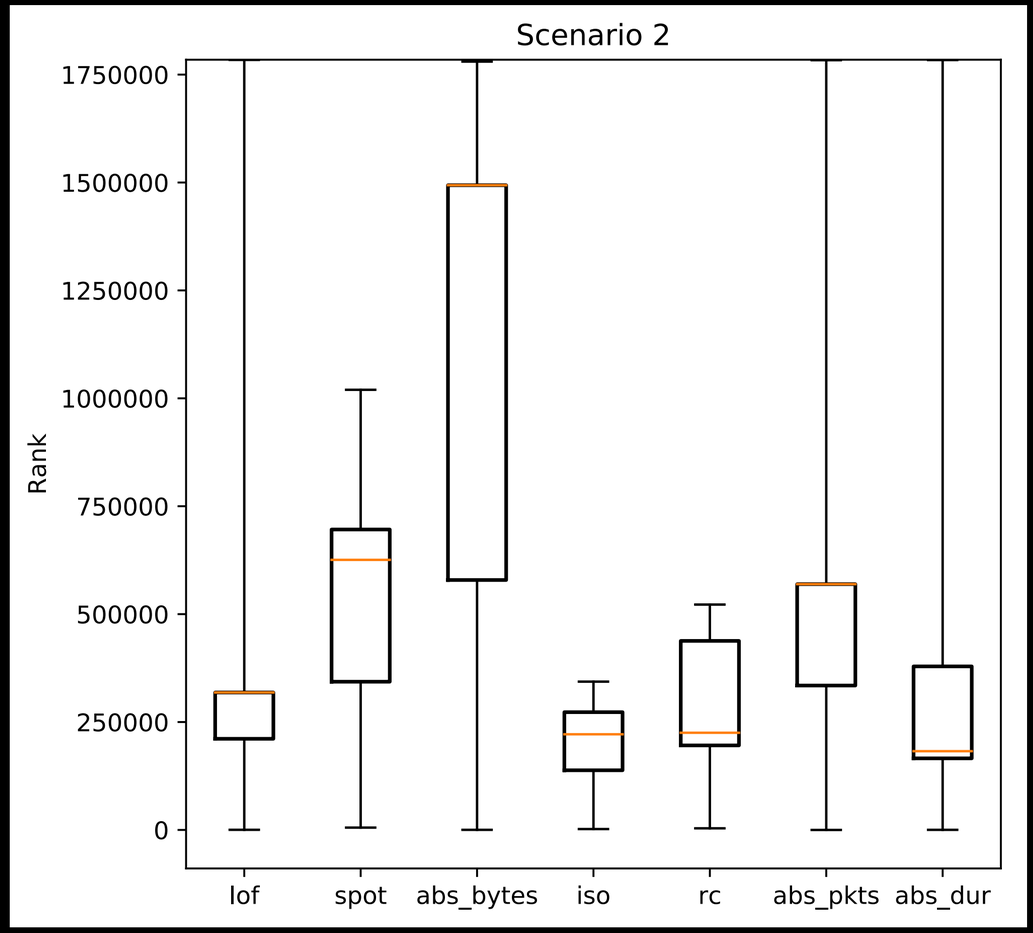
CounterFlowai

# Creating Dimensions

- Use 1-d anomaly scores for numerical values
- Percentage of overall traffic to/from an IP
- Percentage of traffic using a particular protocol or port grouped by Source and Destination IP
- Anything else you can compute!

# Evaluation Criteria

- Minimum rank of all malicious flows (Lower)
  - Evaluates how quickly the first flow shows up
- Precision at 1000 (Higher)
  - Evaluates effort at a small workload
- Precision at 10000 (Higher)
  - Evaluates effort at a larger workload
- Area Under the ROC Curve (AUC) (Higher)
  - Evaluates ranking across the entire scenario



# Number of Wins

- The table shows the number of scenarios in which each algorithm had the best result
- Overall, Local Outlier Factor placed the most malicious flows near the top of the list
- However, Isolation Forest produced the lowest variance overall

| | Min | n=1000 | n=10000 | AUC |
|---|---|---|---|---|
| **Duration** | 1 | 1 | 0 | 2 |
| **Bytes** | 0 | 0 | 1 | 0 |
| **Packets** | 9 | 2 | 0 | 0 |
| **Spot** | 0 | 0 | 0 | 2 |
| **LOF** | 2 | 8 | 11 | 1 |
| **ISO** | 1 | 2 | 1 | 7 |
| **RC** | 0 | 0 | 0 | 1 |
| **Winner** | *Packets* | *LOF* | *LOF* | *ISO* |

# Number of Threats Identified (n=10000)

- Local Outlier Factor (LOF) performed many multiples better than than the alternatives in 12 of 13 scenarios
- We considered two baselines:
  - Baseline 1: **Random**
    - How many malicious flows would you find by chance alone in a sample of 1000 flows?
  - Baseline 2: **Next Best**
    - The best performer of Spot, Packets, Bytes, and Duration?

| Scenario | LOF | Num Expected | Next Best | LOF Lift Over Random | LOF Lift Over Next Best |
|---|---|---|---|---|---|
| 1 | **467** | 147.07 | 21 | 3.18 | **22.24** |
| 2 | **557** | 117.35 | 53 | 4.75 | **10.51** |
| 3 | **102** | 58.53 | 1 | 1.74 | **102.00** |
| 4 | **34** | 16.19 | 2 | 2.10 | **17.00** |
| 5 | 186 | 70.52 | **214** | 2.64 | 0.87 |
| 6 | **261** | 84.03 | 2 | 3.11 | **130.50** |
| 7 | **34** | 5.60 | 28 | 6.07 | **1.21** |
| 8 | **147** | 21.08 | 21 | 6.97 | **7.00** |
| 9 | **1687** | 895.47 | 518 | 1.88 | **3.26** |
| 10 | **34** | 3.19 | 10 | 10.67 | **3.40** |
| 11 | **21** | 2.22 | 9 | 9.44 | **2.33** |
| 12 | **333** | 68.00 | 119 | 4.90 | **2.80** |
| 13 | **389** | 211.69 | 93 | 1.84 | **4.18** |

# Operationalizing ML for Threat Hunting

Theory

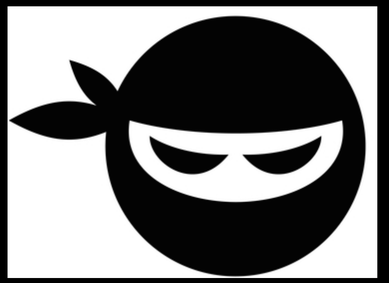
Application

Training

Scoring

Data Scientist

Security Analyst



# System Architecture: Training




**Ax=b**

**YAF > Dragonfly ML Engine > Parquet > Scikit.Learn (Python) > Model**

# System Architecture: Scoring

YAF

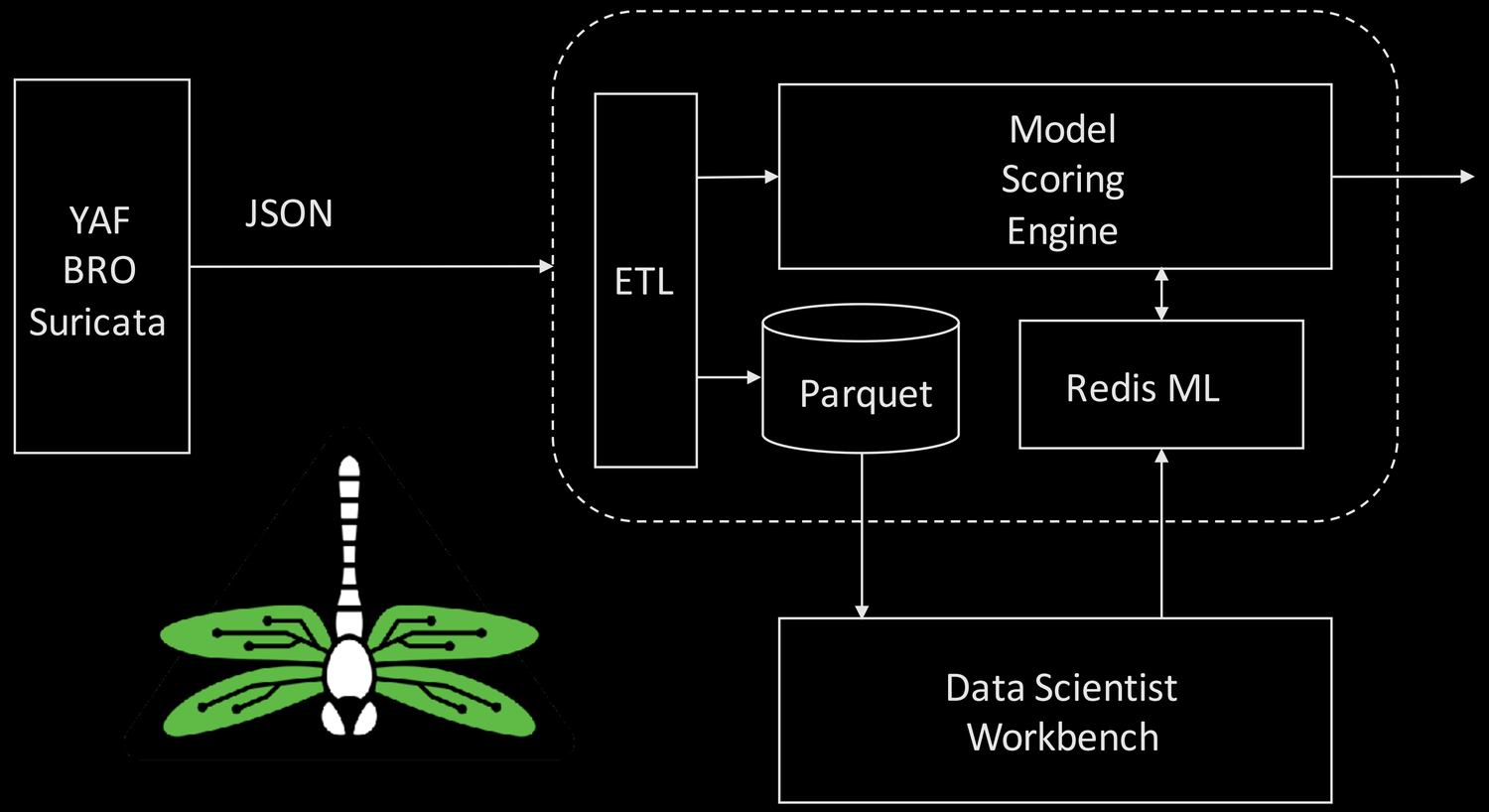

Ax=b

.0778

YAF > Dragonfly ML Engine > Redis Caching > Loadable ML Module > Score

# Dragonfly ML Engine

- Dragonfly ML consumes streaming data from sensor
- Provides ETL for batch data processing
- Provides mechanism for inline scoring of network traffic

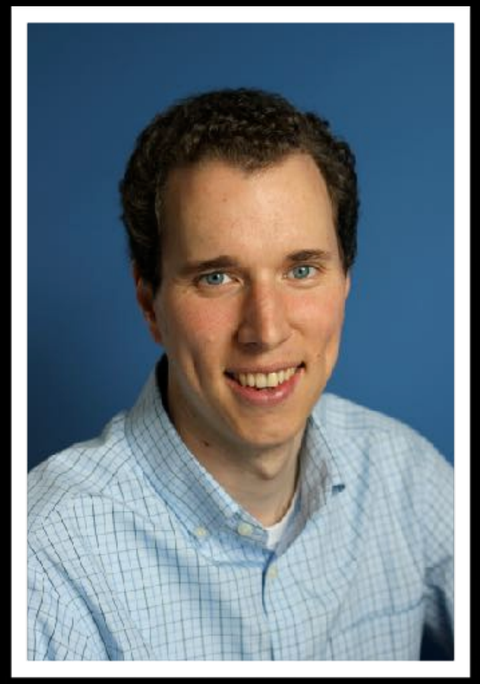


# Summary

- Machine Learning is an ideal threat hunting tool for an analyst
- But it must be trained with the analyst workflow and deployment in mind
- Multi-dimensional anomaly detection techniques provides a powerful approach for threat hunting using multiple kinds of features, leading to improved detection of possible threats
- CounterFlow's Dragonfly ML platform will be available on GitHub:
  https://github.com/orgs/counterflow-ai/

@counterflowai

af@counterflowai.com
rc@counterflowai.com



# Andrew Fast, Ph.D.
## Chief Data Scientist and Co-Founder

CounterFlow AI

CounterFlow AI enables security analysts to hunt more effectively by introducing the next generation of network security sensors. Operating at the tip of the spear, CounterFlow sensors are designed to target threats at wire-speed. The sensors integrate signature-inspection, threat intelligence, and streaming analytics with machine learning to detect known and zero-day attacks, significantly reduce false positives, and drive security automation.

Dr. Andrew Fast is the Chief Data Scientist and co-founder of CounterFlow AI. Previously, he served as the Chief Scientist at Elder Research, Inc., a leading data science consulting firm, where he helped hundreds of companies expand their data science capabilities. Dr. Fast is a frequent author, teacher, and invited speaker on data science topics. In 2012, he co-authored a book titled Practical Text Mining that was published by Elsevier and won the PROSE Award for top book in the field of Computing and Information Sciences for that year. His work analyzing NFL coaching trees was featured on ESPN.com in 2009.

Dr. Fast received Ph.D. and M.S. degrees in Computer Science from the University of Massachusetts Amherst and a B.S. in Computer Science from Bethel University.

*https://counterflow.ai*
*af@counterflowai.com*

# Randy Caldejon
## Co-Founder

CounterFlow AI

CounterFlow AI enables security analysts to hunt more effectively by introducing the next generation of network security sensors. Operating at the tip of the spear, CounterFlow sensors are designed to target threats at wire-speed. The sensors integrate signature-inspection, threat intelligence, and streaming analytics with machine learning to detect known and zero-day attacks, significantly reduce false positives, and drive security automation.

*https://counterflow.ai*
*rc@counterflowai.com*

As CEO, Randy leads the company vision, innovation, and execution. He is a widely-respected authority in network security monitoring and sensor technology. A veteran, engineer, and serial entrepreneur, Randy has over 25 years of technology leadership experience. As cohort and co-founder, he started and successfully exited from two cybersecurity ventures; including nPulse Technologies, which was acquired by FireEye, Inc in 2014.

Randy served honorably in the U.S. Marine Corps. In his spare time, he enjoys biking, fly fishing, and instrumenting his farm with IoT sensors. He holds a B.S. in Computer Science from University of Maryland Baltimore County (UMBC) and a M.Eng. in Computer and Systems Engineering from Rensselaer Polytechnic Institute (RPI).