Secure Because Math: A Deep-Dive on Machine Learning-Based Monitoring (#SecureBecauseMath)

Alex Pinto
Chief Data Scientist | MLSec Project
@alexcpsec
@MLSecProject

whoami

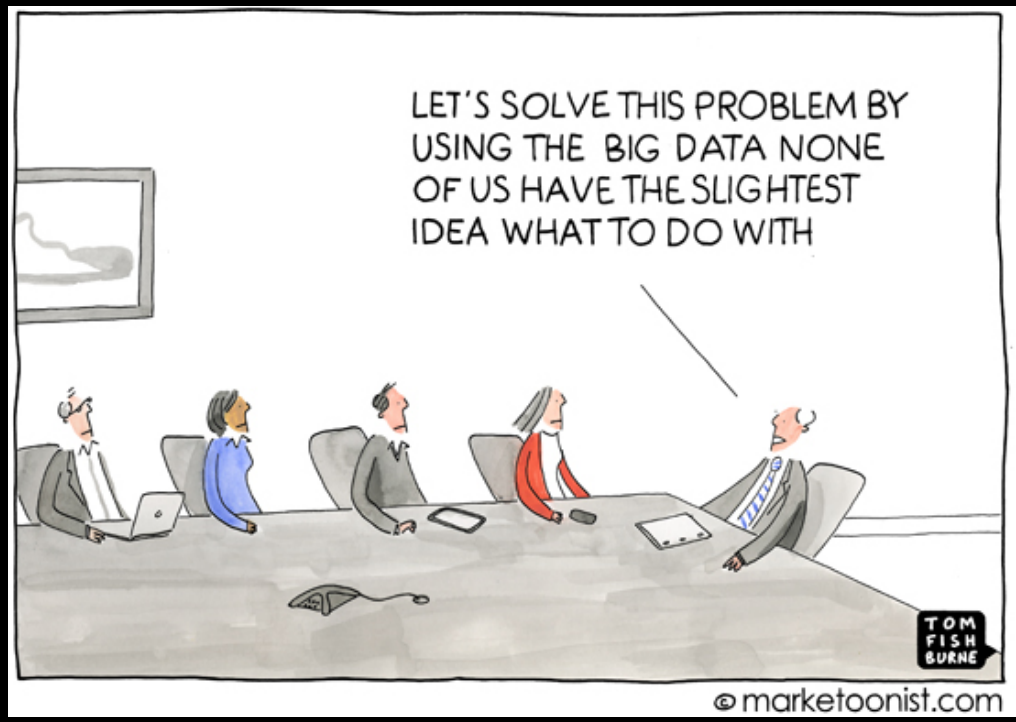
Alex Pinto

- Chief Data Scientist at MLSec Project
- Machine Learning Researcher and Trainer
- Network security and incident response aficionado
- Tortured by SIEMs as a child
- Hacker Spirit Animal™: CAFFEINATED CAPYBARA

(https://secure.flickr.com/photos/kobashi_san/)

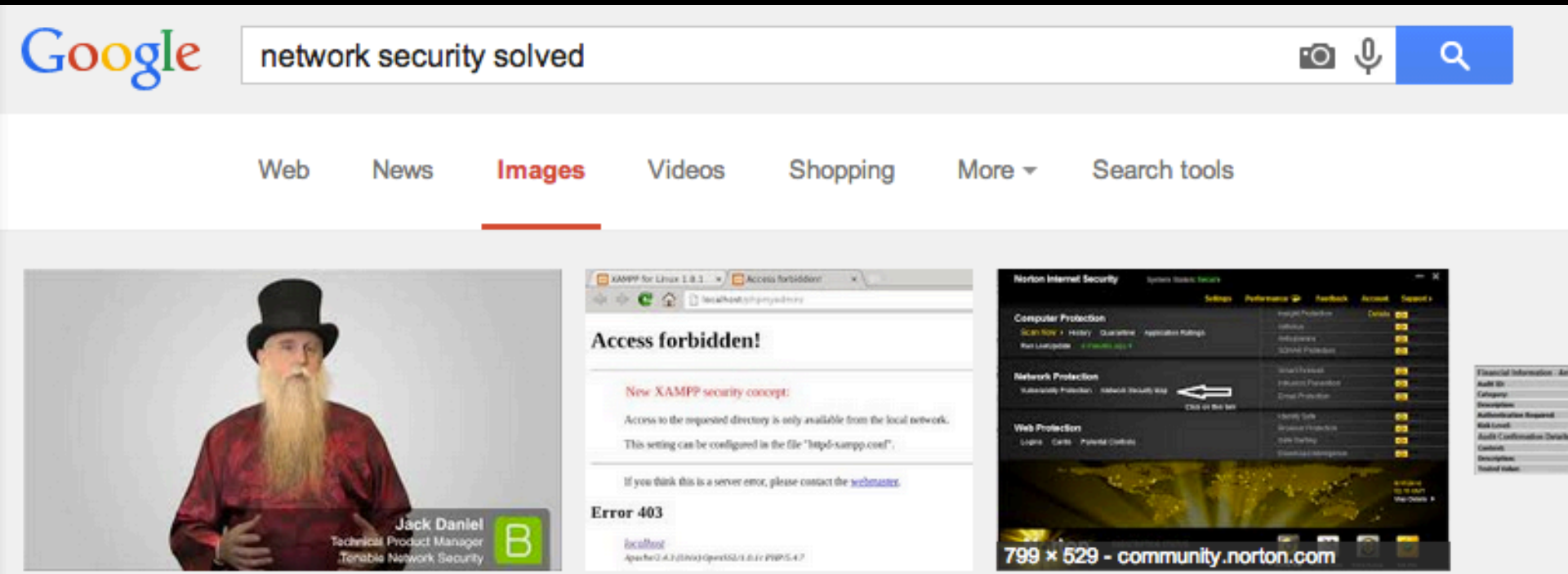
Agenda

- Security Singularity
- Some History
- TLA
- ML Marketing Patterns
- Anomaly Detection
- Classification
- Buyer's Guide
- MLSec Project



Security Singularity Approaches

(Side Note)



First hit on Google images for “Network Security Solved” is a picture of Jack Daniel

Security Singularity Approaches

- “Machine learning / math / algorithms… these terms are used interchangeably quite frequently.”
- “Is behavioral baselining and anomaly detection part of this?”
- “What about Big Data Security Analytics?”

(http://bigdatapix.tumblr.com/)

Are we even trying?

- “Hyper-dimensional security analytics”
- “3rd generation Artificial Intelligence”
- “Secure because Math”

- Lack of ability to differentiate hurts buyers, investors.
- Are we even funding the right things?

Is this a communication issue?

Guess the Year!

- “(…) behavior analysis system that enhances your network intelligence and security by auditing network flow data from existing infrastructure devices”

- "Mathematical models (…) that determine baseline behavior across users and machines, detecting (...) anomalous and risky activities (...)”

- ”(…) maintains historical profiles of usage per user and raises an alarm when observed activity departs from established patterns of usage for an individual.”

A little history

- Dorothy E. Denning (professor at the Department of Defense Analysis at the Naval Postgraduate School)
  - 1986 (SRI) - First research that led to IDS
  - Intrusion Detection Expert System (IDES)
  - Already had statistical anomaly detection built-in
- 1993: Her colleagues release the Next Generation (!) IDES

Three Letter Acronyms - KDD

- After the release of Bro (1998) and Snort (1999), DARPA thought we were covered for this signature thing
- DARPA released datasets for user anomaly detection in 1998 and 1999
- And then came the KDD-99 dataset – over 6200 citations on Google Scholar

Modified Mutual Information-based Feature Selection for Intrusion Detection Systems in Decision Tree Learning [PDF] from academypublisher.com

J Song, Z Zhu, P Scully, C Price - Journal of Computers, 2014 - ojs.academypublisher.com

... This paper proposed a modified mutual information-based feature selection algorithm (MMIFS) for intrusion detection on the **KDD Cup 99** dataset. ... Section 2 introduces the **KDD Cup 99** dataset and reviews the mutual information and the necessary of feature selection. ...

Cite Save

A Hybrid-Based Feature Selection Approach for IDS

P Ahmed - Networks and Communications (NetCom2013), 2014 - Springer

... To reduce the dimen- sionality, without compromising the performance, a new hybrid feature selection method has been introduced and its performance is measured on **KDD Cup'99** dataset by the classifiers Naıve Bayes and C4.5. Three sets of experiments have been ...

Related articles Cite Save

Information theoretic feature extraction to reduce dimensionality of Genetic Network Programming based intrusion detection model

A Arya, S Kumar - Issues and Challenges in Intelligent ..., 2014 - ieeexplore.ieee.org

... Experimentation with **KDD cup 99** shows modified mutual information based feature selection (MMIFS) IS impressive among three [11]. III. ... I 0 No.1, pp. 102-111,2006. [17] **KDD cup 99** dataset, "http://**kdd**.ics.uci.edu/databases Ikddcup99/kddcup99 .html ". ...

Related articles Cite Save

[PDF] Analysis of the Effect of Clustering the Training Data in Naive Bayes Classifier for Anomaly Network Intrusion Detection [PDF] from jacn.net

U Subramanian, HS Ong - Journal of Advances in Computer Networks, 2014 - jacn.net

... classifier. **KDD cup 99** benchmark dataset is used in this research. The training set is clustered using k means clustering algorithm into 5 clusters. ... The **KDD cup 99** data set consists of 4 types of attack data and normal data. Denial ...

Related articles Cite Save More

A Large-Scale Network Data Analysis via Sparse and Low Rank Reconstruction [PDF] from hindawi.com

LF Lu, ZH Huang, MA Ambusaidi... - Discrete Dynamics in ..., 2014 - hindawi.com

The main objective of Discrete Dynamics in Nature and Society is to foster links between basic and applied research relating to discrete dynamics of complex systems encountered in the natural and social sciences. The journal intends to stimulate publications directed to the analyses ...

Related articles All 2 versions Cite Save More

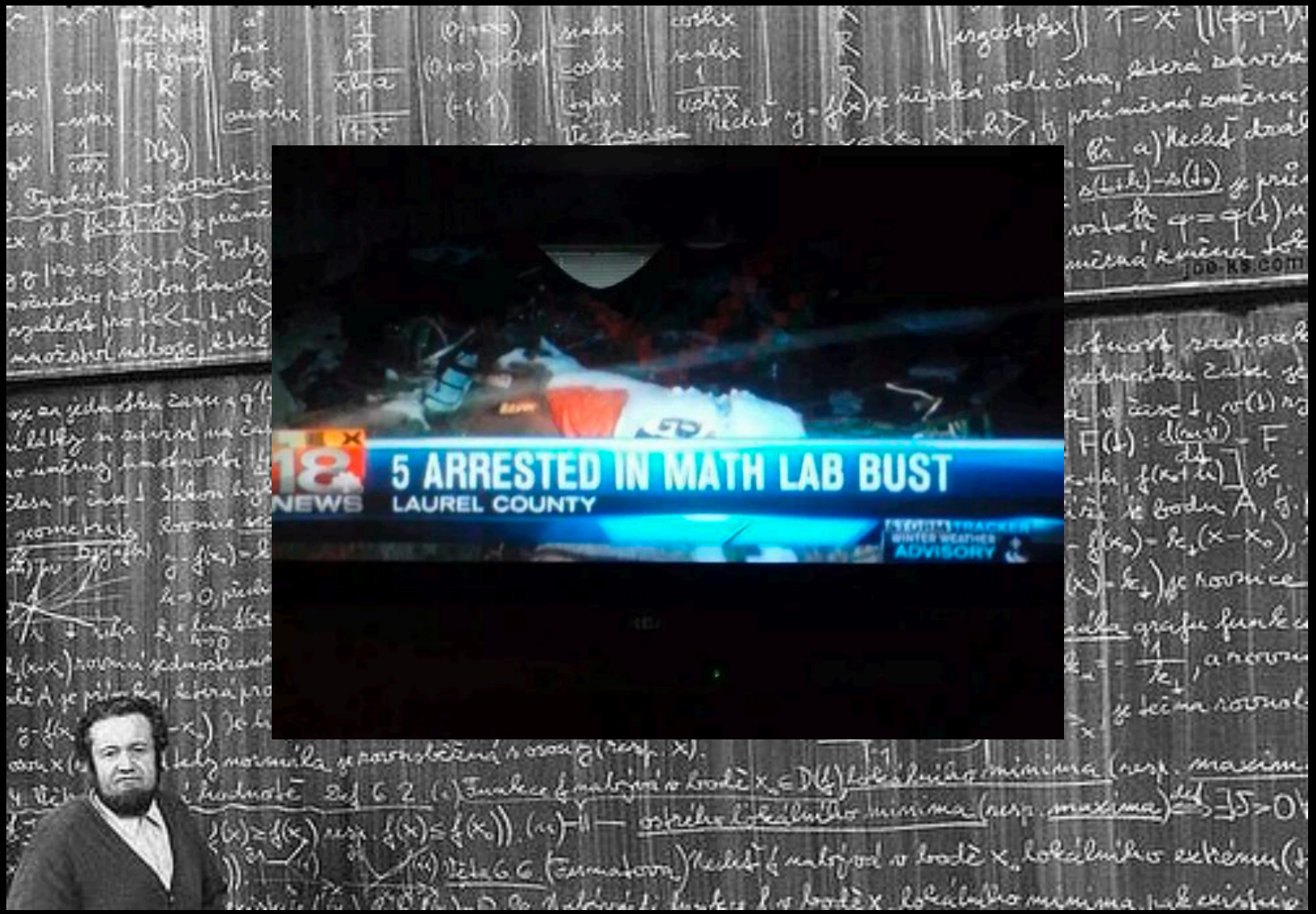
Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations as Performed by Lincoln Laboratory

JOHN McHUGH
Carnegie Mellon University

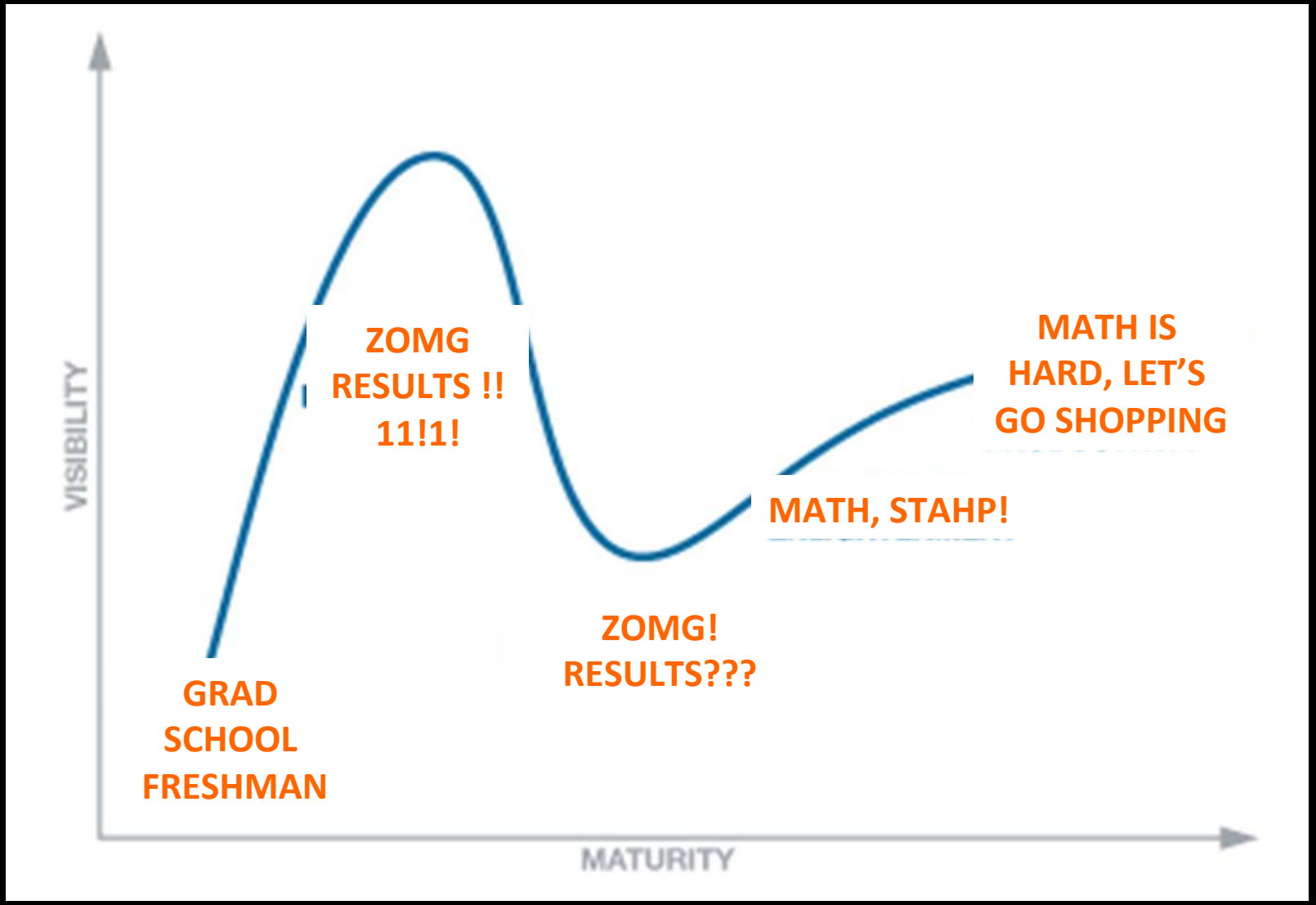
In 1998 and again in 1999, the Lincoln Laboratory of MIT conducted a comparative evaluation of intrusion detection systems (IDSs) developed under DARPA funding. While this evaluation represents a significant and monumental undertaking, there are a number of issues associated with its design and execution that remain unsettled. Some methodologies used in the evaluation are questionable and may have biased its results. One problem is that the evaluators have published relatively little concerning some of the more critical aspects of their work, such as validation of their test data. The appropriateness of the evaluation techniques used needs further investigation. The purpose of this article is to attempt to identify the shortcomings of the Lincoln Lab effort in the hope that future efforts of this kind will be placed on a sounder footing. Some of the problems that the article points out might well be resolved if the evaluators were to publish a detailed description of their procedures and the rationale that led to their adoption, but other problems would clearly remain.

Trolling, maybe?

Not here to bash academia



A Probable Outcome



ML Marketing Patterns

- The “Has-beens”
  - Name is a bit harsh, but hey, you hardly use ML anymore, let us try it

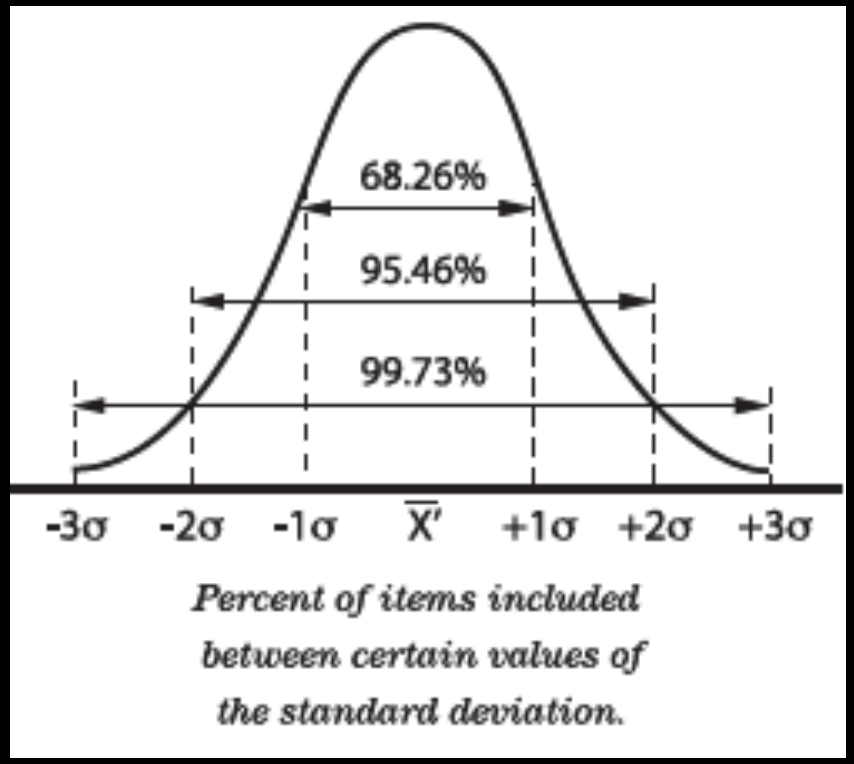
- The “Machine Learning ¯_(ツ)_/¯”
  - Hey, that sounds cool, let’s put that in our brochure

- The “Sweet Spot”
  - People that actually are trying to do something
  - Anomaly Detection vs. Classification

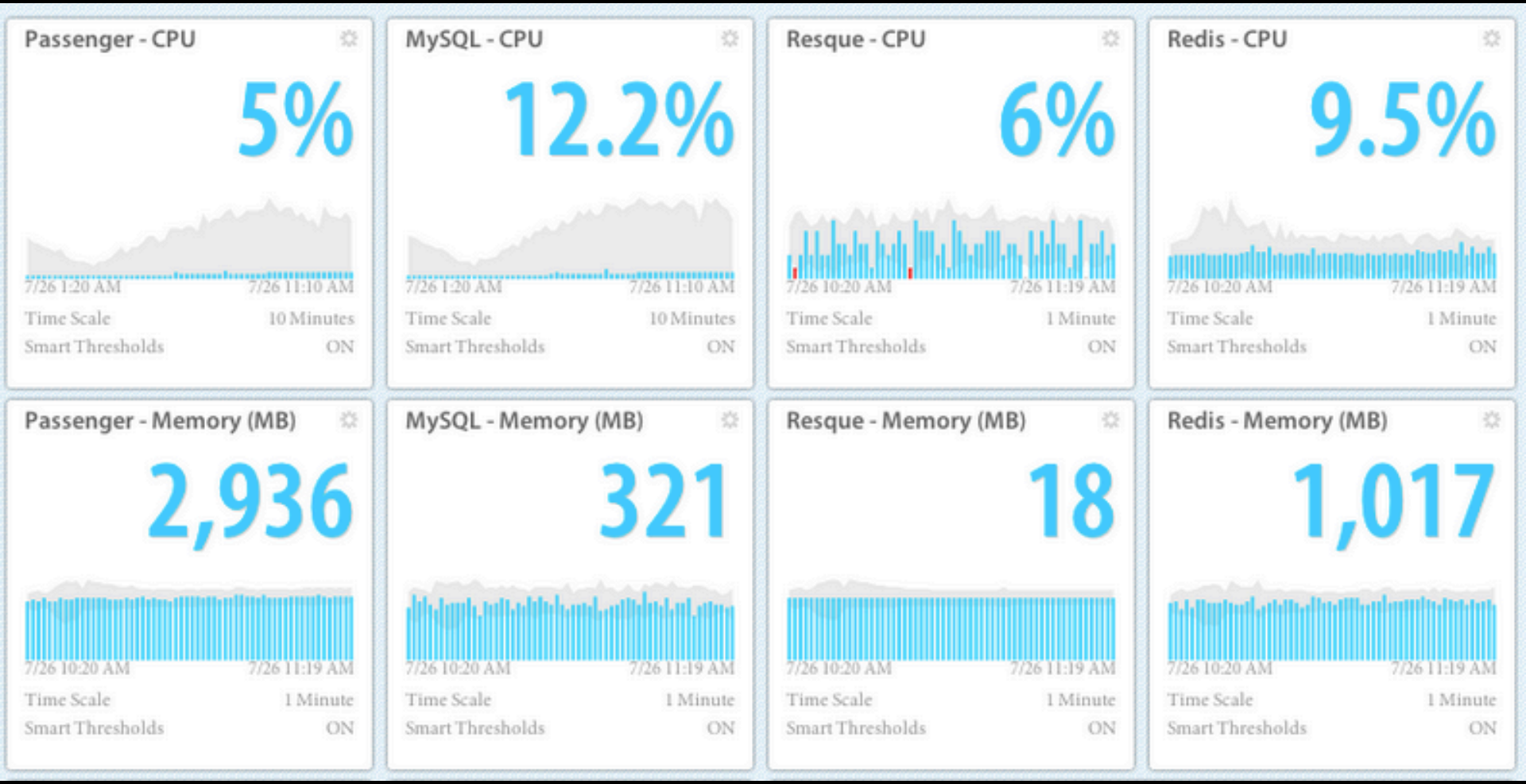
Anomaly Detection

Anomaly Detection

- Works wonders for well defined “industrial-like” processes.
- Looking at single, consistently measured variables
- Historical usage in financial fraud prevention.



Anomaly Detection



Anomaly Detection

- What fits this mold?
  - Network/Netflow behavior analysis
  - User behavior analysis

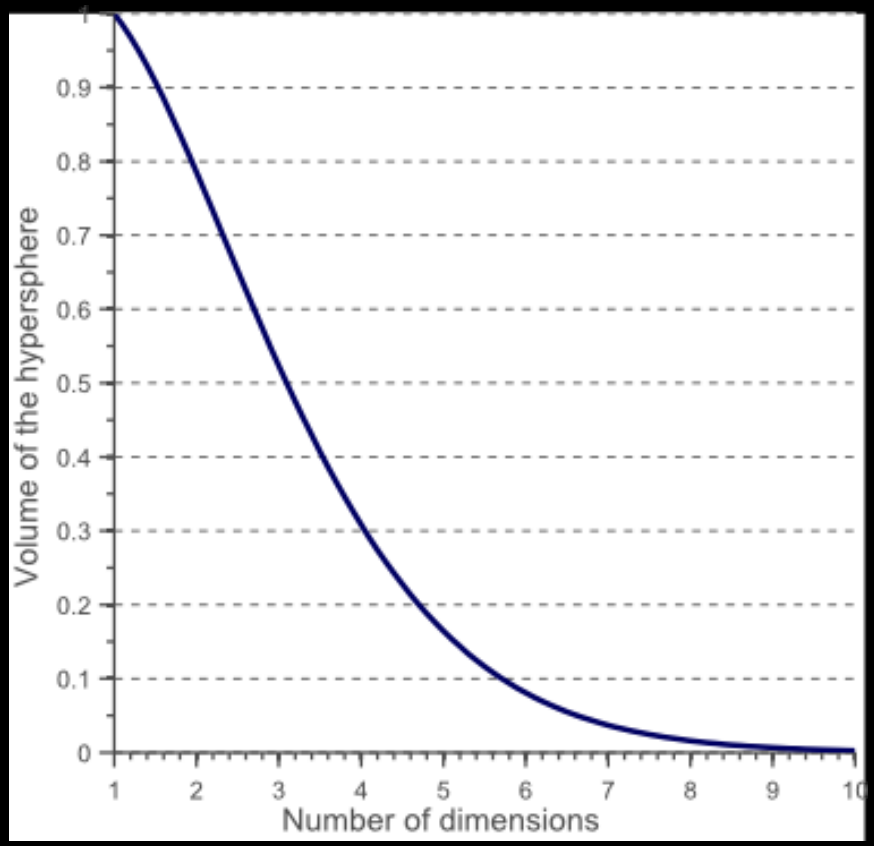
- What are the challenges?
  - Curse of Dimensionality
  - Lack of ground truth and normality poisoning
  - Hanlon's Razor

AD: Curse of Dimensionality

- We need “distances” to measure the features/variables
- Usually Manhattan or Euclidian
- For high-dimensional data, the distribution of distances between all pairwise points in the space becomes concentrated around an average distance.

AD: Curse of Dimensionality

- The volume of the high dimensional sphere becomes negligible in relation to the volume of the high dimensional cube.
- The practical result is that everything just seems too far away, and at similar distances.



(http://www.datasciencecentral.com/m/blogpost?id=6448529%3ABlogPost%3A175670)

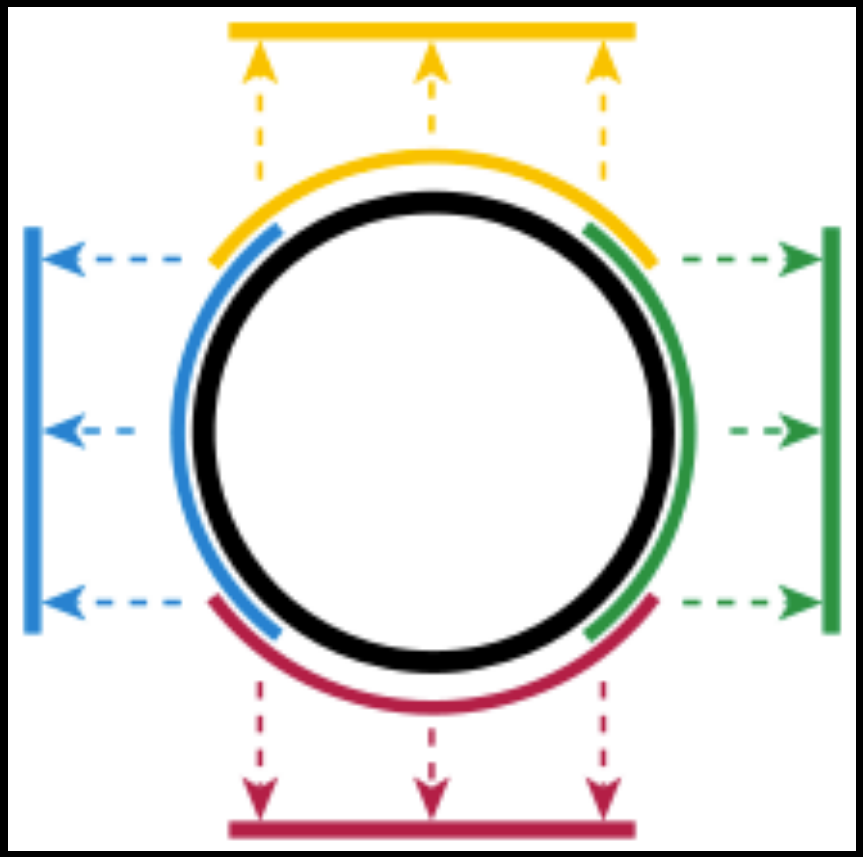
A Practical example

- NetFlow data, company with n internal nodes.
  - 2(n^2 - n) communication directions
  - 2*2*2*65535(n^2 - n) measures of network activity
  - 1000 nodes -> Half a trillion possible dimensions

Breaking the Curse

- Different / creative distance metrics
- Organizing the space into sub-manifolds where Euclidean distances make more sense.
- Aggressive feature removal

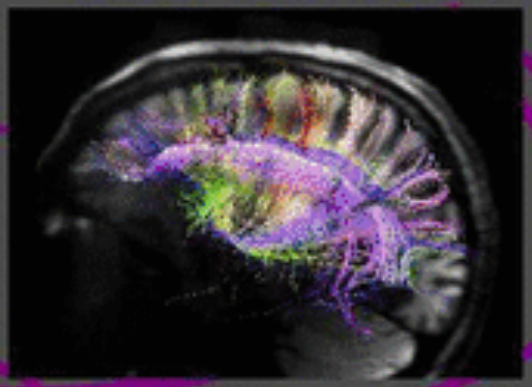
- A few interesting results available



Breaking the Curse

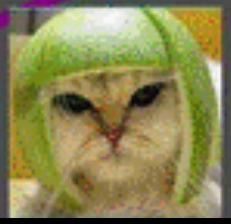
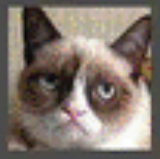
Pennsylvania Learners Being Ripped Off By Not Knowing this One Weird Kernel Trick

The President has ordered all learners to minimize their empirical risk. Learn how to minimize your empirical risk for less by following this simple rule.



PICK YOUR VC DIMENSION

1-10	11-50	51-100	100-1000	1001+

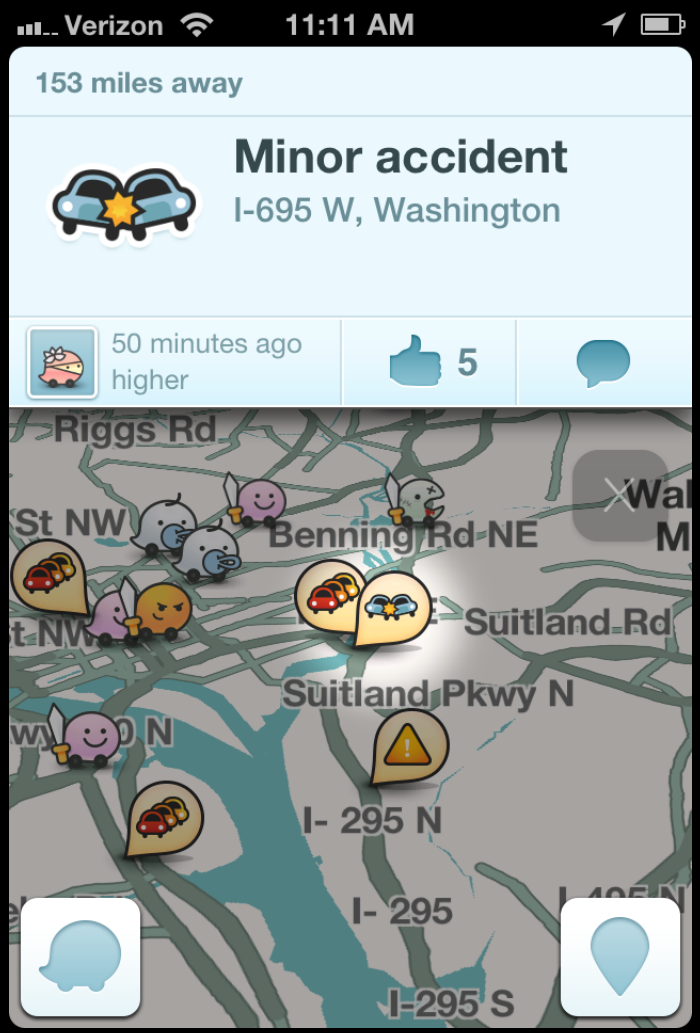

AD: Normality-poisoning attacks

- Ground Truth (labels) >> Features >> Algorithms

- There is no (or next to none) Ground Truth in AD
- What is “normal” in your environment?
  - Problem asymmetry
  - Solutions are biased to the prevalent class

- Very hard to fine-tune, becomes prone to a lot of false negatives or false positives

AD: Normality-poisoning attacks



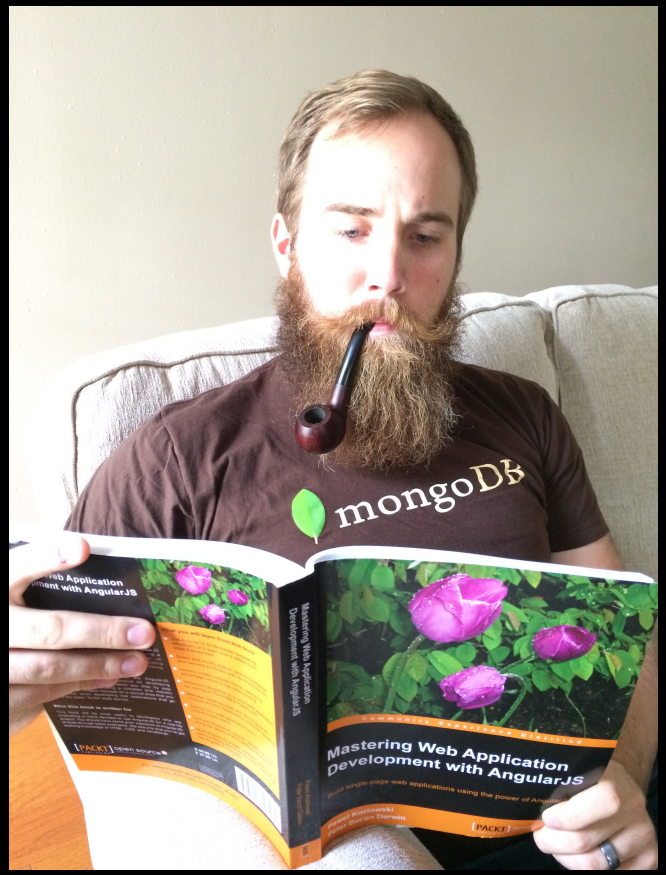
AD: Hanlon's Razor

Never attribute to malice that which is adequately explained by stupidity.

AD: Hanlon's Razor

Evil Hacker

vs



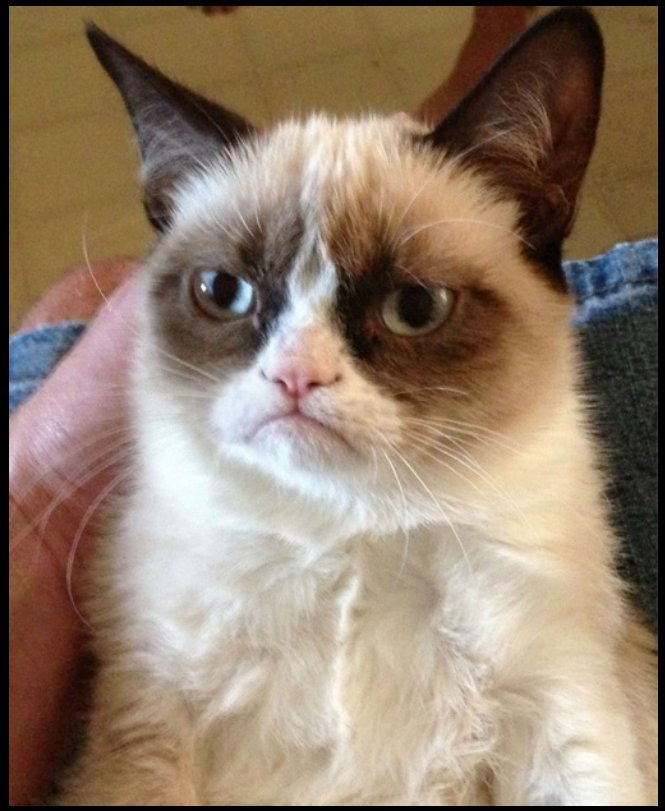
Hipster Developer
(a.k.a. Matt Johansen)

What about User Behavior?

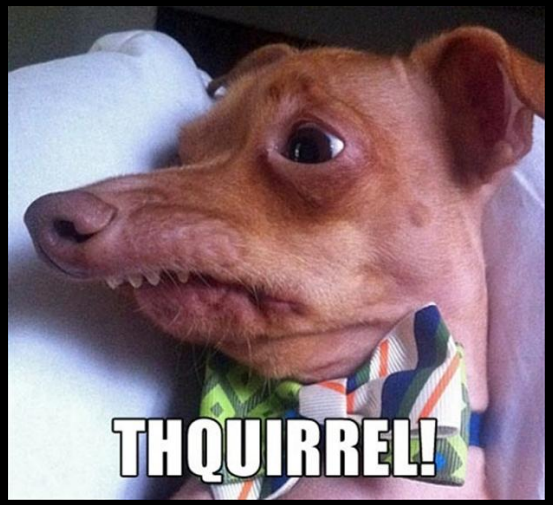
- Surprise, it kinda works! (as supervised, that is)
  - As specific implementations for specific solutions
  - Good stuff from Square, AirBnB
  - Well defined scope and labeling.

- Can it be general enough?
  - File exfiltration example (roles/info classification are mandatory?)
  - Can I “average out” user behaviors in different applications?

Classification!



VS



Lots of Malware Activity

- Lots of available academic research around this
- Classification and clustering of malware samples

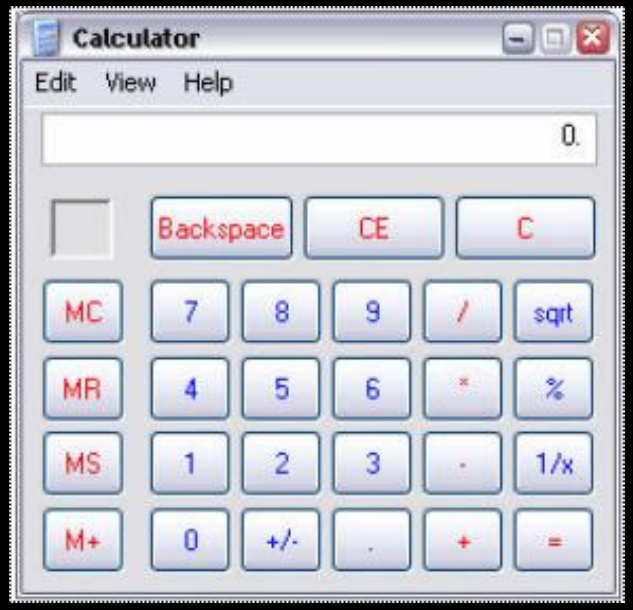
- More success into classifying artifacts you already know to be malware then to actually detect it. (Lineage)

- State of the art? My guess is AV companies!
  - All of them have an absurd amount of samples
  - Have been researching and consolidating data on them for decades.

Lots of Malware Activity

- Can we do better than “AV Heuristics”?
- Lots and lots of available data that has been made public
- Some of the papers also suffer from potentially bad ground truth.

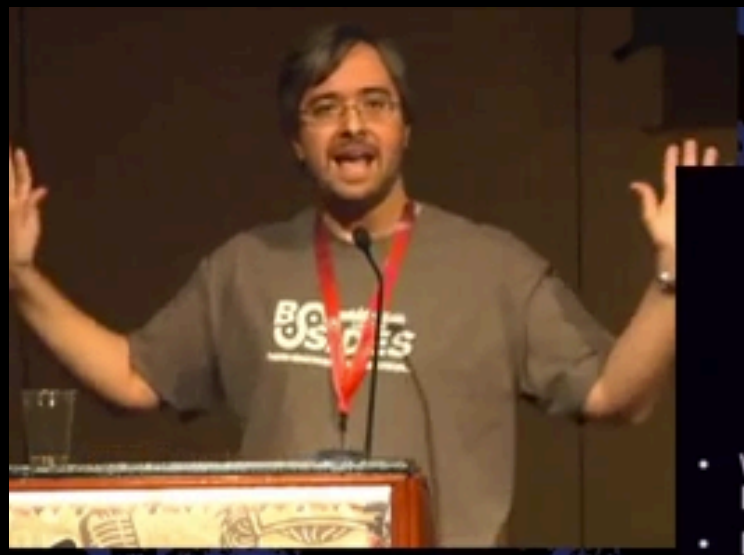
VS



Lots of Malware Activity

VS

Everyone makes mistakes!



Negative and Positive Observations

- We also require "non-malicious" IPs!
- If we just feed the algorithms with one label, they will get lazy.
- CHEAP TRICK: Everything is "malicious" - trivial solution
- Gather "non-malicious" IP addresses from Alexa and Chromium Top 1m Sites.

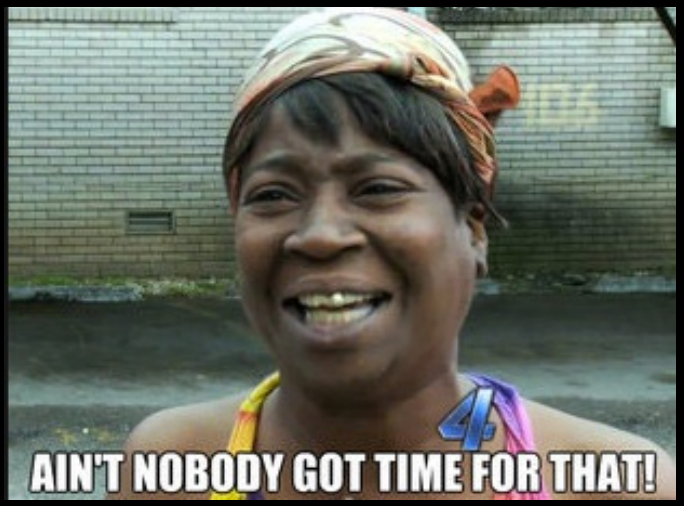
How is it going then, Alex?

- Private Beta of our Threat Intelligence-based models:
  - Some use TI indicator feeds as blocklists
  - More mature companies use the feeds to learn about the threats (Trained professionals only)

- Our models extrapolate the knowledge of existing threat intelligence feeds as those experienced analysis would.
  - Supervised model w/same data analyst has
  - Seeded labeling from TI feeds

Yeah, but why should I care?

- Very effective first triage for SOCs and Incident Responders
  - Send us: log data from firewalls, DNS, web proxies
  - Receive: Report with a short list of potential compromised machines
- Would you rather download all the feeds and integrate it yourself?
  - MLSecProject/Combine
  - MLSecProject/TIQ-test



What about the Ground Truth (labels)?

- Huge amounts of TI feeds available now (open/commercial)
- Non-malicious samples still challenging, but we have expanded to a lot of collection techniques from different sources.
  - Very high-ranked Alexa / Quanticast / OpenDNS Random domains as seeds for search of trust
  - Helped by the customer logs as well in a semi-supervised fashion

But what about data tampering?

- Vast majority of features are derived from structural/ intrinsic data:
  - GeoIP, ASN information, BGP Prefixes
  - pDNS information for the IP addresses, hostnames
  - WHOIS information

- Attacker can't change those things without cost.
- Log data from the customer, can, of course. But this does not make it worse than human specialist.

And what about false positives?

- False positives / false negatives are an intrinsic part of ML.
- “False positives are very good, and would have fooled our human analysts at first.”
- Their feedback helps us improve the models for everyone.
- Remember it is about initial triage. A Tier-2/Tier-3 analyst must investigate and provide feedback to the model.

Buyer's Guide

- 1) What are you trying to achieve with adding Machine Learning to the solution?
- 2) What are the sources of Ground Truth for your models?
- 3) How can you protect the features / ground truth from adversaries?
- 4) How does the solution/processes around it handle false positives?

Buyer's Guide

#NotAllAlgorithms

MLSec Project

- Don’t take my word for it! Try it out!!
- Help us test and improve the models!
- Looking for participants and data sharing agreements

- Limited capacity at the moment, so be patient. :)

- Visit **https://www.mlsecproject.org** , message @MLSecProject or just e-mail me.

Thanks!

- Q&A?
- Don’t forget the feedback!

Alex Pinto
@alexcpsec
@MLSecProject

”We are drowning on information and starved for knowledge"

- John Naisbitt