MLSec

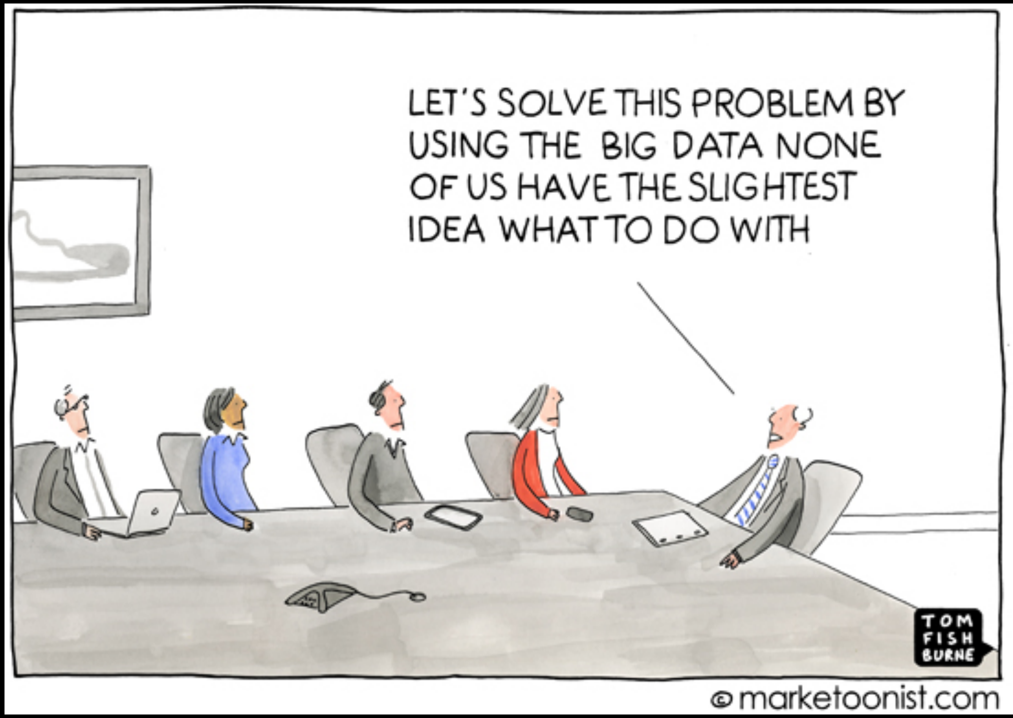
MACHINE LEARNING SECURITY

Secure Because Math: Understanding ML-based Security Products (#SecureBecauseMath)

Alex Pinto
Chief Data Scientist | Niddel / MLSec Project
@alexcpsec
@MLSecProject
@NiddelCorp

Agenda

- Security Singularity
- Some History
- Anomaly Detection
- Classification
- Buyer's Guide


LET'S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH
TOM FISH BURNE
© marketoonist.com

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?”

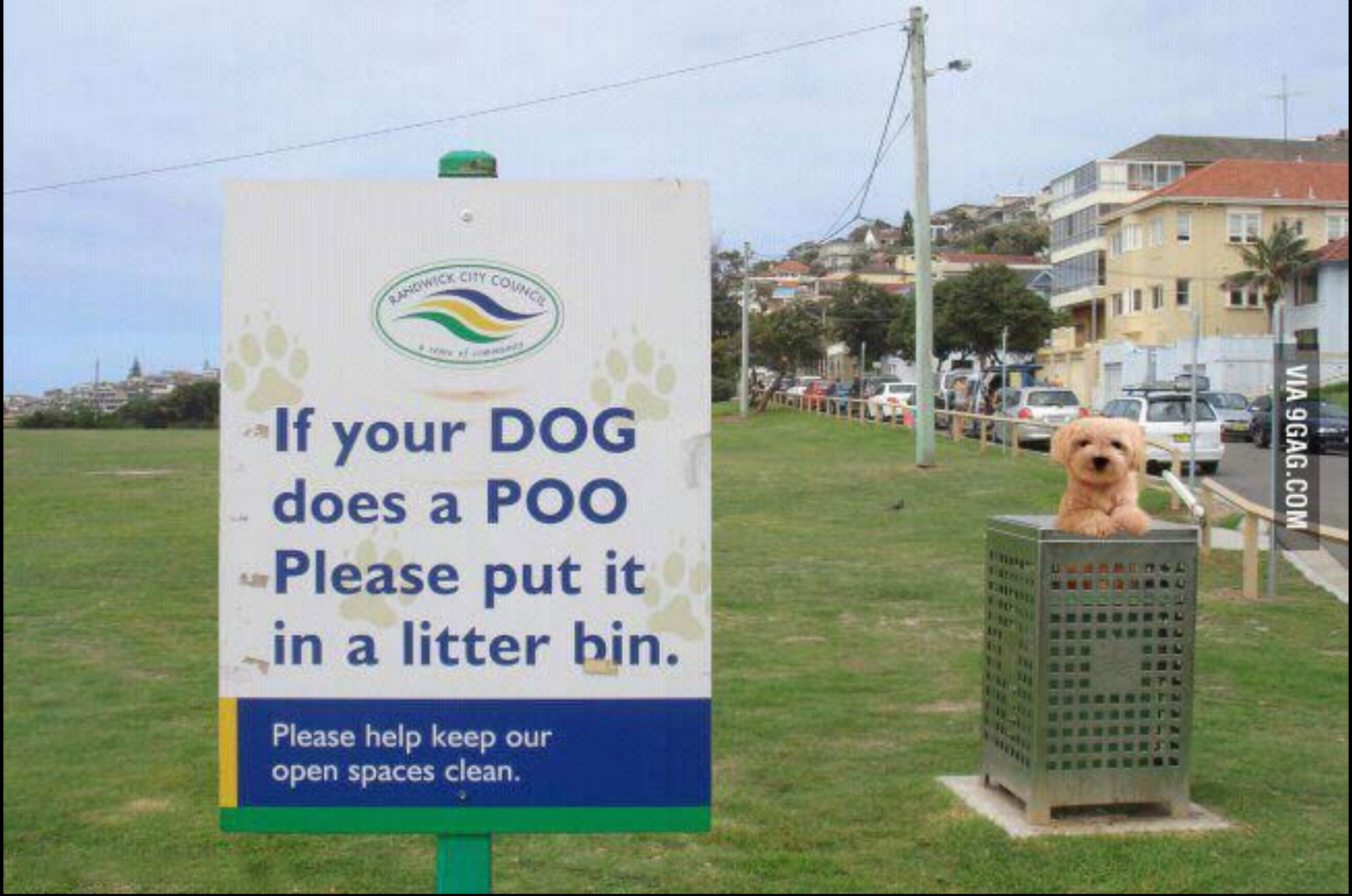
(http://bigdatapix.tumblr.com/)

Are we even trying?

- “Hyper-dim Security Analytics”
- “3rd gen Artificial Intelligence”
- “Secure Because Math”
- “Math > Malware”

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

Is this a communication issue?


RANDWICK CITY COUNCIL
If your DOG does a POO Please put it in a litter bin.
Please help keep our open spaces clean.
VIA 9GAG.COM

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 (...)”

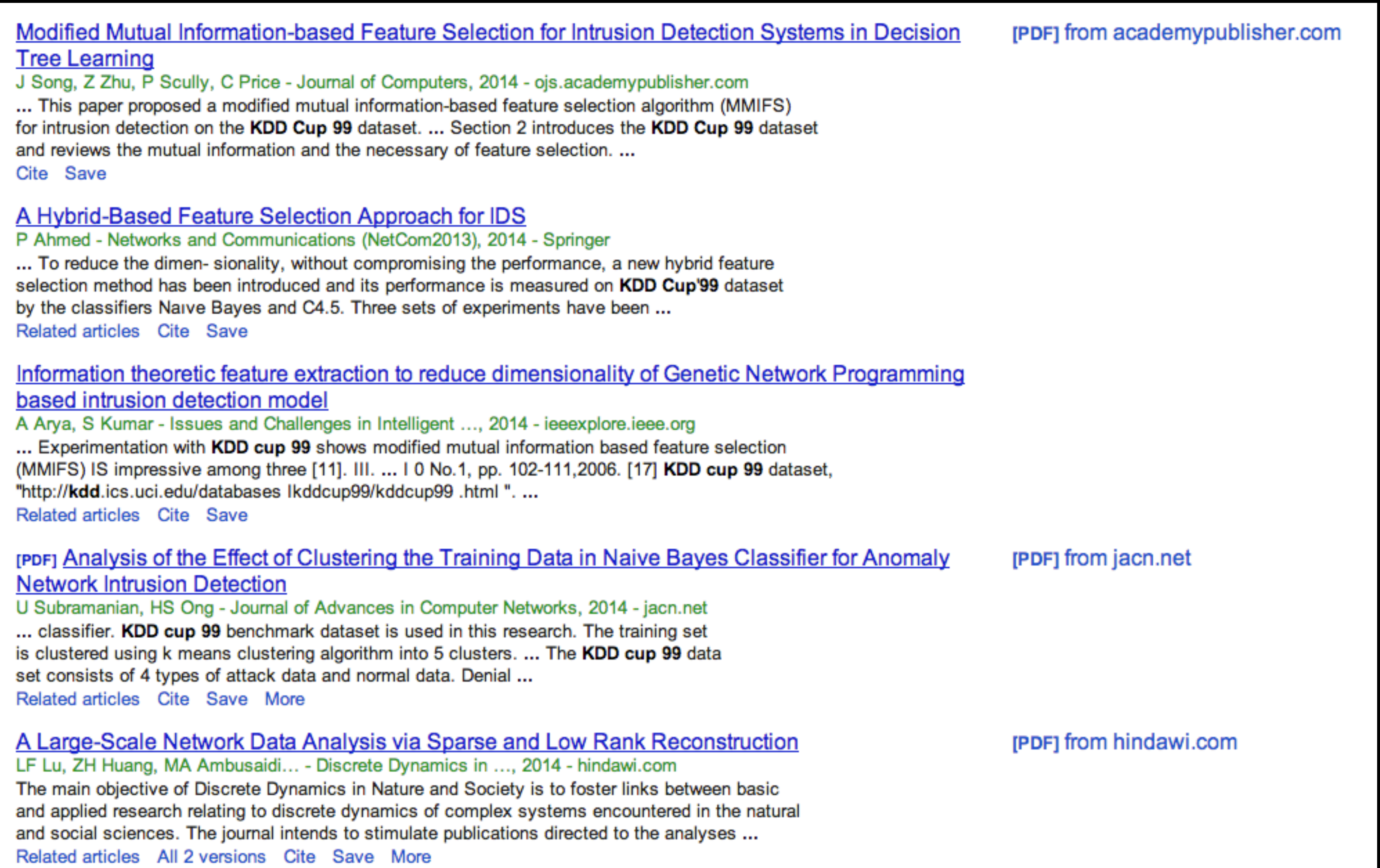
- ”(…) maintains historical profiles of usage per user and raises an alarm when observed activity departs from established patterns of usage for an individual.”

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. ...

[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 ...

[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 ". ...

[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

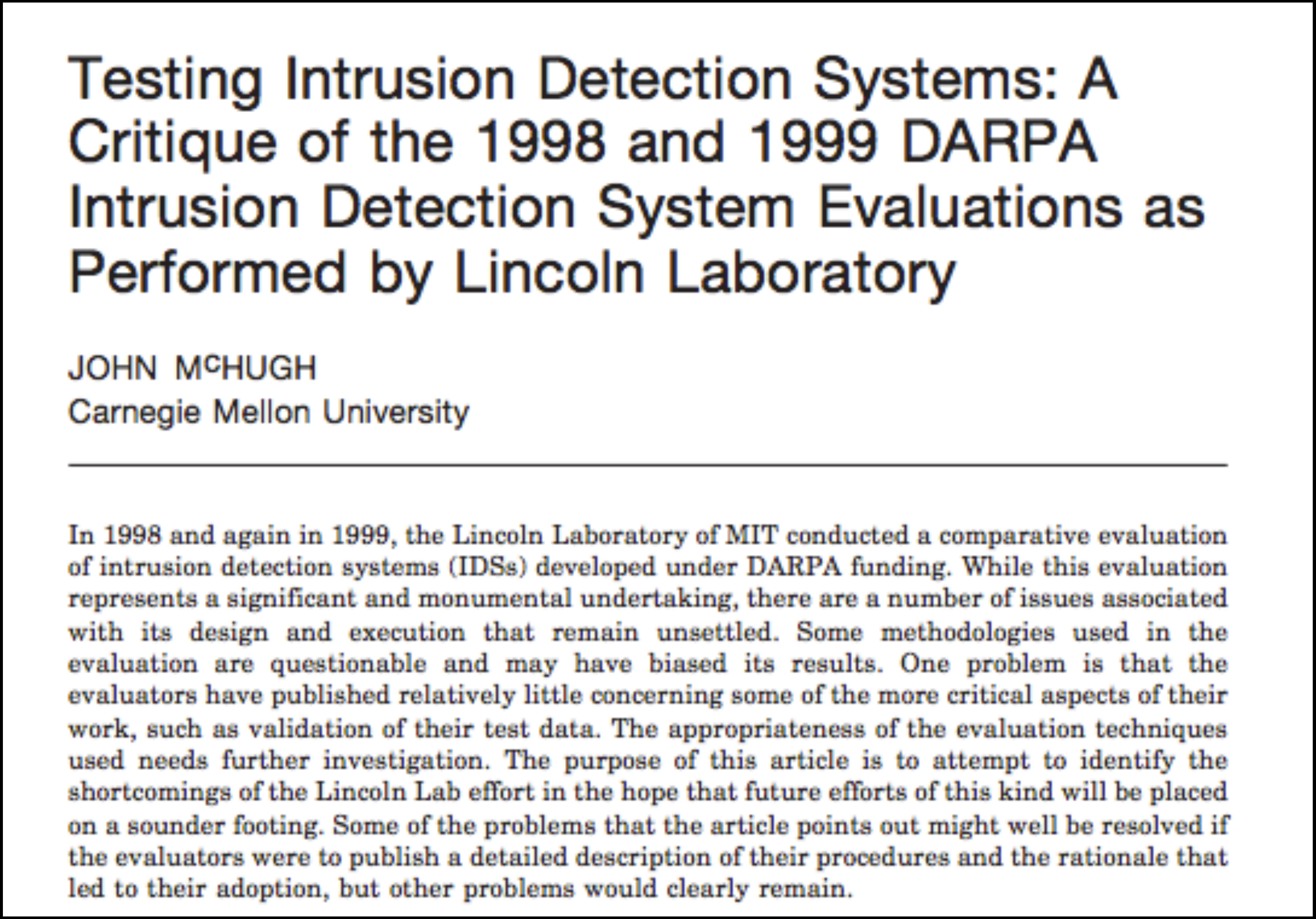
... classifier. **KDD cup 99** benchmark dataset is used in this research. The training set is clustered using k means clustering algorithm into 5 clusters. ... The **KDD cup 99** data set consists of 4 types of attack data and normal data. Denial ...

[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 ...



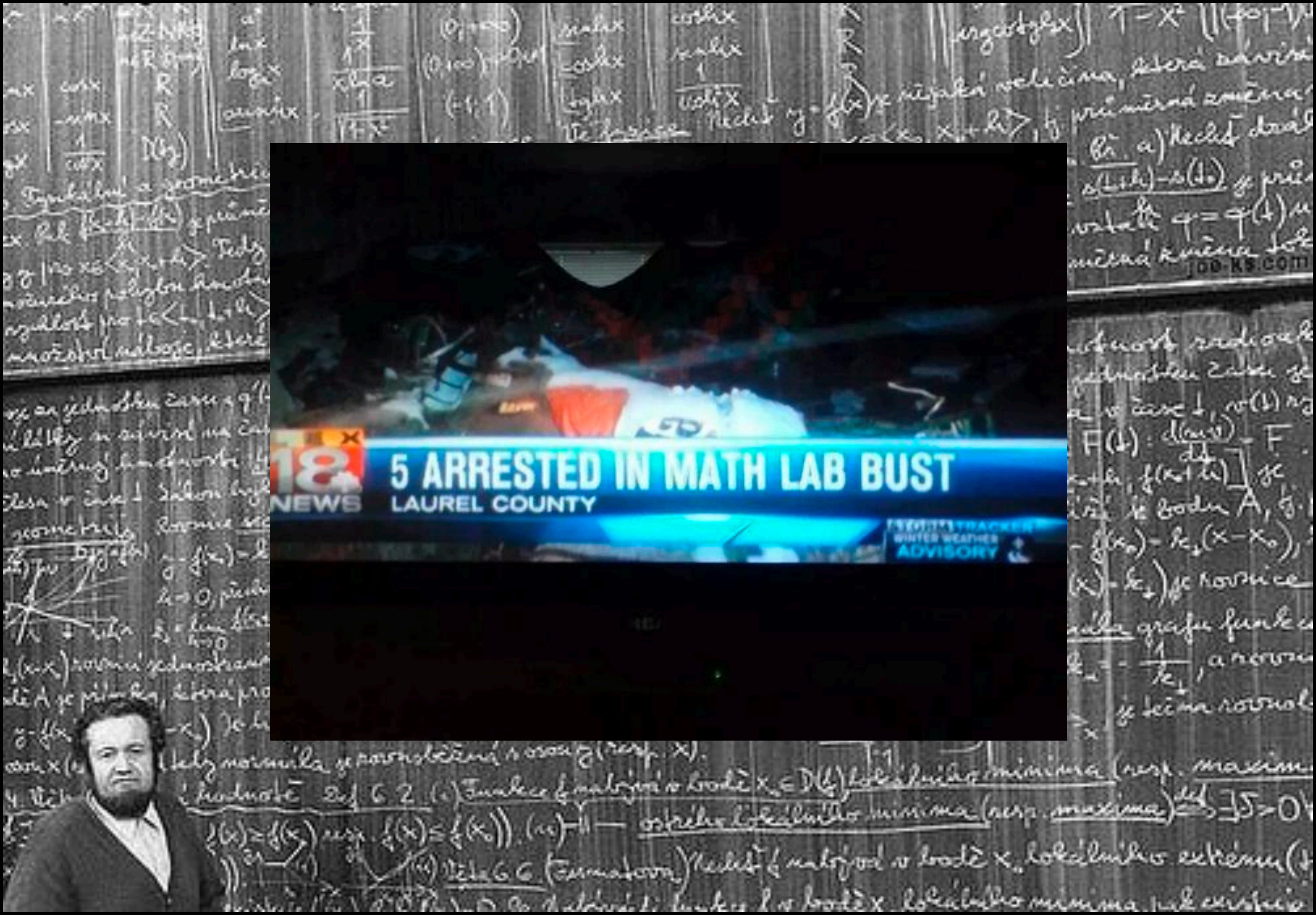
Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations as Performed by Lincoln Laboratory

JOHN McHUGH
Carnegie Mellon University

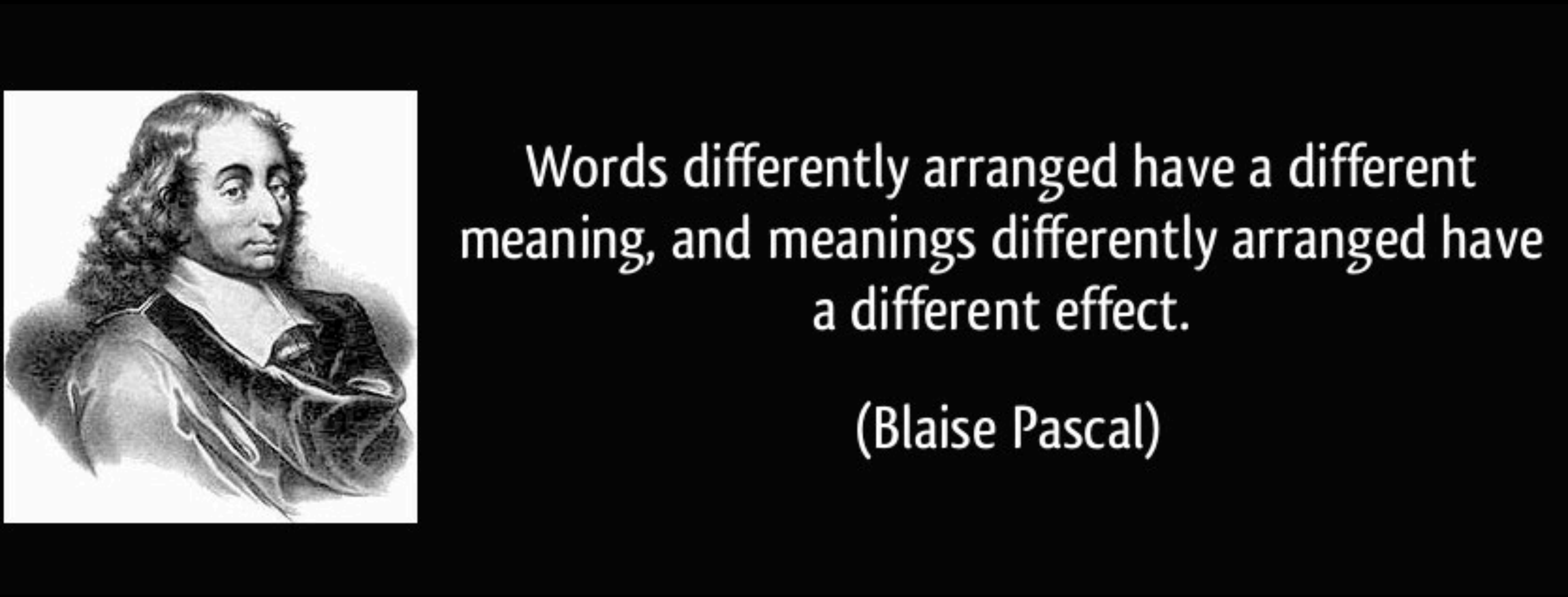
In 1998 and again in 1999, the Lincoln Laboratory of MIT conducted a comparative evaluation of intrusion detection systems (IDSs) developed under DARPA funding. While this evaluation represents a significant and monumental undertaking, there are a number of issues associated with its design and execution 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



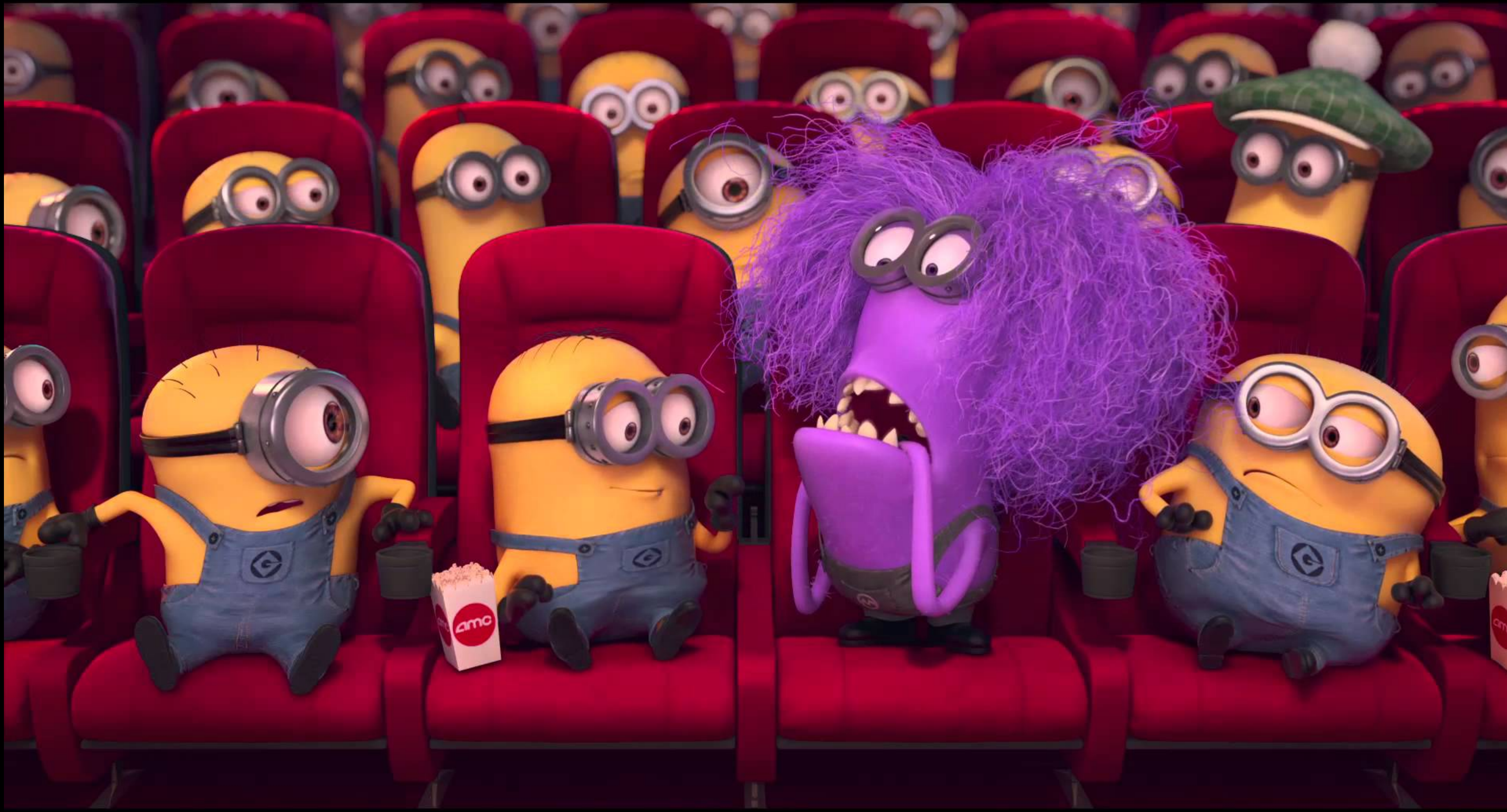
Bringing some meaning to words



Words differently arranged have a different meaning, and meanings differently arranged have a different effect.

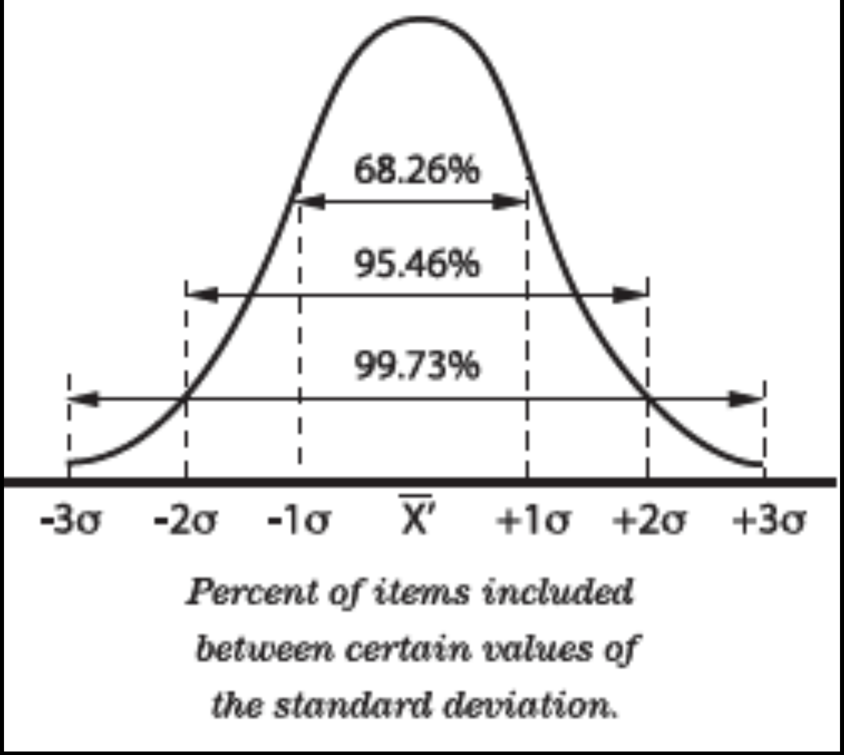
(Blaise Pascal)

Anomaly Detection

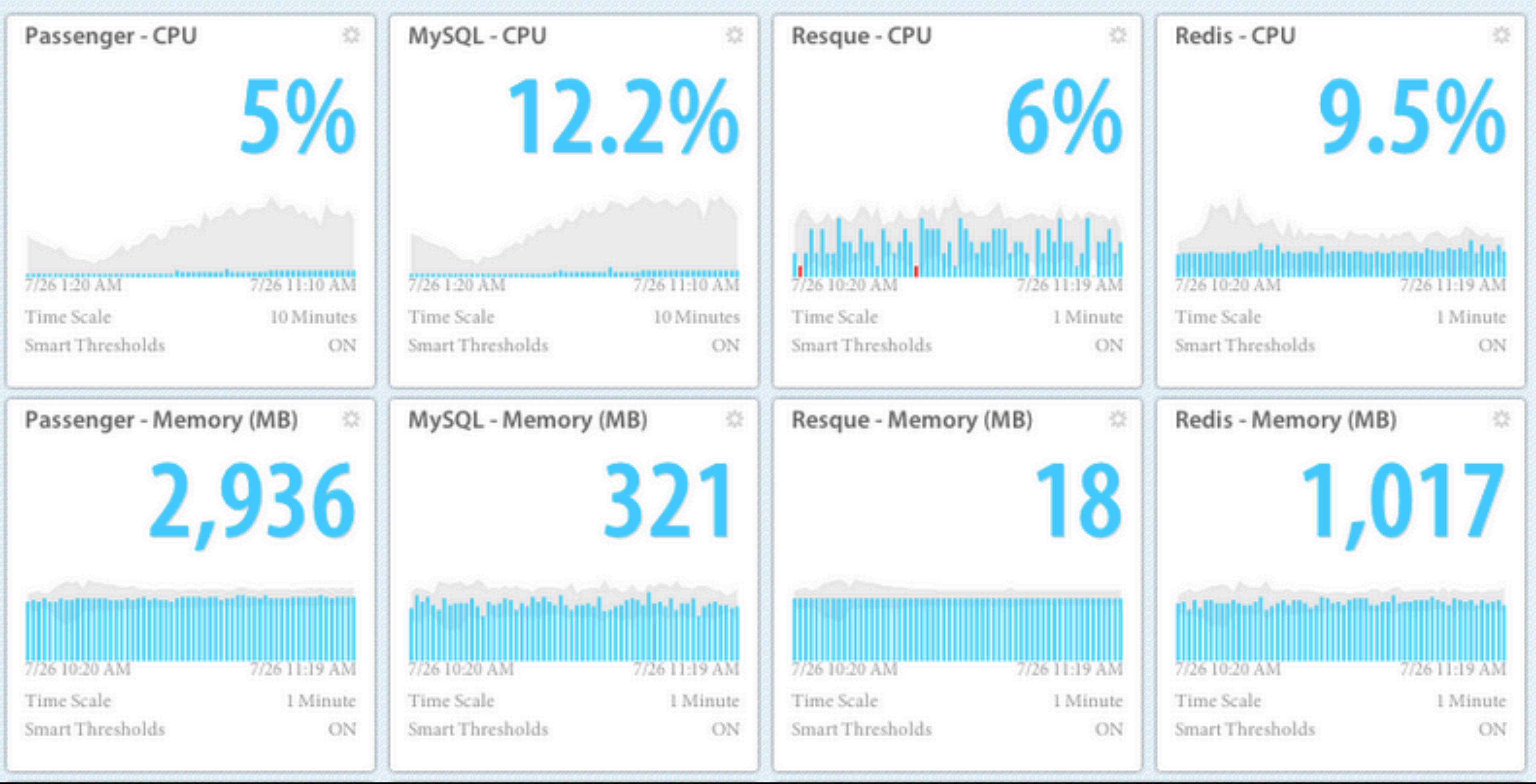


Anomaly Detection

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


68.26%
95.46%
99.73%
-3σ
-2σ
-1σ
X̄'
+1σ
+2σ
+3σ
Percent of items included between certain values of the standard deviation.

Anomaly Detection



Passenger - CPU

5%

7/26 1:20 AM 7/26 11:10 AM

Time Scale 10 Minutes

Smart Thresholds ON

MySQL - CPU

12.2%

7/26 1:20 AM 7/26 11:10 AM

Time Scale 10 Minutes

Smart Thresholds ON

Resque - CPU

6%

7/26 10:20 AM 7/26 11:19 AM

Time Scale 1 Minute

Smart Thresholds ON

Redis - CPU

9.5%

7/26 10:20 AM 7/26 11:19 AM

Time Scale 1 Minute

Smart Thresholds ON

Passenger - Memory (MB)

2,936

7/26 10:20 AM 7/26 11:19 AM

Time Scale 1 Minute

Smart Thresholds ON

MySQL - Memory (MB)

321

7/26 10:20 AM 7/26 11:19 AM

Time Scale 1 Minute

Smart Thresholds ON

Resque - Memory (MB)

18

7/26 10:20 AM 7/26 11:19 AM

Time Scale 1 Minute

Smart Thresholds ON

Redis - Memory (MB)

1,017

7/26 10:20 AM 7/26 11:19 AM

Time Scale 1 Minute

Smart Thresholds ON

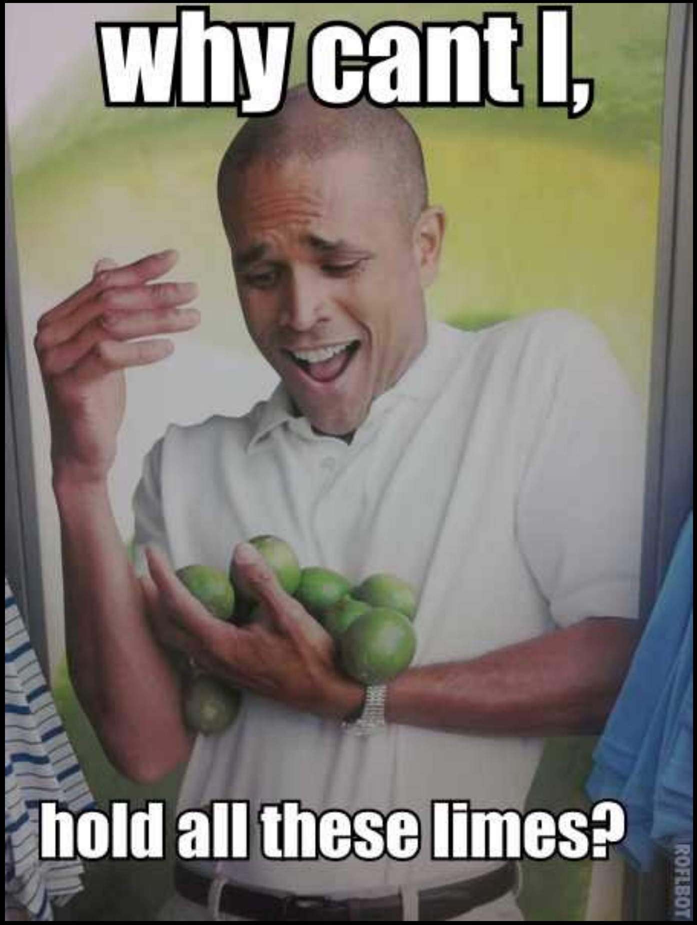
Anomaly Detection

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

- "Sommer, Robert and Paxson, Vern - Outside the Closed World: On Using Machine Learning For Network Intrusion Detection (2010)"

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.

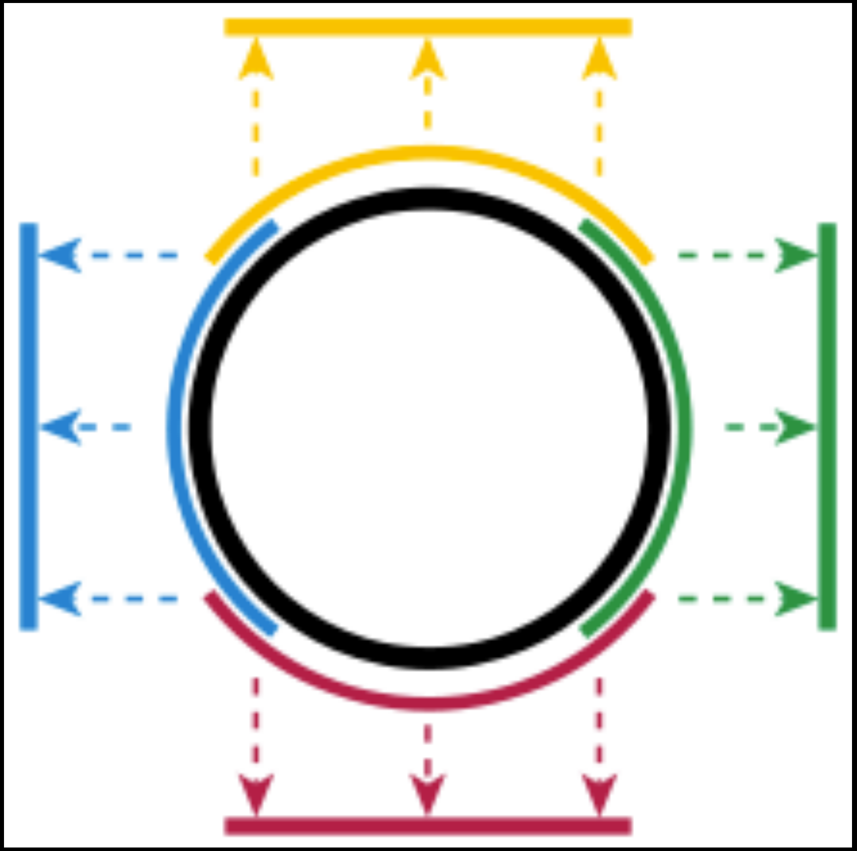


A Practical Example

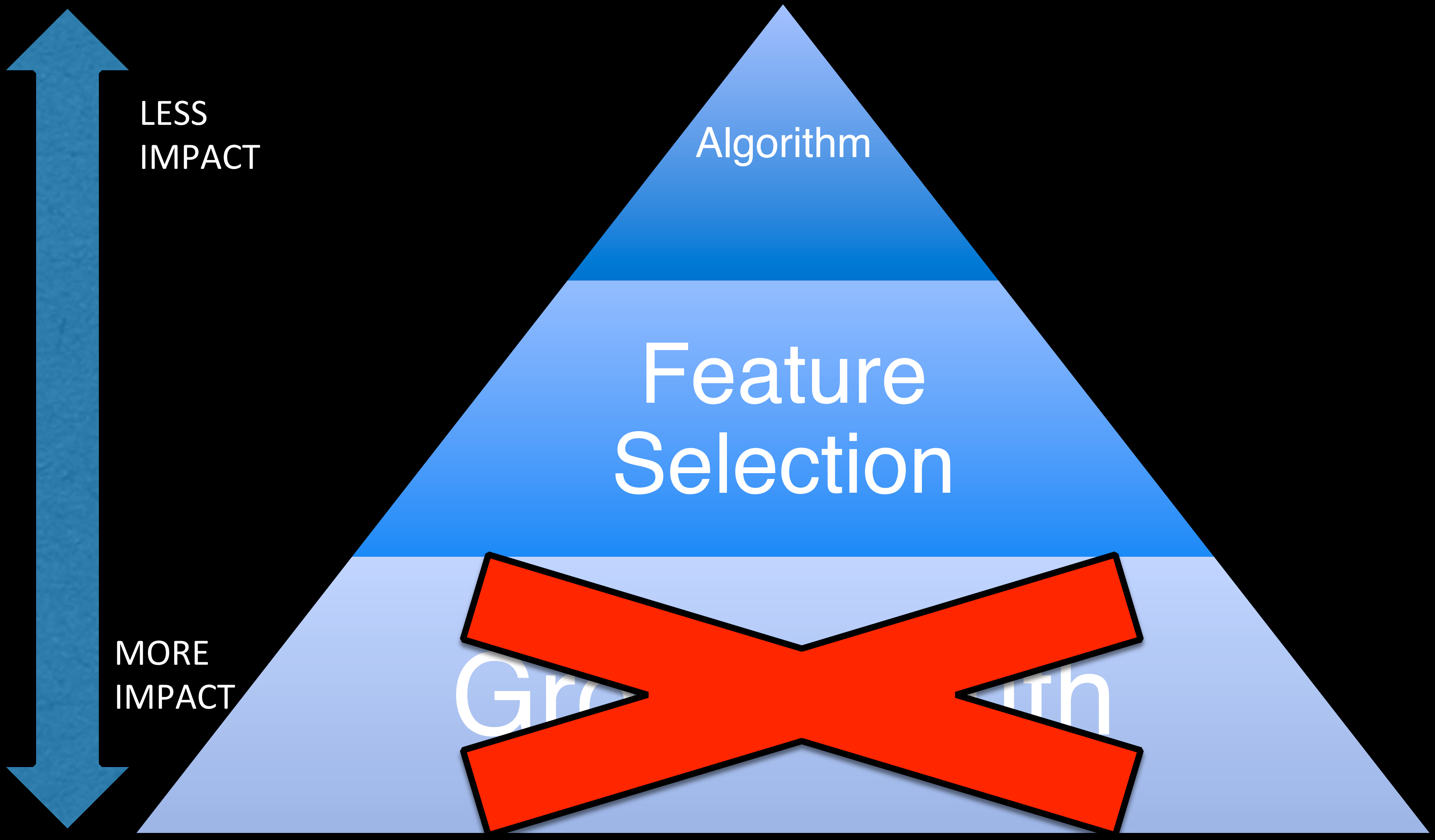
- NetFlow data, company with n internal nodes.
  - 2(nˆ2 - n) communication directions
  - 2*2*2*65535(nˆ2 - n) measures of network activity
  - 1000 nodes -> Half a trillion possible dimensions

Breaking the Curse

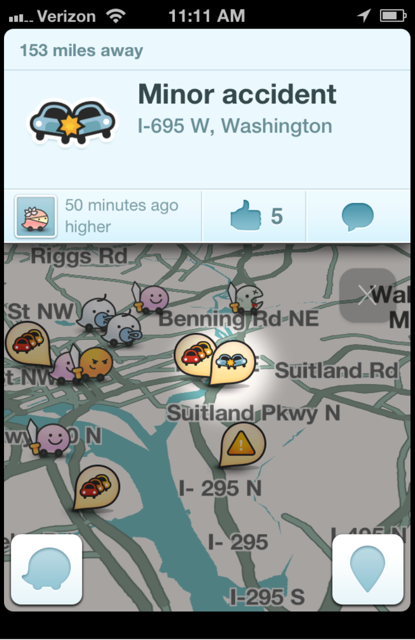
- Different / creative distance metrics
- Organizing the space into sub-manifolds where Euclidean distances make more sense.
- Aggressive feature removal
- A few (very few) interesting results available

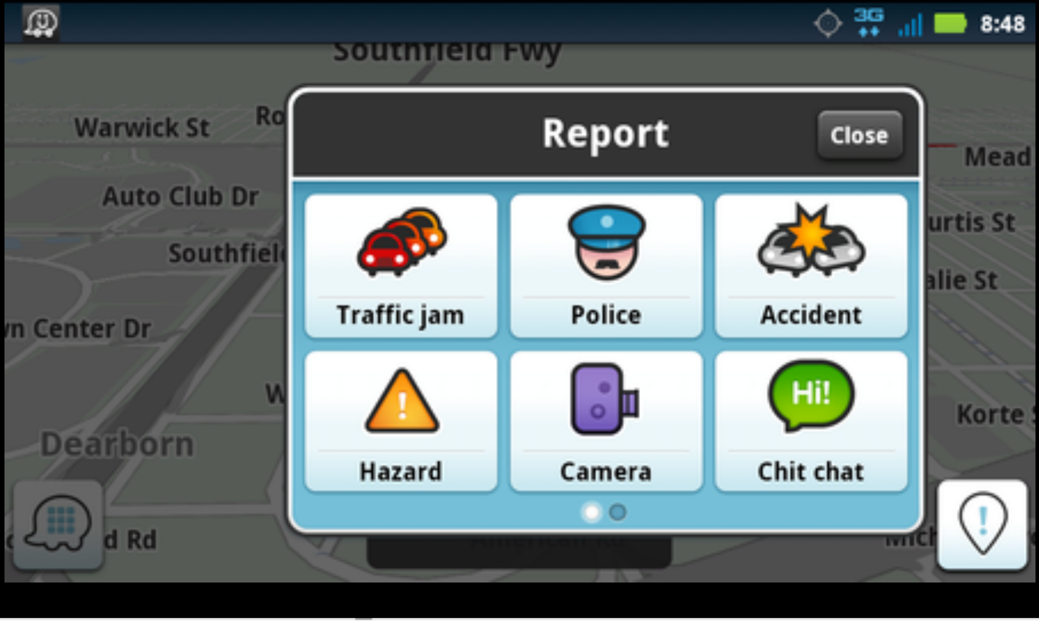


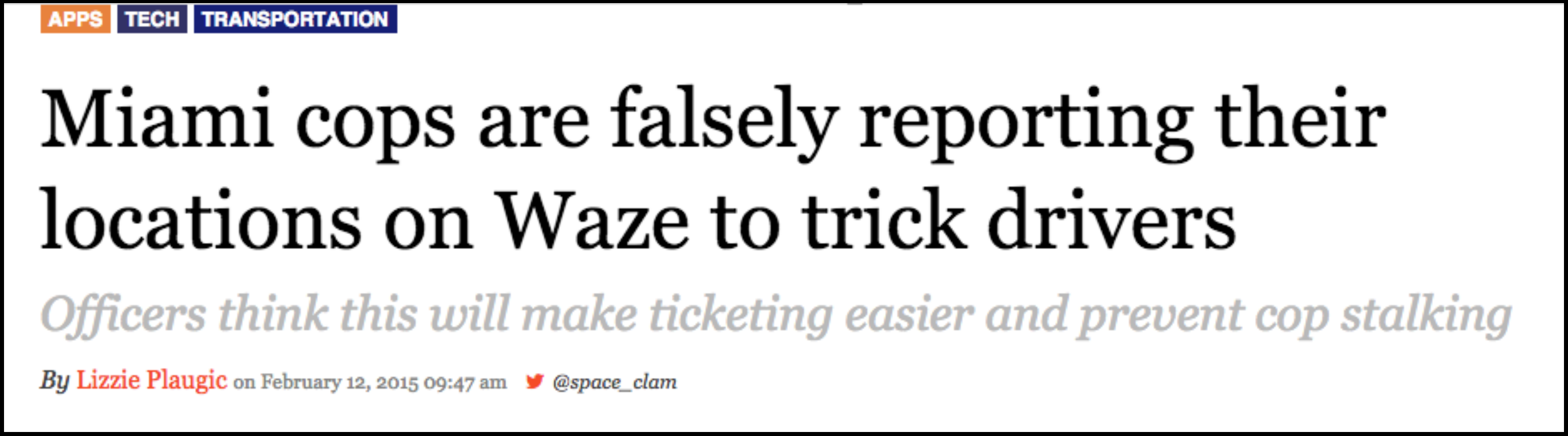
AD: Normality-poisoning attacks



AD: Normality-poisoning attacks






APPS TECH TRANSPORTATION

Miami cops are falsely reporting their locations on Waze to trick drivers

Officers think this will make ticketing easier and prevent cop stalking

By Lizzie Plaugic on February 12, 2015 09:47 am *@space_clam*

AD: Hanlon's Razor (aka Semantic Gap)

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

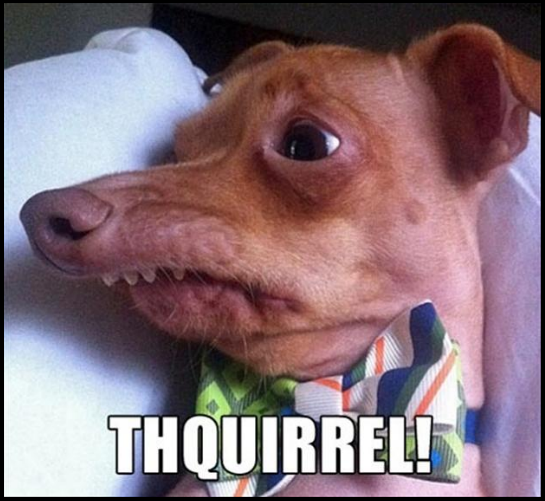
What about User Behavior?

- Behavior analysis != Fraud Detection / Product Security
  - Works as supervised
  - Specific implementations for specific solutions
  - Well defined scope and labeling

- Can user behavior (AD analysis, VPN connections) be generalized enough?
- Should I "average out" user behaviors in different applications?

Classification!

VS



Why Supervised?

- If you can establish "ground truth" for your security data, by all means, use it
  - Good sources of "malicious" data (mostly paid)
  - Vetted "non-malicious" is still a challenge

  (Irena Damsky's – CheckPoint - talk on "whitelist" data)

- Timeliness is also a big challenge. Your predictions affect the world, and you have to keep updating the models

- When you create a classification model, you are forced to think about what you are modeling.

Lots of Malware Research

- GROUND TRUTH (PBs of malware)!!
- Lots of available academic research around this
- Classification and clustering of malware samples

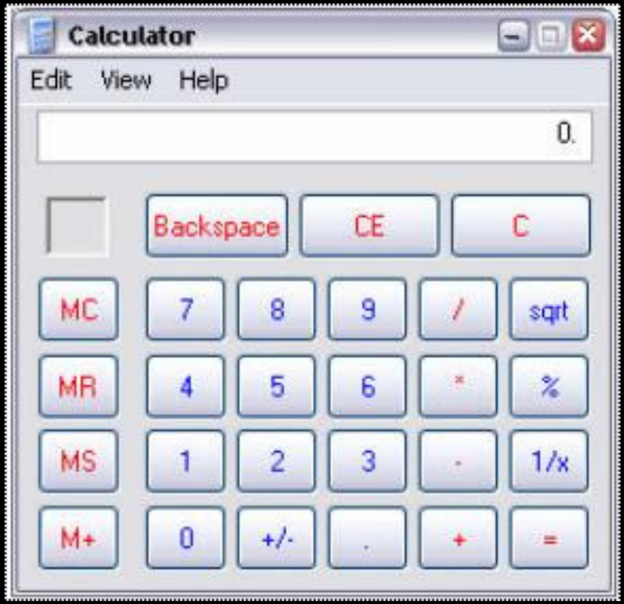
- More success into classifying artifacts you already know to be malware then to actually detect it. (Lineage)

- 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 bad ground truth.

VS

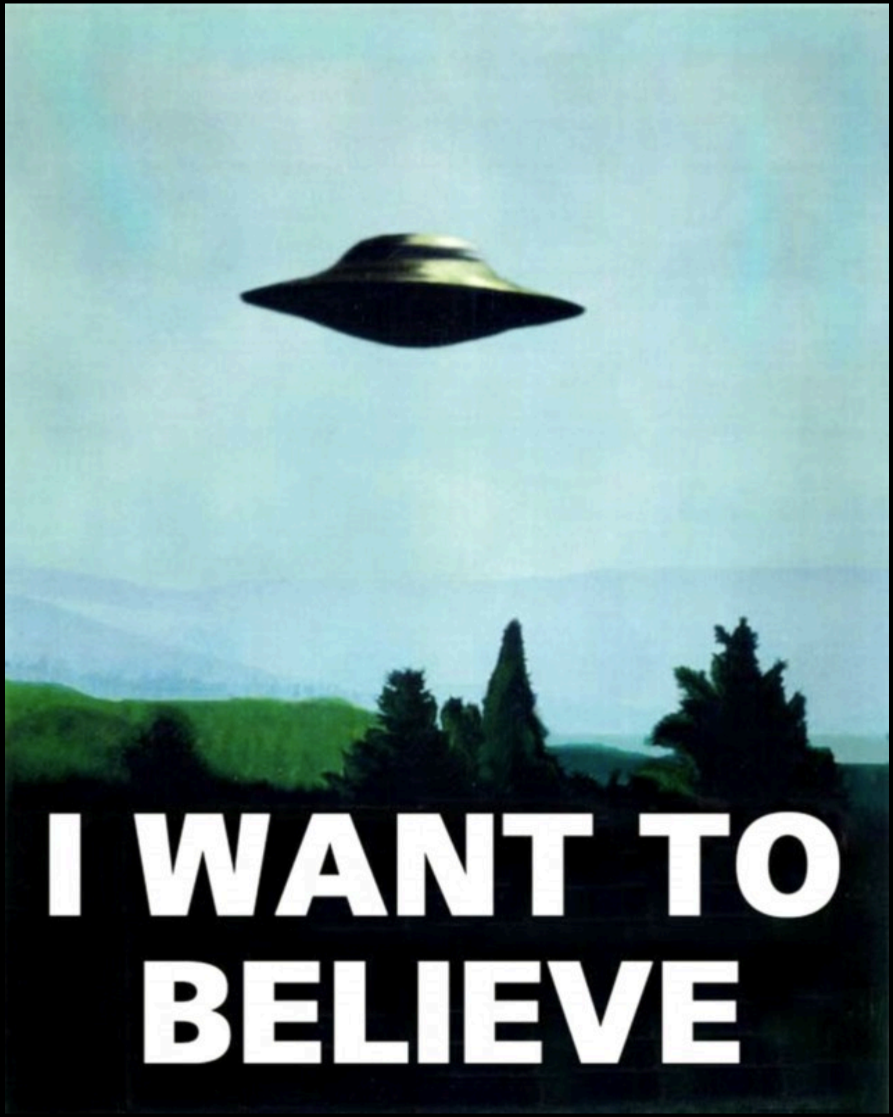


Network Data Classification

- "IP reputation" was the first step in this direction
- Lots of published papers about DGA classification and domain classification in general
- Terry Nelms' – Damballa - talk has good examples

- Greatest change in the last 2 years – Threat Intelligence
  - Network indicators can be used as malicious labels
  - Feature extraction from internet topology, pDNS and WHOIS infrastructure
  - This is what Niddel's technology is based on

Can we fix security with ML?



Security Machine Learning 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 and ground truth from adversaries?
- 4) How does the solution and processes around it handle false positives?

MLSec Project / Niddel

- MLSec Project – research-focused branch of Niddel for open-source tools and community building
- Niddel builds Magnet, a machine learning-based Threat Intelligence Platform, that applies threat intelligence in a fully personalized and hassle-free way.
- Looking for beta testers and research collaboration

- https://www.niddel.com
- https://www.mlsecproject.org

MLSec
MACHINE LEARNING SECURITY

NIDDEL™

Thanks!

- Q&A?

Alex Pinto
@alexcpsec
@MLSecProject
@NiddelCorp

DOGBERT, THE VP OF MARKETING
DESCRIBE YOUR PRODUCT IN TECHNICAL TERMS AND I'LL TURN IT INTO MARKETING LANGUAGE.
WELL, IT TENDS TO OVER-HEAT.
"HOTTEST PRODUCT ON THE MARKET!"
ALL THE PARTS ARE KNOWN CARCINOGENS.
"MAKES YOU APPRECIATE LIFE!"
www.dilbert.com
scottadams@aol.com
©2007 Scott Adams, Inc./Dist. by UFS, Inc.
9-17-07

© Scott Adams, Inc./Dist. by UFS, Inc.

"We are drowning on information and starved for knowledge"

- John Naisbitt