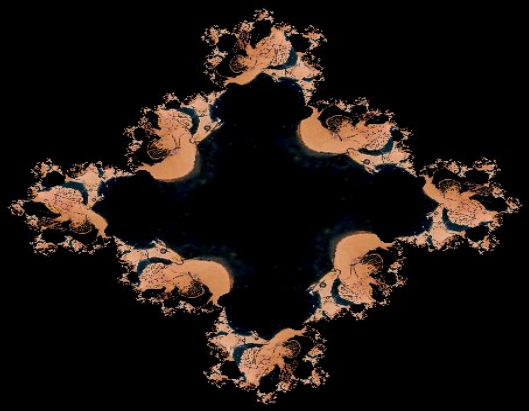
Redteam Village 2020

Aktaion: Hands on with Open Source ML for Exploit Detection

Joseph Zadeh @josephzadeh

Rod Soto @rodsoto



Learning Objectives and Requirements

1. **Building an end to end intrusion detection workflow for detecting exploit behavior**
2. **Analysis of exploit attack data for meaningful encoding of statistics and behaviors to build detections with**
3. **Examples of orchestration workflows tying in an upstream action into a detection event**
4. **The Do's and Don'ts of using Machine Learning (and complex software paradigms) for Cybersecurity use cases**

Audience: network and security engineers and architects, or anyone with an interest in hacking and building tools that use behavior for defensive use cases.

Students are encouraged to use their own laptop and latest versions of python/pip tested and working. It is recommended to use PyCharm Community Edition and have a working version of Bash or Zsh installed if you want to follow along with the training examples interactively.

Training Info

- Class Format
- Speaker will be on video and audio, as well as chatting over Discord.
  - You will only be able to ask questions in Discord (not via voice to the speaker).
  - Join the [Red Team Village Discord Server](https://RedTeamVillage.io/discord),
- ***#Hands-on-labs*** channel during the training
- Discord Related Stuff: starting point for Defcon Redteam Village:

  https://discord.gg/redteamvillage

Outline:

- ***Part 1: Theory and Key Concepts***
- ***Part 2: Hands On Demo***

$Whoami (Project Contributors)

Rod Soto (@RodSoto Twitter, @Trajan Discord)

Principal Security Research Engineer, Splunk. Former , AKAMAI, Prolexic PLXSert Principal Researcher. Like to break things, p0wn botnets and play CTFs.

Joseph Zadeh @JosephZadeh Twitter, @MathMakesMeDance Discord (Class Instructor)

Principal Security Research Engineer, Splunk. Silicon Valley career building behavioral intrusion detection technologies at scale. Enjoy working on defense projects that combine security, artificial intelligence and distributed systems.

Aktaion Current Project Home Page:

https://github.com/jzadeh/aktaion

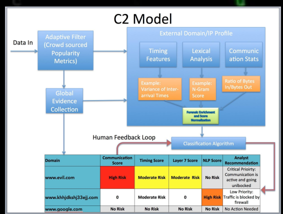
https://github.com/jzadeh/aktaion2

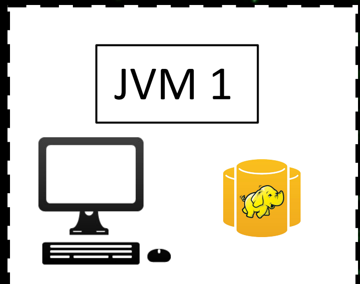
Thanks to @MykeylxKnight (Discord + Twitch), John Pierce @Splunk, Alex Zadeh, et al

Fun Security Research Examples

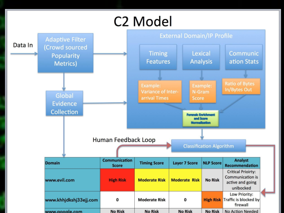
- Splunks Threat Research Team Open Source
  - https://github.com/splunk/security-content (detections)
  - https://github.com/splunk/attack_range (check out Rod and Jose's talk/training in red team village this year)
- Older Presentation Examples
  - Defcon 2018 Recon Village: Fighting Human Trafficking and modeling on the darkweb : https://www.youtube.com/watch?v=S2nOj8r0i5Y
  - Dynamic Population Discovery For Lateral Movement Detection Joseph Zadeh and Rod Soto (Defcon / Hackmiami 2016): https://www.youtube.com/watch?v=LzzbbX5a3J0
- Accepted Patents
  - A graph based method to detect Malware Command and Control Infrastructure US 9195826 B1
  - Malware Communications Detection , SPO139.18US (8035.US 01), 112509-8034.US01, 14/929,204
  - Fingerprinting entities based on activity in an information technology environment, 112509-8096.US01

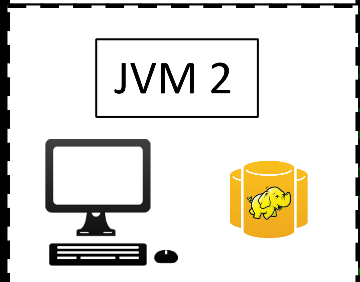
Today's Research Discussion



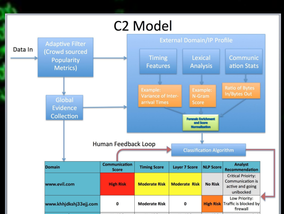


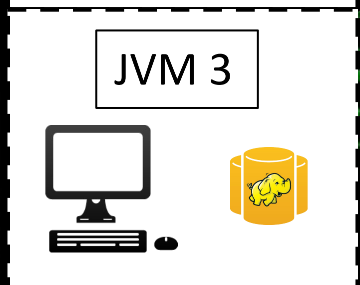
1. GET http://forbes.com/gels-contrariness-domain-punchable/"
2. GET http://portcullisesposturen.europartsplus.org/
3. POST http://dpckd2ftmf7lelsa.jjeyd2u37an30.com/





1. GET http://youtube.com/
2. GET http://avazudsp.net/
3. GET http://betradar.com/
4. GET http://displaymarketplace.com/



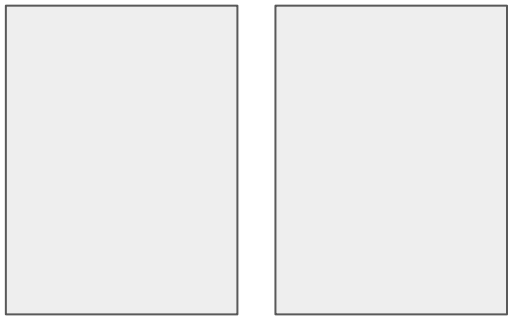


1. GET http:/clickable.net/
2. GET http://vuiviet.vn/
3. GET http://homedepotemail.com/
4. GET http://css-tricks.com/

ML first steps in adaptive security: Commoditization of Compute

1 Trillion Letters of DNA = 800 MB (all of nature's local adaptation information encoded in a string < 1 GB in Size). The real value in learning is Time!!

1 Terabyte of Data

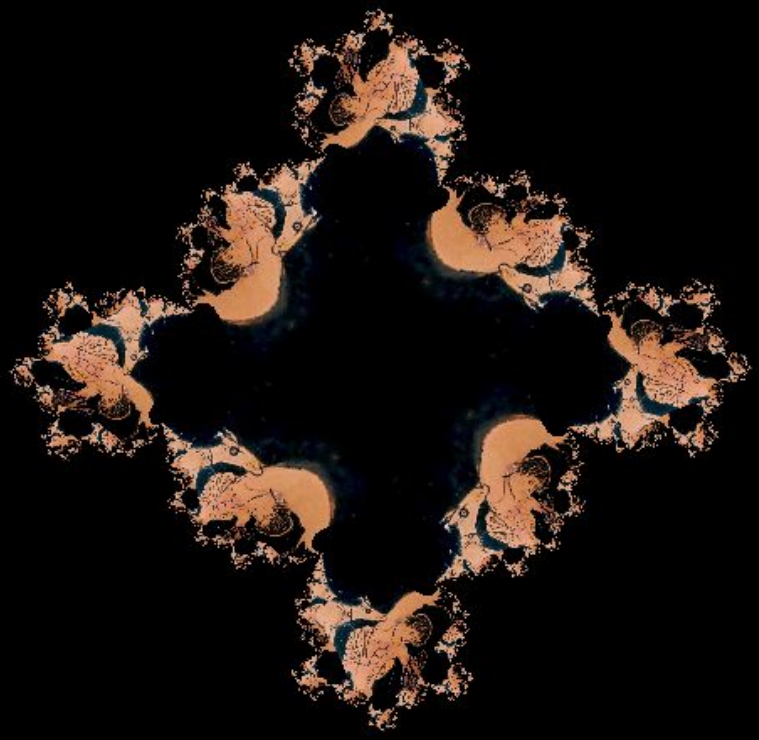


1 Kilobyte of Output

>(41.107678, -3.635677, 0.007473, -1.784944)

PART 1:

Theory and Concepts



Training Objectives

1. Part 1
   a. Why Aktaion ? history and ransomware discussion
   b. Ransomware: use case motivation
   c. ML: Tribal knowledge for security
   d. Theoretical Comments on Adaptation -> pattern recognition immune system paper
2. Part 2
   a. Building the project and running the demo
   b. Entry point into building a new behavior based “feature”
   c. simple character count example?
   d. Extra credit: orchestration example for taking action on GPO automation given classifier output

INTRODUCTION:
Ransomware Use Case

Expressing Microbehaviors in Code: Aktaion V2

https://github.com/jzadeh/aktaion2/tree/blackhat_eu_2017
(careful not merged into master until EOW for now use the following for access)

git clone https://github.com/jzadeh/aktaion2.git

pip install virtualenv && virtualenv -p python3 venv && source venv/bin/activate && pip install -r requirements.txt

Migrated Core Logic and Modules to Python3 Implementation

aktaion2/python

./run_ml.sh : builds machine learning ready data sets given input directories and window sizes and config parameters

./run_aktaion2.sh : use to reproduce what you see in the arsenal demo and score individual files

aktaion2/python/research_dev/random_forest:

- micro_beahvior_core_logic.py : key abstraction used for building important characteristics about attack/benign in a form useful for AI to learn from
- rf_v2.py, svm.py, calibration.py : main machine learning based tools using Random Forest, Support Vector Machine and Classifier performance analysis

Introduction

- Crypto Ransomware has become an increasing attack vector used by malicious actors to quickly turn infections into profits.
- Current state of threat detection technologies is based on static signature approach applied to executing binaries.
- This approach is insufficient and inefficient protecting victims against this type of threat as malicious actors will apply obfuscation and user deceiving techniques that easily bypass static signature based defense technologies

Introduction

A novel approach against this threat using Machine Learning algorithms provides a framework to approach ransomware without depending on static signature, basing its detection on contextual indicators and micro behaviors of such type of malware.

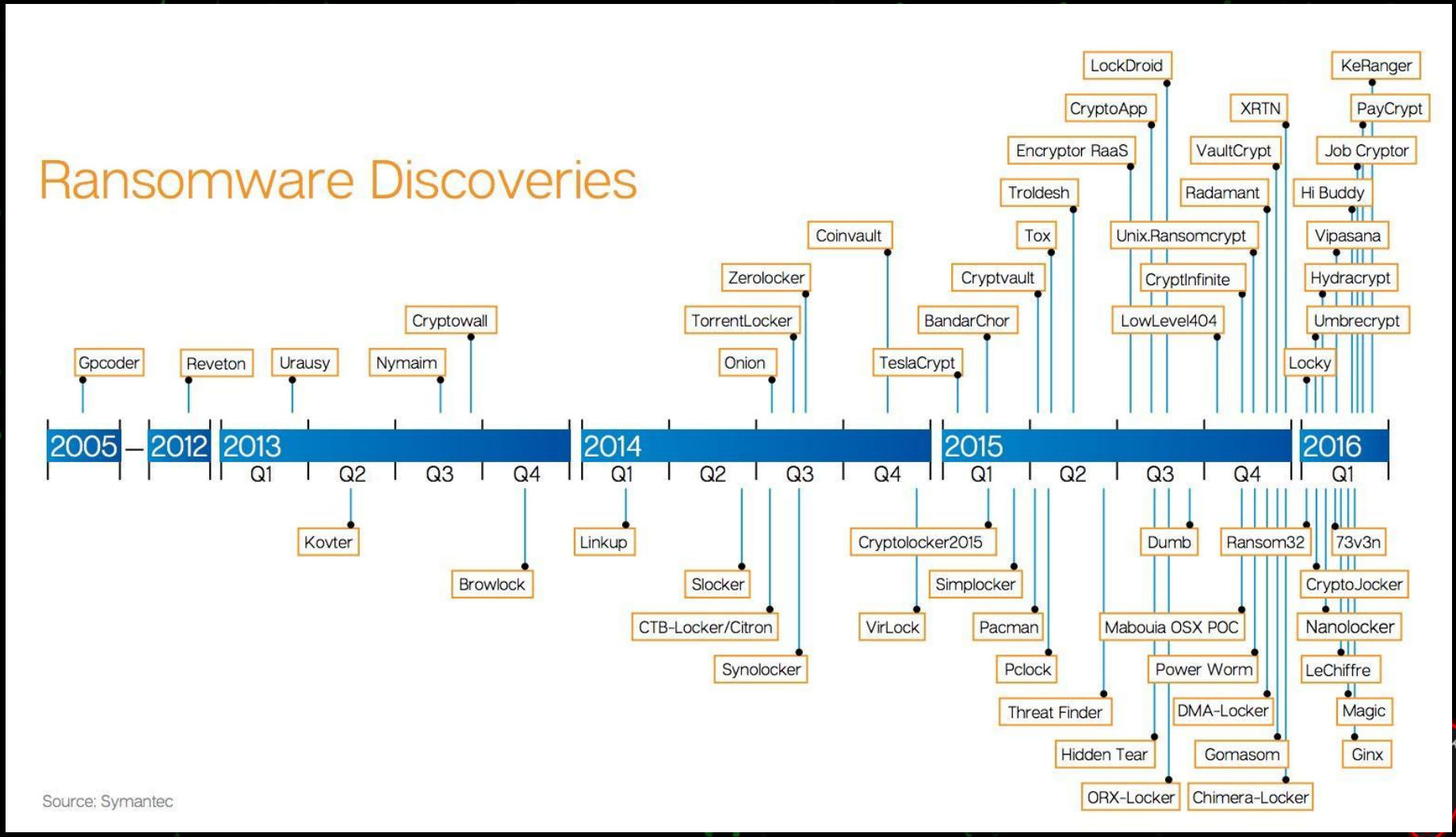
What is Ransomware?

What is Ransomware

Ransomware is a type of malware that can be covertly installed on a computer without knowledge or intention of the user that restricts access to the infected computer system in some way,[1] and demands that the user pay a ransom to the malware operators to remove the restriction.

* Wikipedia

What is Ransomware



Ransomware IOCs

- The modification of the registry keys (Most associated with persistence. I.E execute after reboot).
- Renames and encrypts file extensions of files (Targets User's docs. I.E .doc, xls, ppt, mp3, wallet).
- Modifies Master Boot Record to prevent rebooting, usually encrypting it relocating it and placing a replacement.
- Removal of Volume Snapshot Service files (VSS) or volume shadow files, use for system restoration and backup
- Encryption of map and unmapped network shares with write permission.
- Some variants show outbound connection to Command & Control Server (C2). In some cases TOR traffic is observed. Notice that not all variants of ransomware present C2 communications.
- Use of RSA encryption (2048,9046) *, AES encryption algorithm.

What is Ransomware and why are victims paying?

BUSINESS INSIDER

LEARN MORE ABOUT OUR

The FBI says you may need to pay up i your computer with ransomware

Tess Danielson
Oct. 26, 2015, 2:56 PM 15,487 4

FACEBOOK LINKEDIN TWITTER

2016 Big Data Trends
tableau.com/big-data
Top 8 Trends in Big Data for 2016. Get the Whitepaper!

If a hacker hijacks your computer with malware and holds your data

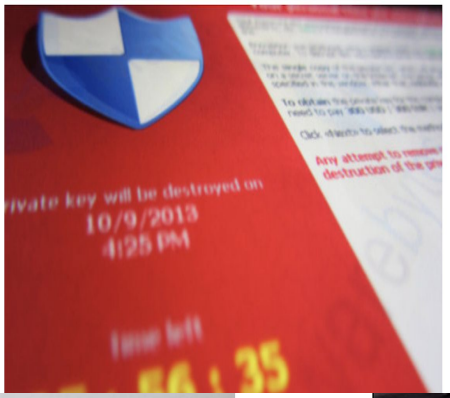
CSO FROM IDG

INSIDER Sign In | Register

NEWS

Ransomware damage costs predicted to hit $11.5B by 2019

Cyber crime rises sharply amidst a decline in the percentage of victims willing to pay ransoms.



MORE LIKE THIS

How K-12 schools can protect against ransomware attacks

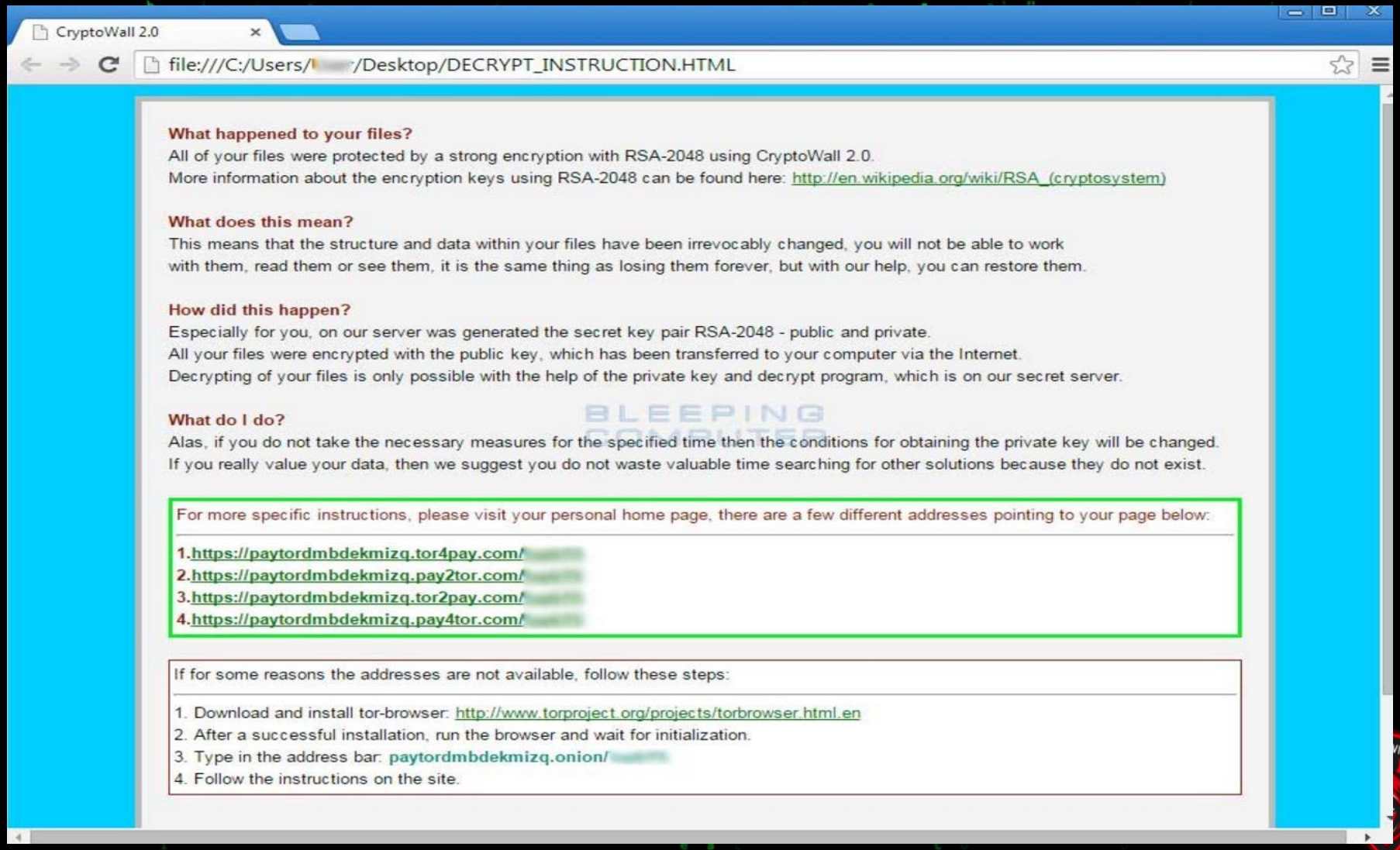
Kevin Mitnick's ransomware defense firm lands $30M investment

What's new in ransomware?

Current Ransomware Threatscape

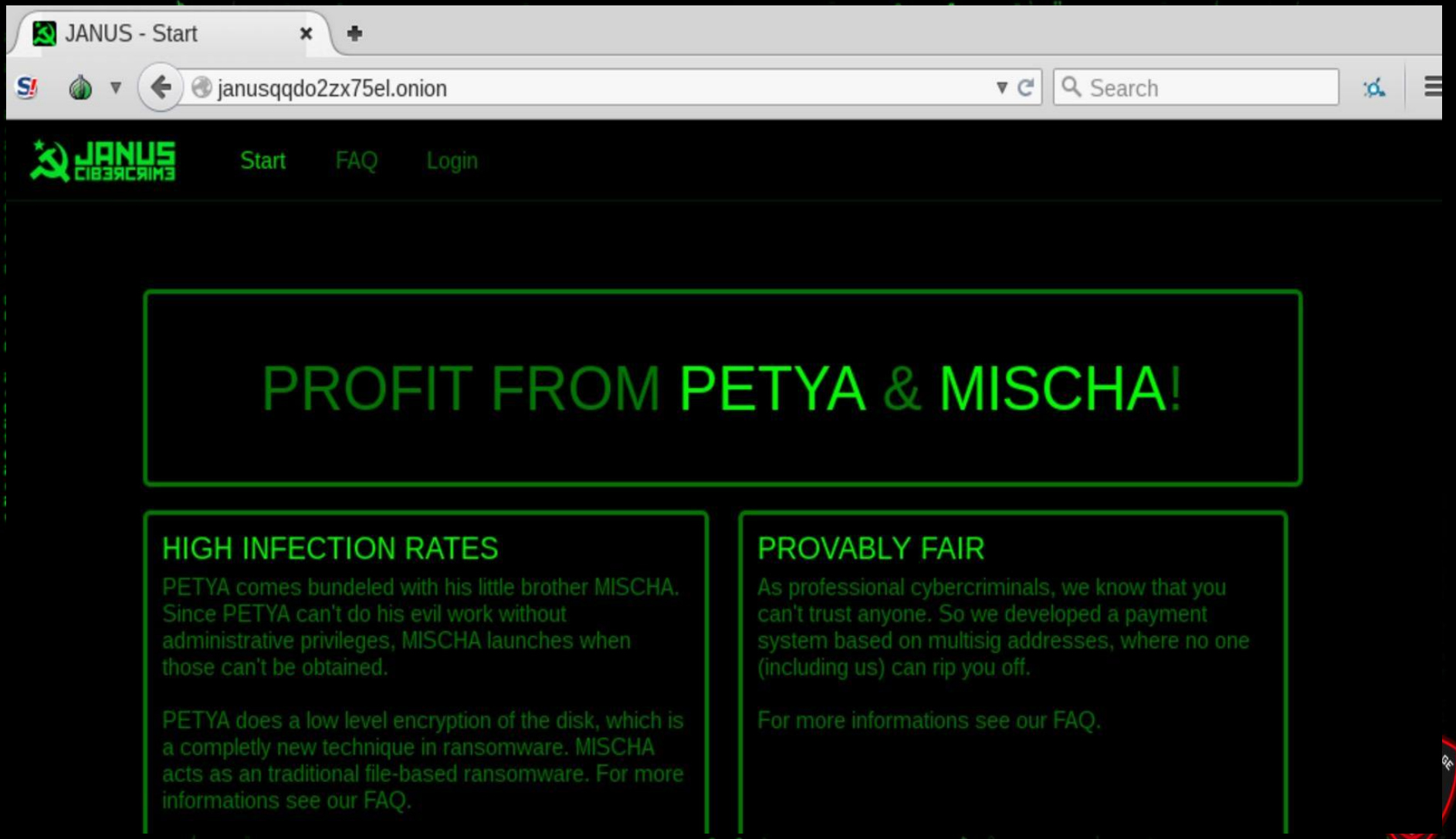
- Ransomware malware cost $18 million in loses according to FBI in 2015 (Each time cost between $200 & 10K). Fast forward half a decade and 11.5 Billion by 2019
  - Update: https://www.infosecurity-magazine.com/news/ransomware-costs-may-have-hit-170/
- Attacks targeting individuals and organizations, HealthCare & Utilities present a concerning example of real life consequences on human well beings.
- Attacks continue increasing due to successful infestation and ransom cash out.
  - Adversarial shift shows current defense technologies are insufficient.
- Crypto ransomware payload being used as post exploitation payloads in EK such as Zeus, Drydex, Neutrino, etc.
  - Monetization expanding further (Ransomware as a Service)
  - Bitcoin/TOR unfortunately an enablers of this modus operandi
- In some cases they are used as a SHROUD to cover destructive targeted campaigns (NotPetya, BadRabbit)

TOR unfortunately a crime enabler

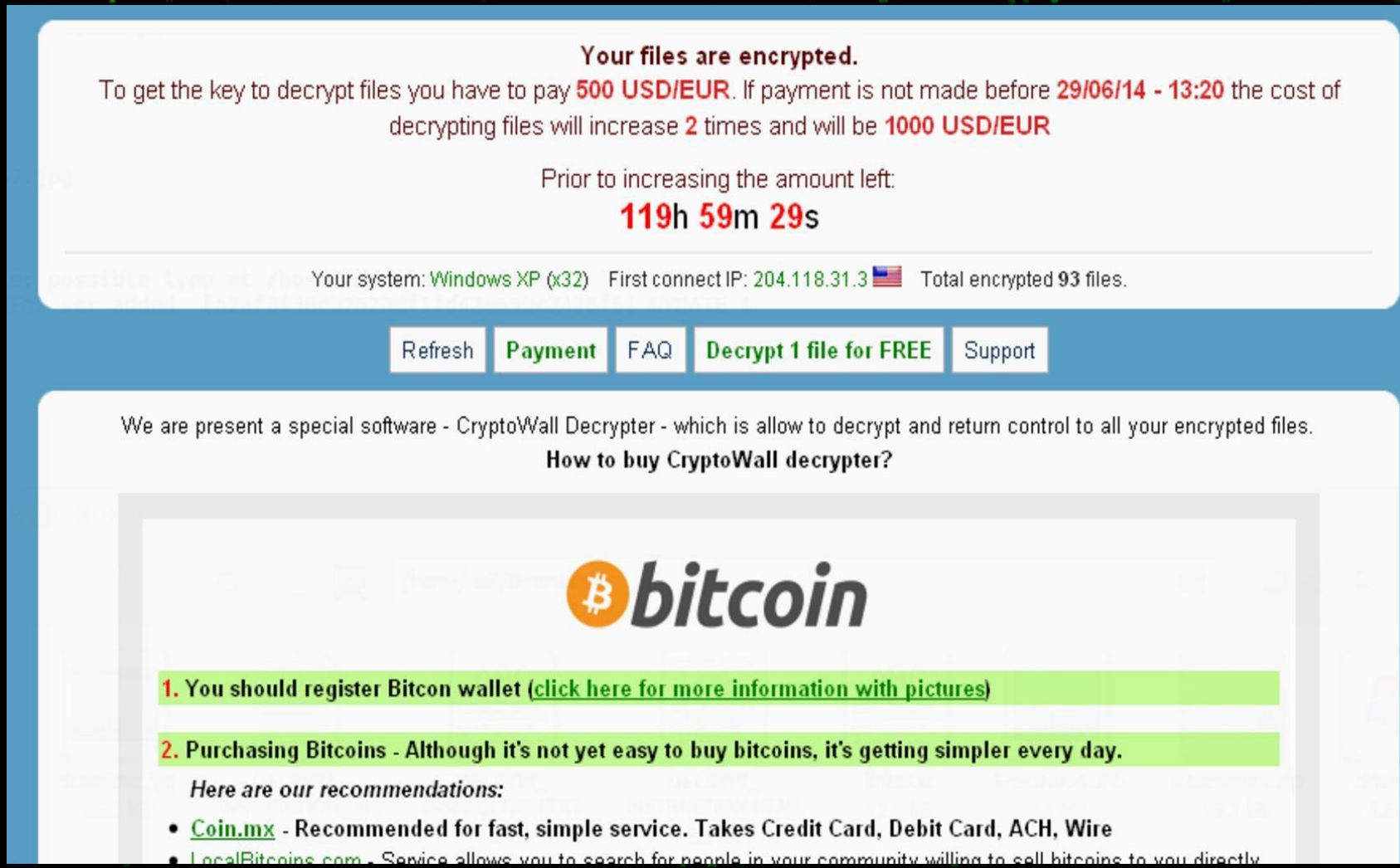


RASS: Ransomware as a Service

Underground crime ecosystem



Why Bitcoin?

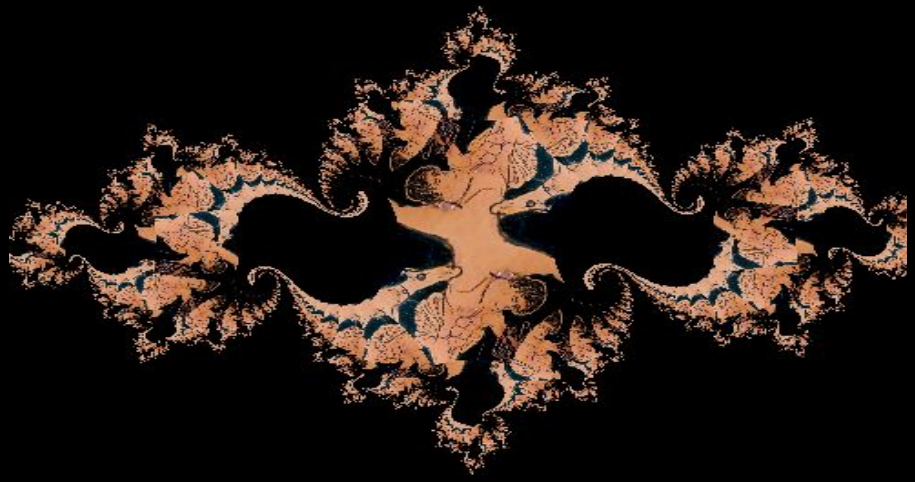

Your files are encrypted.
To get the key to decrypt files you have to pay 500 USD/EUR. If payment is not made before 29/06/14 - 13:20 the cost of decrypting files will increase 2 times and will be 1000 USD/EUR
Prior to increasing the amount left:
119h 59m 29s
Your system: Windows XP (x32) First connect IP: 204.118.31.3 Total encrypted 93 files.
Refresh
Payment
FAQ
Decrypt 1 file for FREE
Support
We are present a special software - CryptoWall Decrypter - which is allow to decrypt and return control to all your encrypted files.
How to buy CryptoWall decrypter?
bitcoin
1. You should register Bitcon wallet (click here for more information with pictures)
2. Purchasing Bitcoins - Although it's not yet easy to buy bitcoins, it's getting simpler every day.
Here are our recommendations:
Coin.mx - Recommended for fast, simple service. Takes Credit Card, Debit Card, ACH, Wire

Why Bitcoin?

- Provides a level of anonymity in transactions
- Acceptance worldwide with relative value higher than mainstream fiat currencies
- Bitcoin is not subjected to the controls and regulations of fiat currencies allowing malicious actors to exchange and transfer with practically no oversight from government or international regulatory body.
- Bitcoin allows transfers of currency value much higher than using other traditional crime related schemes such as prepaid cards or Moneypak.

TOR

- In many ransomware attacks, it is usual to observe instructions to victims on how to access the TOR network for further negotiations.
- TOR used as covert channel as well for some C2 and exploitation operations.
- The combination of the use of TOR network with covert communications using SSL, makes it even more difficult to detect infected hosts.



MACHINE LEARNING: TRIBAL KNOWLEDGE FOR CYBERSECURITY

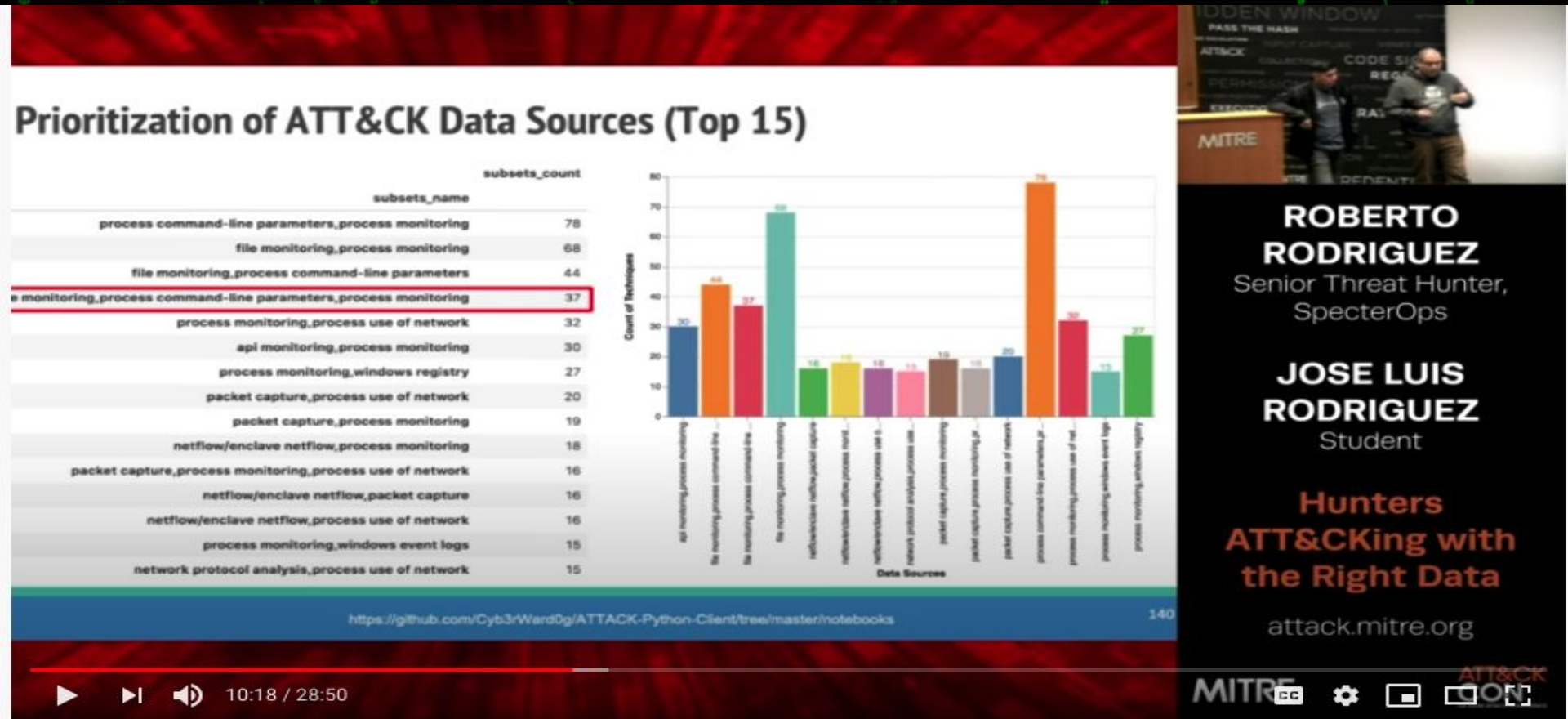
Reddit Ask me Anything: Micheal Jordan, I. Pehong Chen Distinguished Professor in the Department of Electrical Engineering and Computer Science and the Department of Statistics at the University of California, Berkeley: ***https://www.reddit.com/r/MachineLearning/comments/2fxi6v/ama_michael_i_jordan/***

“Q: If you got a billion dollars to spend on a huge (ML based) research project that you get to lead, what would you like to do?”

“A: I'd invest in human-intensive labeling processes...”

ML = Good Data

In security good data is very hard to come by it turns out. See for example this great talk example. MITRE ATT&CKcon 2018: Hunters ATT&CKing with the Data https://www.youtube.com/watch?v=QCDBjFJ_C3g



Exploit Delivery Detection: Layer 7 Data

1. **Initial Redirect From Poisoned Domain:** [29/Apr/2015:16:52:23 -0700] "Nico Rosberg" 192.168.122.177 69.162.78.253 1500 200 TCP_HIT "GET http://forbes.com/gels-contrariness-domain-punchable/1.html/548828415920276748 HTTP/1.1" "Internet Services" "low risk" "text/html" 604 142 "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.1; WOW64; Trident/4.0; SLCC2; .NET CLR 2.0.50727; .NET CLR 3.5.30729; .NET CLR 3.0.30729; Media Center PC 6.0)" "http://forbes.com/gels-contrariness-domain-punchable/1.html" "-" "0" "" "-"

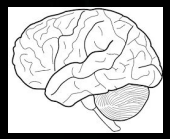
2. **Flash Exploit:** [29/Apr/2015:16:52:26 -0700] "Nico Rosberg" 192.168.122.177 69.162.78.253 1500 200 TCP_HIT "GET http://portcullisesposturen.europartsplus.org/IMvOBBZKDLqAJYIDe02t5hMMNyzBLN_q4kafJkVNqJVTnTmd HTTP/1.1" "Internet Services" "low risk" "application/x-shockwave-flash" 518 821 "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.1; WOW64; Trident/4.0; SLCC2; .NET CLR 2.0.50727; .NET CLR 3.5.30729; .NET CLR 3.0.30729; Media Center PC 6.0)" "http:///forbes.com/gels-contrariness-domain-punchable/1.html/548828415920276748" "-" "0" "" "-"

3. **Payload:** [29/Apr/2015:16:52:27 -0700] "Nico Rosberg" 192.168.122.177 69.162.78.253 1500 200 TCP_HIT "GET http://portcullisesposturen.europartsplus.org/UX7n1YkbNn8FUV6QVtEZLj-p-gLvRKlWEWmz3r7Ug8suRiY_ HTTP/1.1" "Internet Services" "low risk" "application/octet-stream" 136 915 "" "" "-" "0" "" "-"

4. **Command and Control:** [29/Apr/2015:16:52:33 -0700] "Nico Rosberg" 192.168.122.177 104.28.28.165 1500 200 TCP_HIT "GET http://dpckd2ftmf7lelsa.jjeyd2u37an30.com/tsdfewr2.php?U3ViamVj49MCZpc182ND0xJmlwPTIxMy4yMjkuODcuMjgmZXhlX3R5cGU9MQ== HTTP/1.1" "Internet Services" "low risk" "text/html; charset=UTF-8" 566 5 "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; .NET CLR 2.0.50727; .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729)" "" "-" "0" "" "-"

Learning = Compression?

There is a duality between learning and compression



Primary Key	Time	UserID	Count
Row 1	…	…	…
Row 2	…	…	…
Row 3	…	…	…
…	…	…	…
Row N	…	…	…

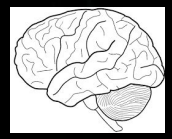
Learning Machine: R Linear Model

C1	C2	C3	C4	C5

Learning = Compression?

Example of Linear Regression in R

```
> mtcars
                     mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4           21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710          22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout   18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant             18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360          14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D           24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230            22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280            19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C           17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
Honda Civic         30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
```

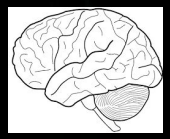


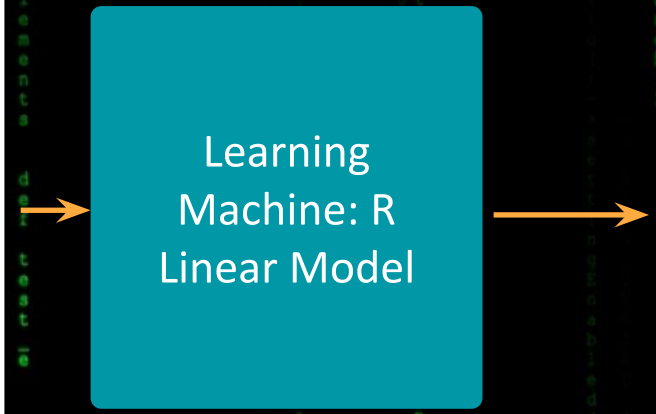
Learning Machine: R Linear Model

Learning = Compression?

Train a model to predict mpg as a function of car weight, number of cylinders and displacement

```
> mtcars
                     mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4           21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710          22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout   18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant             18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360          14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D           24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230            22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280            19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C           17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
Honda Civic         30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
```



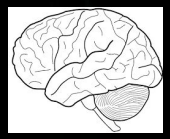


Learning = Compression?

Train a model to predict mpg as a function of car weight, number of cylinders and displacement

```
> mtcars
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1



Learning Machine: R Linear Model

```
> mfit = lm(mpg ~ wt + disp + cyl, data=mtcars)
```

Learning = Compression?

The overall input data is reduced in a “compressed form” to use in future predictions

```
> mtcars
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1

Learning Machine: R Linear Model

```
> (41.107678, -3.635677, 0.007473,-1.784944)
```

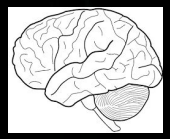
```
> mfit = lm(mpg ~ wt + disp + cyl, data=mtcars)
```

Learning = Compression?

This process is extremely brittle in terms of modeling a changing signal or an adversary that changes patterns over time

```
> mtcars
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1



Learning Machine: R Linear Model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.107678   2.842426  14.462 1.62e-14 ***
wt          -3.635677   1.040138  -3.495  0.00160 **
disp         0.007473   0.011845   0.631  0.53322
cyl         -1.784944   0.607110  -2.940  0.00651 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.595 on 28 degrees of freedom
Multiple R-squared:  0.8326,    Adjusted R-squared:  0.8147
F-statistic: 46.42 on 3 and 28 DF,  p-value: 5.399e-11

> y = c_1 x_1 + c_2 x_2 + c_3 x_3 + b
```

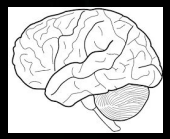
```
> mfit = lm(mpg ~ wt + disp + cyl, data=mtcars)
```

Learning = Compression?

The simple linear model gives us output that separates the **Signal** from the **Noise** (this is not always possible with a model)

```
> mtcars
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1



Learning Machine: R Linear Model

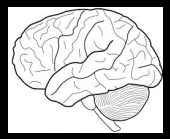
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.107678   2.842426  14.462 1.62e-14 ***
wt          -3.635677   1.040138  -3.495  0.00160 **
disp         0.007473   0.011845   0.631  0.53322
cyl         -1.784944   0.607110  -2.940  0.00651 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.595 on 28 degrees of freedom
Multiple R-squared:  0.8326,    Adjusted R-squared:  0.8147
F-statistic: 46.42 on 3 and 28 DF,  p-value: 5.399e-11

> y = c_1 x_1 + c_2 x_2 + c_3 x_3 + b
```

```
> mfit = lm(mpg ~ wt + disp + cyl, data=mtcars)
```

Learning = Compression?

Real example of random forest trained on C2 traffic

Domain Name	TotalCnt	RiskFactor	AGD	SessionTime	RefEntropy	NullUa
europartsplus.org	144	6.05	1	1	0	0
jjeyd2u37an30.com	6192	5.05	0	1	0	0
cdn4s.steelhousemedia.com	107	3	0	0	0	0
log.tagcade.com	111	2	0	1	0	0
go.vidprocess.com	170	2	0	0	0	0
statse.webtrendslive.com	310	2	0	1	0	0
cdn4s.steelhousemedia.com	107	1	0	0	0	0
log.tagcade.com	111	1	0	1	0	0

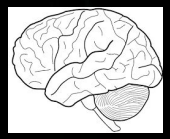


Learning Machine: MLLib Random Forest

```
Random Forest:
TreeEnsembleModel classifier with 6 trees

 Tree 0:
   If (feature 7 <= 0.0)
    If (feature 11 <= 0.0)
     If (feature 80 <= -51.45518112)
      Predict: 1.0
     Else (feature 80 > -51.45518112)
      If (feature 76 <= 7.0)
       If (feature 60 <= -48.02338409)
        If (feature 81 <= 0.0)
         If (feature 84 <= 2.0)
          Predict: 1.0
         Else (feature 84 > 2.0)
          Predict: 0.0
        Else (feature 81 > 0.0)
         Predict: 0.0
       Else (feature 60 > -48.02338409)
```

Learning = Compression?

Real example of random forest trained on C2 traffic

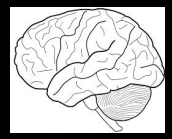


Domain Name	TotalCnt	RiskFactor	AGD	SessionTime	RefEntropy	NullUa
europartsplus.org	144	6.05	1	1	0	0
jjeyd2u37an30.com	6192	5.05	0	1	0	0
cdn4s.steelhousemedia.com	107	3	0	0	0	0
log.tagcade.com	111	2	0	1	0	0
go.vidprocess.com	170	2	0	0	0	0
statse.webtrendslive.com	310	2	0	1	0	0
cdn4s.steelhousemedia.com	107	1	0	0	0	0
log.tagcade.com	111	1	0	1	0	0

Learning Machine: MLLib Random Forest

Learning = Compression?

We really “learn” a function we can call in batch or real time



Domain Name	TotalCnt	RiskFactor	AGD	SessionTime	RefEntropy	NullUa
europartsplus.org	144	6.05	1	1	0	0
jjeyd2u37an30.com	6192	5.05	0	1	0	0
cdn4s.steelhousemedia.com	107	3	0	0	0	0
log.tagcade.com	111	2	0	1	0	0
go.vidprocess.com	170	2	0	0	0	0
statse.webtrendslive.com	310	2	0	1	0	0
cdn4s.steelhousemedia.com	107	1	0	0	0	0
log.tagcade.com	111	1	0	1	0	0

```
import org.apache.spark.mllib.tree.model.RandomForestModel
```

```
val rfModel: RandomForestModel = RandomForest.trainClassifier(trainData.union(cvData),
  numClasses, categoricalFeaturesInfo, loopNumTrees, featureSubsetStrategy, loopImpurity, loopDepth, loopBins)
```

```
//decorate the (domain,feature vector) pair with predicted output of the model
//use pair RRD's to recover the full data we need for anomaly generation by a inner join
val predictedOutput = vectorizeFeatures.map(x => (x._1, randomForestModel.predict(x._2)))
```

Key to ML: Label Your Analysis

Label output of every investigation in a consistent manner!!!

Domain Name	TotalCnt	RiskFactor	AGD	SessionTime	RefEntropy	NullUa
yyfaimjmocdu.com	144	6.05	1	1	0	0
jjeyd2u37an30.com	6192	5.05	0	1	0	0
cdn4s.steelhousemedia.com	107	3	0	0	0	0
log.tagcade.com	111	2	0	1	0	0
go.vidprocess.com	170	2	0	0	0	0
statse.webtrendslive.com	310	2	0	1	0	0
cdn4s.steelhousemedia.com	107	1	0	0	0	0
log.tagcade.com	111	1	0	1	0	0

Key to ML: Label Your Analysis

This is how the algorithms will "learn" from human expertise and help support a common security workflow

Domain Name	TotalCnt	RiskFactor	AGD	SessionTime	RefEntropy	NullUa	Outcome
yyfaimjmocdu.com	144	6.05	1	1	0	0	Malicious
jjeyd2u37an30.com	6192	5.05	0	1	0	0	Malicious
cdn4s.steelhousemedia.com	107	3	0	0	0	0	Benign
log.tagcade.com	111	2	0	1	0	0	Benign
go.vidprocess.com	170	2	0	0	0	0	Benign
statse.webtrendslive.com	310	2	0	1	0	0	Benign
cdn4s.steelhousemedia.com	107	1	0	0	0	0	Benign
log.tagcade.com	111	1	0	1	0	0	Benign

Human Expertise is manually encoded into a format computers understand: Sometimes this process is called Labeling or "Truth-ing" the data

MODELING RANSOMWARE DELIVERY BEHAVIORS

Ransomware Delivery and The Modern Threat Surface

- Stages of A Typical Enterprise Campaign
- Phishing Campaign/Watering Hole Established
- **Exploitation (Focus Of Aktaion)**
- File System Modification (Too Late – Much Of Current Research Focuses Here)
- Ransom Note (DNS IPS Monitoring Can be Helpful Here)

Post Exploit Resources/Research

- CryptoLock (and Drop It): Stopping Ransomware Attacks on User Data
  - http://www.cise.ufl.edu/~traynor/papers/scaife-icdcs16.pdf
- Ransomware Overview
  - This initial list has been composed by Mosh @nyxbone and transformed into this Google Docs format by @cyb3rops (Ransomware Overview is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.)
  - https://docs.google.com/spreadsheets/d/1TWS238xacAto-fLKh1n5uTsdijWdCEsGIM0Y0Hvmc5g/pubhtml
- Something Called Cryptowall prevention kit (now is pay to play it seems)

Exploit Research

Detecting malicious HTTP redirections using trees of user browsing activity, Hesham Mekky et. Al.

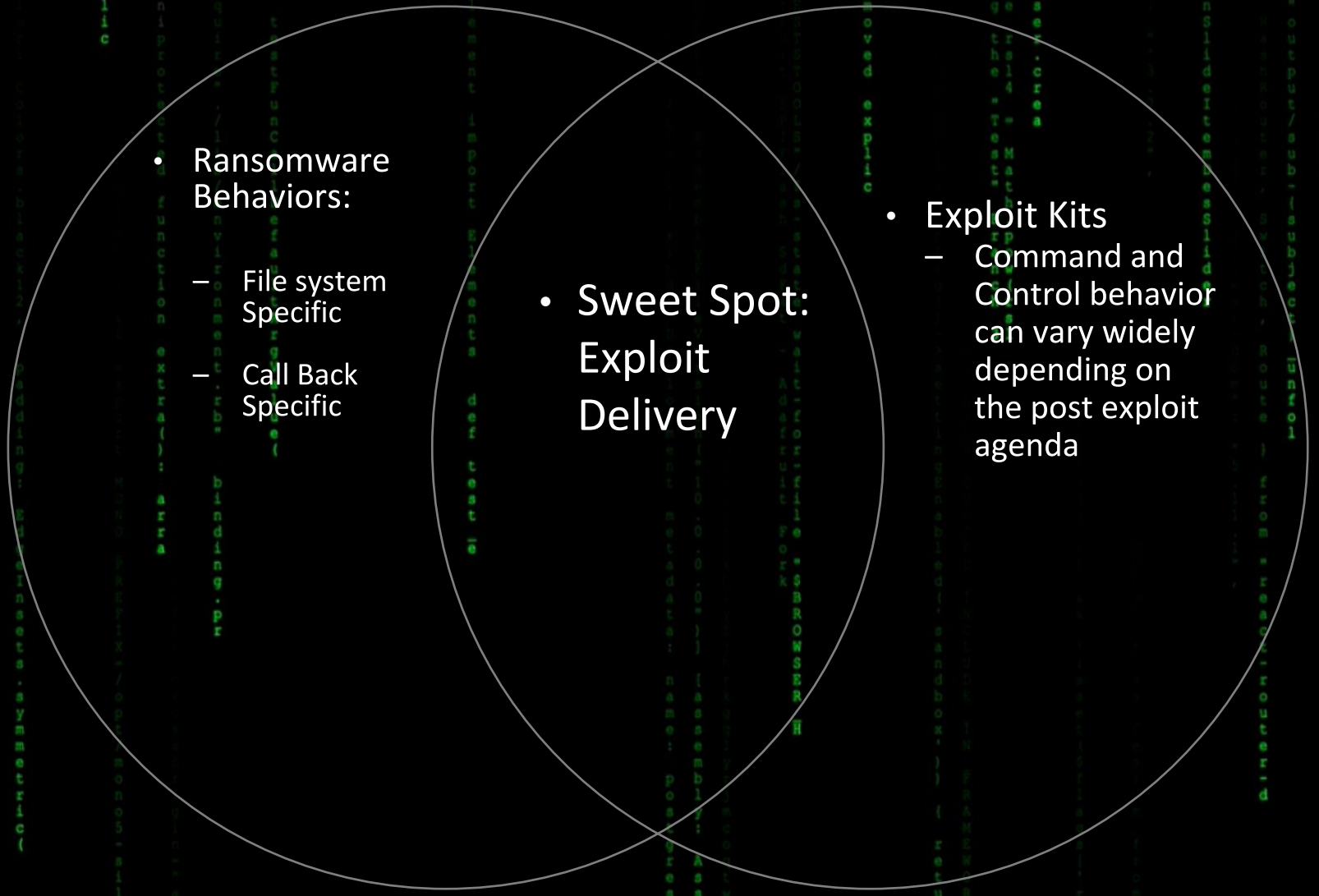
> "...We build per-user chains from passively collected traffic and extract novel statistical features from them, which capture inherent characteristics from malicious redirection cases. Then, we apply a supervised decision tree classifier to identify malicious chains. Using a large ISP dataset, with more than 15K clients, we demonstrate that our methodology is very effective in accurately identifying malicious chains, with recall and precision values over 90% and up to 98%"

IEEE INFOCOM 2014 - IEEE Conference on Computer Communications

Our Original Exploit Data

- 386 Labeled Exploit chain examples from Contagio (PCAP extracts into a generic proxy format)
- **CRIMEb** from DeepEnd Research (DeepEnd Research)
  - https://www.dropbox.com/sh/7fo4efxhpenexqp/AADHnRKtL6qdzCdRlPmJpS8Aa/CRIME?dl=0

Plan of Attack



ML For Security = Labeled Malicious or Benign Events

Samples located at

https://github.com/jzadeh/aktaion2/data

https://github.com/jzadeh/aktaion/data (deprecated version)

Ransomware Samples

Small amount of mixed call back/file system level indicators

Exploit Samples

348 PCAPs converted to a Proxy Format

Thanks to the hard work of Contagio and Mila Parkour

http://contagiodump.blogspot.com/

Benign Bro Traffic Samples

Multiple independent user sessions

Expressing Microbehaviors in Code: Aktaion V2

https://github.com/jzadeh/aktaion2/tree/blackhat_eu_2017
(careful not merged into master until EOW for now use the following for access)

git clone https://github.com/jzadeh/aktaion2.git

git checkout origin blackhat_eu_2017

Migrated Core Logic and Modules to Python3 Implementation

aktaion2/python

./run_ml.sh : builds machine learning ready data sets given input directories and window sizes and config parameters

./run_aktaion2.sh : use to reproduce what you see in the arsenal demo and score individual files

aktaion2/python/research_dev/random_forest:

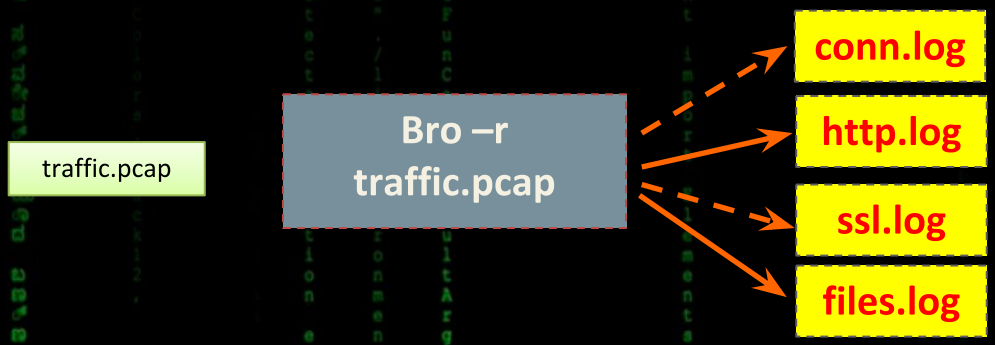
micro_beahvior_core_logic.py : key abstraction used for building important characteristics about attack/benign in a form useful for AI to learn from

rf_v2.py, svm.py, calibration.py : main machine learning based tools using Random Forest, Support Vector Machine and Classifier performance analysis

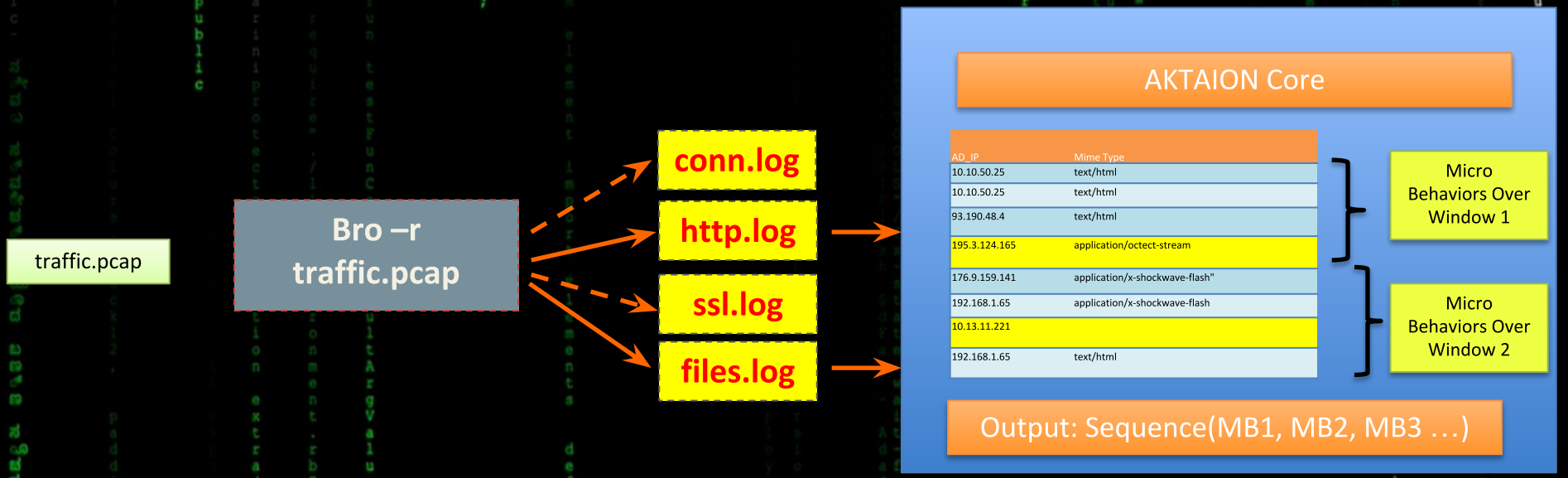
Workflow

1. Take PCAPs of known (labeled) exploits and known (labeled) benign behavior and convert them to bro format
2. Convert each Bro log to a sequence of micro behaviors (machine learning input)
3. Compare the sequence of micro behaviors to a set of known benign/malicious samples using a Random Forest Classifier (http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/RandomForest.html)
4. Derive a list of indicators from any log predicted as malicious
5. Pass the list of IOCs (JSON) to a GPO generation script (https://github.com/jzadeh/Aktaion/tree/master/python)

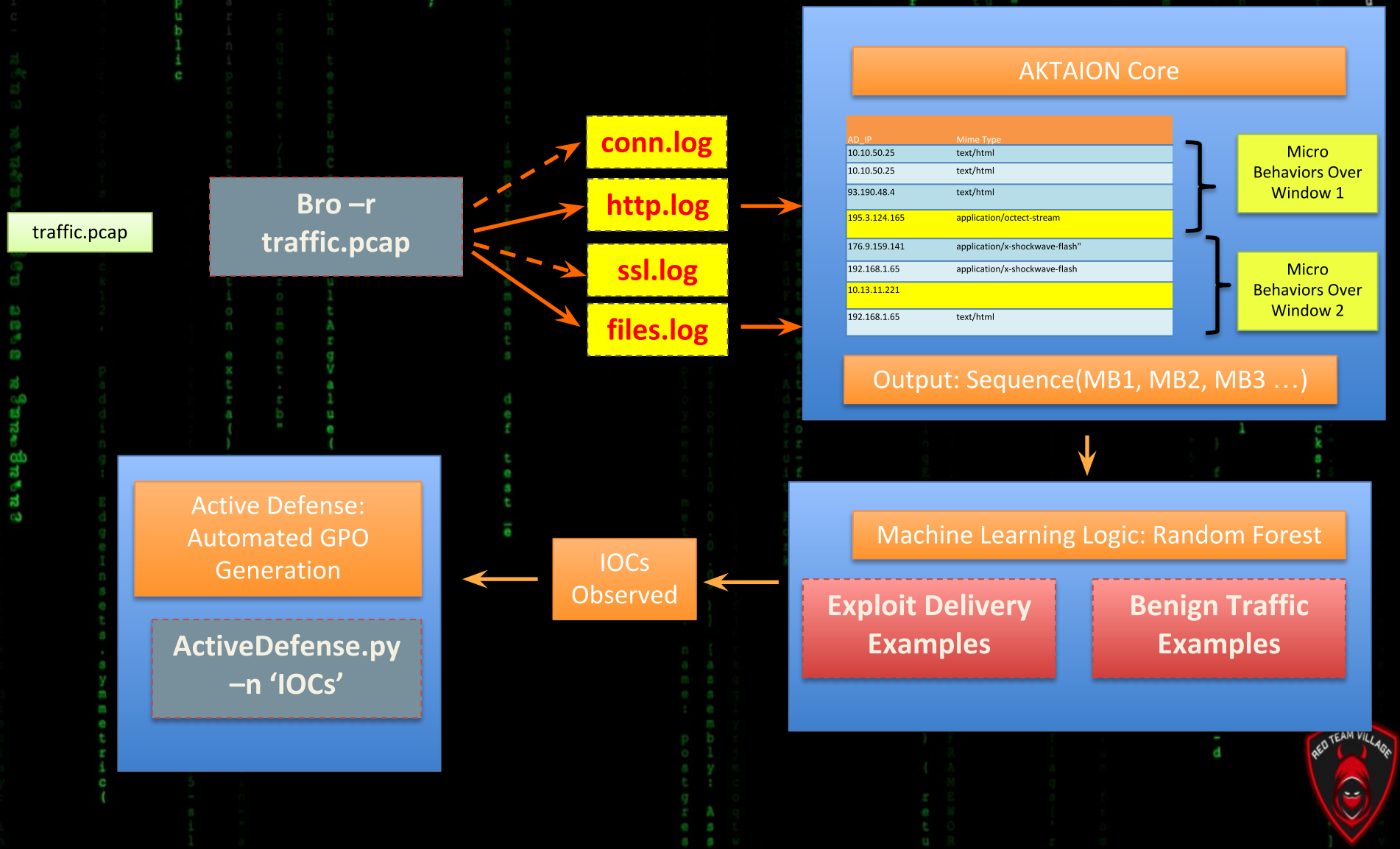
Aktaion Logical Workflow



Aktaion Logical Workflow



Aktaion Logical Workflow



Finding the Initial Exploit

1. **Initial Redirect From Poisoned Domain:** [29/Apr/2015:16:52:23 -0700] "Nico Rosberg" 192.168.122.177 69.162.78.253 1500 200 TCP_HIT "GET http://forbes.com/gels-contrariness-domain-punchable/1.html/548828415920276748 HTTP/1.1" "Internet Services" "low risk" "text/html" 604 142 "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.1; WOW64; Trident/4.0; SLCC2; .NET CLR 2.0.50727; .NET CLR 3.5.30729; .NET CLR 3.0.30729; Media Center PC 6.0)" "http://forbes.com/gels-contrariness-domain-punchable/1.html" "-" "0" "" "-"

Finding the Initial Exploit

1. **Initial Redirect From Poisoned Domain:** [29/Apr/2015:16:52:23 -0700] "Nico Rosberg" 192.168.122.177 69.162.78.253 1500 200 TCP_HIT "GET http://forbes.com/gels-contrariness-domain-punchable/1.html/548828415920276748 HTTP/1.1" "Internet Services" "low risk" "text/html" 604 142 "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.1; WOW64; Trident/4.0; SLCC2; .NET CLR 2.0.50727; .NET CLR 3.5.30729; .NET CLR 3.0.30729; Media Center PC 6.0)" "http://forbes.com/gels-contrariness-domain-punchable/1.html" "-" "0" "" "-"

2. **Flash Exploit:** [29/Apr/2015:16:52:26 -0700] "Nico Rosberg" 192.168.122.177 69.162.78.253 1500 200 TCP_HIT "GET http://portcullisesposturen.europartsplus.org/IMvOBBZKDLqAJYIDe02t5hMMNyzBLN_q4kafJkVNqJVTnTmd HTTP/1.1" "Internet Services" "low risk" "application/x-shockwave-flash" 518 821 "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.1; WOW64; Trident/4.0; SLCC2; .NET CLR 2.0.50727; .NET CLR 3.5.30729; .NET CLR 3.0.30729; Media Center PC 6.0)" "http:///forbes.com/gels-contrariness-domain-punchable/1.html/548828415920276748" "-" "0" "" "-"

Finding the Initial Exploit

1. **Initial Redirect From Poisoned Domain:** [29/Apr/2015:16:52:23 -0700] "Nico Rosberg" 192.168.122.177 69.162.78.253 1500 200 TCP_HIT "GET http://forbes.com/gels-contrariness-domain-punchable/1.html/548828415920276748 HTTP/1.1" "Internet Services" "low risk" "text/html" 604 142 "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.1; WOW64; Trident/4.0; SLCC2; .NET CLR 2.0.50727; .NET CLR 3.5.30729; .NET CLR 3.0.30729; Media Center PC 6.0)" "http://forbes.com/gels-contrariness-domain-punchable/1.html" "-" "0" "" "-"

2. **Flash Exploit:** [29/Apr/2015:16:52:26 -0700] "Nico Rosberg" 192.168.122.177 69.162.78.253 1500 200 TCP_HIT "GET http://portcullisesposturen.europartsplus.org/IMvOBBZKDLqAJYIDe02t5hMMNyzBLN_q4kafJkVNqJVTnTmd HTTP/1.1" "Internet Services" "low risk" "application/x-shockwave-flash" 518 821 "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.1; WOW64; Trident/4.0; SLCC2; .NET CLR 2.0.50727; .NET CLR 3.5.30729; .NET CLR 3.0.30729; Media Center PC 6.0)" "http:///forbes.com/gels-contrariness-domain-punchable/1.html/548828415920276748" "-" "0" "" "-"

3. **Payload:** [29/Apr/2015:16:52:27 -0700] "Nico Rosberg" 192.168.122.177 69.162.78.253 1500 200 TCP_HIT "GET http://portcullisesposturen.europartsplus.org/UX7n1YkbNn8FUV6QVtEZLj-p-gLvRKlWEWmz3r7Ug8suRiY_ HTTP/1.1" "Internet Services" "low risk" "application/octet-stream" 136 915 "" "" "-" "0" "" "-"

Finding the Initial Exploit

1. **Initial Redirect From Poisoned Domain:** [29/Apr/2015:16:52:23 -0700] "Nico Rosberg" 192.168.122.177 69.162.78.253 1500 200 TCP_HIT "GET http://forbes.com/gels-contrariness-domain-punchable/1.html/548828415920276748 HTTP/1.1" "Internet Services" "low risk" "text/html" 604 142 "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.1; WOW64; Trident/4.0; SLCC2; .NET CLR 2.0.50727; .NET CLR 3.5.30729; .NET CLR 3.0.30729; Media Center PC 6.0)" "http://forbes.com/gels-contrariness-domain-punchable/1.html" "-" "0" "" "-"

2. **Flash Exploit:** [29/Apr/2015:16:52:26 -0700] "Nico Rosberg" 192.168.122.177 69.162.78.253 1500 200 TCP_HIT "GET http://portcullisesposturen.europartsplus.org/IMvOBBZKDLqAJYIDe02t5hMMNyzBLN_q4kafJkVNqJVTnTmd HTTP/1.1" "Internet Services" "low risk" "application/x-shockwave-flash" 518 821 "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.1; WOW64; Trident/4.0; SLCC2; .NET CLR 2.0.50727; .NET CLR 3.5.30729; .NET CLR 3.0.30729; Media Center PC 6.0)" "http:///forbes.com/gels-contrariness-domain-punchable/1.html/548828415920276748" "-" "0" "" "-"

3. **Payload:** [29/Apr/2015:16:52:27 -0700] "Nico Rosberg" 192.168.122.177 69.162.78.253 1500 200 TCP_HIT "GET http://portcullisesposturen.europartsplus.org/UX7n1YkbNn8FUV6QVtEZLj-p-gLvRKlWEWmz3r7Ug8suRiY_ HTTP/1.1" "Internet Services" "low risk" "application/octet-stream" 136 915 "" "" "-" "0" "" "-"

4. **Command and Control:** [29/Apr/2015:16:52:33 -0700] "Nico Rosberg" 192.168.122.177 104.28.28.165 1500 200 TCP_HIT "GET http://dpckd2ftmf7lelsa.jjeyd2u37an30.com/tsdfewr2.php?U3ViamVj49MCZpc182ND0xJmlwPTIxMy4yMjkuODcuMjgmZXhlX3R5cGU9MQ== HTTP/1.1" "Internet Services" "low risk" "text/html; charset=UTF-8" 566 5 "Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; .NET CLR 2.0.50727; .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729)" "" "-" "0" "" "-"

Expressing Microbehaviors in Code: Aktaion V1

```scala
package com.aktaion.ml.behaviors

import com.aktaion.DebugLoggingLogic

/**
  * Simple abstraction for numeric data associated to
  * some statistic we are computing about a behavior for example
  * the number of unique MIME types served by a single domain
  */
trait MicroBehaviorNumericLike {...}

/**
  * Abstraction for non numerical data associated to a behavior
  *
  */
trait MicroBehaviorCategoricalLike {...}

/**
  * Micro Behavior: Main abstraction for individual unit of behavior
  * that we want to model in the upstream workflow
  *
  */
case class MicroBehaviorData(behaviorName: String,
                             behaviorDescription: String,
                             var numData: Double = 0.0,
                             var catData: String = "")
  extends MicroBehaviorNumericLike with MicroBehaviorCategoricalLike {
  def valueToCsv = behaviorName + "," + numData.toString + "," + catData
}


/**
  * Represents a single set of behaviors for some input data
  * this can be a single log line or multiple log lines
  * depending on how we implement the upstream logic
  */
trait MicroBehaviorSet extends DebugLoggingLogic {...}

/**
  * In the current implementation we score a window of time
  * and derive a single set of values for each behavior we are modeling
  *
  * @param behaviorVector the computed values for each behavior in that window
  * @param windowSize the window size (number of log lines) the computation took place over
  */
case class MicroBehaviorWindow(behaviorVector: List[MicroBehaviorData], windowSize: Int) extends MicroBehaviorSet
```

Expressing Microbehaviors in Code

TTP = Microbehavior

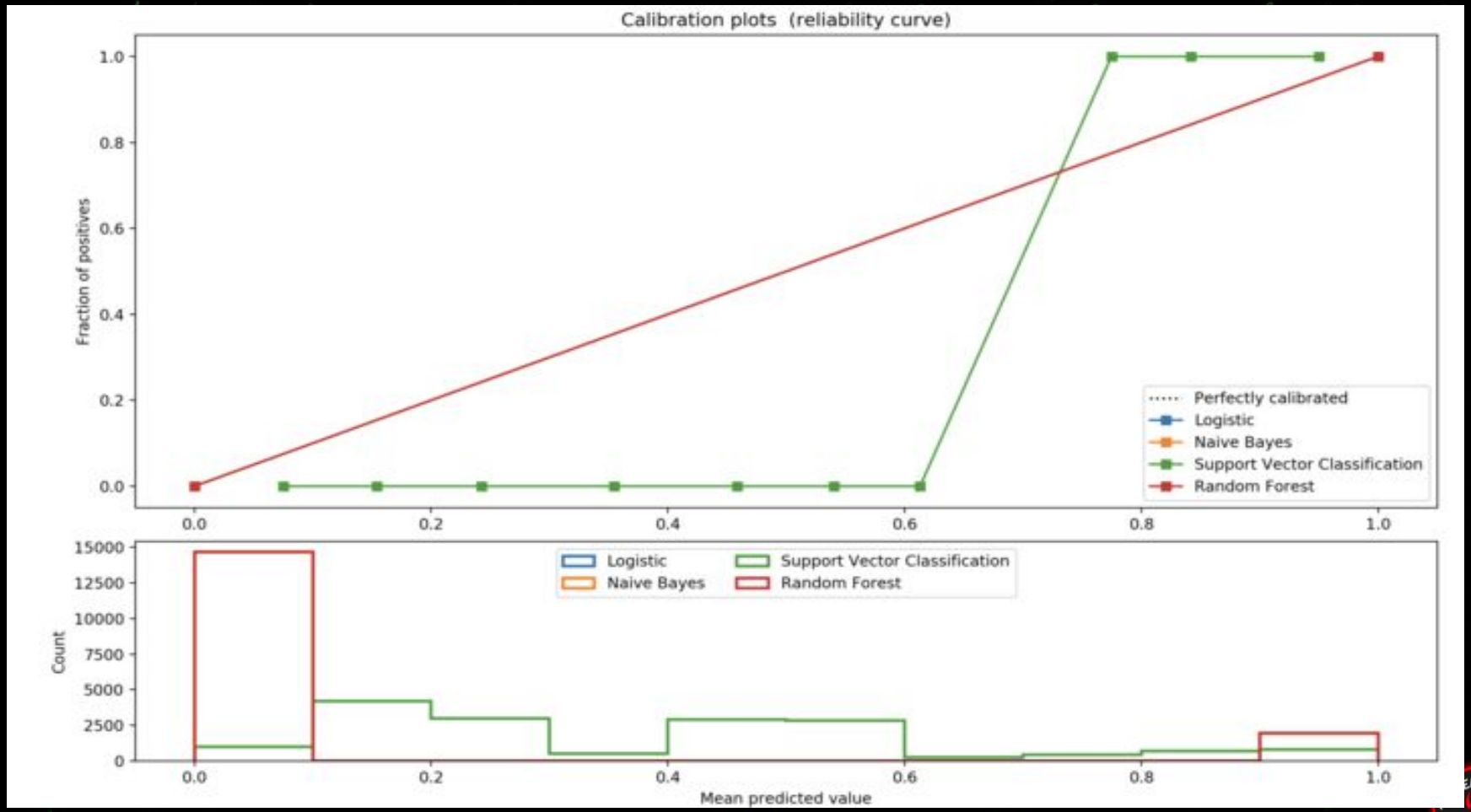
We want a flexible enough framework for blending both signatures on a pattern match with more general observations like

- ‘Interesting’ sequences of Mime types of length < 5
- Referrer Sequences with Small inter-arrival times
- Mime type distribution in a time window across a single source IP
- Comparison of probability of observed (MIME Type, Extension) to overall enterprise population: example (octect-stream, .mp3) is extremely rare

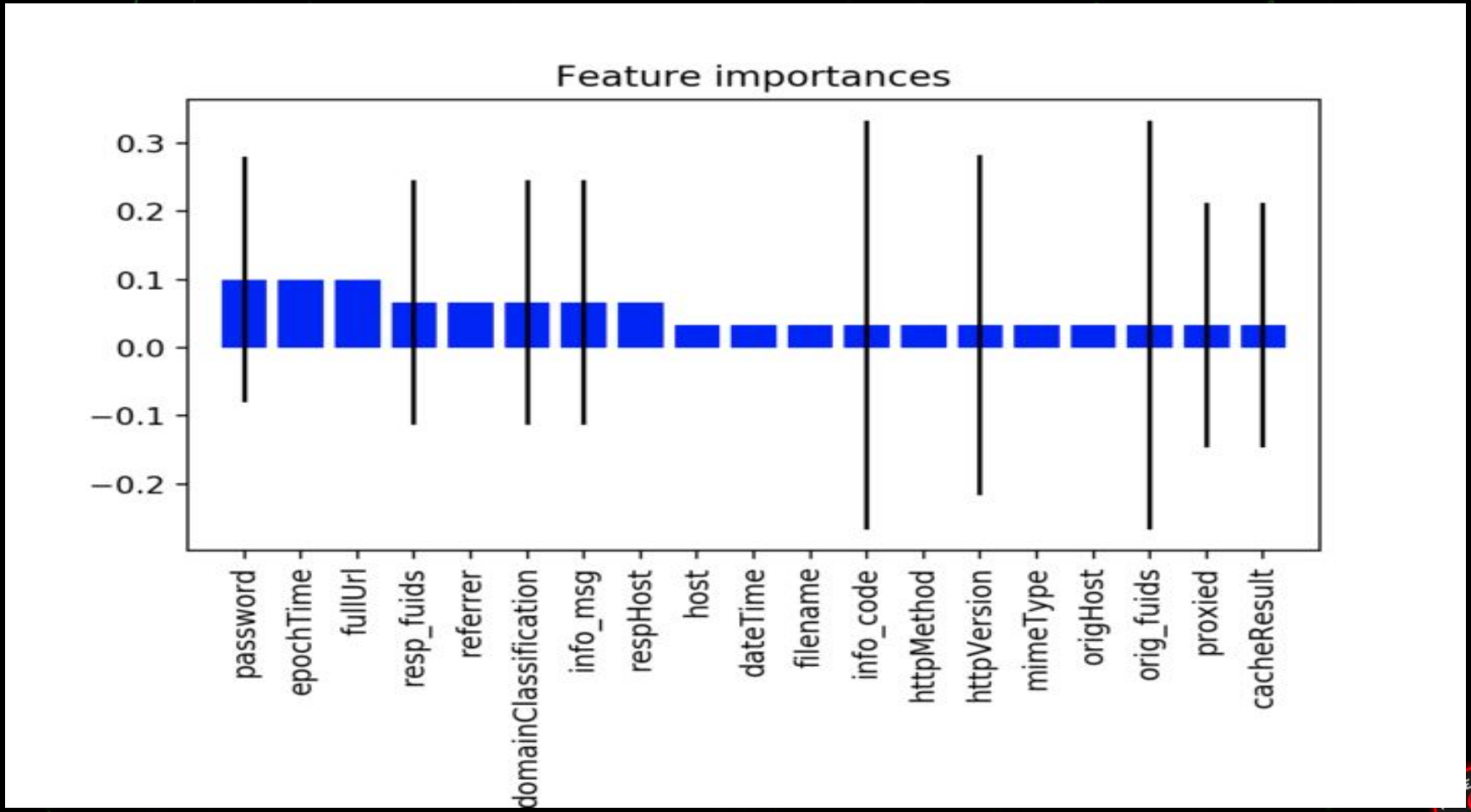
Expressing Microbehaviors in Code: Aktaion V1

```
package com.aktaion.ml.behaviors

import scala.util.matching.Regex

class ExploitationUriBehaviors extends MicroBehaviorSet {
  val uriMaxPathDepth = MicroBehaviorData("MaxPathDepth", "Maximum path length of all URI's observed")
  val uriMinPathDepth = MicroBehaviorData("MinPathDepth", "Minimum path length of all URI's observed")
  val uriMaxLength = MicroBehaviorData("MaxUriLength", "Maximum length of all URI's observed")
  val uriMinLength = MicroBehaviorData("MinUriLength", "Minimum length of all URI's observed")
  val uriDistinct = MicroBehaviorData("UniqNumberOfUri", "Unique count of URI's in window")
  val uriMaxEntropy = MicroBehaviorData("UriMaxEntropy", "Maximum entropy of all URI's observed")
  val uriMinEntropy = MicroBehaviorData("UriMinEntropy", "Minimum entropy of all URI's observed")
  val base64Match = MicroBehaviorData("UriBase64", "URI's obseved contain large number of base 64 strings")
  val percentEncondingMatch = MicroBehaviorData("UriBase64", "URI's obseved contain large number of percent encoded strings")

  val behaviorVector = List(...)

  val encodingBase64 = new Regex( """^(?:[A-Za-z0-9+/]{4})*(?:[A-Za-z0-9+/]{2}==|[A-Za-z0-9+/]{3}=)?$""")
  val encodingBase64modified = new Regex( """^([A-Za-z0-9+/]{4})*([A-Za-z0-9+/]{4}|[A-Za-z0-9+/]{3}=|[A-Za-z0-9+/]{2}==)$""")
  val encodingPercentSimple = new Regex( """%[A-Fa-f0-9]{2}""")
}

class ExploitationTimingBehaviors extends MicroBehaviorSet {
  val maxTimeIntervalA = MicroBehaviorData("MaxTimeIntervalA", "Difference in timestamp between event 1 and event 2")
  val maxTimeIntervalB = MicroBehaviorData("MaxTimeIntervalB", "Difference in timestamp between event 2 and event 3")
  val maxTimeIntervalC = MicroBehaviorData("MaxTimeIntervalC", "Difference in timestamp between event 3 and event 4")
  val maxTimeIntervalD = MicroBehaviorData("MaxTimeIntervalD", "Difference in timestamp between event 4 and event 5")

  val minTimeIntervalA = MicroBehaviorData("MinTimeIntervalA", "Difference in timestamp between event 1 and event 2")
  val minTimeIntervalB = MicroBehaviorData("MinTimeIntervalB", "Difference in timestamp between event 2 and event 3")
  val minTimeIntervalC = MicroBehaviorData("MinTimeIntervalC", "Difference in timestamp between event 3 and event 4")
  val minTimeIntervalD = MicroBehaviorData("MinTimeIntervalD", "Difference in timestamp between event 4 and event 5")

  val intervalLength = MicroBehaviorData("TimeLengthInWindow", "Difference last and first time in the window")

  val ratioOfDeltasA = MicroBehaviorData("DeltaRatioA", "Count of time delta < 1 second over window size")
  val ratioOfDeltasB = MicroBehaviorData("DeltaRatioB", "Count of time delta < 5 second over window size")
  val ratioOfDeltasC = MicroBehaviorData("DeltaRatioC", "Count of time delta < 10 second over window size")
  val ratioOfDeltasD = MicroBehaviorData("DeltaRatioD", "Count of time delta < 20 second over window size")
  val ratioOfDeltasE = MicroBehaviorData("DeltaRatioE", "Count of time delta >= 100 second over window size")

  val behaviorVector = List(...)
}
```

Classifier Trained On Raw Exploit/Benign HTTP Samples



Classifier Trained On Raw Exploit/Benign HTTP Samples: Random Forest Variable Importance Plot



Ranking the best ML

Nearest Neighbors 0.925
Linear SVM 0.4
RBF SVM 0.875
Gaussian Process 0.9
Decision Tree 0.775
Random Forest 0.825
Neural Net 0.775
AdaBoost 0.825
Naive Bayes 0.7
QDA 0.725

TTPs and Machine Learning

David Bianco
http://detect-respond.blogspot.com/2013/03/the-pyramid-of-pain.html

- Pyramid of Pain Paraphrase "1st level IOC's can be modified easily by an adversary (IP address, File Hash) whereas higher level TTPs (Techniques, Tactics, Procedure) are expressions of an adversaries behavior at a higher level and are harder for the adversary to modify (the attackers training to use a specific tool, exfil sequence etc..)

TTP We Focus on Primarily: Initial Exploit Delivery

We use this idea to reduce the problem of detecting Ransomware in the environment to an early stage in the life cycle of the attacker

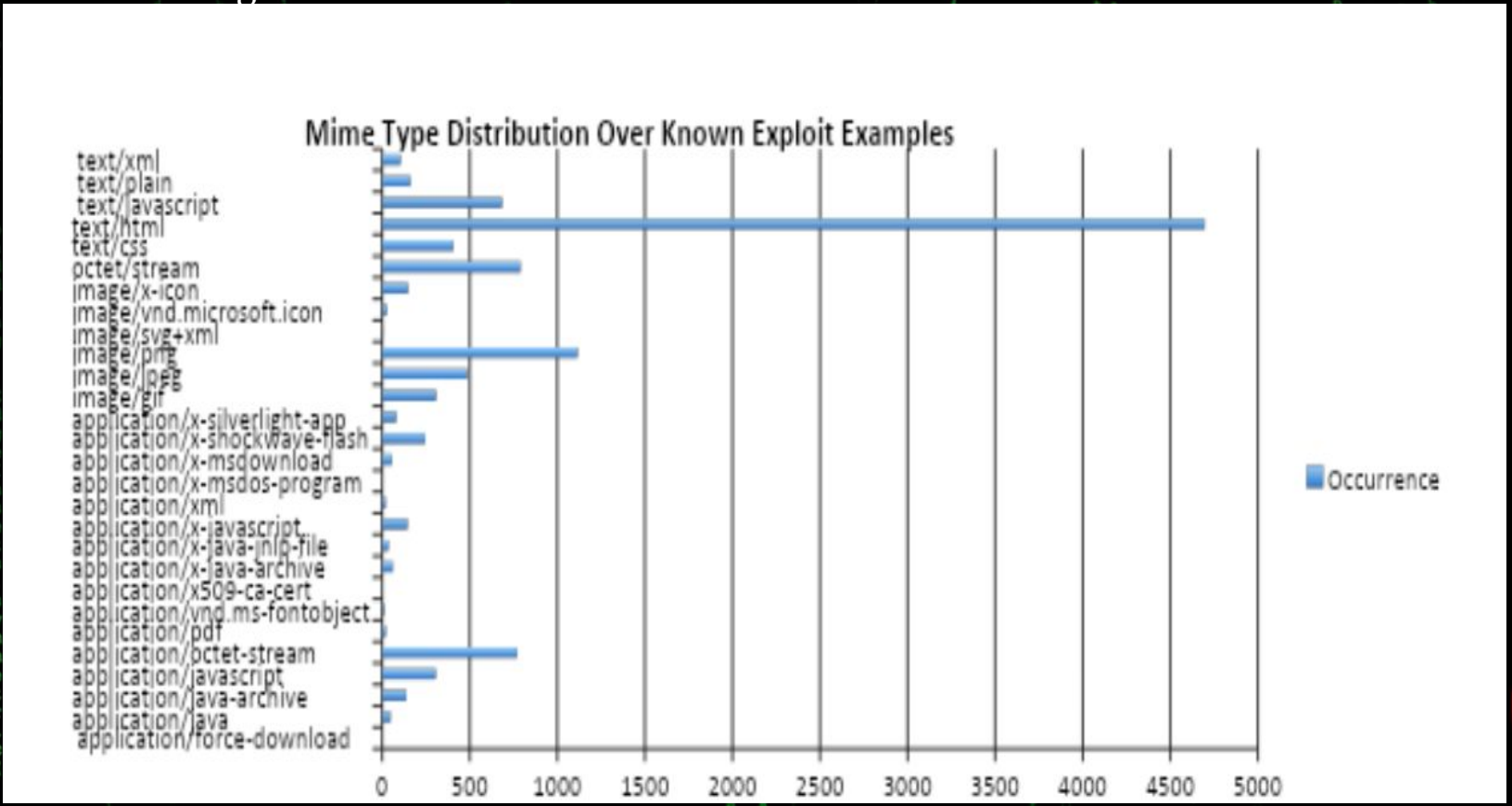
- Lots of re

More Micro Behavior Mining

- Payload delivery. Focused on traffic to malicious sites and the related indicators when malicious code is served. Including things such as URI entropy, redirects, domain generated by algorithms (DGAs), types and sequences of MIME content presented to victim during payload delivery. A reputation feed from ransomware domains and IP was used as ground truth (http://ransomwaretracker.abuse.ch/), as well as PCAP samples from sites such as http://www.malware-traffic-analysis.net/ , http://contagiodump.blogspot.com/ , and data collected during field research.
- Call backs (Phone home) patterns, including user agent , URI strings, HTTP “GET” or “POST” requests, DNS queries, URI strings, frequency of call backs, periodicity of connections.
- Covert Channel indicators, such as non HTTP traffic (HTTPS), and non DNS traffic present during such communications.

Key Micro Behaviors

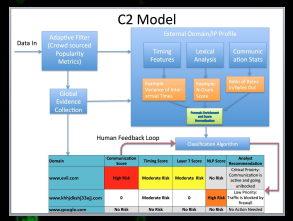
Mime Type Sequences Occurring in a Single Sequence of Small Time length from a Single Host

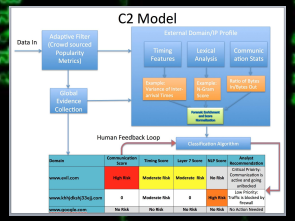


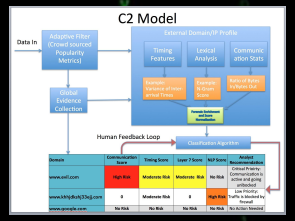
SCALING THE APPROACH TO OTHER PROBLEMS

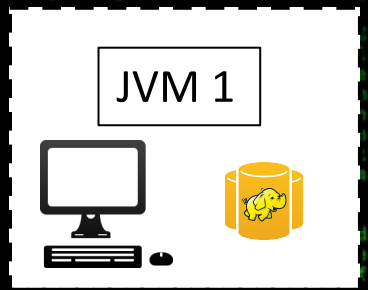
Original Workflow High Level

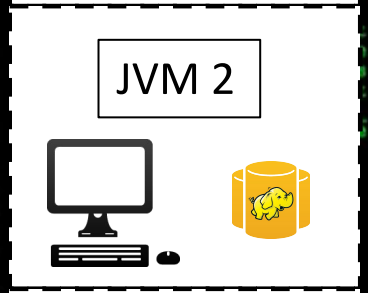
In theory we can perform this detection in parallel

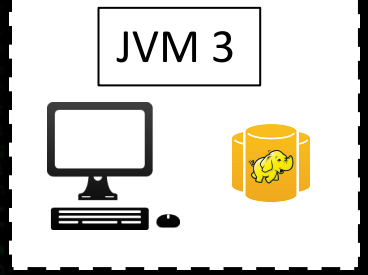










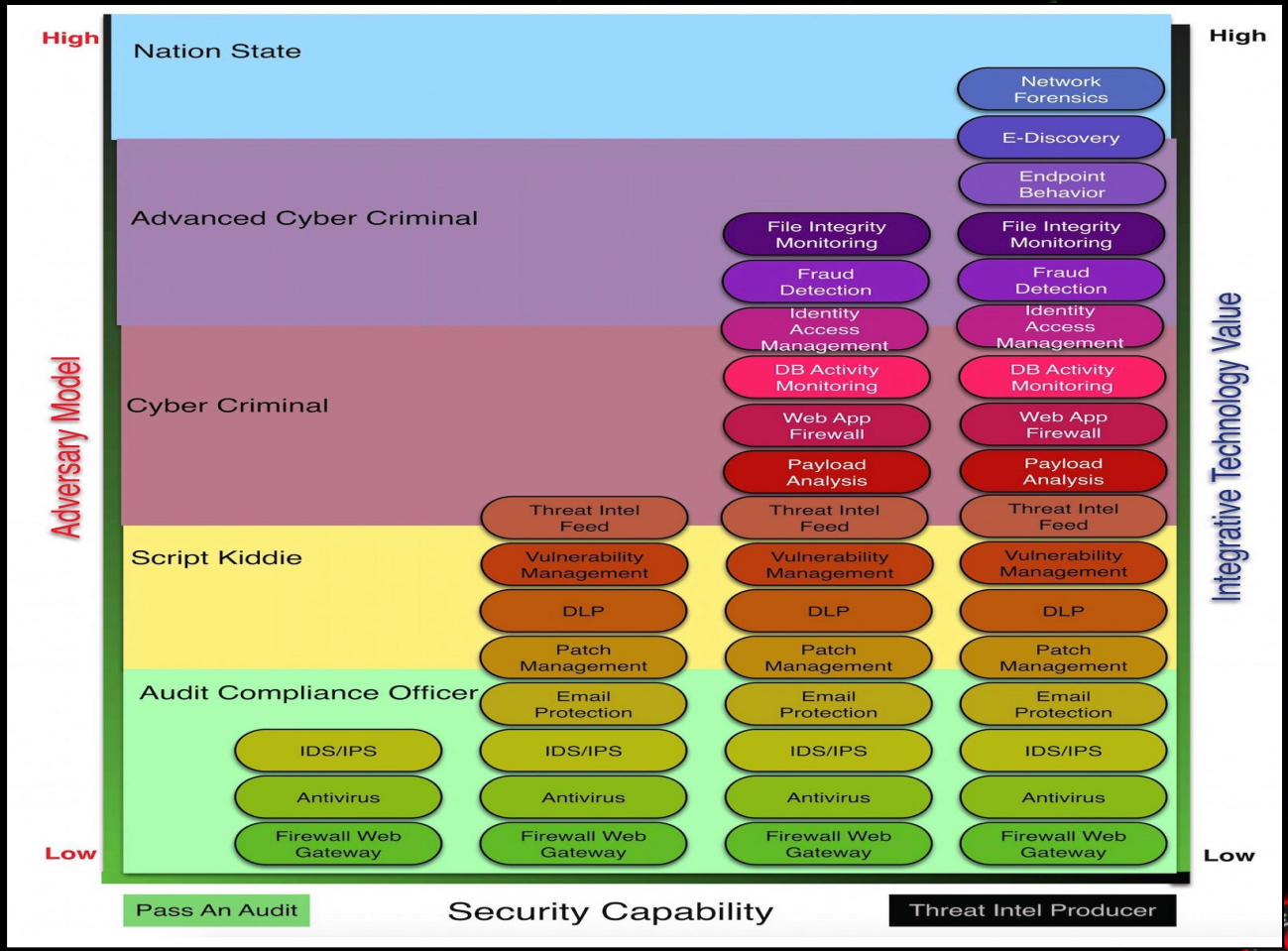


1. GET http://forbes.com/gels-contrariness-domain-punchable/"
2. GET http://portcullisesposturen.europartsplus.org/
3. POST http://dpckd2ftmf7lelsa.jjeyd2u37an30.com/

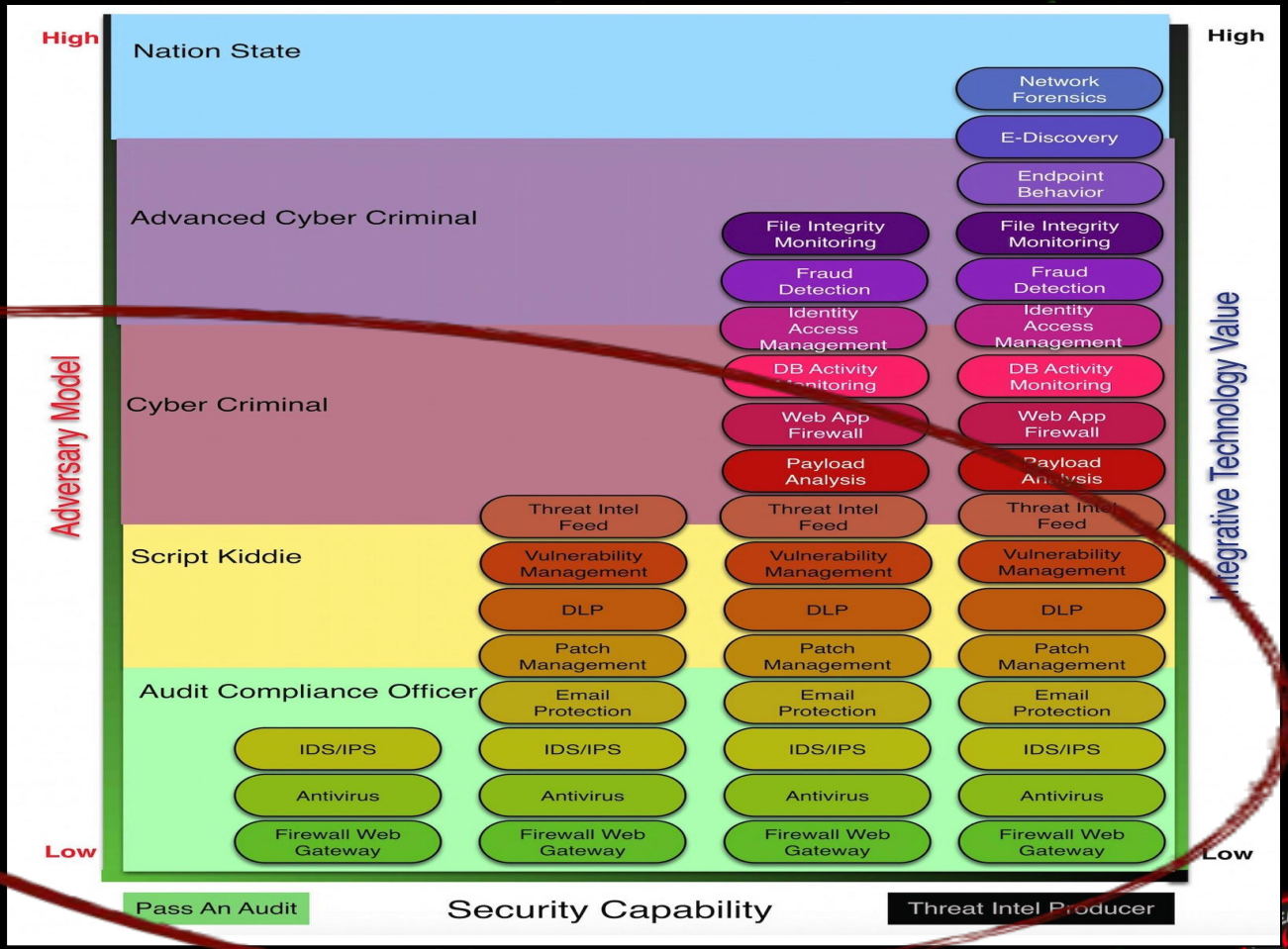
1. GET http://youtube.com/
2. GET http://avazudsp.net/
3. GET http://betradar.com/
4. GET http://displaymarketplace.com/

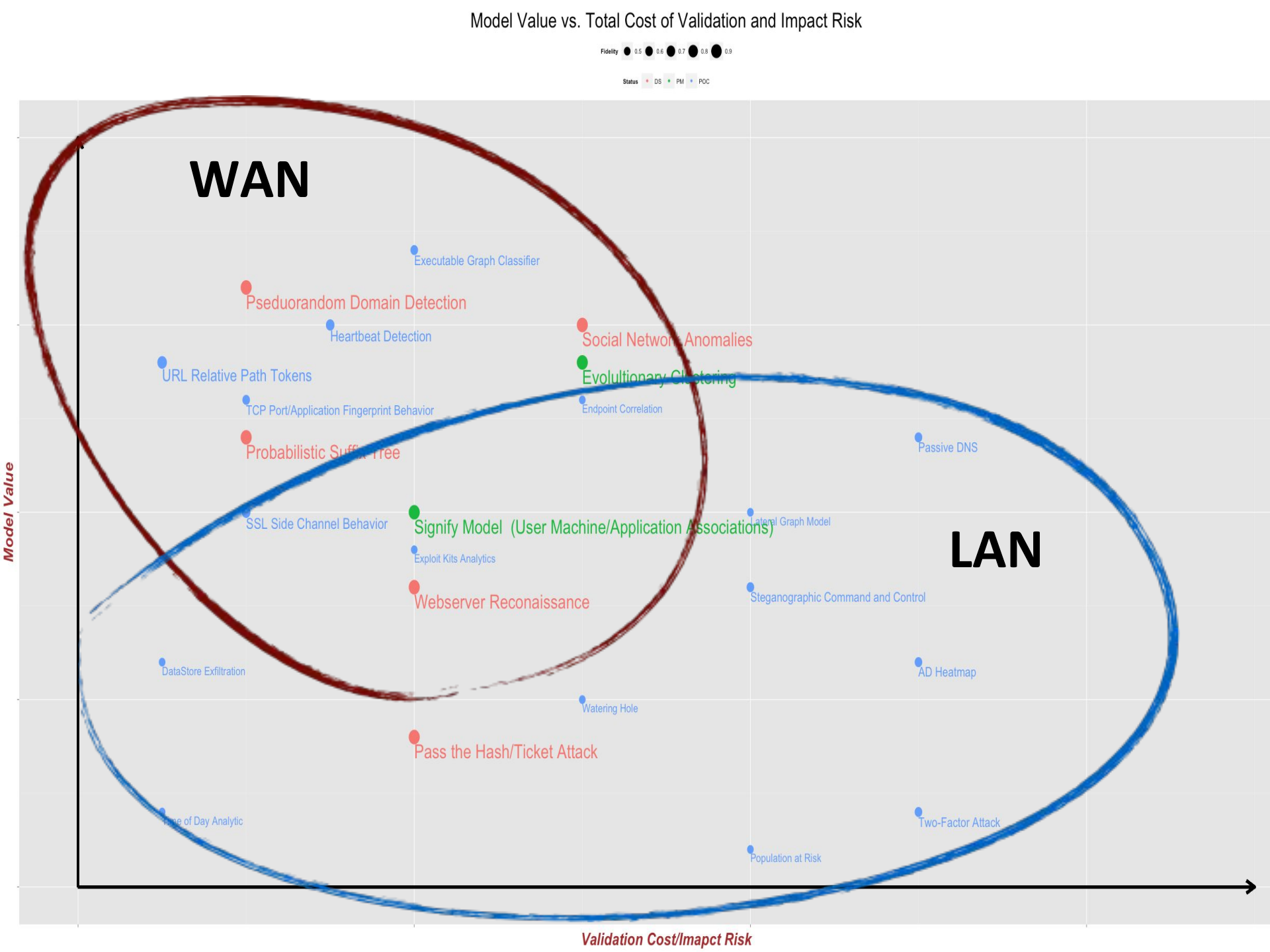
1. GET http:/clickable.net/
2. GET http://vuiviet.vn/
3. GET http://homedepotemail.com/
4. GET http://css-tricks.com/

Adversarial Models



Adversarial Models




Model Value vs. Total Cost of Validation and Impact Risk
Fidelity 0.5 0.6 0.7 0.8 0.9
Status DS PM POC
WAN
LAN
Executable Graph Classifier
Pseduorandom Domain Detection
Heartbeat Detection
Social Network Anomalies
URL Relative Path Tokens
Evolultionary Clustering
TCP Port/Application Fingerprint Behavior
Endpoint Correlation
Probabilistic Suffix Tree
Passive DNS
SSL Side Channel Behavior
Signify Model (User Machine/Application Associations)
Lateral Graph Model
Exploit Kits Analytics
Webserver Reconaissance
Steganographic Command and Control
DataStore Exfiltration
AD Heatmap
Watering Hole
Pass the Hash/Ticket Attack
Time of Day Analytic
Two-Factor Attack
Population at Risk
Model Value
Validation Cost/Imapct Risk

Lambda Security

splunk>

DHCP

IMS/IPAM

FW

AD

Proxy

VPN

Data Ingest

APACHE STORM™

Real Time Layer

Real Time Identity Resolution

IP	DHCP.MAC	DHCP_Lasteventtime	AD_FQDN
10.100.1.23	58:5c:35:c3:6e:a4	2014-03-11T14:00:00	joe.eng.acme.com
10.13.11.221	12:3a:74:b2:6a:22	2014-03-12T14:30:00	ad.hr.acme.com

Distributed ETL

Username = select coallesce(user_name, hostname, IP) from Active_ID_Table where IP = '10.10.100.23)

Sequential Models and IOC's

Hybrid View (Batch + Real Time)

Batch Layer

Spark™

Data-Parallel

Table

Row

Row

Row

Row

Result

Large Scale Models and Non-Sequential IOC's

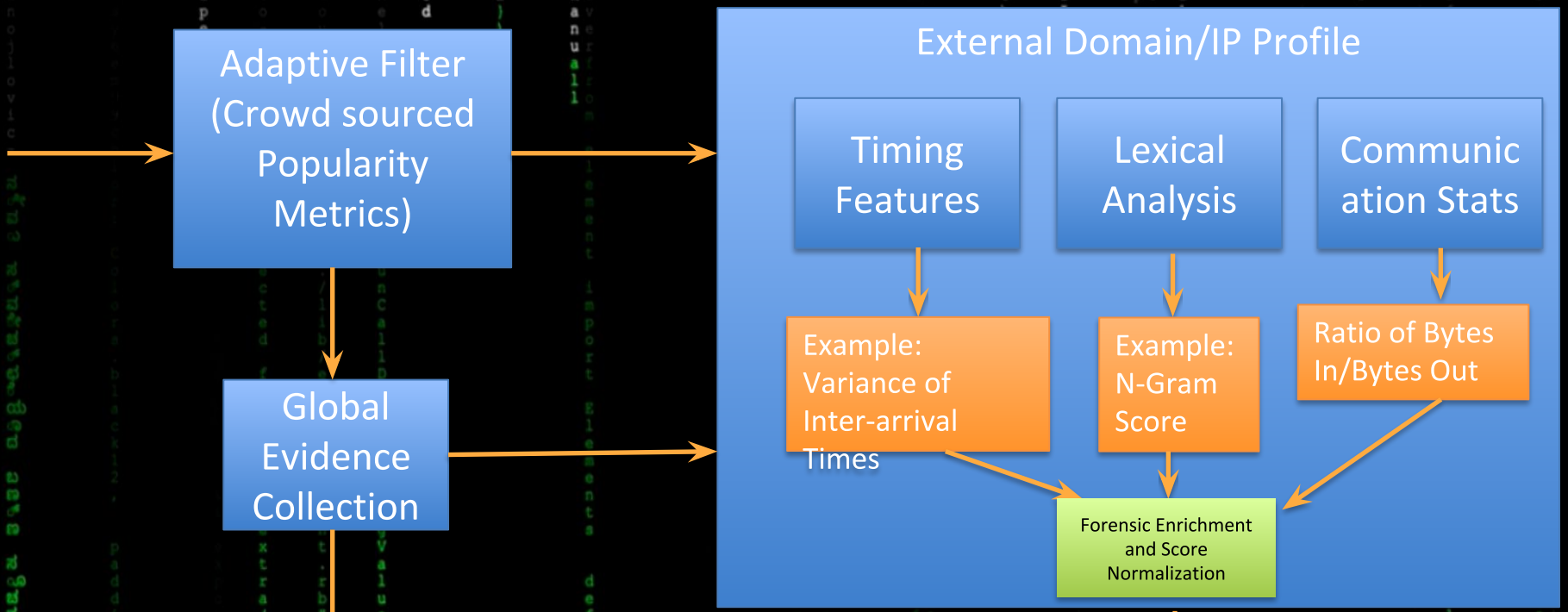
GraphX

SELECT date, COUNT(*) FROM product GROUP BY date

Map

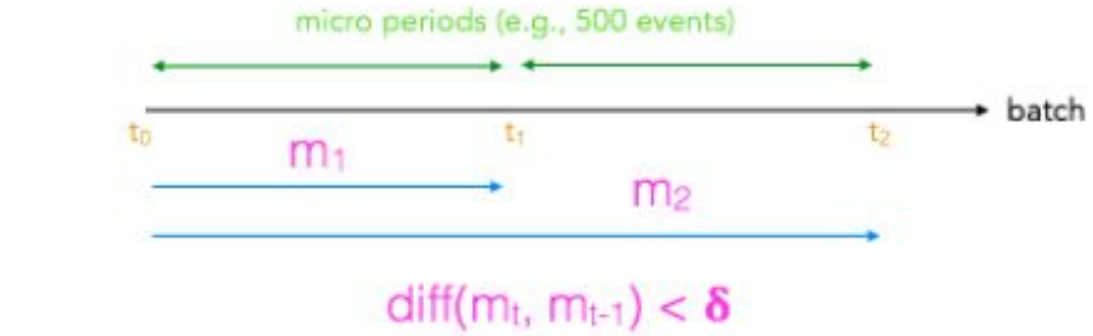
Reduce

C2 Model



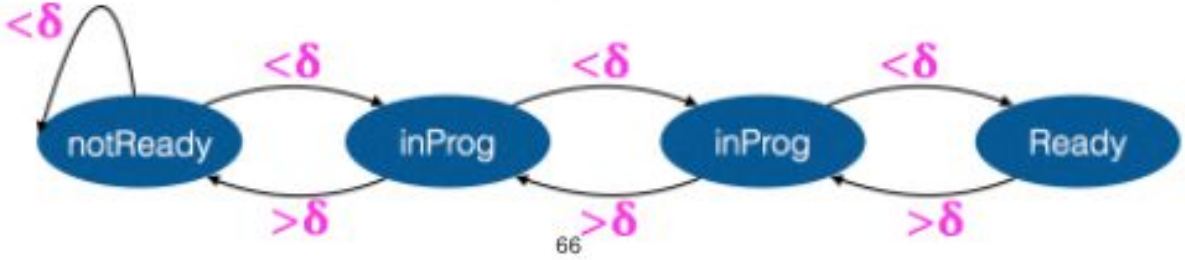
Domain	Communication Score	Timing Score	Layer 7 Score	NLP Score	Analyst Recommendation
www.evil.com	**High Risk**	**Moderate Risk**	**Moderate Risk**	**No Risk**	Critical Priority: Communication is active and going unblocked
www.khhjdkshj33ejj.com	**0**	**Moderate Risk**	**0**	**High Risk**	Low Priority: Traffic is blocked by firewall
www.google.com	**No Risk**	**No Risk**	**No Risk**	**No Risk**	No Action Needed

When is a model ready?



$$diff(m_t, m_{t-1}) < \delta$$

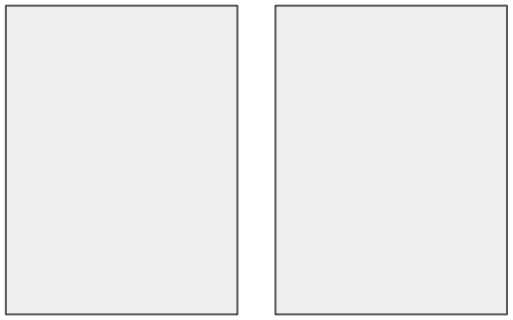
A model stabilization algorithm:



ML first steps in adaptive security: Commoditization of Compute

1 Trillion Letters of DNA = 800 MB (all of nature's local adaptation information encoded in a string < 1 GB in Size). The real value in learning is Time!!

1 Terabyte of Data



1 Kilobyte of Output

>(41.107678, -3.635677, 0.007473, -1.784944)

Analytic Use Cases: Stories and Fun Facts

1. Introduction

The immune system is a highly evolved biological system whose function is to identify and eliminate foreign material. In order to do this, it must be able to distinguish between foreign molecules (or *antigens*) and the molecules that constitute self. A prerequisite for the performance of this task is a powerful capability for learning, memory, and pattern recognition. In order to accomplish this, the immune system employs genetic mechanisms for change similar to those used in biological evolution. In the immune system, however, these processes function on a time scale that can be as short as a few days, making the immune system an ideal candidate for the study and modeling of adaptive processes.

THE IMMUNE SYSTEM, ADAPTATION, AND MACHINE LEARNING

J. Doyne FARMER and Norman H. PACKARD*
The Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

and

Alan S. PERELSON
Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

The immune system is capable of learning, memory, and pattern recognition. By employing genetic operators on a time scale fast enough to observe experimentally, the immune system is able to recognize novel shapes without preprogramming. Here we describe a dynamical model for the immune system that is based on the network hypothesis of Jerne, and is simple enough to simulate on a computer. This model has a strong similarity to an approach to learning and artificial intelligence introduced by Holland, called the classifier system. We demonstrate that simple versions of the classifier system can be cast as a nonlinear dynamical system, and explore the analogy between the immune and classifier systems in detail. Through this comparison we hope to gain insight into the way they perform specific tasks, and to suggest new approaches that might be of value in learning systems.

Analytic Use Cases: Stories and Fun Facts

2. A brief summary of the properties of the immune system

Given the rich chemical environment present in a highly evolved organism, it is inevitable that foreign organisms will attempt to invade in an effort to make use of these resources. To counteract this, in vertebrates the immune system has evolved to identify and dispose of foreign material. This is done in part by antibody molecules that tag foreign material and mark it for eventual removal by lymphocytes, phagocytic cells, and the complement system. A typical mammal such as a mouse or a human is thought to contain on the order of 10^7–10^8 different antibody types, each with its own unique chemical composition. The specialized portion of the antibody molecule used for identifying other molecules is called the *antibody combining region* or *paratope* (see fig. 1). The

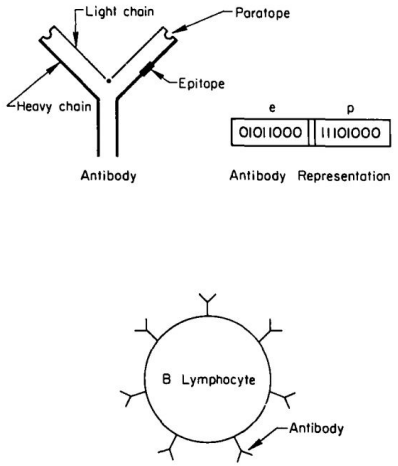


Fig. 1. A schematic representation of the structure of an antibody, an antibody as we represent it in our model, and a B-lymphyocyte with antibodies on its surface that act as antigen detectors.

RED TEAM VILLAGE

Great Job You Made it This Far!

Bathroom Break + Q&A Time

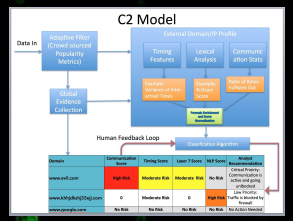
Questions **->#Hands-on-labs**

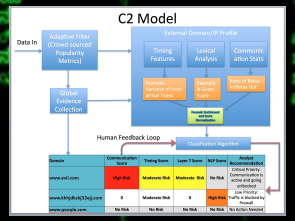
PART 2:

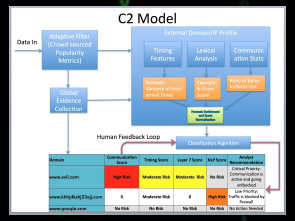
Hands On

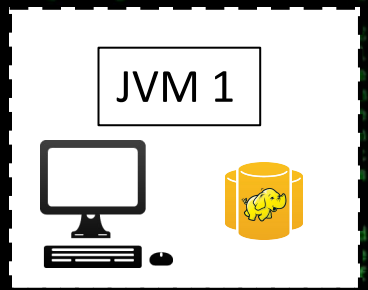
Original Workflow High Level

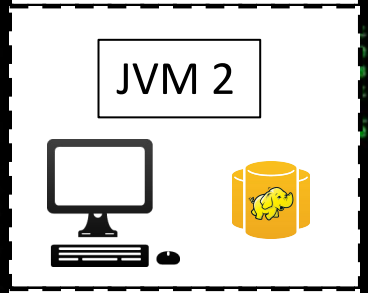
In theory we can perform this detection in parallel

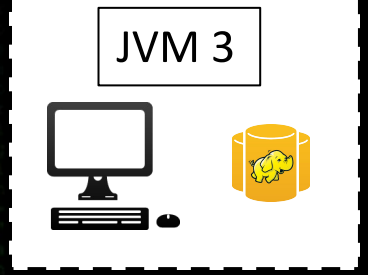












1. GET http://forbes.com/gels-contrariness-domain-punchable/"
2. GET http://portcullisesposturen.europartsplus.org/
3. POST http://dpckd2ftmf7lelsa.jjeyd2u37an30.com/

1. GET http://youtube.com/
2. GET http://avazudsp.net/
3. GET http://betradar.com/
4. GET http://displaymarketplace.com/

1. GET http:/clickable.net/
2. GET http://vuiviet.vn/
3. GET http://homedepotemail.com/
4. GET http://css-tricks.com/

Results

```
File for analysis :  data/logs_proxy_format/exploit/2014-01-02-neutrino-exploit-traffic.webgateway
Successfully monitoring data/logs_bro_format/exploit/http.log...
{'max_path_depth': 9, 'min_path_depth': 9, 'max_length': 47, 'min_length': 47, 'uri_Distinct': 29, 'max_entropy': 6.453848993519228
5, 'min_entropy': 6.4538489935192285, 'base64Match': 7, 'percentEncoded': [False, False, False, False, False, False, False, False,
False, False, False, False, False, False, False, False, False, False, False, False, False, False, False, False, False, False, False
, False, False, False, False, False], 'time_interval': [5.153927, 0.583675, 20.504216, -61668972.641135, 0.0, 1.0, 1.0, 0.0, 1.0, 1
.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 6.0, 0.0, 0.0, 0.0, 0.0, 0.0, 4.0, 1.0, 0.0, 7.0, 2.0, 6.0, 0.0, 1.0, 1.0], 'max_deltas': [20.504
216, 7.0, 6.0, 6.0, 5.153927, -61668972.641135, 0.0, 0.0, 0.0, 0.0], 'min_deltas': [-61668972.641135, 0.0, 0.0, 0.0, 0.0, 20.504216
, 7.0, 6.0, 6.0, 5.153927], 'ratio_of_deltas_A': 8.107802227636874e-07, 'ratio_of_deltas_B': 8.107802227636874e-07, 'ratio_of_delta
s_C': 9.72936267316425e-07, 'ratio_of_deltas_D': 4.864681336582125e-07, 'ratio_of_deltas_E': 4.864681336582125e-07, 'max_time_inter
val': -61668952.136919, 'min_time_interval': -61668952.136919, 'interval_length': 41.008432000437736, 'short_url': False, 'risk_ext
ention': False}
Completed!
```

POC

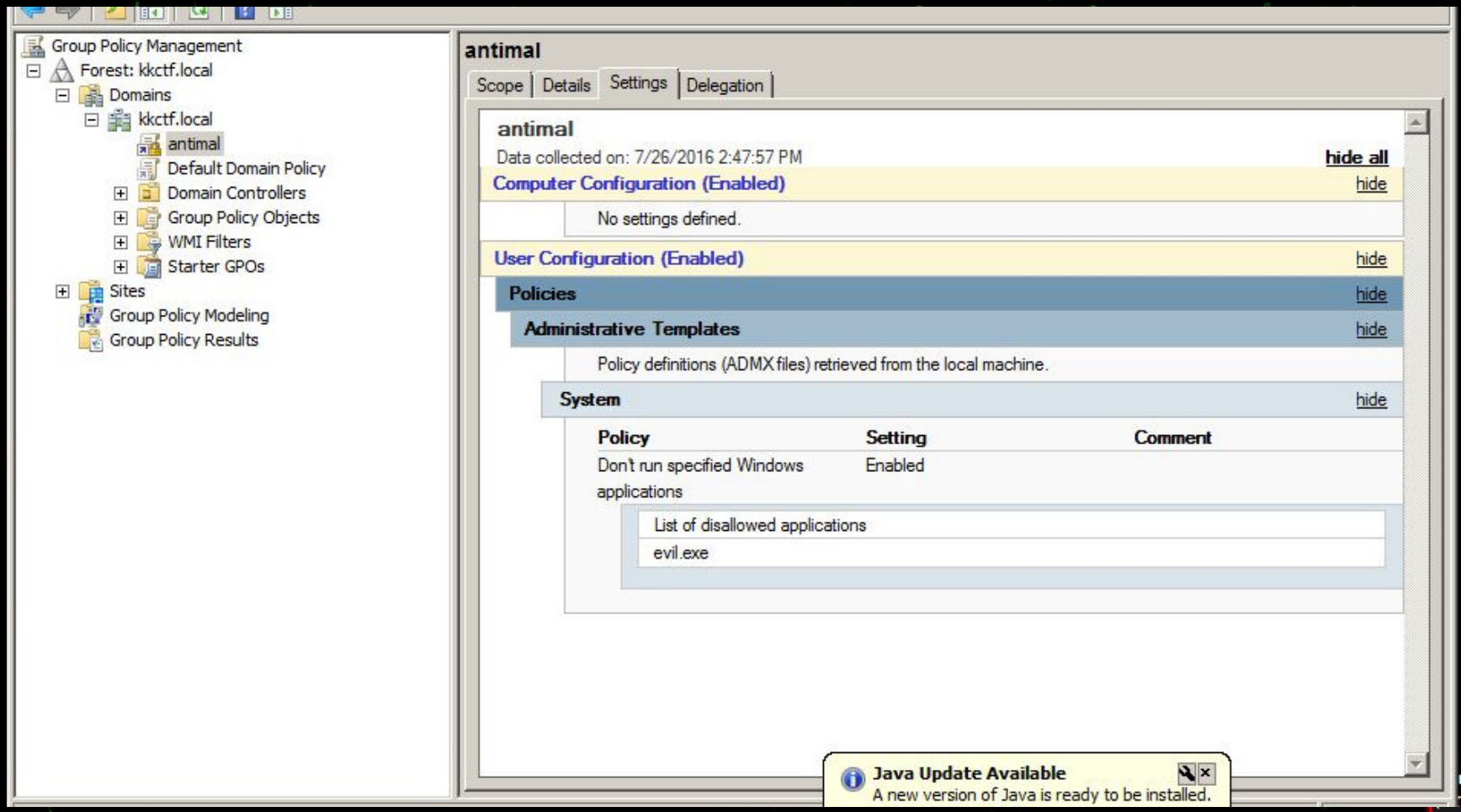
The Tool


scala-2.11 — java -jar Aktaion-assembly-0.1.jar — 101×53
[caspidian:scala-2.11 rsoto$ java -jar Aktaion-assembly-0.1.jar
****
****
** Aktaion Version 0.1 **
****
****
1: Analyze Bro HTTP Sample
2: Analyze PCAP Sample (Bro must be installed)
3: Demo
Enter Execution Choice [1-3]:

- Ransomware prevention items (Risk Scores --> GPO, ACLs)

```
research@securityonion14:~/Desktop$ python aktaionAD.py -f event.json
Executing command- C:\Windows\System32\WindowsPowerShell\v1.0\powershell -Inpu
tFormat none -OutputFormat TEXT -command "Import-Module grouppolicy; Set-GPReg
istryValue -Name antimal -Key HKCU\Software\Microsoft\Windows\CurrentVersion\P
olicies\Explorer\DisallowRun -ValueName 1 -Type String -Value evil.exe"
olicies\Explorer\DisallowRun -ValueName 1 -Type String -Value evil.exe"



DisplayName       : antimal
DomainName        : kkctf.local
Owner             : KKCTF\Domain Admins
Id                : 30b3a3cb-76bf-4345-a3b3-3ac91c21f916
GpoStatus         : AllSettingsEnabled
Description       :
CreationTime      : 7/25/2016 4:54:44 PM
ModificationTime : 7/29/2016 2:35:22 PM
UserVersion       : AD Version: 15, SysVol Version: 15
ComputerVersion   : AD Version: 0, SysVol Version: 0
WmiFilter         :
```

Ransomware prevention items (Risk Scores --> GPO, ACLs)



Active Defense

Output can be used to further application of defensive measures such as:

- ACL
- Group Policy in Windows Active Directory
- Blacklists
- Snort
- Bro signatures

- Conclusion

Use of Machine Learning techniques can enhance detection and defense technologies.

Signature less approach is the way of the present/future as conventional static signature based approach is insufficient

Use of Behavioral Analytic techniques provides an expansion into multiple items previously not considered in detection and defense technologies.

Combination of Machine Learning/Analytics streamlines analysis and detection when processing multiple sources of events or multiple micro behaviors present in individual events.

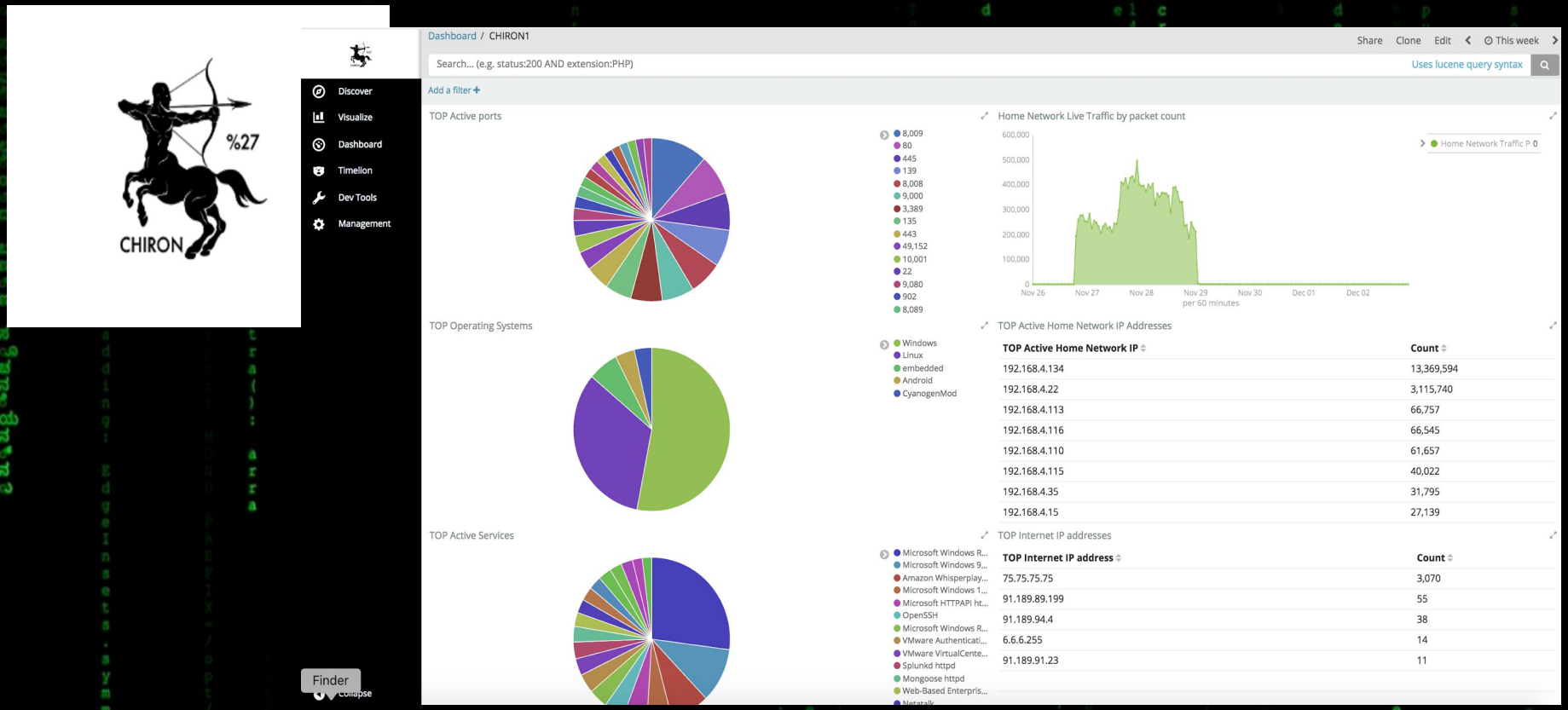
THE FUTURE
CHIRON HOME BASED ML IDS

Chiron combines analytics with machine learning.

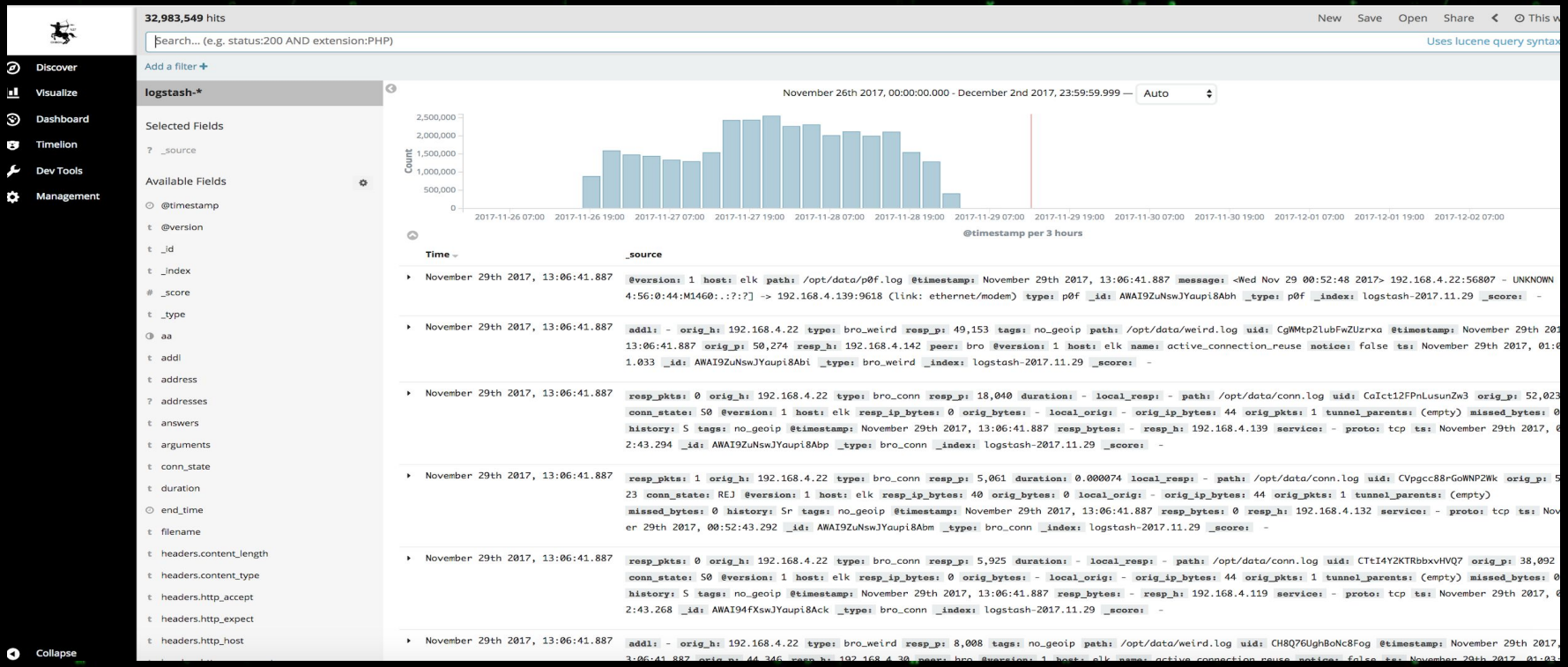
Designed for home networks, place with no or limited vision and very vulnerable

Aktaionv2 is will be the ML engine behind threat detection

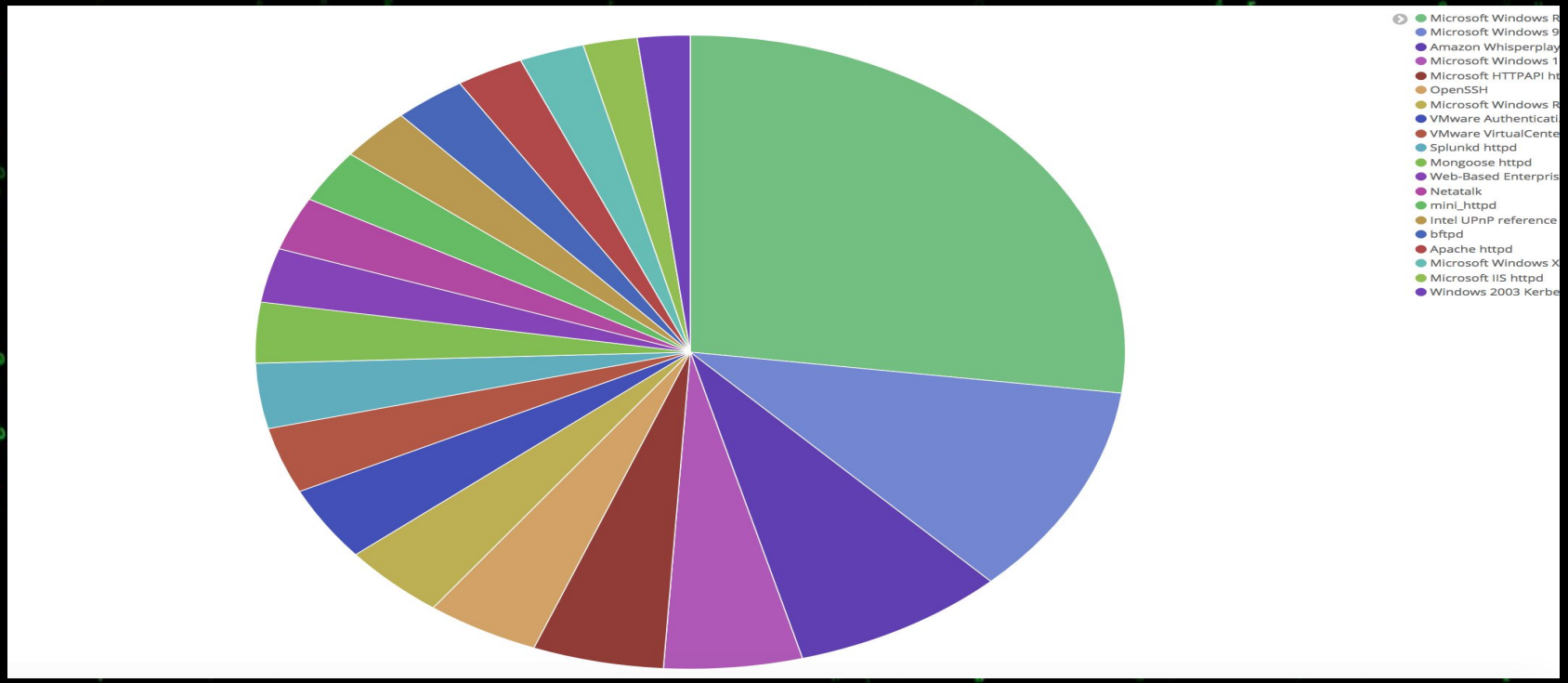
POC - Screenshots



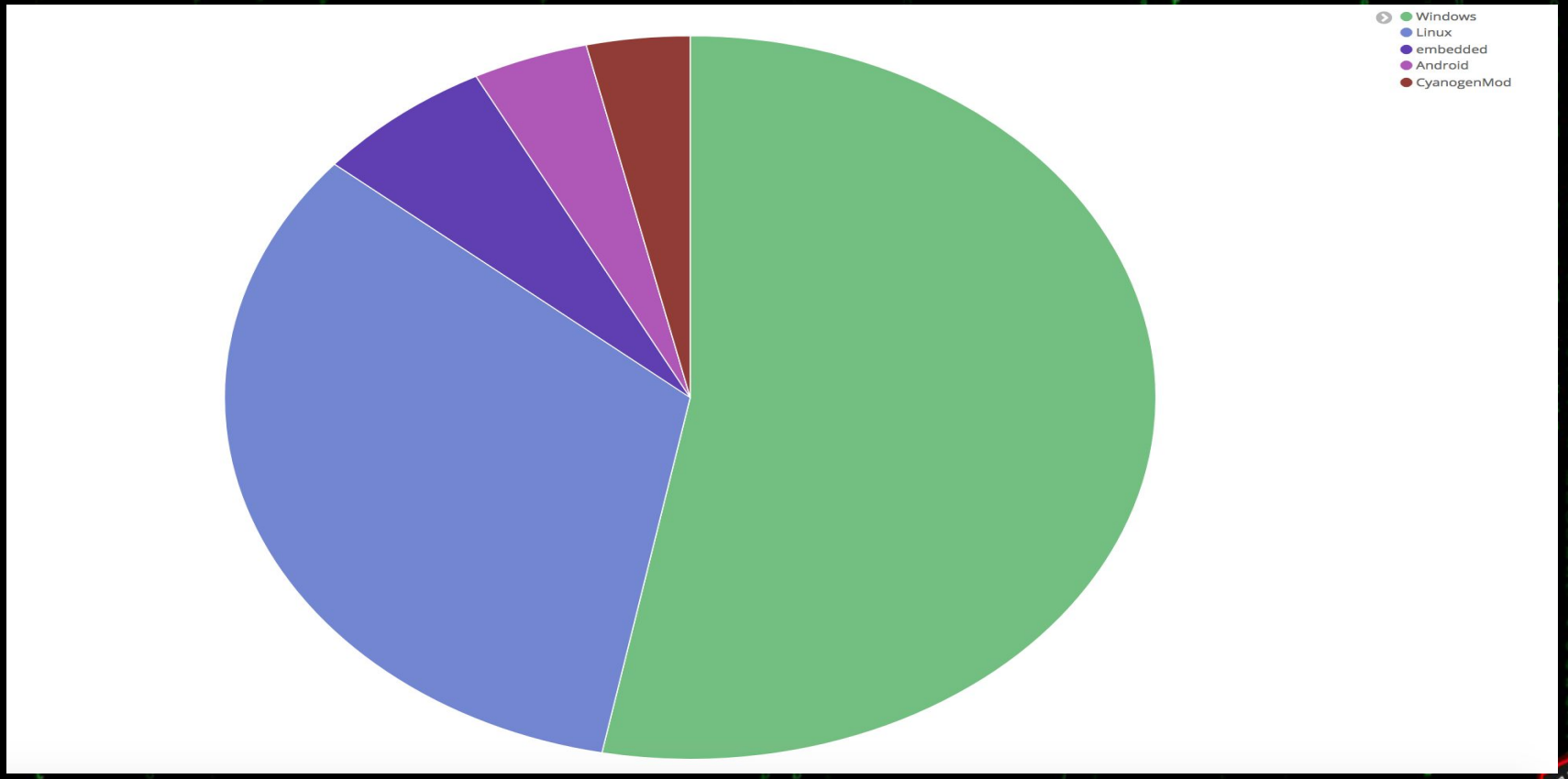
POC Screenshots



POC Screen shots



POC Screenshots



Q&A

Joseph Zadeh @josephZadeh

Rod Soto @rodsoto