No Silver Bullet
Multi contextual threat detection via Machine Learning.

By @rodsoto @jozephzadeh

$:Whoami..

- Rod Soto
  - Researcher at Splunk UBA, former AKAMAI, Prolexic PLXSert. Like to break things, p0wn botnets and play CTFs.
- Joseph Zadeh
  - Data Scientist at Splunk UBA, building behavioral intrusion detection technologies at scale. Enjoy working on defense projects that combine security, artificial intelligence and distributed systems.

Introduction

Agenda

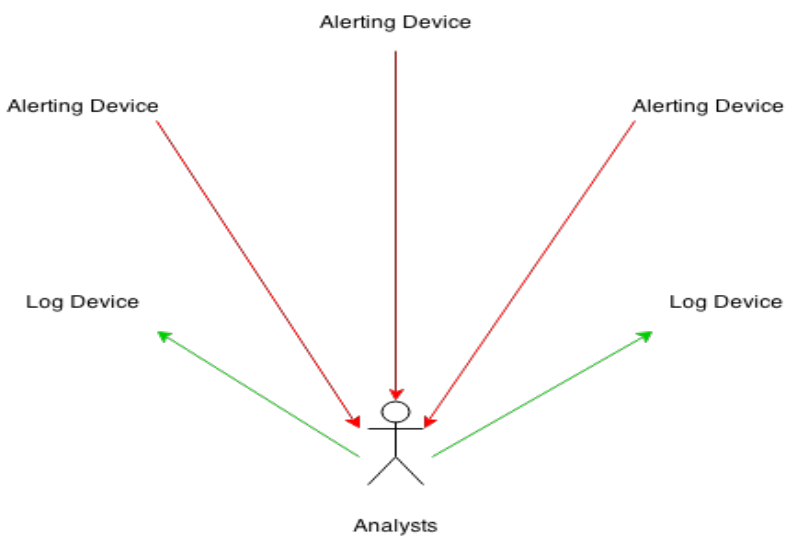
- Introduction: Big Data and Machine Learning
- Machine Learning in security workflows and how it can help and limitations
- Describe central nervous system approach to behavioral security: Lambda Defense

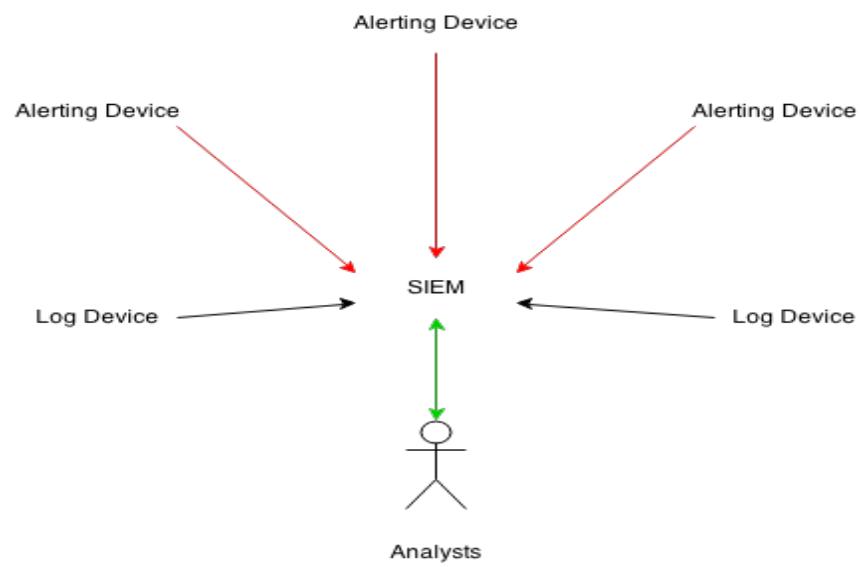
Big Data

BIG
DATA
ALSO
TOOLS
DISK
TARGET
APPLIED
SENSOR
SHARED
DEFINITION
MAY
MOVING
WITHIN
ARCHIVES
MANAGE
TECHNOLOGIES
EXAMPLES
EVERY
CURRENT
PRACTITIONERS
CAPTURE
BUSINESS
INTERNET
SETS
DISTRIBUTED
GENOMICS
NEEDED
TIME
ONE
LARGE
INFORMATION
MANAGEMENT
TERABYTES
PETABYTES
INCLUDE
SOFTWARE
PERFORMANCE
STORE
SYSTEMS
TOLERABLE
ELAPSED
ABILITY
SIZE
RESEARCH
STORAGE
LOGS
PARALLEL
MASSIVELY
GROW
SAN
GARTNER
WORKING
DIFFICULTY
SEARCH
RECORDS
NETWORKS
DATABASES
HUNDREDS
CAPACITY
BIOLOGICAL
PROCESSING
LARGER
CONTINUES
SOCIAL
RELATIONAL
SET
USE
COMPLEX
CITATION
TENS
ANALYTICS

Challenges in Current Threat Indicator Technologies

- Many devices generating logs and alerts
- Data distributed in too many places slows analysis, preventing analysts from effectively analyzing all alerts
- SIEM makes life somewhat easier, giving analysts one place to collect data but still deal with needle in haystack issues





The Big Data Challenge

- " It costs organizations an average of $1.27 million annually in time wasted responding to erroneous or inaccurate malware alerts. According to respondents, an average of 395 hours is wasted each week detecting and containing malware because of false positives and/or false negatives. The extrapolated average value of lost time is estimated at approximately $25,000 per week or $1.27 million each year for participating organizations." Ponemon Institute

The Big Data Challenge

- SOCs are challenged and limited in the scope of detection, analysis and action.
- Constant required training, updates and turnover of SOCs present a challenge for organizations.
- As of now People vs People model has proven to be more effective as current threat detecting/ prevention technologies do not seem sufficient nor effective against malicious actors. The numbers speak for themselves.

Big Data challenge, presents a new opportunity as well. Enter Machine Learning

- Machine learning is a subfield of computer science[1] that evolved from the study of pattern recognition and computational learning theory in artificial intelligence.[1] Machine learning explores the study and construction of algorithms that can learn from and make predictions on data.[2] Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions,[3]:2 rather than following strictly static program instructions.

*Wikipedia

Machine Learning & Big Data Technologies

- The ability to process very large sets of data through distributed computing plus the ability to apply algorithms that can learn based on these large datasets, will provide analysts with more meaningful detection and actionable items.

Learning Algorithms

- “a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.” *Wikipedia
- These learners can be designed and develop to scale against all these sources of data and produce meaningful detection of anomalies.
- By applying these learners we can build models that can approach threats from a multi contextual, dynamic perspective, thus going beyond the concept of static signature based security technologies.

Sequencing the Security DNA

- The next gen paradigm:
  - 1:1 Correspondence between users data footprint and Compute Resources
- Commoditization of compute means for 300,000 User Accounts means assign 300,000 individual threads + memory + disk to run learning algorithms per individual log footprint simultaneously

Adversarial Drift

- Current status quo, is driven by adversaries developing and introducing changes in their TTPs, bypassing all current detection technologies.

Spammers

Organised crime

Script kiddies/crackers

Hacktivists/pressure groups

Disgruntled employees

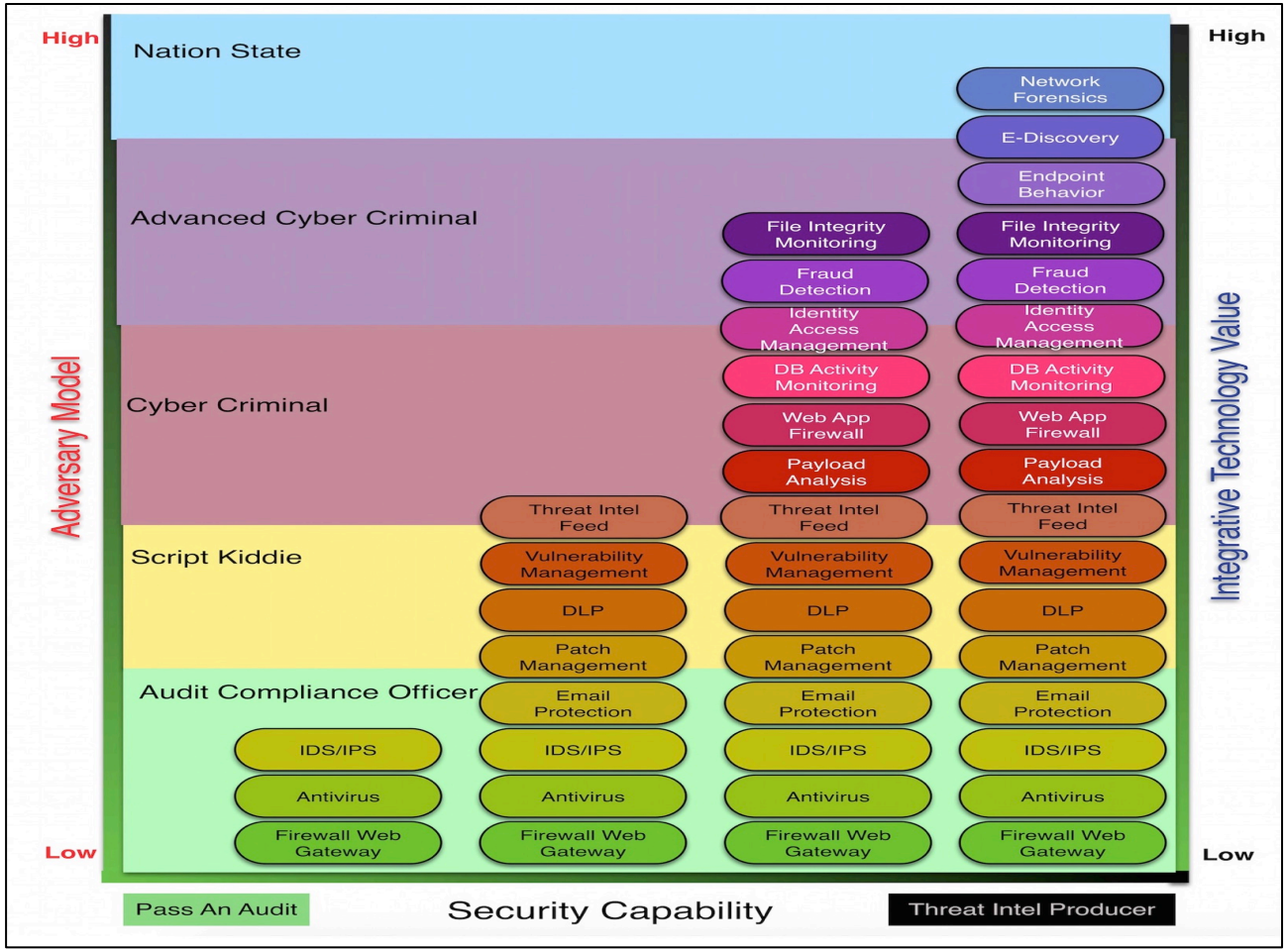
Foreign government agencies

The competition

Domestic government agencies

Advesarial Models

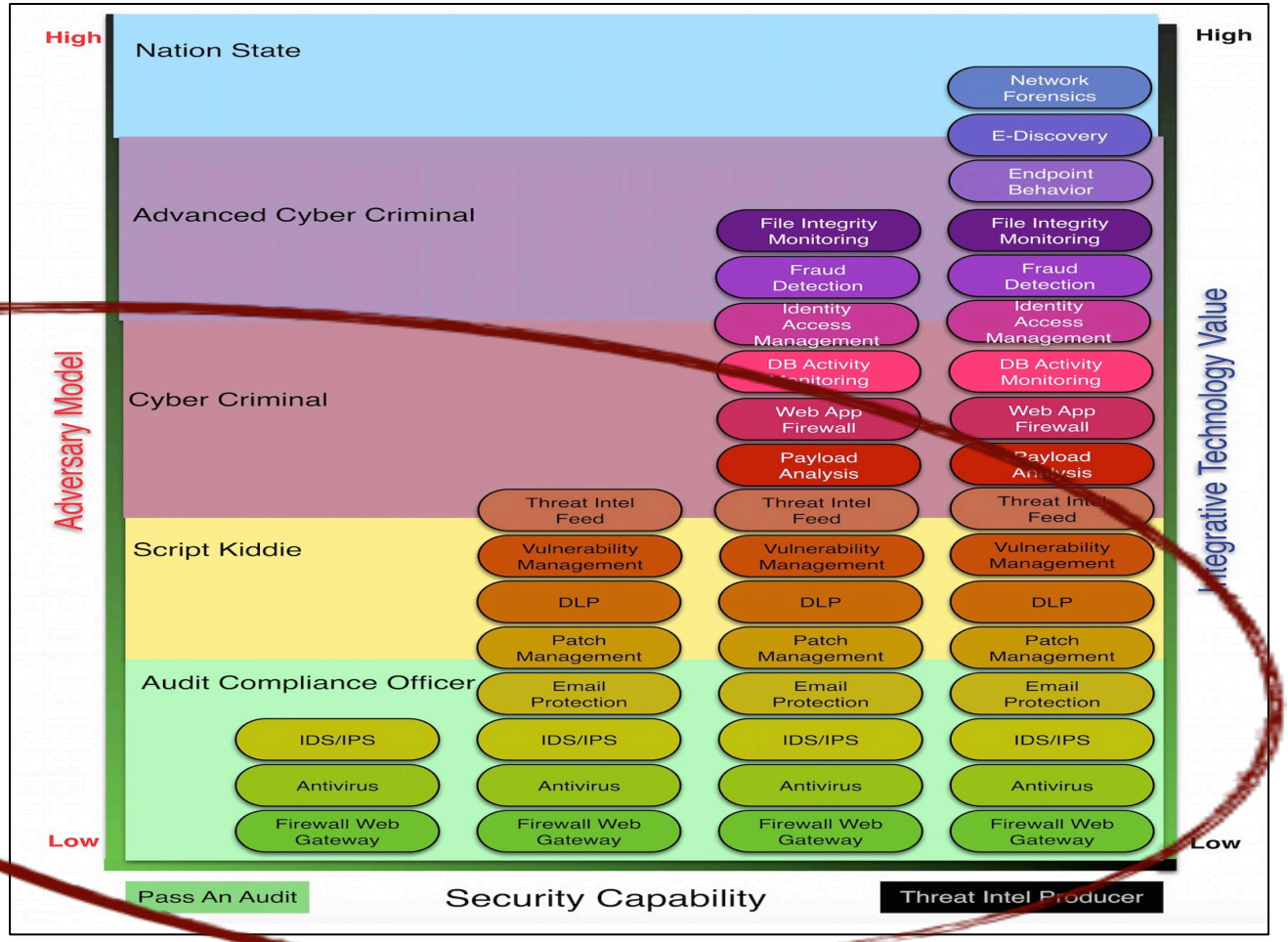
- Machine Learning Looses Effectiveness the more complex the adversary



Advesarial Models

Non-Automatable Actions: Hybrid Human/Computer Analysis

Automatable Actions: Good for ML



Learning Algorithms in Security

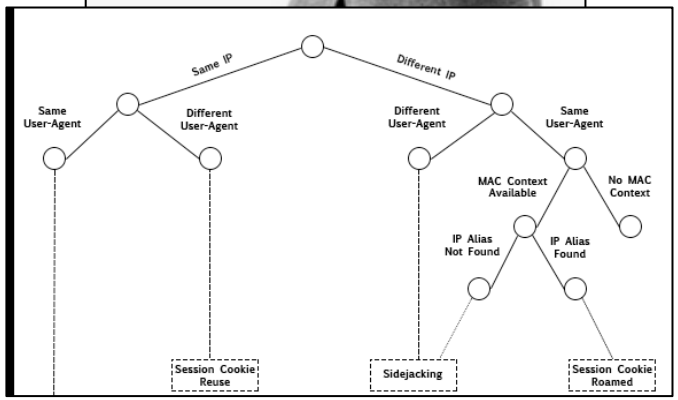
Advantages of using ML

- Using ML allows us to put together very large and distinct sources of data into a platform for analysis, interpretation and prediction.
- ML allows us to go beyond of static signature based technologies.
- ML creates an scenario where detection of threats based on dynamic and multi contextual indicators is possible.

Automating the Forensic Workflow

- Incident Response Is Hard Work! What can we automate?

A security analyst is an oracle whose input is evidence and whose output is True Positive, False Positive, True Negative or False Negative



  - The list of possible questions is large but typically the flow is a type of decision tree for example

ML as a tool to make your job easier

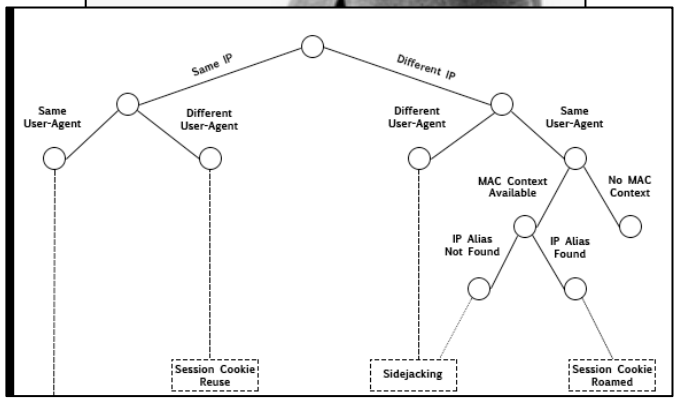
Security Oracle Workflow

Example 1:

Evidence => Periodic Communication => LAN to WAN Data =>WAN URL has Bad Reputation => Correlate with VT => True Positive

Example 2:

Evidence => Potential C2 Domain => LAN to WAN Data => WAN URL is new Google IP => False Positive

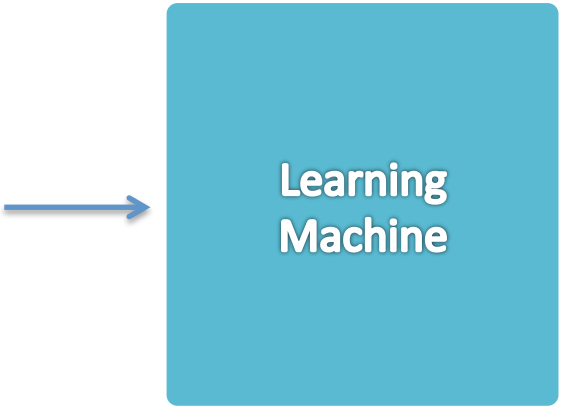


Learning = Compression?

- There is a duality between learning and compression

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

Input Data Total Size = 1 GB



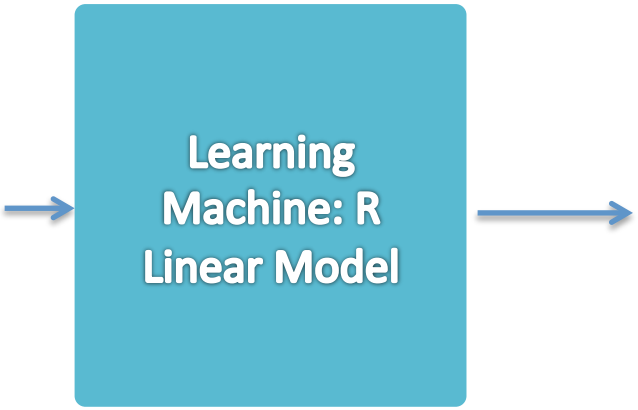
Learned output is a set of "coefficients"

Total Output Size = 1K

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
```

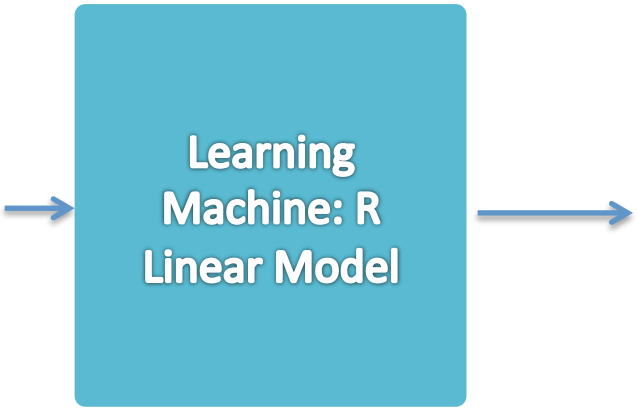


Learning = Compression?

- 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?

- 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)))
```

ML Challenges

- Over fitting/Under fitting
- Technology still in early stages
- “Operationalization”
- Advesarial drift and changing TTP’s means models have to change over time (retraining)

Lambda Defense

Decomposing Behaviors for Intrusion Detection

DEFINITION 1. ***The Lambda Defense Pattern***

1. $|S| = |\mathbb{R}|$: *Start with the threat surface S for a fixed adversarial environment. (The examples in this paper S will usually be the threat surface associated to a large enterprise network). The space of possible behaviors is a continuum at this point.*

2. *Enumerate the threat surface: How do we reduce the space of possible behaviors to actionable units that we can process in an upstream functional workflow? To answer this question we project down to modeling only a finite set $S \approx \{TTP_1, \ldots TTP_n\}$. For a given threat surface the art of the initial design is to identify the core attack patterns needed to be modeled and to anticipate which of those parts are most likely to "drift" [19]. Better yet assign metrics to this finite approximation of the threat surface (see figure 1).*

3. *Functional Decomposition: For each TTP_i we model the behavior as $TTP_i := B_i + NB_i$. Where B_i is what we call the sequential component of the behavior and NB_i is called the non-sequential component. Formally speaking each TTP_i is actually a stochastic process (sequence of random variable) $TTP_i = \{TTP_i(t) : t \geq 0\}$.*

4. *Behavioral Model: For each sub component B_i, NB_i we assign one or more models $M_1(B_i), \ldots, M_k(B_i), M_{k+1}(NB_i), M_{k+2}(NB_i), \ldots$ and we evolve these models over time in the next step. We keep the definition of model intentionally vague here but in practice each model is usually machine learning based (supervised/semi-supervised/unsupervised) or heuristically engineered around a use case that is not driven by a pattern recognition problem. It is important to note in that we are looking for the right tool for the job and machine learning is sometimes only used as a pre-filtering step or not at all. In some cases we combine graph processing or techniques from signal analysis where it helps with overall problem of a specific model implementation.*

5. *Map/Reduce Based Model Life Cycle: We can update simultaneous models in a scalable way with a simple map and reduce operation given by a signature of the form [41]: ModelRDD.aggregate[M](seqOp: (M, Data) => M(Data), combOp: (M_{t_0},M_{t_1}) => M_{t_2})*

Behaviors: Sequential + “Unordered”

- Sequential Behaviors
  - Exploit Chains
  - Timing Analysis (Periodicity)
  - Active Directory Sequence
  - Authentication Graph
- Non Sequential Behaviors
  - Fingerprinting
  - Grouping Behaviors
  - Application Counts
  - Rare file extension counts for Webshell detection

Mapping Behaviors to Code

- Easy to Parallelize
  - Count()
  - Average()
  - Time series()
  - Local state computations
    - Per user/IP/account/…
- Hard to Parallelize (NC Complete Complexity)
  - Rank()
  - Median
  - …
  - Anything that keeps track of global state

In complexity theory, the notion of **P-complete** decision problems is useful in the analysis of both:

1. which problems are difficult to parallelize effectively, and;
2. which problems are difficult to solve in limited space.

Lambda Security

- Lambda architecture provides a design paradigm for a “Scalable Central Nervous System” for the SOC whose components include
  - Machine learning based ETL(Extract/Transform/Load)
  - Distributed crawlers
  - Automated identity/session resolution and fingerprinting
  - Formal evidence collection protocol for automated labeling of incident response data
  - Analytics Metrics and establishing benchmarks for heterogeneous data

Batch Features + Real Time Features

- Keep in mind all work is done on a cluster (distributed system)
  - Concepts: groupBy (User,Domain, “arbitrary field”)
- Batch Example
  - Data driven domain popularity
- Real time example
  - Exploit chain content types
- Lambda => Immutable/Functional data structures
  - Spark RDD’s (abstraction for a distributed computation as opposed to result of a distrubted computation)

Lambda + Central Nervous System

- Augment “in memory” lightweight signal from the point with large scale processing platforms that can “sequence the security DNA”
  - Classical IDS/FW/Point solutions have significant limitations in terms of sharing state and being able to correlate across nodes

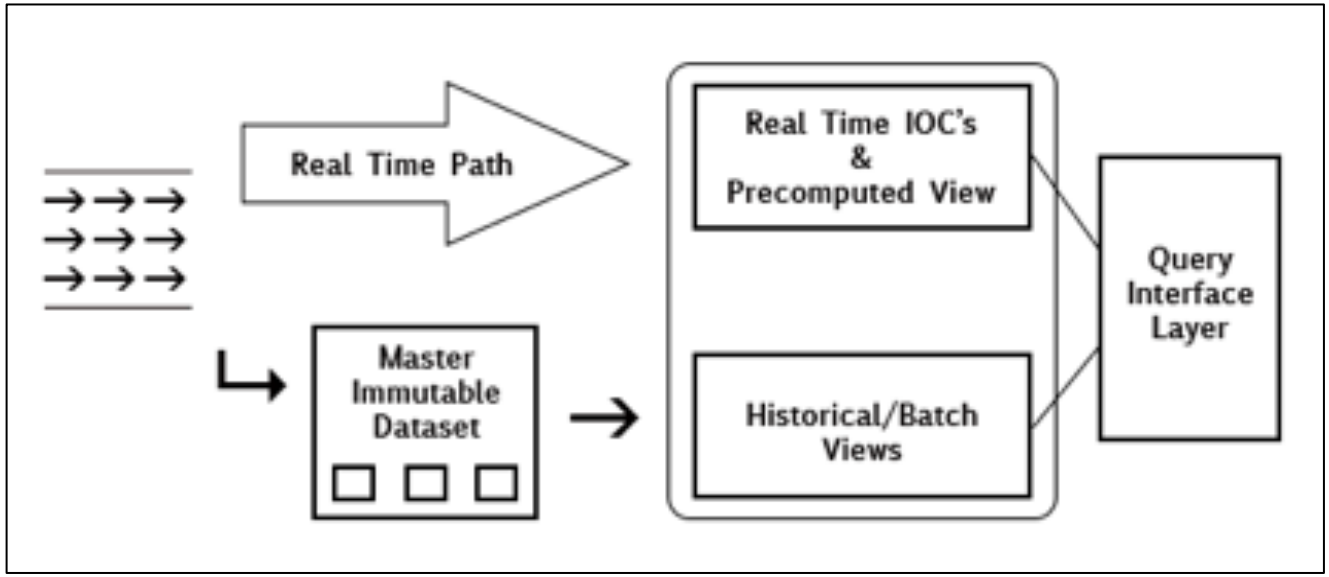
Lambda Architecture

- Architecture is described by three simple equations:

batch view = function(all data)

realtime view = function(realtime view, new data)

query = function(batch view, realtime view)



When is a model ready?

- Q1: When should we re-train?
- Q2: How is new data weighed over older data?
- Q3: How do we know when a model is ready?

When is a model ready?

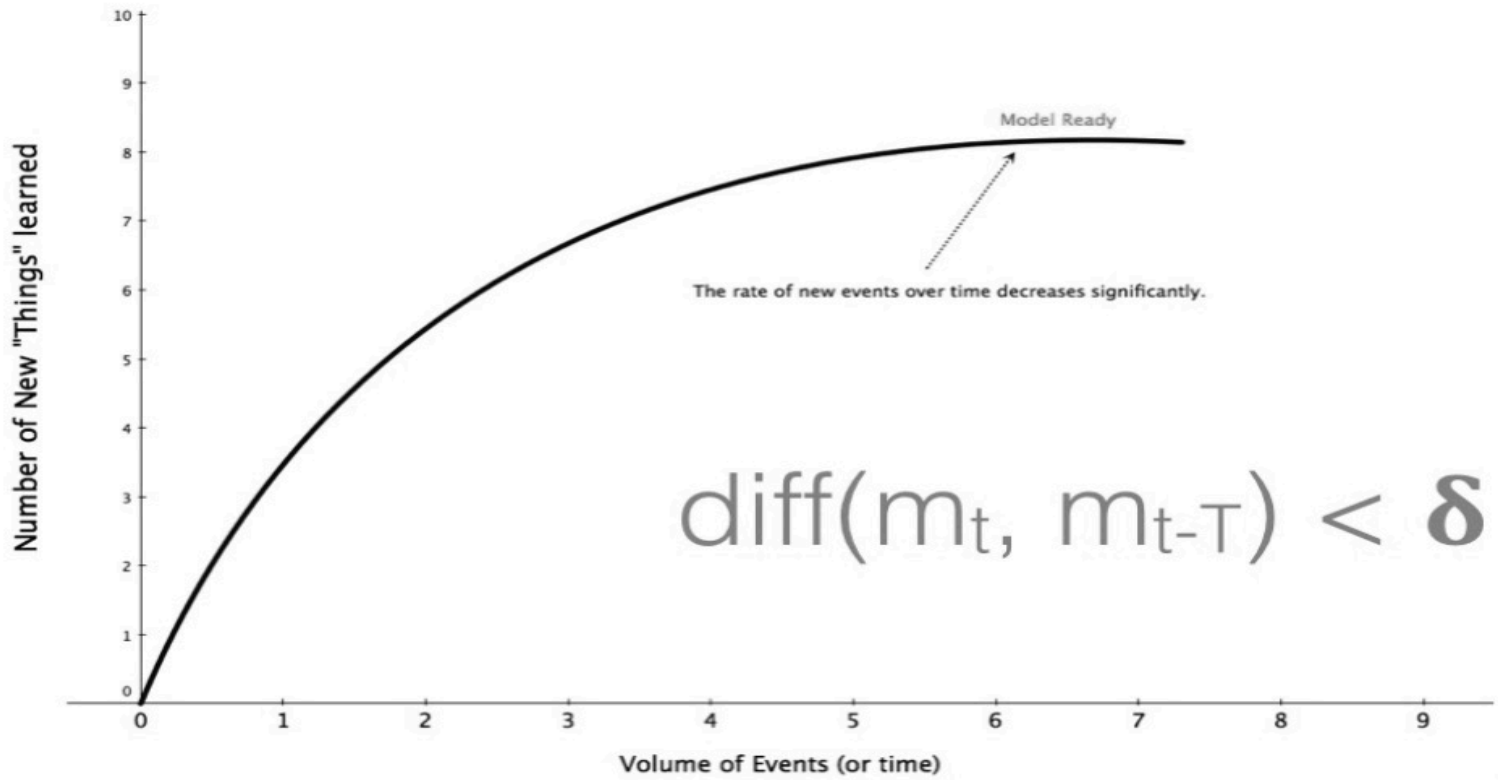
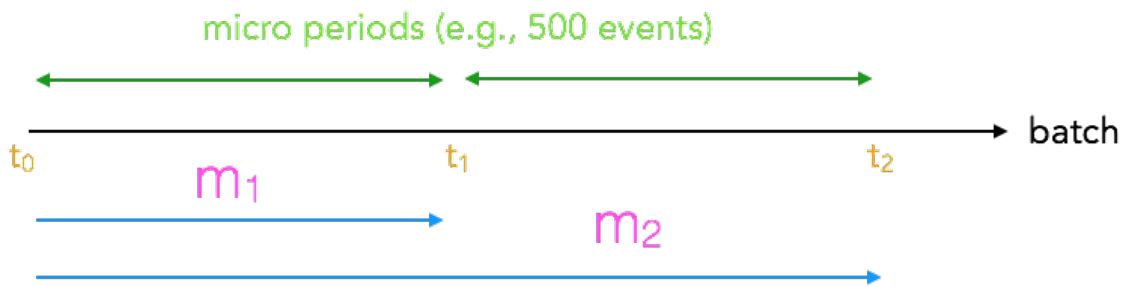


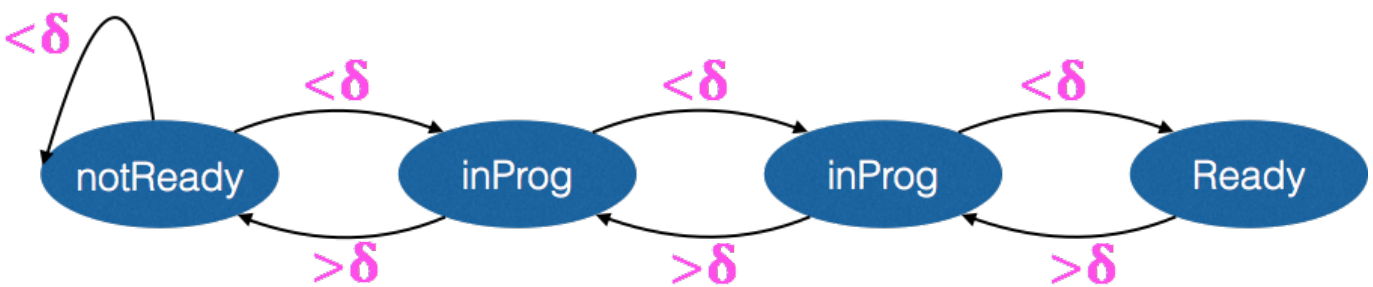
Figure 3: Model Stabilization Over Time

When is a model ready?



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

A model stabilization algorithm:



Model Life Cycle Implementation

The flow we described has a natural breakdown in terms of Map and Reduce on RDD's. In our context we implement a reduce on two models: reduce(M_1, M_2) => NewModel, and map from data and a model to Model(data): map(M_1, Data) => M_1(Data). We use the following definition for aggregate from the Spark AP [36]:

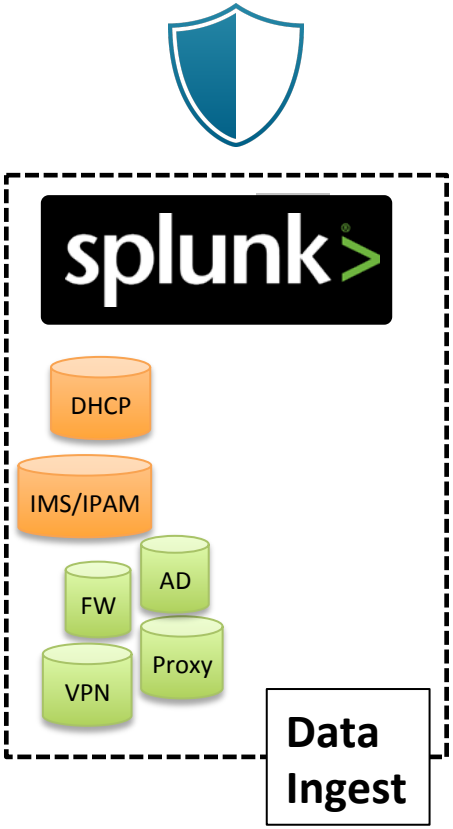
```
aggregate[U](zeroValue: U)(seqOp: (U, T) => U, combOp:
      (U, U) => U): U

Aggregate the elements of each partition, and then the
      results for all the partitions, using given combine
       functions and a neutral ''zero value''. This
      function can return a different result type, U,
      than the type of this RDD, T. Thus, we need one
      operation for merging a T into an U and one
      operation for merging two U's, as in scala.
      TraversableOnce. Both of these functions are
      allowed to modify and return their first argument
      instead of creating a new U to avoid memory
      allocation.

seqOp = an operator used to accumulate results within a
      partition (The Map operation in our case).

combOp = an associative operator used to combine results
       from different partitions (The Reduce operation in
       our case).
```

In our applications using the basic functional approach each model definition implements a map to update the model with data: map(M, Data) => M(Data), and a reduce to merge models reduce(M_i,M_j) = $M_{\text{merge}(i,j)}$:

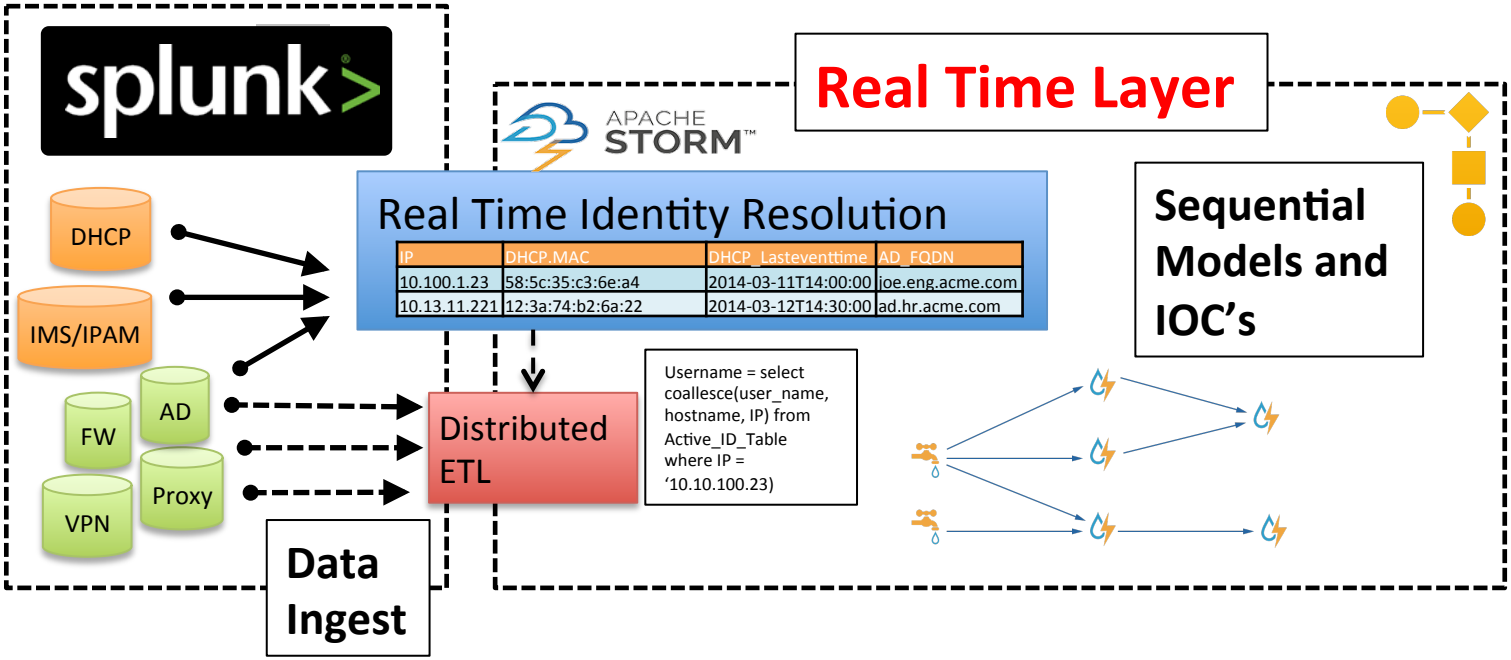
SIGNATURE 1. *Map/Reduce Model Life Cycle*

```
aggregate[M](seqOp: (M, Data) => M, combOp: (M,M) => M) =
      { ... }
```

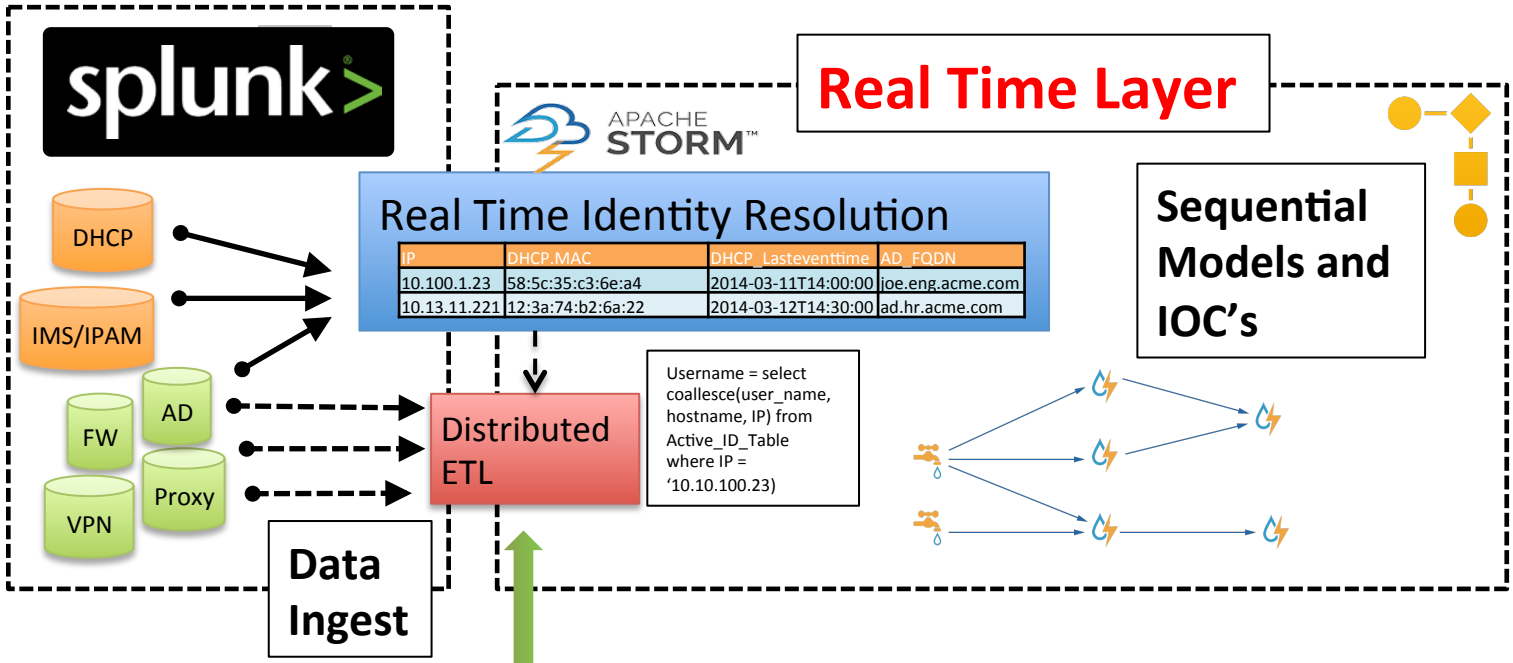
Lambda Security

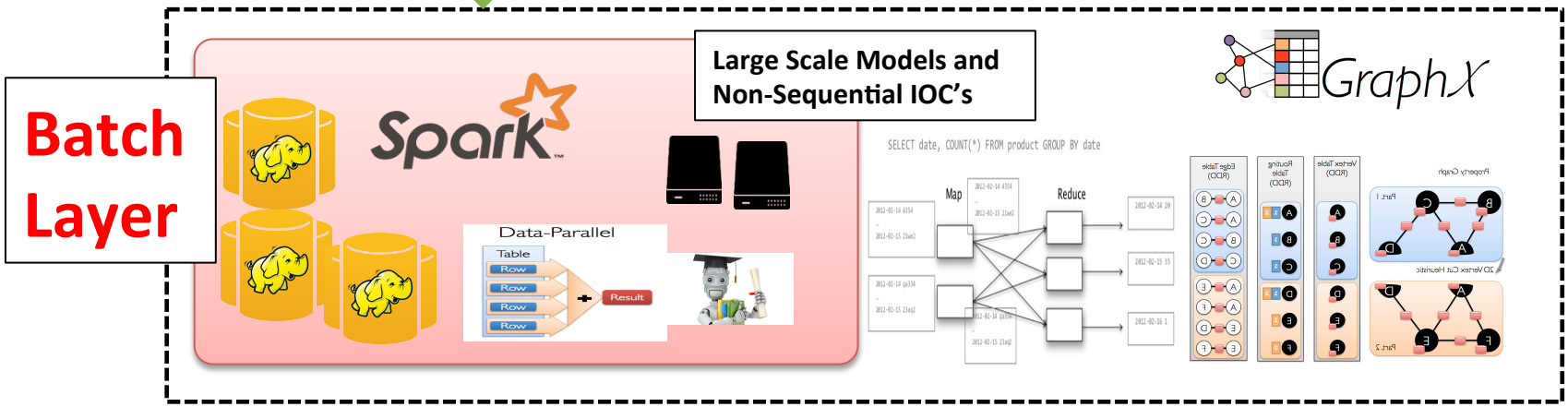


Lambda Security

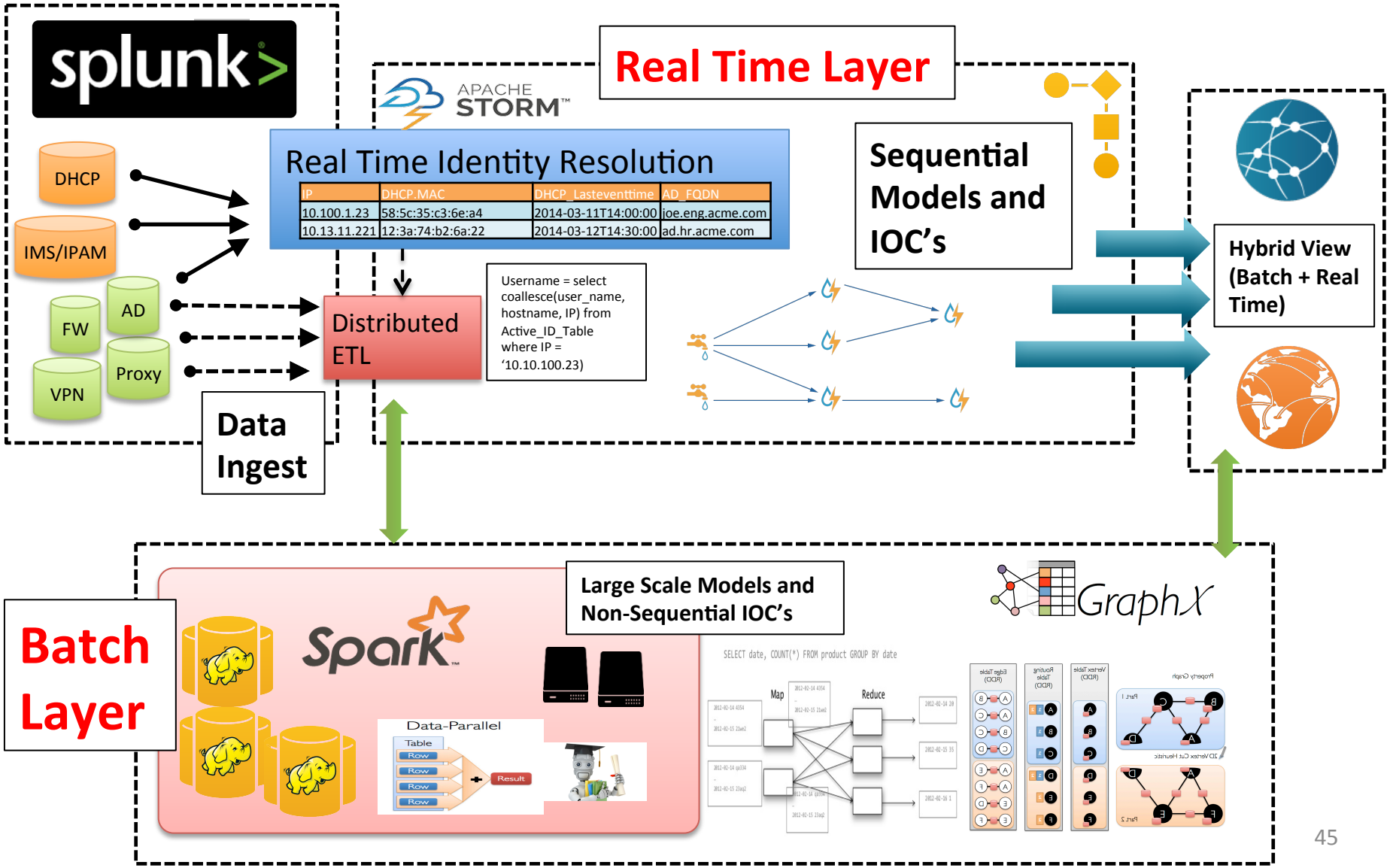


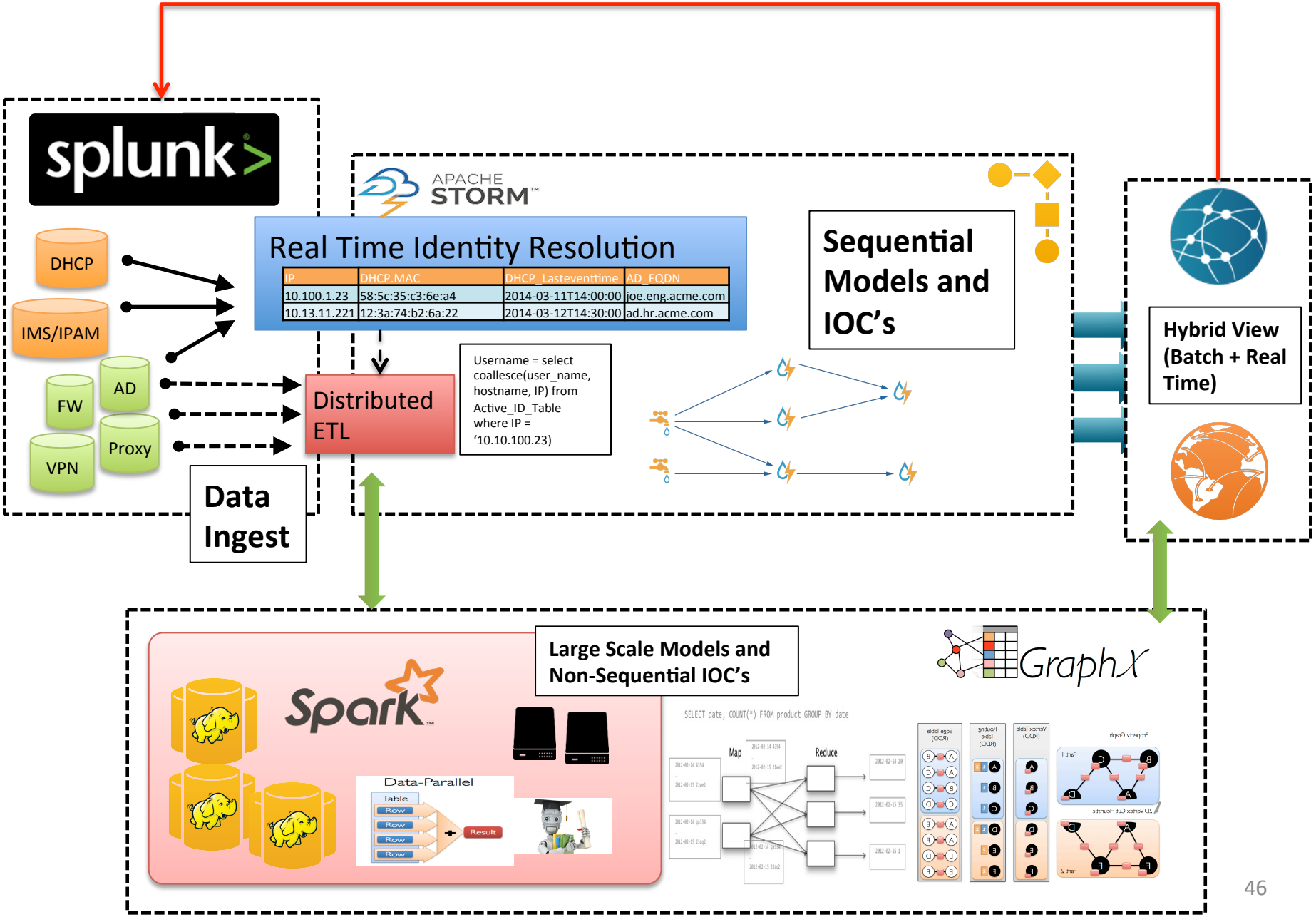
Lambda Security



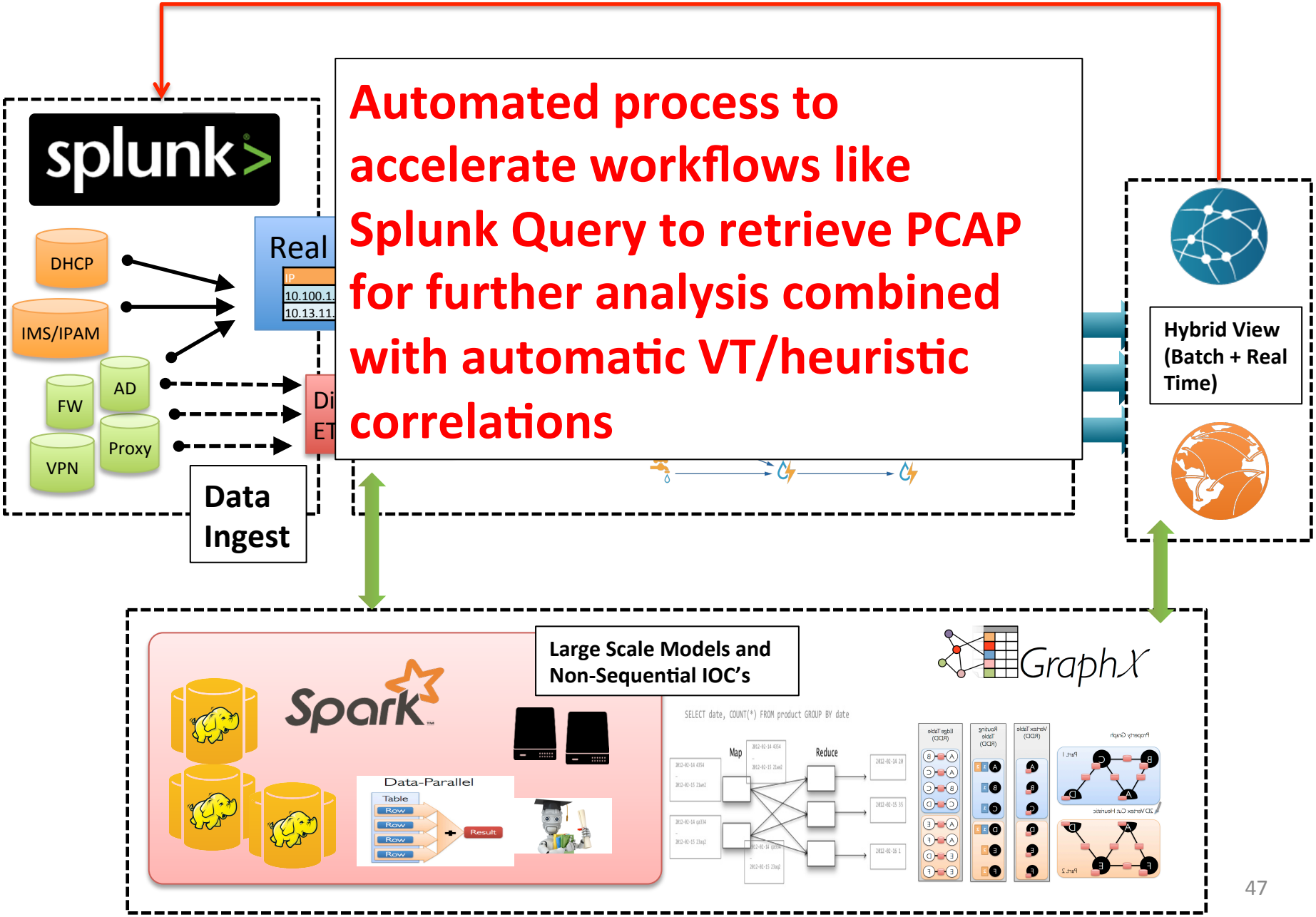


Lambda Security



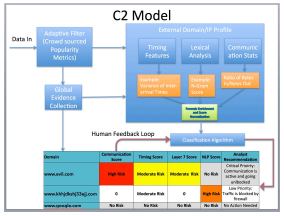

splunk>
DHCP
IMS/IPAM
AD
FW
Proxy
VPN
Data Ingest
APACHE STORM
Real Time Identity Resolution
IP | DHCP.MAC | DHCP_Lasteventtime | AD_FQDN
10.100.1.23 | 58:5c:35:c3:6e:a4 | 2014-03-11T14:00:00 | joe.eng.acme.com
10.13.11.221 | 12:3a:74:b2:6a:22 | 2014-03-12T14:30:00 | ad.hr.acme.com
Distributed ETL
Username = 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)
Large Scale Models and Non-Sequential IOC's
Spark
Data-Parallel
Table
Row
Result
GraphX
SELECT date, COUNT(*) FROM product GROUP BY date
Map
Reduce

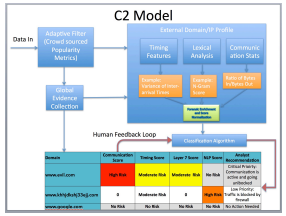
Automated process to accelerate workflows like Splunk Query to retrieve PCAP for further analysis combined with automatic VT/heuristic correlations

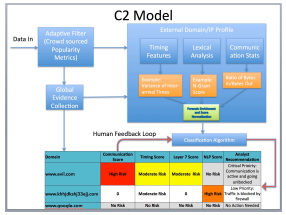


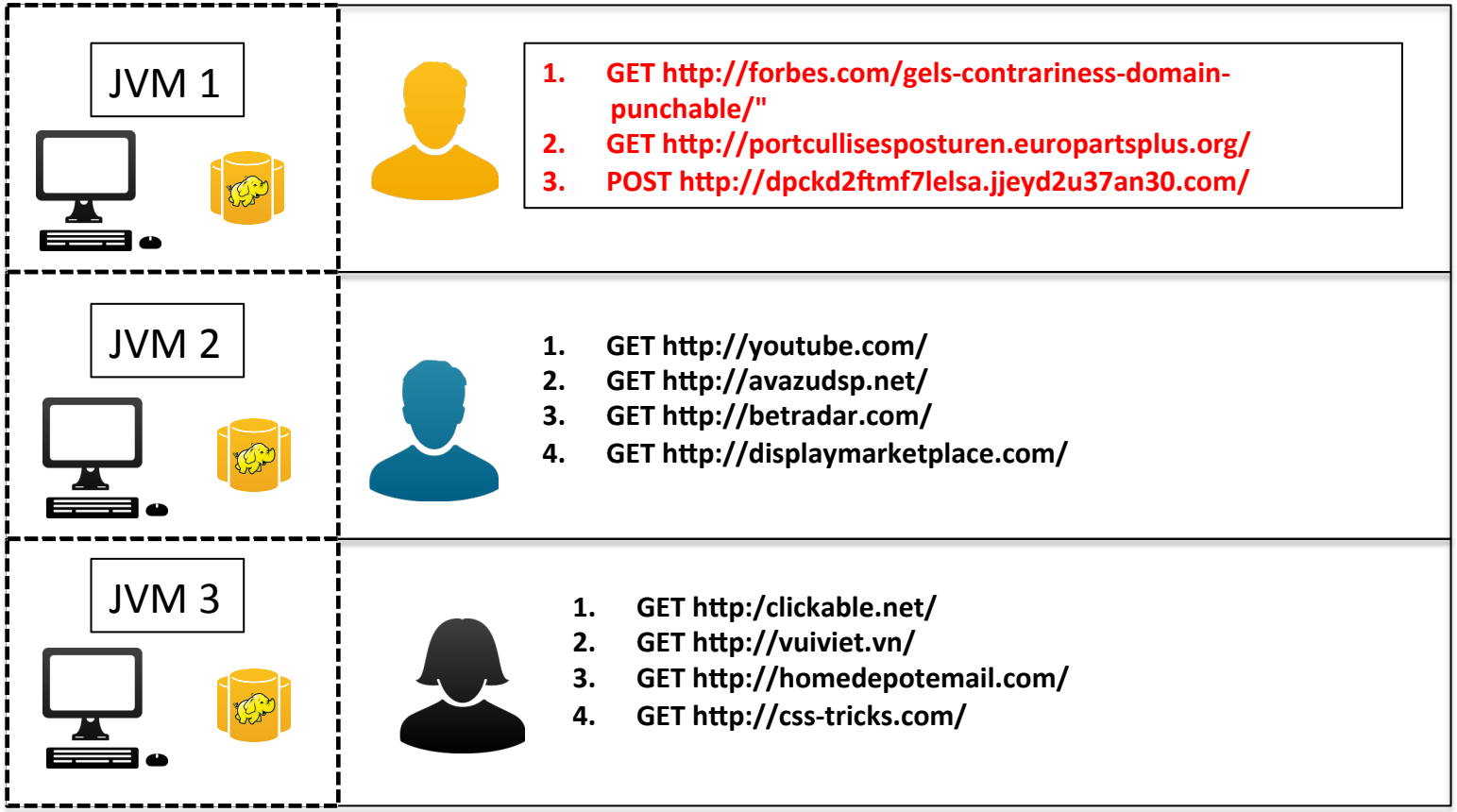
ML + Sequencing the Security DNA

- We parallelize across many nodes (JVMs) and use both real time and batch computations





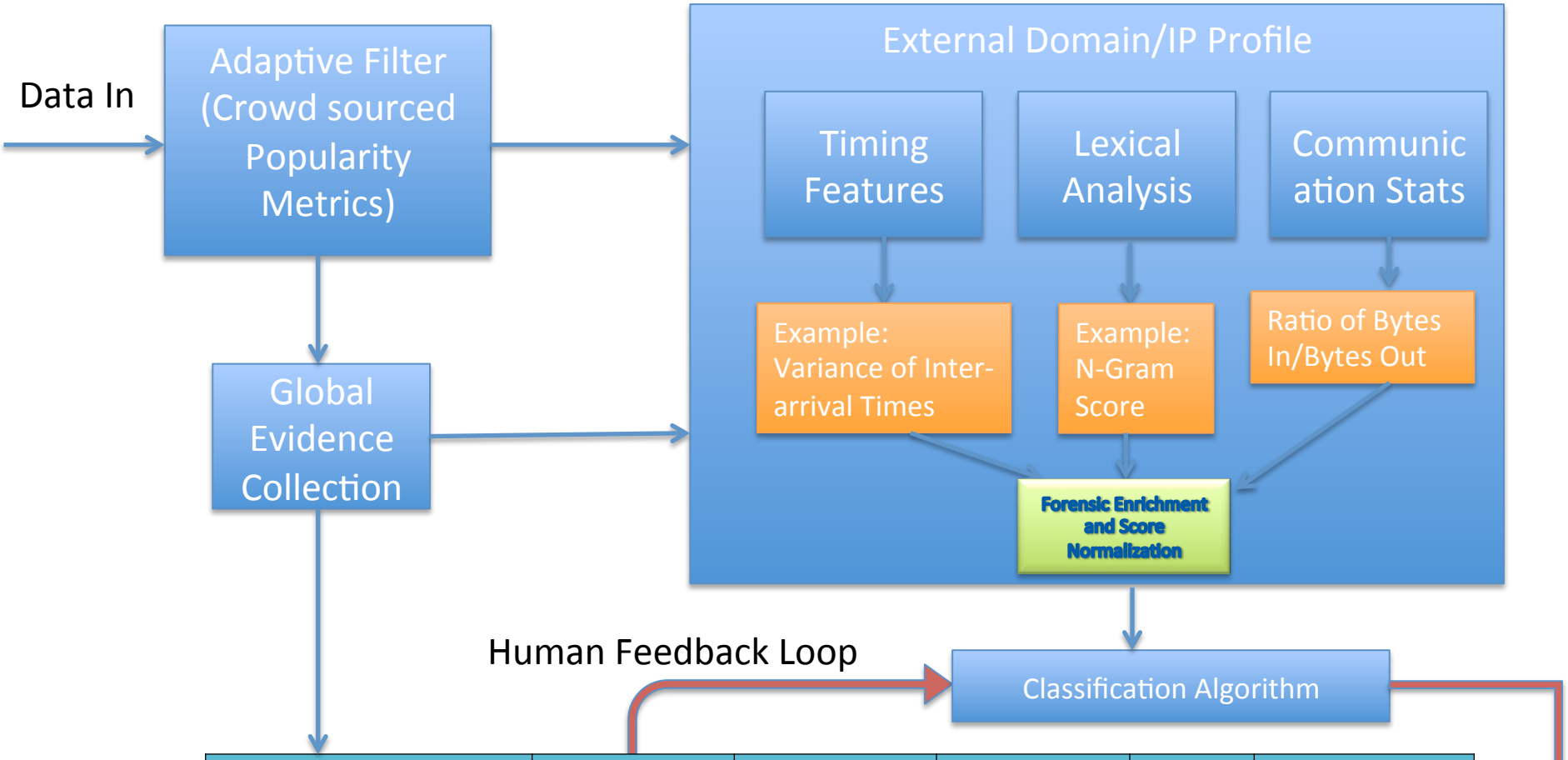




ML applied to Malware Research Dridex, Zeus

- Malware uses covert command and control techniques to evade detection
- Malware communication leaves footprints of anomalous behaviors
  - Domain Generation Algorithms
  - SSL command and control
  - Twitter/Facebook/Gmail based steganography
  - RFC Compliant DNS backdoor

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 Priorty: Communication is active and going unlbocked
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

Key to ML: Label Your Analysis

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

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

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

Sequential Behaviors: Exploit Chain

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" "" "-"

Sequencing data by account name is a great way to catch certain attacks over http data that are otherwise very expensive to compute downstream

Sequential Behaviors: Exploit Chain

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" "" "-”

Sequential Behaviors: Exploit Chain

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" "" "-"

Sequential Behaviors: Exploit Chain

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" "" "-"

Sequential Behaviors: Exploit Chain

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" "" "-"

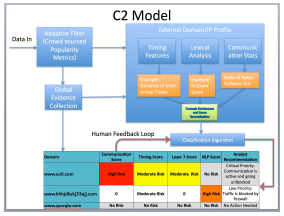
2. **Flash Exploit:** [29/Apr/2015:16:52:26 -0700] "Nico Rosberg" 192.168.122.177 69.162.78.253 1500 200 TCP_HIT "**GET http://portcullisesposturen.europartsplus.org/IMvOBBZKDLqAJYIDe02t5hMMNyzBLN_q4kafJkVNqJVTnTmd** HTTP/1.1" "Internet Services" "low risk" "**application/x-shockwave-flash**" 518 821 "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" "" "-"

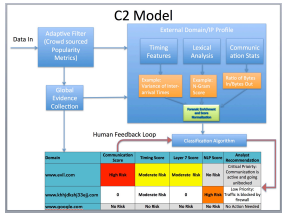
3. **Payload:** [29/Apr/2015:16:52:27 -0700] "Nico Rosberg" 192.168.122.177 69.162.78.253 1500 200 TCP_HIT "**GET http://portcullisesposturen.europartsplus.org/UX7n1YkbNn8FUV6QVtEZLj-p-gLvRKlWEWmz3r7Ug8suRiY_** HTTP/1.1" "Internet Services" "low risk" "**application/octet-stream**" 136 915 "" "" "-"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

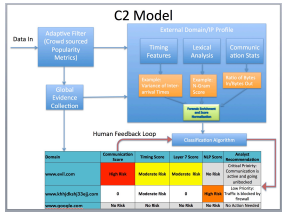
4. **Command and Control:** [29/Apr/2015:16:52:33 -0700] "Nico Rosberg" 192.168.122.177 104.28.28.165 1500 200 TCP_HIT "GET **http://dpckd2ftmf7lelsa.jjeyd2u37an30.com/tsdfewr2.php?U3ViamVj49MCZpc182ND0xJmlwPTIxMy4yMjkuODcuMjgmZXhlX3R5cGU9MQ==** HTTP/1.1" "Internet Services" "low risk" "**text/html; 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" "" "-"

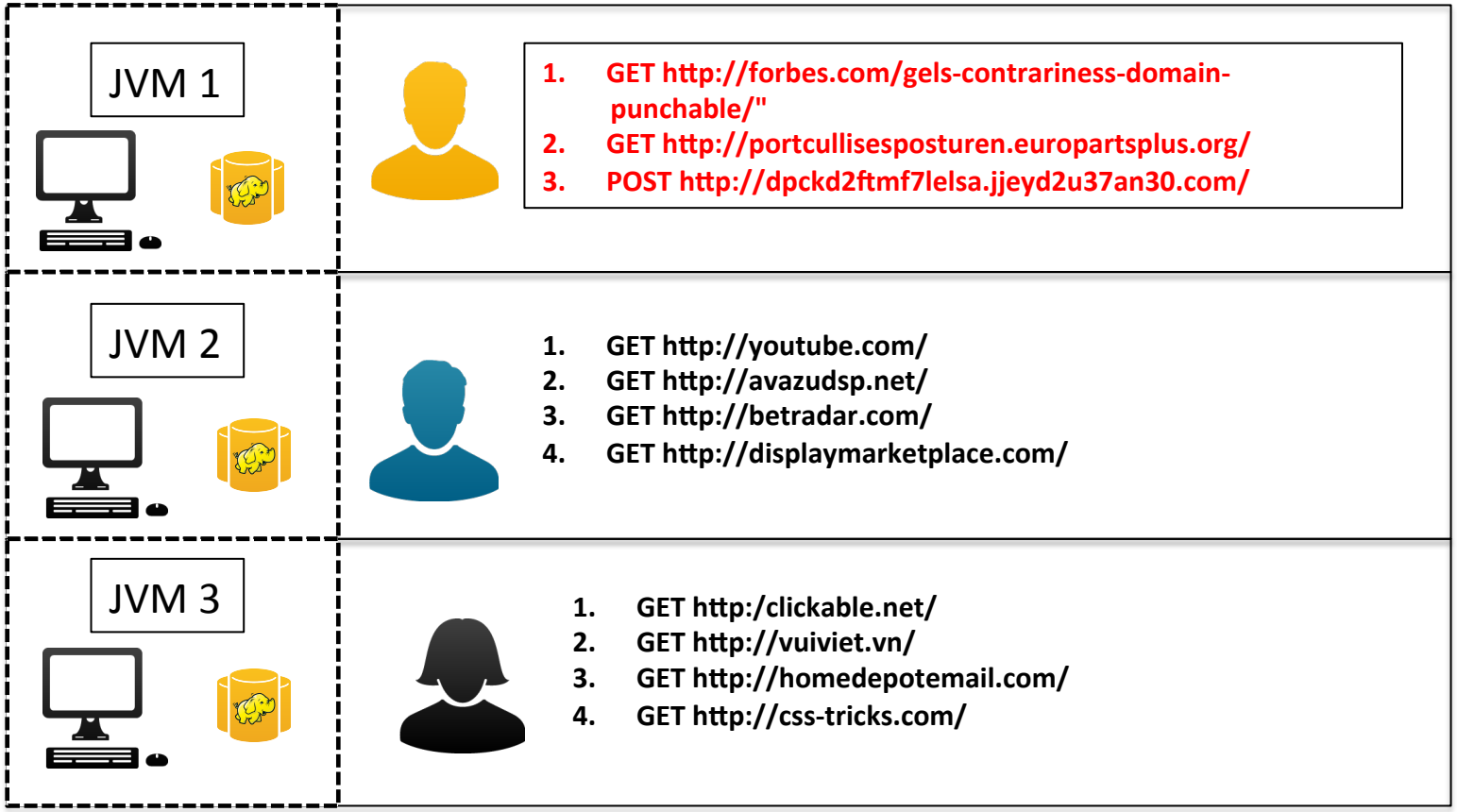
ML + Sequencing the Security DNA

- We parallelize across many nodes (JVMs) and use both real time and batch computations









Conclusion

- ML can potentially become a milestone technology in Cybersecurity
- Upcoming advances in hardware and distributed computing will accelerate development in ML: Lambda Security
- Need to industry standard to share behavioral indicators and labels
- NO SKYNET in the foreseeable future ☹

Thank you

- Rod Soto

  rsoto@splunk.com @rodsoto

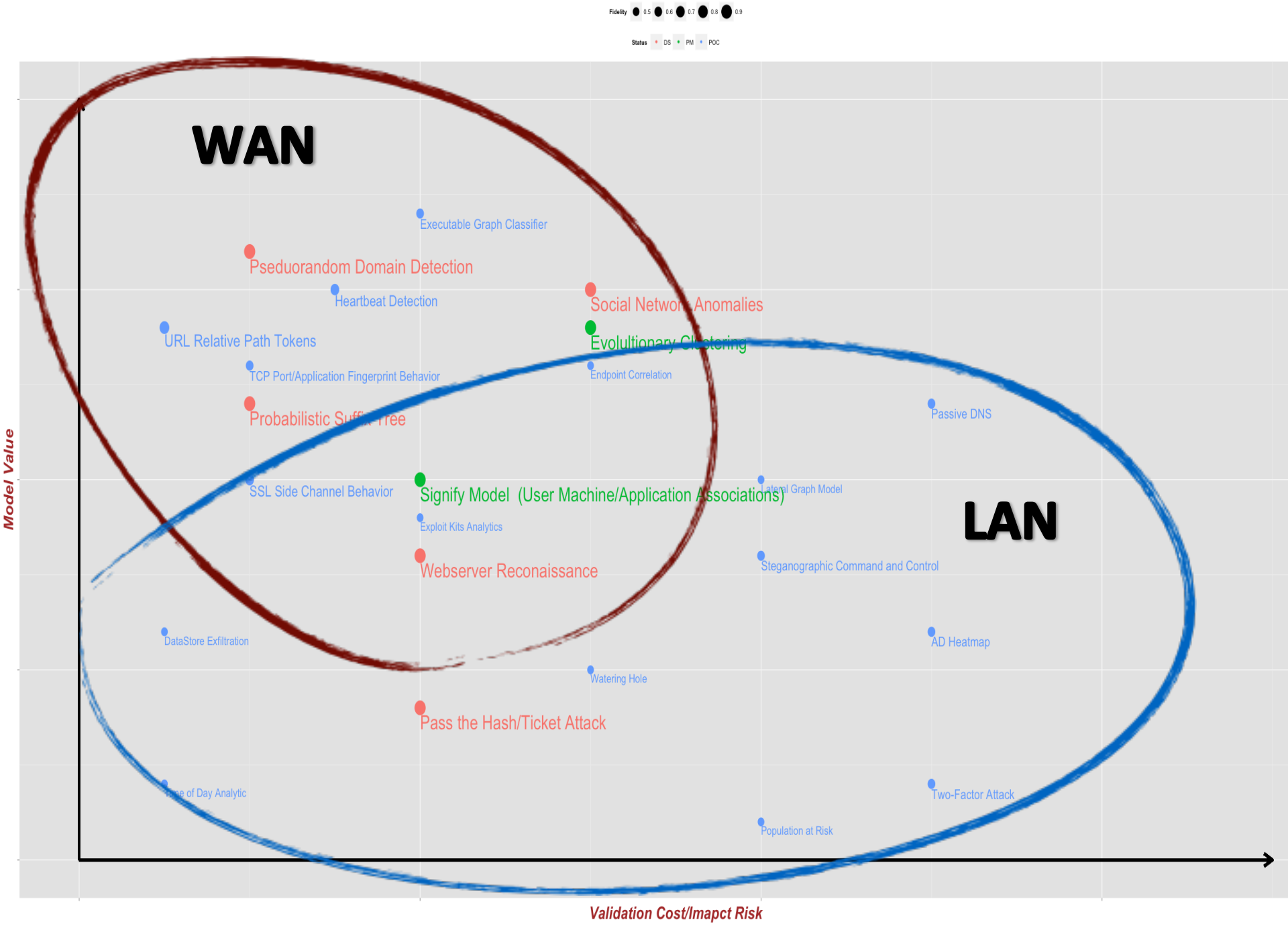
- Joseph Zadeh

  jzadeh@splunk.com @josephzadeh

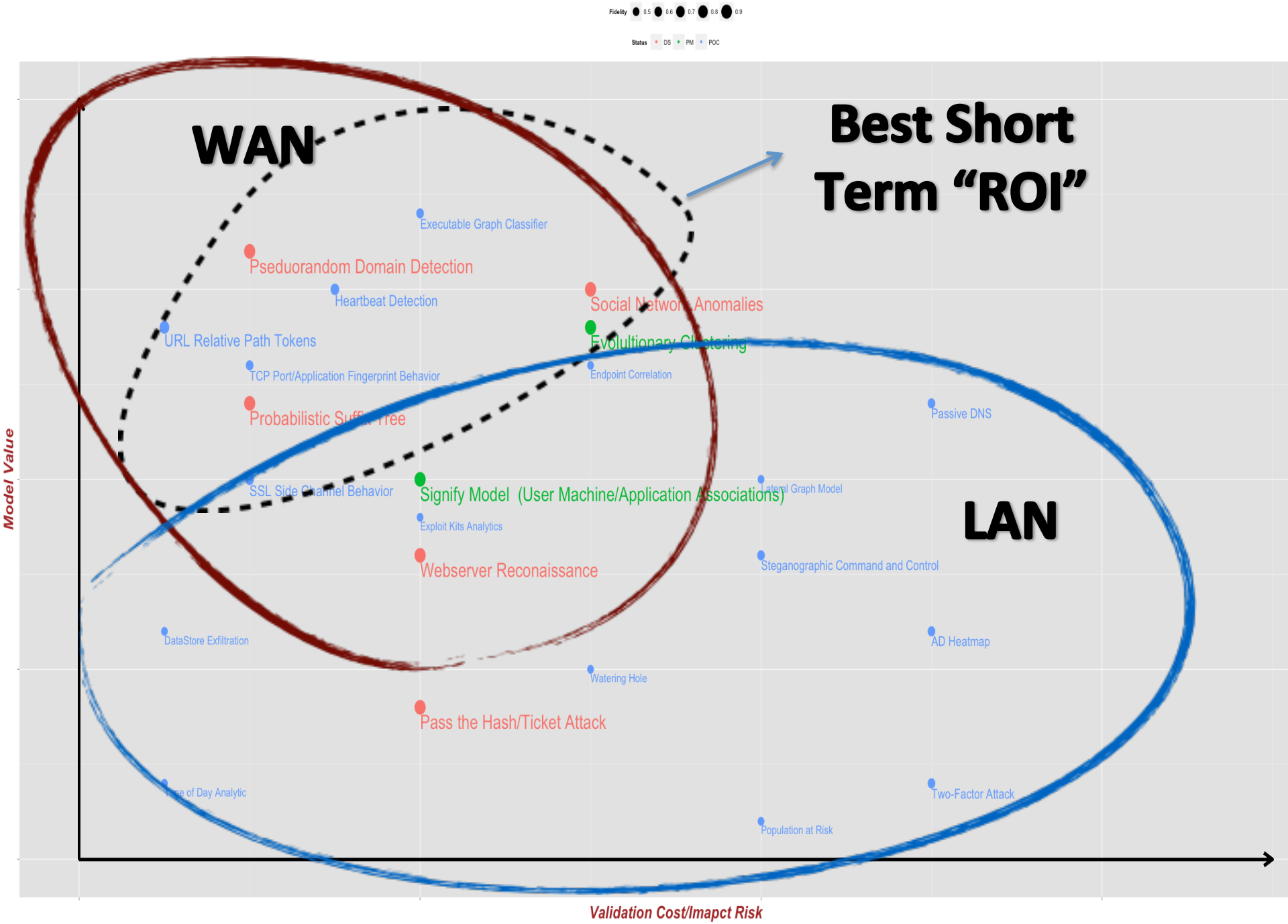
Appendix

Model Value vs. Total Cost of Validation and Impact Risk

Model Value vs. Total Cost of Validation and Impact Risk



Model Value vs. Total Cost of Validation and Impact Risk

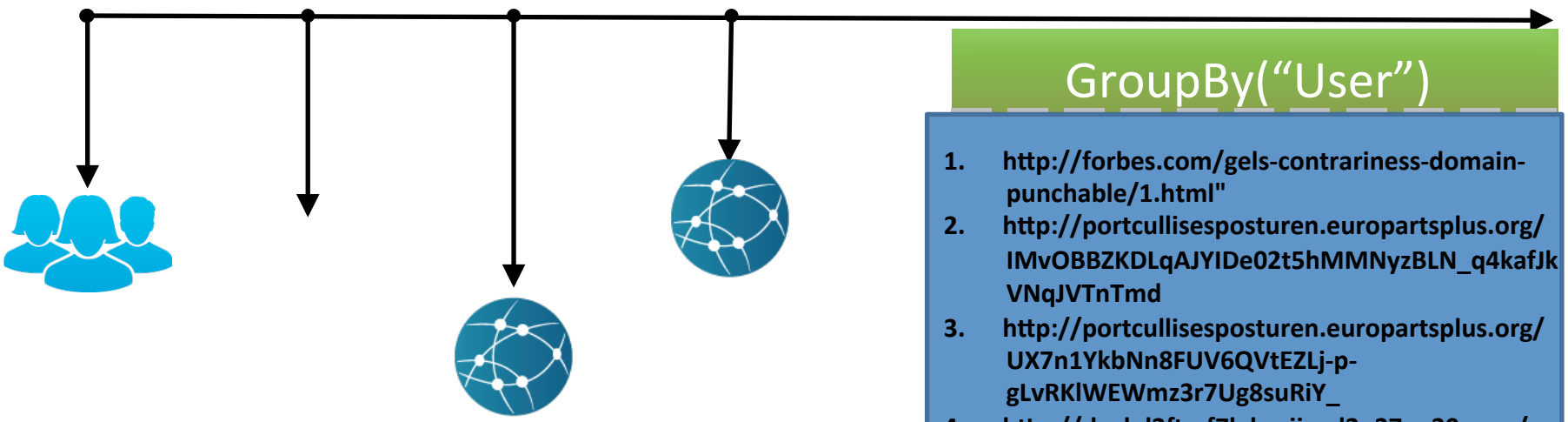


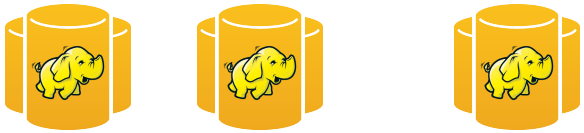
Lambda Firewalls?!

Manage the paths accordingly start building lambda workflows into Everything!!!

- Lambda firewall
  - Statistical whitelist computation aspect (fuzzy ACL's)
  - Path for signatures and sequential behaviors that is more expressive than PCRE
- Central nervous system approach to blending signals
  - Defense should scale up and down the size of organization: a properly engineered central nervous system should be able to protect SMB market as well as large scale deployments
- Difference between a classical firewall and a lambda firewall

Parallel Sequencing of Behaviors



GroupBy("User")

1. http://forbes.com/gels-contrariness-domain-punchable/1.html"
2. http://portcullisesposturen.europartsplus.org/IMvOBBZKDLqAJYIDe02t5hMMNyzBLN_q4kafJkVNqJVTnTmd
3. http://portcullisesposturen.europartsplus.org/UX7n1YkbNn8FUV6QVtEZLj-p-gLvRKlWEWmz3r7Ug8suRiY_
4. http://dpckd2ftmf7lelsa.jjeyd2u37an30.com/tsdfewr2.php?U3ViamVj49MCZpc182ND0xJmlwPTIxMy4yMjkuODcuMjgmZXhlX3R5cGU9MQ==