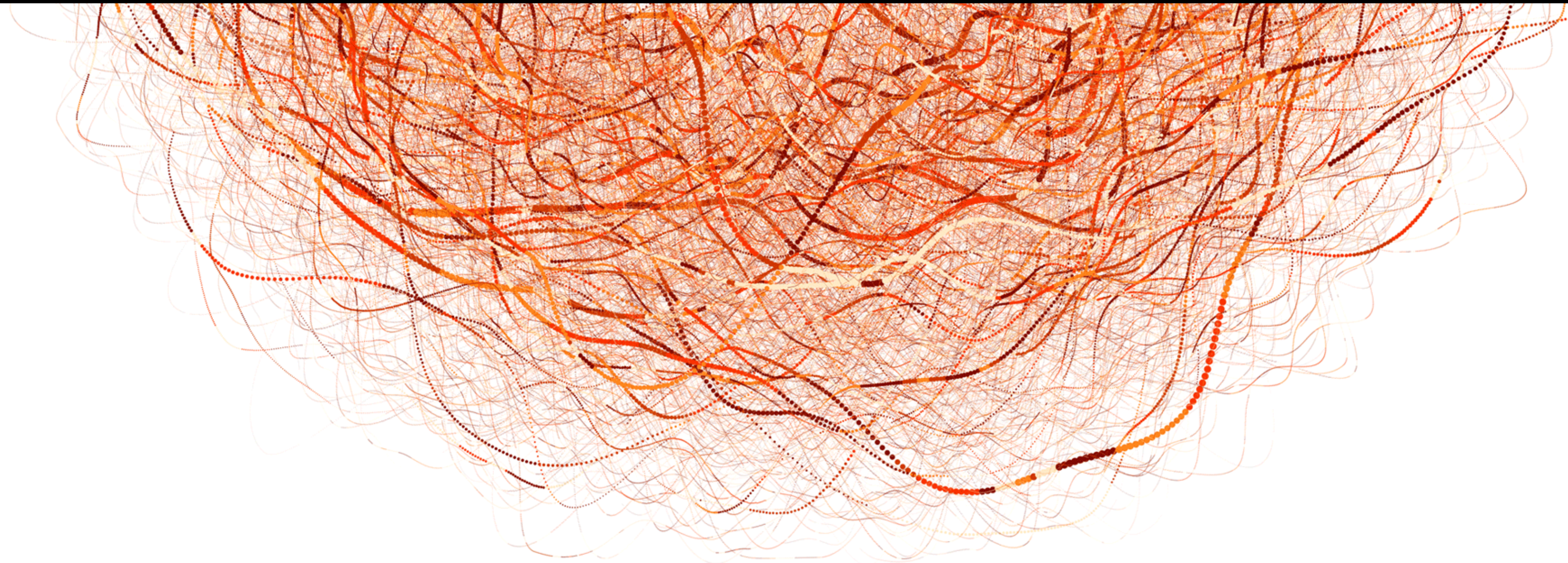


Optimal Machine Learning Algorithms

for Cyber Threat Detection

FloCon 2018

A Presentation by Hafiz Farooq, Saudi Aramco

Hafiz Farooq

Senior Cyber Security Consultant, Saudi Aramco

ECC (EXPEC Computer Center) SOC

MS Data Communication Networks, Aston University, United Kingdom
BE Computer Engineering, NUST, Pakistan
DELL Secureworks - Worked as Senior SOC Architect
SANS Forensic Examiner, SANS Exploit Researcher
Splunk Big Data Architect, Qradar Deployment Professional
Juniper Networks – JNCIE Security and JNCIP-Service Provider Routing

Why we moved to Machine Learning

- Post-Shamoon Scenario
- Machine Learning vs Orthodox Cyber Security
- Big Data Analytics & Machine Learning

STATISTICAL APPROACH

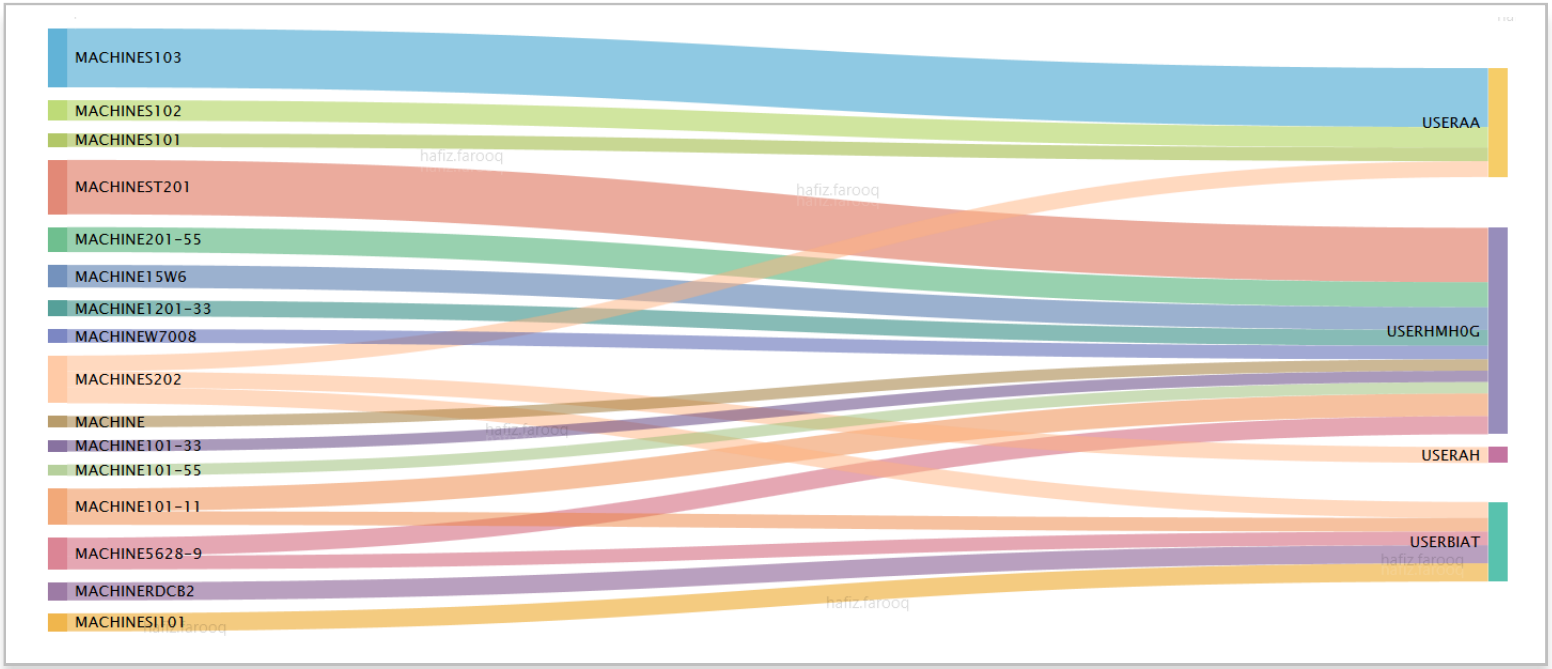
MACHINE
LEARNING

Optimal Machine Learning Algorithms

for Cyber Security

ANOMALY DETECTION – PRIVILEGED ACCOUNTS

BIG DATA STATISTICAL ANALYSIS



Feature Space: MachineID, UserID, EventCount, Severity, Multihoming

SANKEY VISUALIZATION

http://www.sankey-diagrams.com/

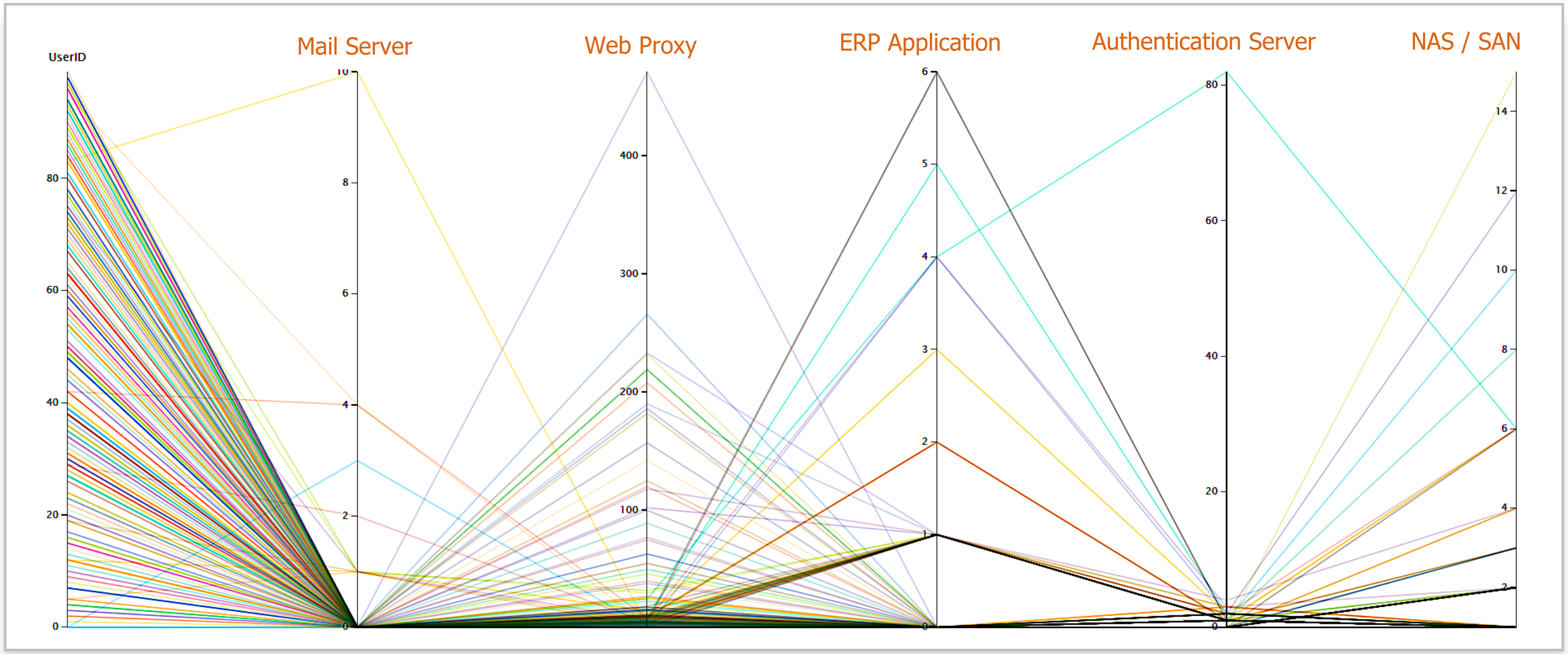
QUERY

```
source=windows AND ( usertype=Administrator*  OR usertype=root*)
| stats count by host user
| sort count desc
| head 20
```

ANOMALY DETECTION – TOP TALKERS

BIG DATA STATISTICAL ANALYSIS

n-dimensional feature space & n-parallels



PARALLEL COORDINATES

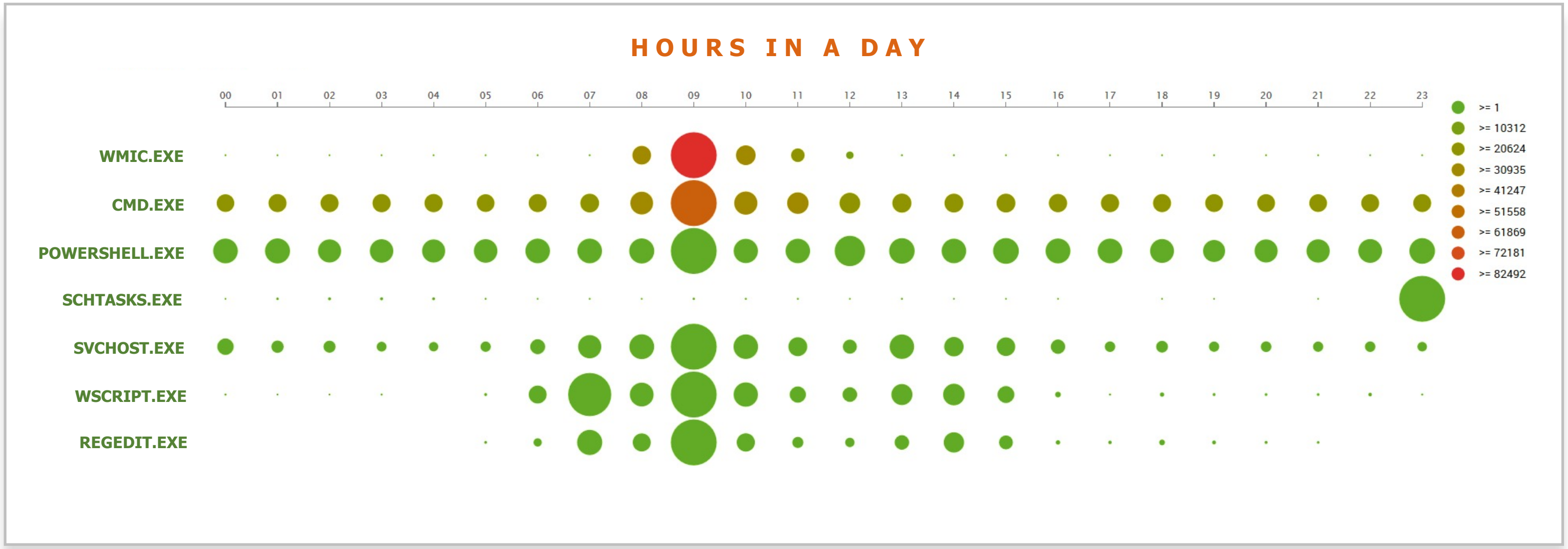
https://datavizcatalogue.com/methods/parallel_coordinates.html

QUERY

index*=firewall dest=Authentication Server |* ***stats*** *count by src*
| ***appendcols*** *[search index=juniper dest=Mail Server | stats count by src*
| appendcols [search index=juniper dest=NAS/SAN | stats count by src
| appendcols [search index=juniper dest=ERP | stats count by src
| appendcols [search index=juniper dest=Web | stats count by src

ANOMALY DETECTION – CRITICAL PROCESSES

BIG DATA STATISTICAL ANALYSIS



Discrete / Continuous Time Series Analytics

PUNCHCARD VISUALIZATION

http://bl.ocks.org/kaezarrex/10122633

QUERY

```
index=wineventlog AND (New_Process_Name IN (*\\powershell*, *\\wscript* ,*\\wmic* ,*\\svchost*,*\\regedit*, *\\cmd.*)
| eval WorkTime=strftime(_time,"%H")
| rex field=New_Process_Name ".*\\\(?<executable>.*)$"
| stats count by WorkTime executable
```

OPTIMAL ML ALGORITHMS

MACHINE

LEARNING

Optimal Machine Learning Algorithms

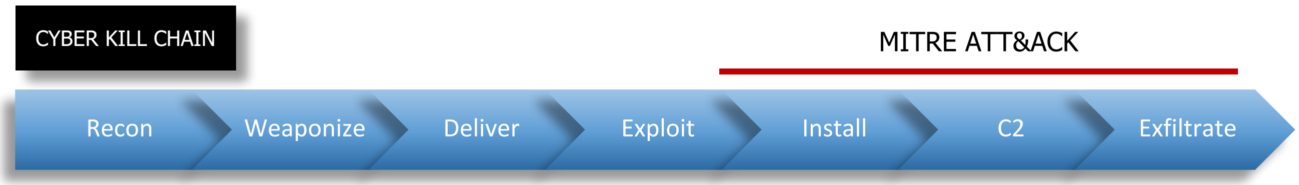
for Cyber Security

أرامكو السعودية
saudi aramco

Standards Used for ML based Threat Detection

CYBER THREAT STANDARDIZATION

- MITRE Standards for Post-Compromise Detection
  - ATT&CK | Adversarial Tactics, Techniques, and Common Knowledge
  - CAPEC | Common Attack Pattern Enumerations and Classification
  - MAEC | Malware Attribute Enumeration and Characterization
- Lockheed Martin's Cyber Kill Chain



CYBER KILL CHAIN

MITRE ATT&ACK

Recon → Weaponize → Deliver → Exploit → Install → C2 → Exfiltrate

MITRE ATT&CK CATEGORIES

Persistence	Privilege Escalation	Defense Evasion	Credential Access	Discovery
Lateral Movement	Executions	Collection	Exfiltration	Command & Control

IMPORTANT USE CASES

BASED ON MITRE ATT&CK MATRIX

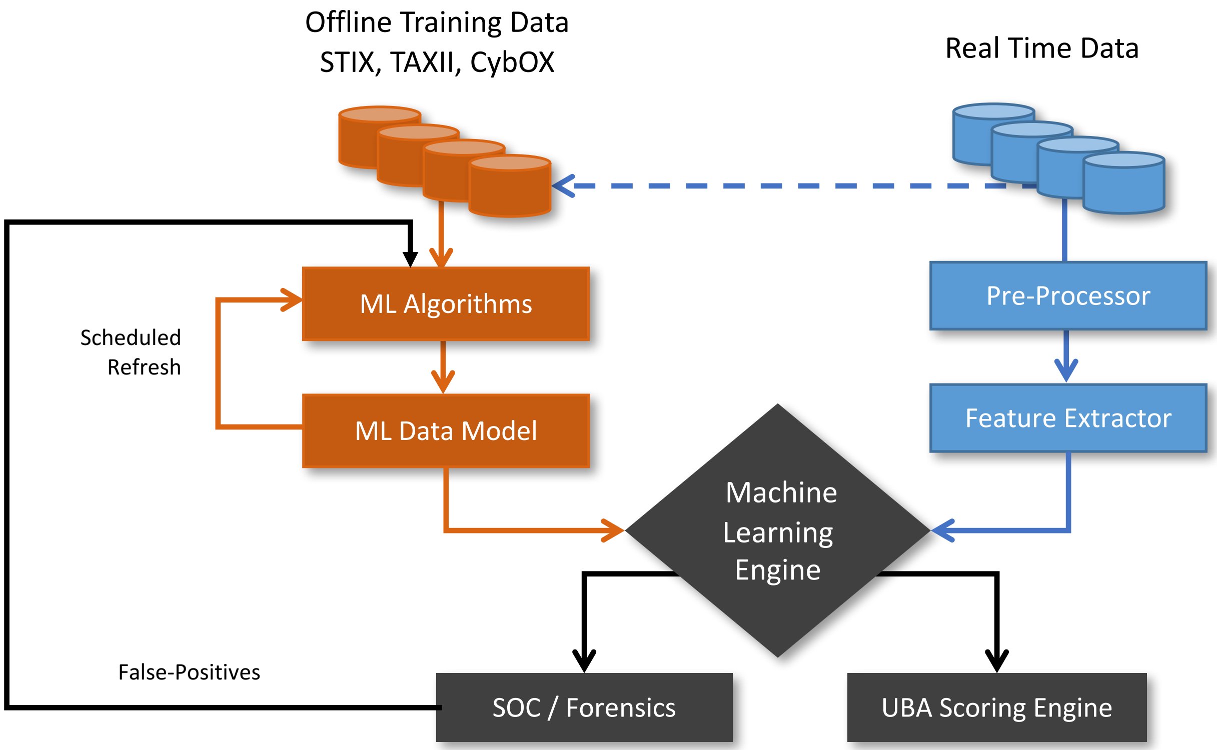
Threat Use Cases	Pre-Processing	ML based Detector Algorithms	ATT&CK Category
Exfiltration over C2 Channels	Standard Scaler / PCA	KMeans / X-Means	Exfiltration
Service Scanning Analysis	PCA, KMeans	Linear, RF, DT Regressors	Discovery
PowerShell Anomaly Detection	PCA	One-Class SVM with Linear Kernel	Execution
DLL Injection Anomaly Detection	PCA/Kernel-PCA	One-Class SVM with Linear Kernel	Privilege Escalation
Process Hollowing via System Calls	TFIDF (Logarithmic)	LR with SGD Detector	Defense Evasion
Web URLs Analysis	Levenshtein Distance	Shannon Entropy	Command & Control
Email Spam Classification	TFIDF	RF Classifier	Execution
Analyzing Web Proxy Logs	BM25	SGD with Naïve Bayesian	Command & Control

MITRE ATT&CK *https://attack.mitre.org/wiki*

Persistence	Privilege Escalation	Defense Evasion	Credential Access	Discovery
Lateral Movement	Executions	Collection	Exfiltration	Command & Control

Machine Learning Workflow

CYBER THRET DETECTION & MACHINE LEARNING



SUPERVISED & UNSUPERVISED WORKFLOWS

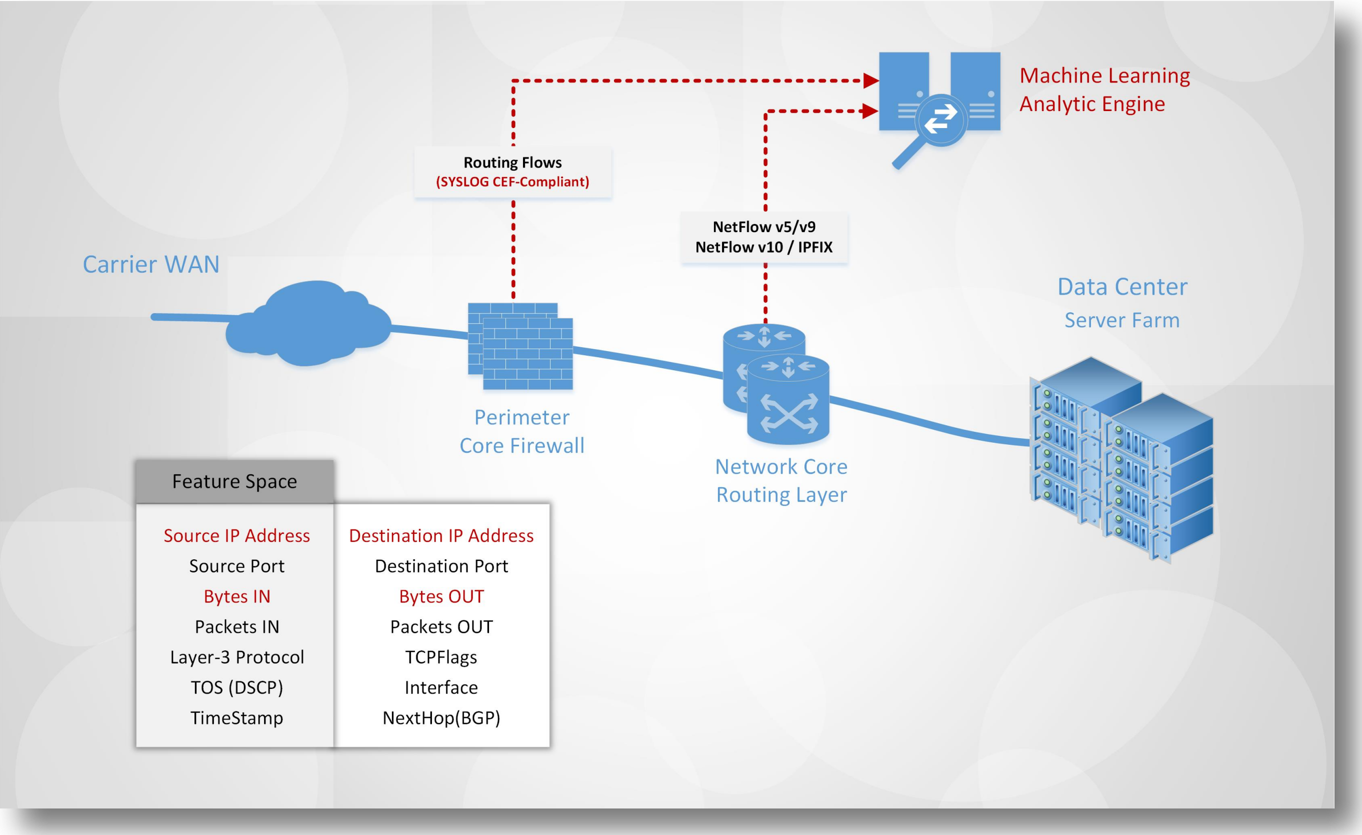
Curses of Dimensionality in Cyber Security ML

FEATURE ENGINEERING & BAGGING

- Feature Engineering is Critical in Cyber Security
- More Categorical Data than Numerical
- Important Algorithms
  - Feature Extraction | PCA/Kernel-PCA, TF-IDF/BM25
  - Normalization | StandardScaler (Z-Score), Normalizer (Min-Max)
  - Feature Selection |Sampling, SubSampling, OverSampling, KMeans

Upload/Download Analytic using Numerical Clustering

MACHINE LEARNING – USE CASE NO - 1



Feature Space	
Source IP Address	Destination IP Address
Source Port	Destination Port
Bytes IN	Bytes OUT
Packets IN	Packets OUT
Layer-3 Protocol	TCPFlags
TOS (DSCP)	Interface
TimeStamp	NextHop(BGP)

K-Means Clusters

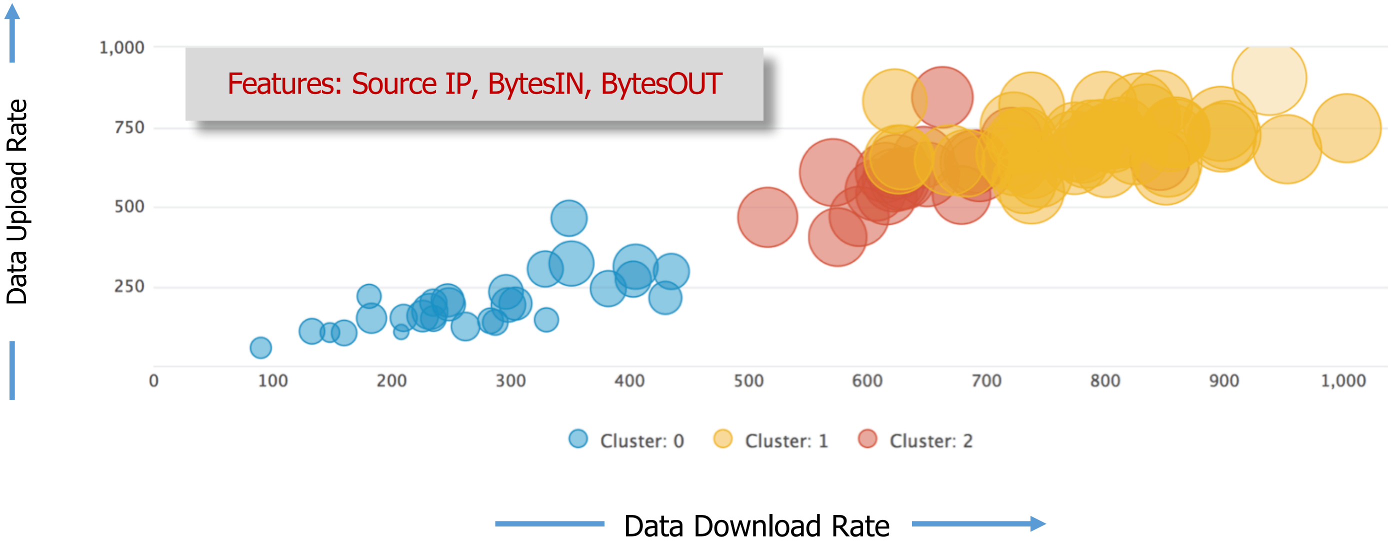
MacQueen, 1976: Some Methods for Classification and Analysis of Mulivariate Observations.

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - \mu_k||^2$$

Complexity: O(n . k . Iterations . Attributes)

Upload/Download Analytic using Numerical Clustering

MACHINE LEARNING – USE CASE NO - 1



Features: Source IP, BytesIN, BytesOUT

Firewall Netflow / RT Stats

Feature PreProcess

Standard Scaler/PCA

KMeans Clustering (k=3)

K-Means Clusters

MacQueen, 1976: Some Methods for Classification and Analysis of Mulivariate Observations.

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - \mu_k||^2$$

Complexity: O(n . k . Iterations . Attributes)

Upload/Download Analytic using Numerical Clustering

MACHINE LEARNING – USE CASE NO - 1

- K-Means creates clusters of homogeneous shapes and much faster than hierarchical clustering techniques
- DBSCAN is less accurate here due to the dynamically varying traffic densities and highly scattered data values
- BIRCH clustering is very slow for larger datasets and hence only limited to micro-level clustering, in conjunction with a macro-level algorithm

Clustering Algorithms

Chakraborty, Sanjay, "Performance Comparison of Incremental k-Means and DBScan."

BIRCH

$$N, LS = \sum_{i=1}^{N} X_i, \; SS = \sum_{j=1}^{N} X_j^2$$

DBSCAN

$$N_\varepsilon(p) : \{ \, q | d(p,q) \leq \varepsilon \, \}$$

KMeans

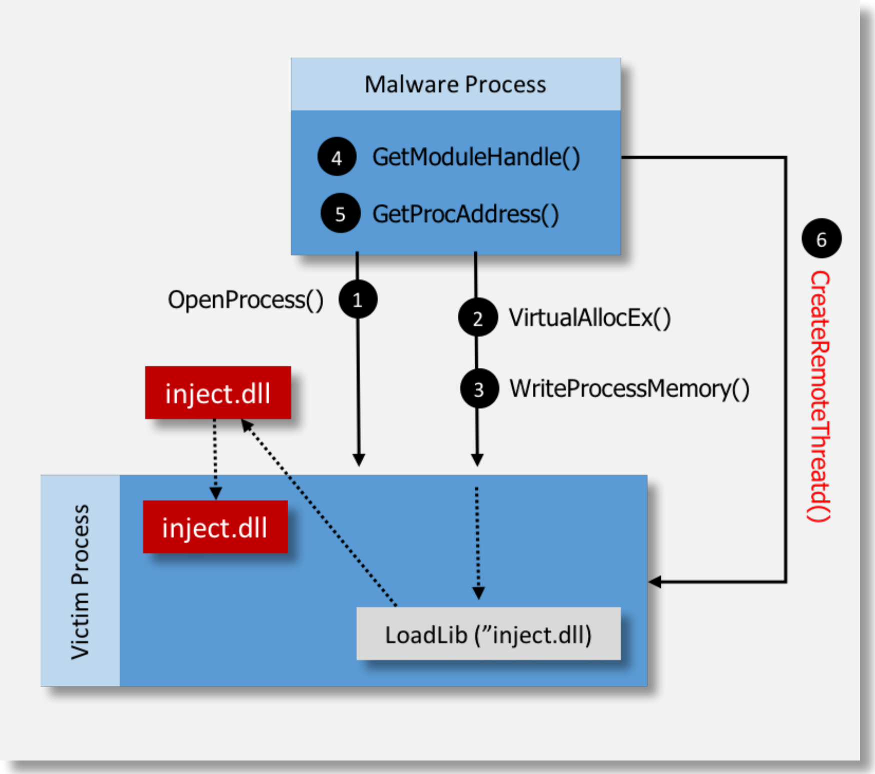
$$J = \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - \mu_k||^2$$

DLL Injection Detection using OneClassSVM (OSVM)

MACHINE LEARNING – USE CASE NO – 2

SYSMON Events	
1	Process Create
2	File Creation Time
3	Network Connection
5	Process Terminated
6	Driver Loaded
7	Image Loaded
8	CreateRemoteThread

```
HANDLE WINAPI CreateRemoteThread(
  _In_  HANDLE                 hProcess,
  _In_  LPSECURITY_ATTRIBUTES  lpThreadAttributes,
  _In_  SIZE_T                 dwStackSize,
  _In_  LPTHREAD_START_ROUTINE lpStartAddress,
  _In_  LPVOID                 lpParameter,
  _In_  DWORD                  dwCreationFlags,
  _Out_ LPDWORD                lpThreadId
);
```



SYSMON Events

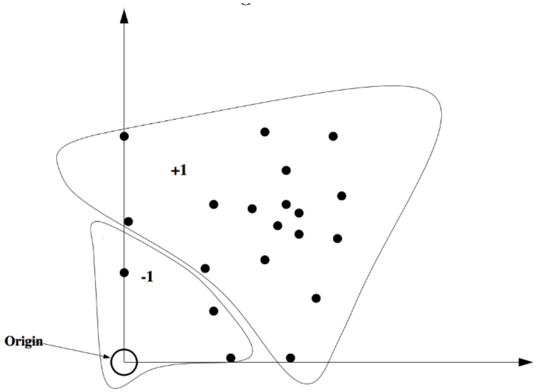
Reference: https://docs.microsoft.com/en-us/sysinternals/downloads/sysmon

QUERY

```
index=sysmon-events EventID=8
sourcetype="XmlWinEventLog:Microsoft-Windows-Sysmon/Operational"
| table host _time, SourceImage, TargetImage
```

Detect DLL Injection using OneClassSVM (OSVM)

MACHINE LEARNING – USE CASE NO - 2



$$min\frac{1}{2}\|w\|^2 + \frac{1}{vl}\sum_i \xi_i - \rho$$

subject to $(w.\Phi(x_i)) \geq \phi - \xi_i, \xi_i \geq 0.$

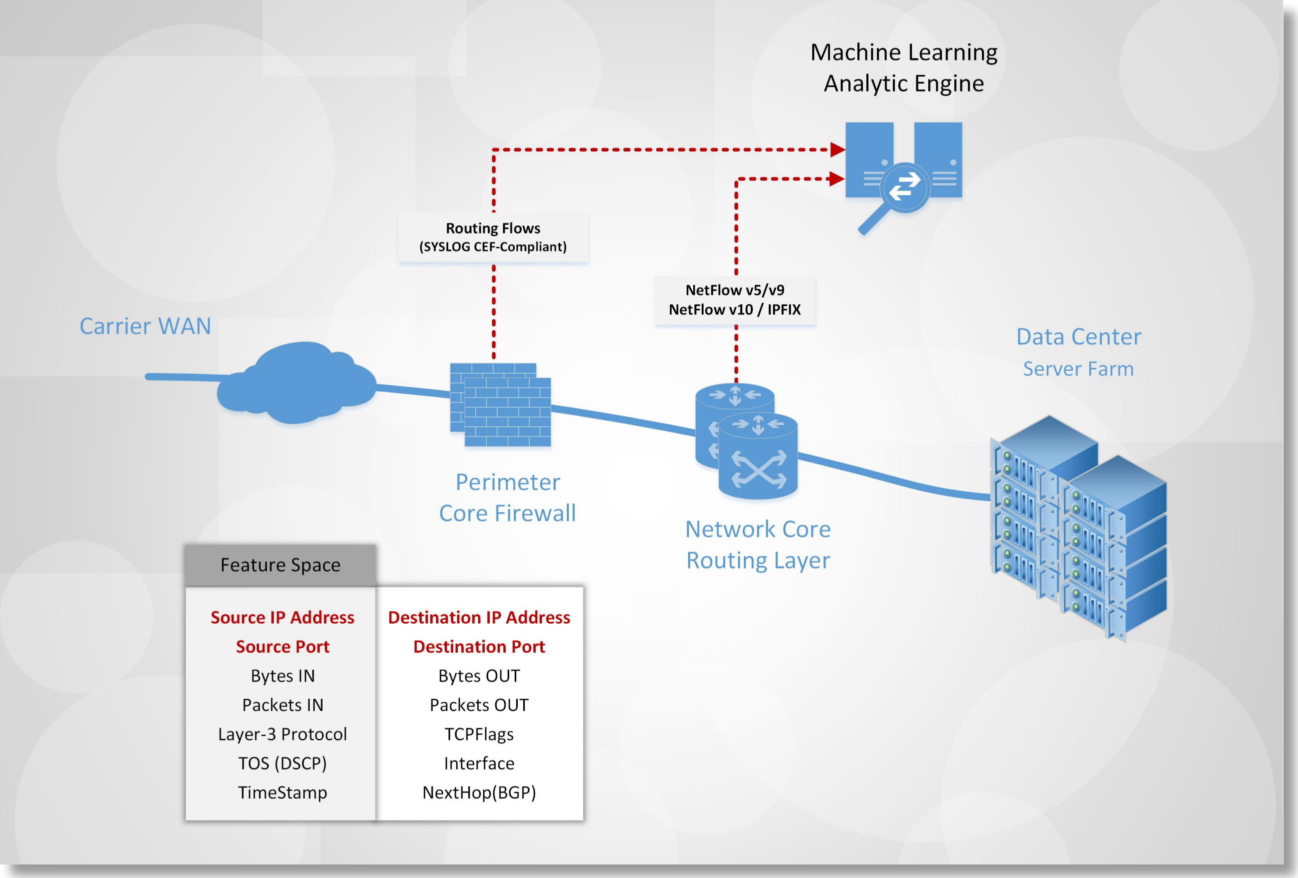
```
DataSource: SYSMON-Logs
if EventID == 8 AND isNormal != 1 then
    do OneClassSVM Source, Target
    set kernel = linear nu = 0.01 coef = 0.5
    set gamma = 0.01 tol = 1 deg = 3 shrinking = f
    save model CreateRemoteThreatOSVM
    do deup Source Target
end if
```

One-Class SVM

Bernhard Schölkopf, "One-Class Support Measure Machines for Group Anomaly Detection"

Detecting Recon using Numerical Prediction

MACHINE LEARNING – USE CASE NO - 3

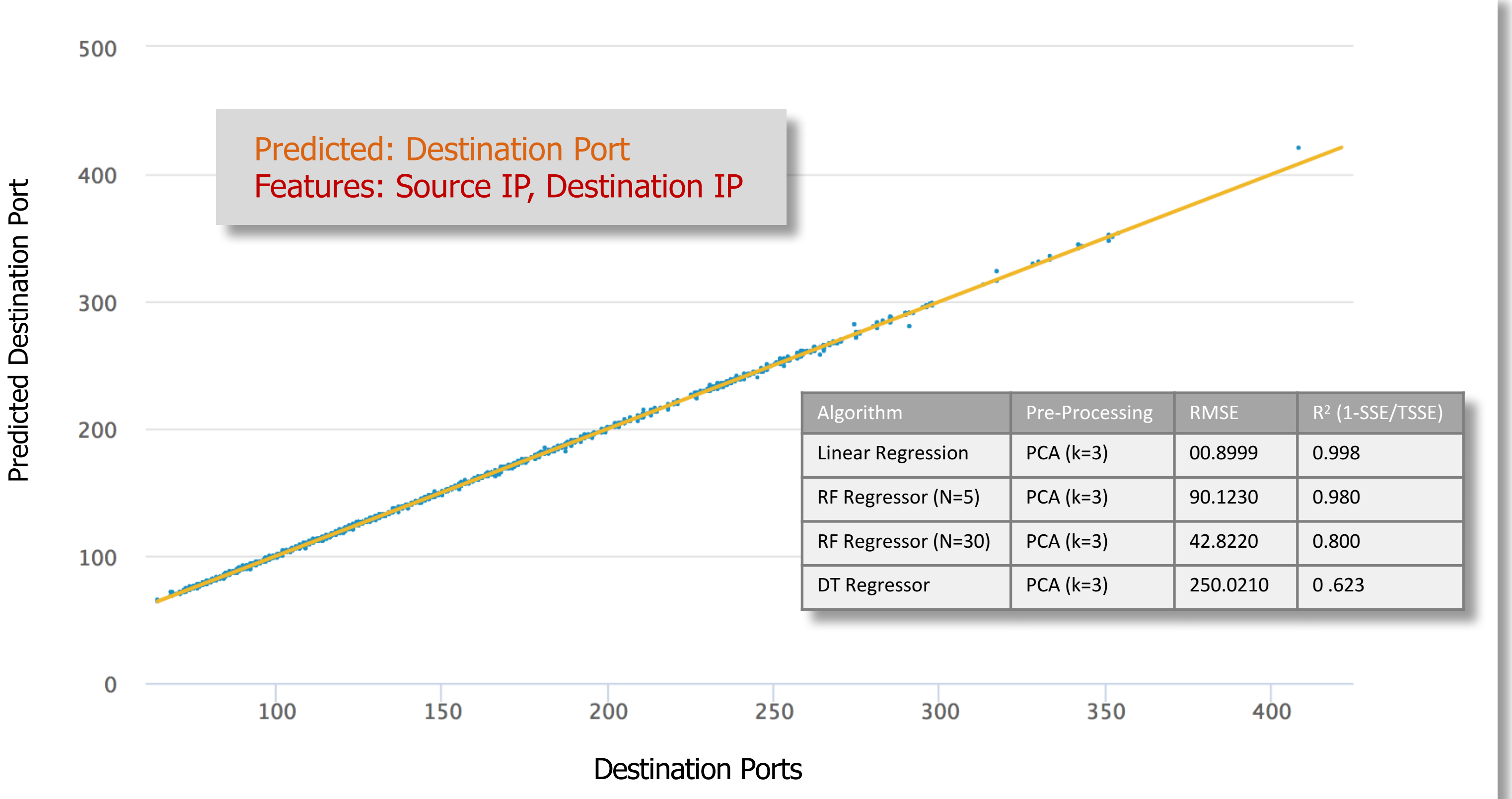


Feature Space	
Source IP Address	**Destination IP Address**
Source Port	**Destination Port**
Bytes IN	Bytes OUT
Packets IN	Packets OUT
Layer-3 Protocol	TCPFlags
TOS (DSCP)	Interface
TimeStamp	NextHop(BGP)

Regression / Prediction

Detecting Recon using Numerical Prediction

MACHINE LEARNING – USE CASE NO - 3



Predicted: Destination Port
Features: Source IP, Destination IP

Algorithm	Pre-Processing	RMSE	R^2 (1-SSE/TSSE)
Linear Regression	PCA (k=3)	00.8999	0.998
RF Regressor (N=5)	PCA (k=3)	90.1230	0.980
RF Regressor (N=30)	PCA (k=3)	42.8220	0.800
DT Regressor	PCA (k=3)	250.0210	0 .623

Numerical Prediction

Linear Regression, Random Forest Regressor, DecisionTree Regressor, LASSO

Detecting Recon Anomaly using Numerical Prediction

MACHINE LEARNING – USE CASE NO - 3

- Logistic Regression (LR) worked well here due to linear dataset and due to the absence of multicollinearity between the independent predictor variables (i.e. time, source, destination).
- RandomForest Ensemble Algorithm (with multiple tree estimators) is also an ideal predictor for this analysis being relatively more accurate on relatively weaker training set.
- DecisionTree required very accurate training set, so was not suitable here.

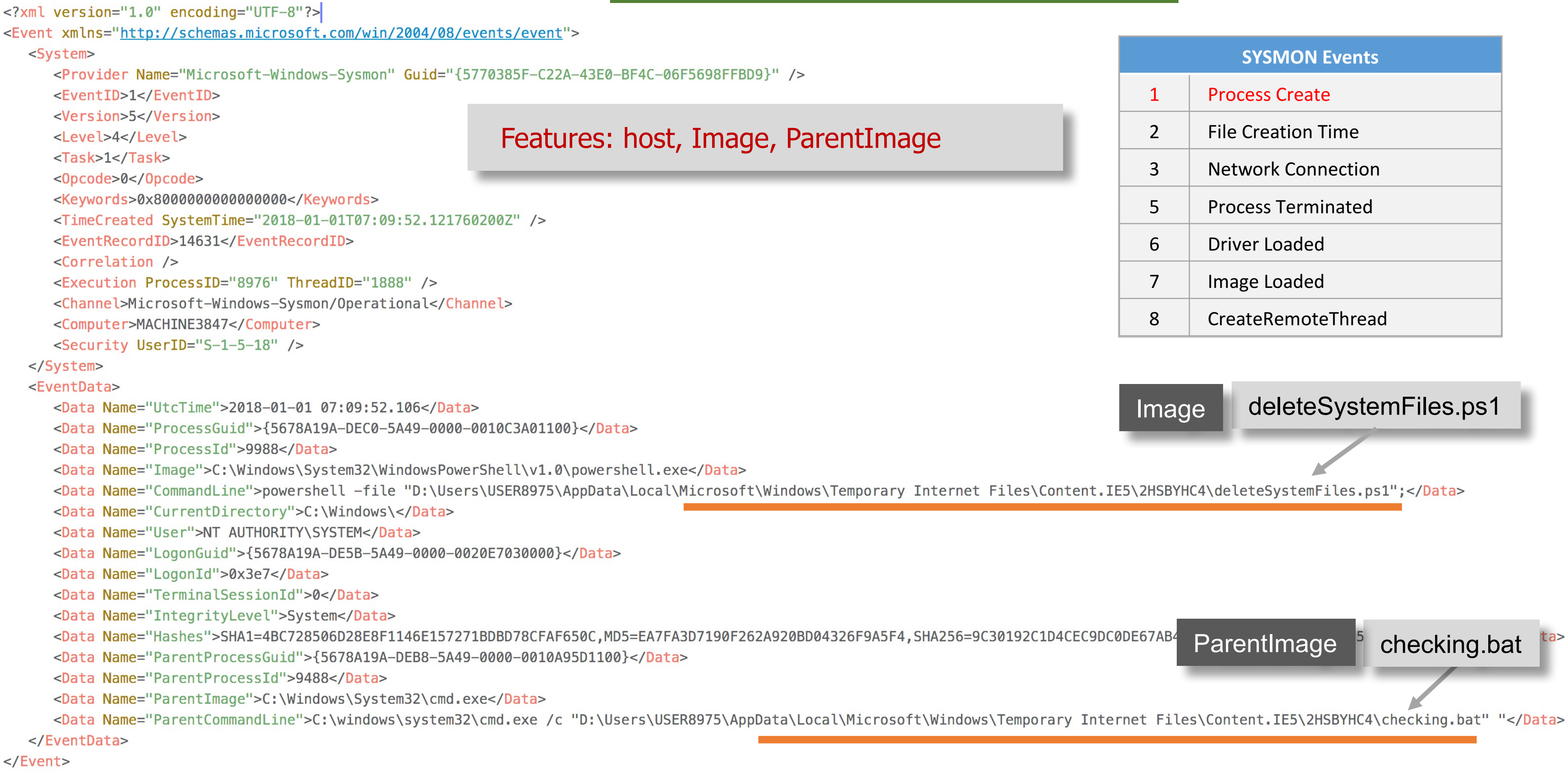
Linear Regression

Bernhard Schölkopf, "One-Class Support Measure Machines for Group Anomaly Detection"

PowerShell Anomaly Detection using OneClassSVM

MACHINE LEARNING – USE CASE NO - 4

Features: host, Image, ParentImage



```
<?xml version="1.0" encoding="UTF-8"?>
<Event xmlns="http://schemas.microsoft.com/win/2004/08/events/event">
   <System>
      <Provider Name="Microsoft-Windows-Sysmon" Guid="{5770385F-C22A-43E0-BF4C-06F5698FFBD9}" />
      <EventID>1</EventID>
      <Version>5</Version>
      <Level>4</Level>
      <Task>1</Task>
      <Opcode>0</Opcode>
      <Keywords>0x8000000000000000</Keywords>
      <TimeCreated SystemTime="2018-01-01T07:09:52.121760200Z" />
      <EventRecordID>14631</EventRecordID>
      <Correlation />
      <Execution ProcessID="8976" ThreadID="1888" />
      <Channel>Microsoft-Windows-Sysmon/Operational</Channel>
      <Computer>MACHINE3847</Computer>
      <Security UserID="S-1-5-18" />
   </System>
   <EventData>
      <Data Name="UtcTime">2018-01-01 07:09:52.106</Data>
      <Data Name="ProcessGuid">{5678A19A-DEC0-5A49-0000-0010C3A01100}</Data>
      <Data Name="ProcessId">9988</Data>
      <Data Name="Image">C:\Windows\System32\WindowsPowerShell\v1.0\powershell.exe</Data>
      <Data Name="CommandLine">powershell -file "D:\Users\USER8975\AppData\Local\Microsoft\Windows\Temporary Internet Files\Content.IE5\2HSBYHC4\deleteSystemFiles.ps1";</Data>
      <Data Name="CurrentDirectory">C:\Windows\</Data>
      <Data Name="User">NT AUTHORITY\SYSTEM</Data>
      <Data Name="LogonGuid">{5678A19A-DE5B-5A49-0000-0020E7030000}</Data>
      <Data Name="LogonId">0x3e7</Data>
      <Data Name="TerminalSessionId">0</Data>
      <Data Name="IntegrityLevel">System</Data>
      <Data Name="Hashes">SHA1=4BC728506D28E8F1146E157271BDBD78CFAF650C,MD5=EA7FA3D7190F262A920BD04326F9A5F4,SHA256=9C30192C1D4CEC9DC0DE67AB4 5 ta>
      <Data Name="ParentProcessGuid">{5678A19A-DEB8-5A49-0000-0010A95D1100}</Data>
      <Data Name="ParentProcessId">9488</Data>
      <Data Name="ParentImage">C:\Windows\System32\cmd.exe</Data>
      <Data Name="ParentCommandLine">C:\windows\system32\cmd.exe /c "D:\Users\USER8975\AppData\Local\Microsoft\Windows\Temporary Internet Files\Content.IE5\2HSBYHC4\checking.bat" "</Data>
   </EventData>
</Event>
```

Image deleteSystemFiles.ps1

ParentImage checking.bat

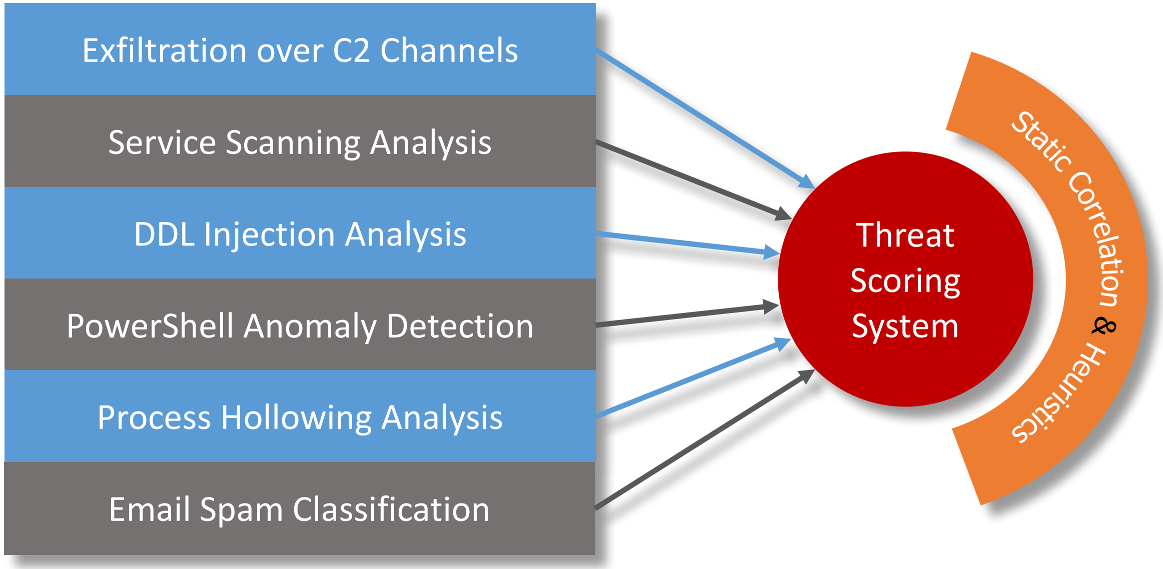
SYSMON Events	
1	Process Create
2	File Creation Time
3	Network Connection
5	Process Terminated
6	Driver Loaded
7	Image Loaded
8	CreateRemoteThread

One-Class SVM

Bernhard Schölkopf, "One-Class Support Measure Machines for Group Anomaly Detection"

User Behavioral Model

Machine Learning & Static Correlation


Exfiltration over C2 Channels
Service Scanning Analysis
DDL Injection Analysis
PowerShell Anomaly Detection
Process Hollowing Analysis
Email Spam Classification
Threat Scoring System
Static Correlation & Heuristics

Distributed Machine Learning Detection System

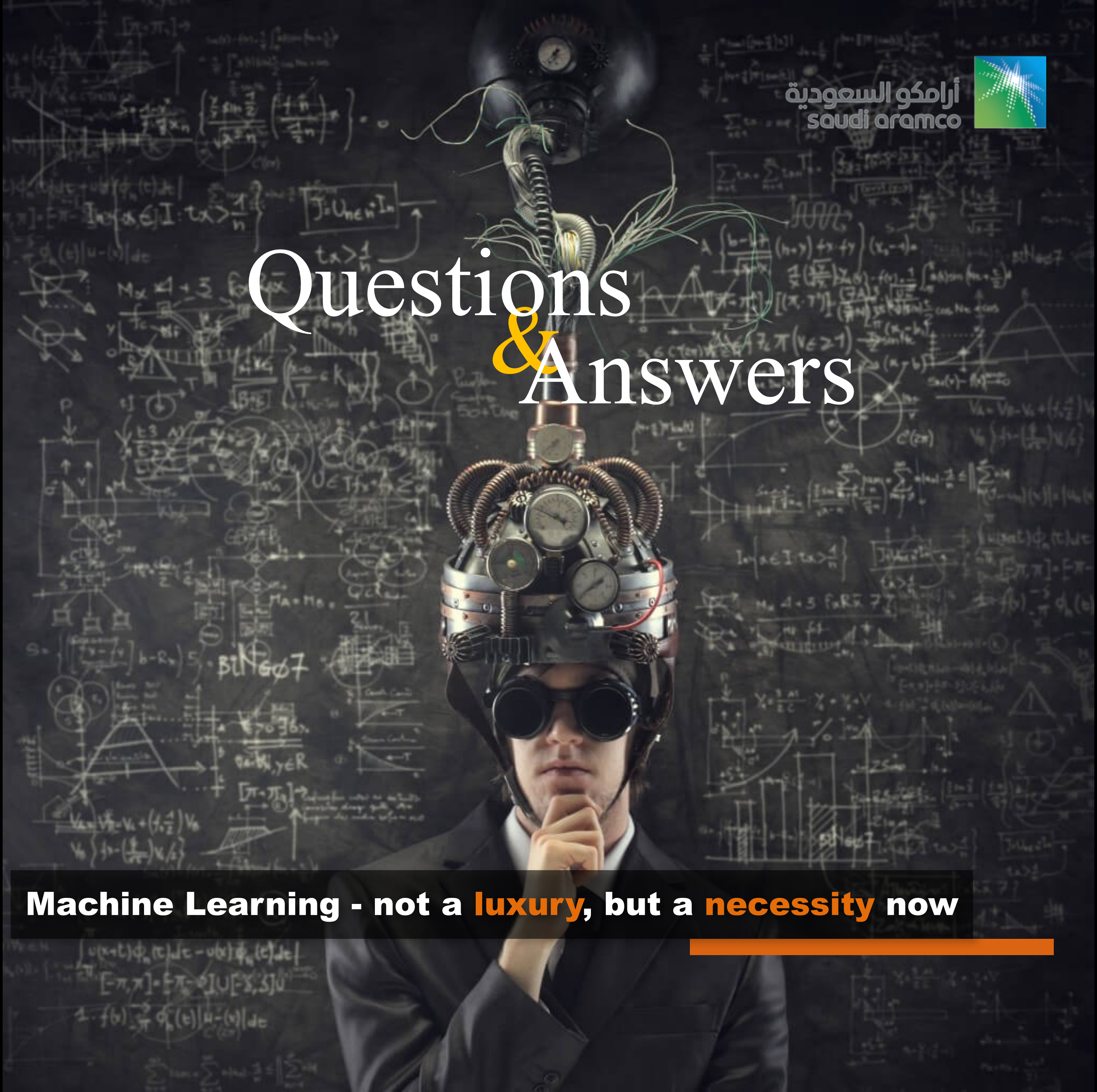
Machine Learning based User Behavioral Model - MLUBA

machine >>>> LEARNING

OPTIMAL ALGORITHMS FOR CYBER THREAT DETECTION

- Preprocessing (Sampling, Conversion, Extraction) is the key
- Scope of OneClassClassification in Cyber Security
- Machine Learning for Routine Operational Intelligence

أرامكو السعودية
saudi aramco



أرامكو السعودية
saudi aramco

Questions & Answers

Machine Learning - not a luxury, but a necessity now

Information is the oxygen of the modern age. It seeps through the walls topped by barbed wire, it wafts across the electrified borders

أرامكو السعودية
saudi aramco