High time to add machine learning to your information security stack?

Minhaz | minhaz@owasp.org | https://blog.minhazav.xyz | twitter.com/minhazav | github.com/mebjas | Hyderabad, India

whoami

- Currently, **Software Engineer II at Microsoft** Azure, Production & Infrastructure Engineering.
- Like to play around with data, statistics, machine learning.
- **OWASP** Project maintainer for CSRF Protector Project, Currently mentoring student as GSOC mentor for OWASP Security Knowledge Framework Project

CODING SOMETHING OR OTHER SINCE 2009

Minhaz

OWASP
Open Web Application Security Project

Some Previous Talks with (以前的一些谈话) OWASP / CSA!

Disclaimer (放弃)

1. This talk is about defending not attacking
2. No IP was damaged to make this presentation.
3. I’m not here to make inferences on what is or not the perfect way to solve issue / or if ML is going to be the solution for everyone
4. I’ll be citing couple of Organizations / Individuals whose work I’ll be using here. I have no formal connection / sponsorship from them – it’s purely based on my personal research.

Outline

High time to add machine learning to your information security stack?

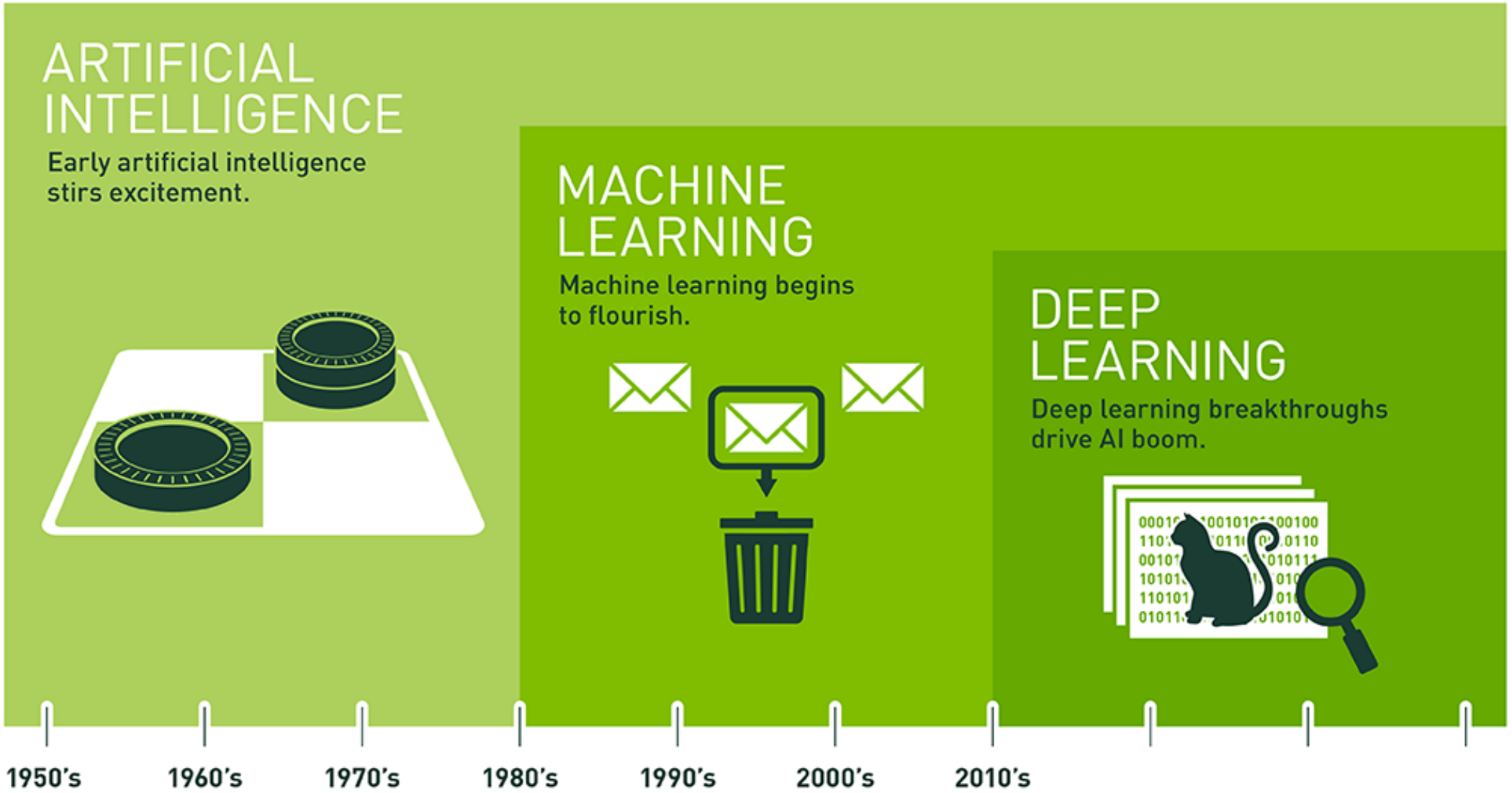
High time to add machine learning to your information security stack?

High time to add machine learning to your information security stack?

Machine Learning


Supervised Learning
Statistics
Big Data
neural networks
Algorithms
Data Mining
Data Science
Machine Learning
Computer Vision
Bayesian Statistics
Confidence Interval
Categorical Variable
Image Processing
One Hot Encoding
Predictor Variable
Biomedical Imaging
Overfitting
Standard Deviation
Frequentist Statistics
F-Score
large scale data mining
Descriptive Statistics
P-Value
Type II error
Range
Classification
Open Source
Mode
data analytics
Decile
Outlier
Neural Network
quantitative imaging
nonlinear dynamical systems
video analysis
Variance
Dummy Variable
K-Means
Robotics
predictive analytics
Degree of Freedom
Semi supervised learning
Artificial Intelligence
Binomial Distribution
Univariate Analysis
Evaluation Metrics
Type I error
Normal Distribution
Lasso Regression
TensorFlow
Hypothesis
Bioinformatics
Social Computing
Inferential Statistics
Regression

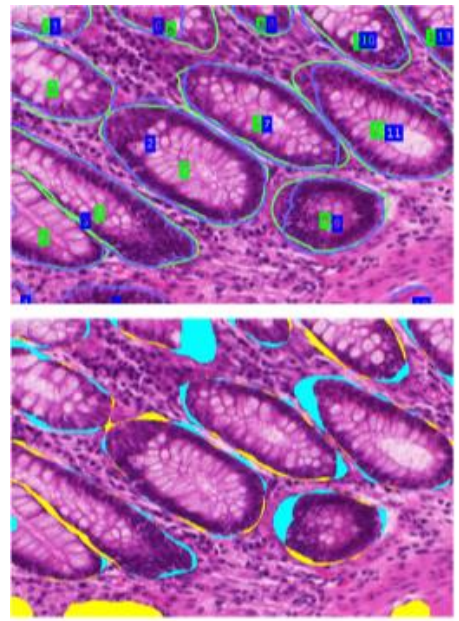
Machine Learning
Deep Learning
Artificial Intelligence



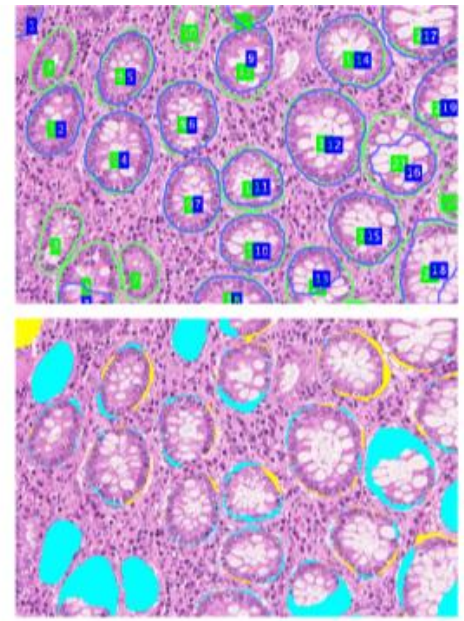
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Problems being solved in the world using these techniques

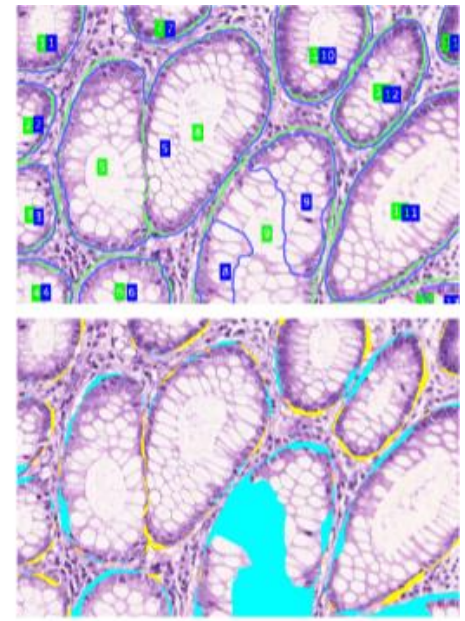
(使用这些技术在世界上解决的问题)



(d) benign

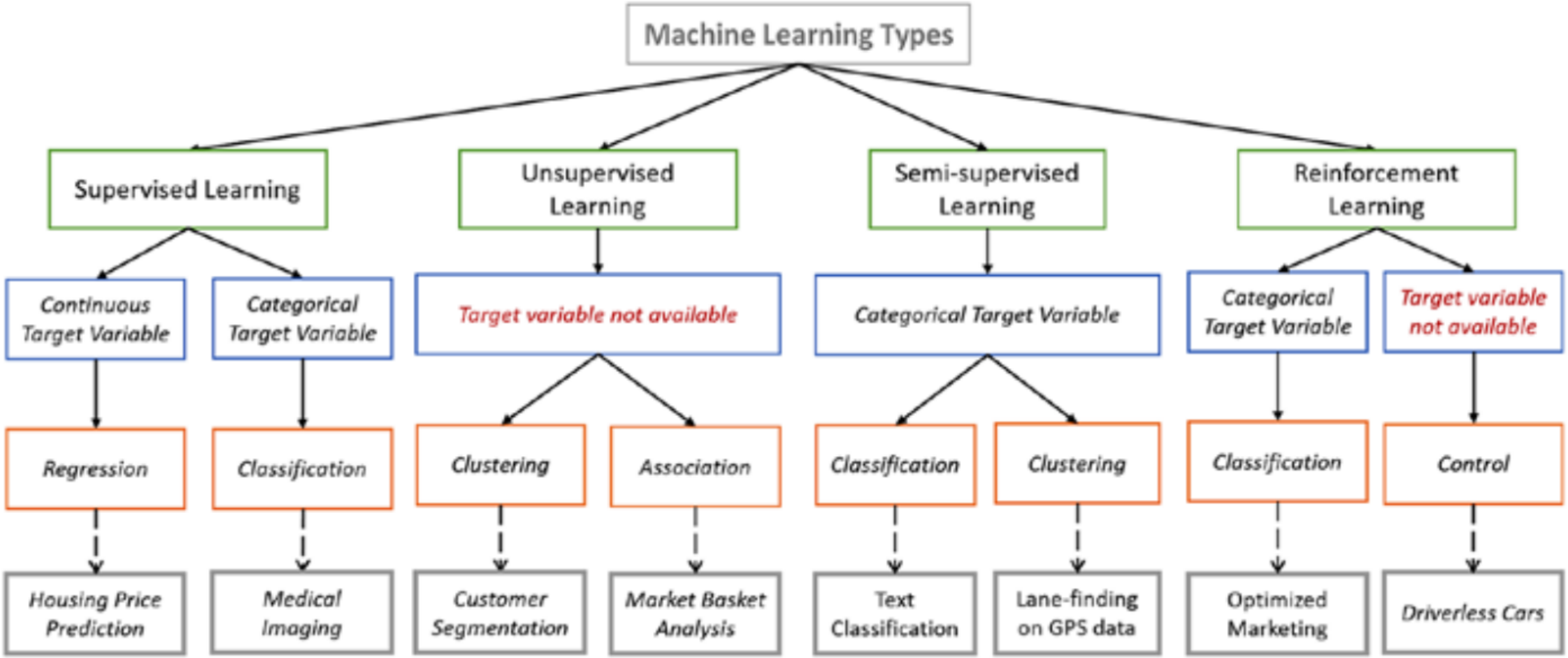


(e) benign



(f) malignant

Different areas of machine learning

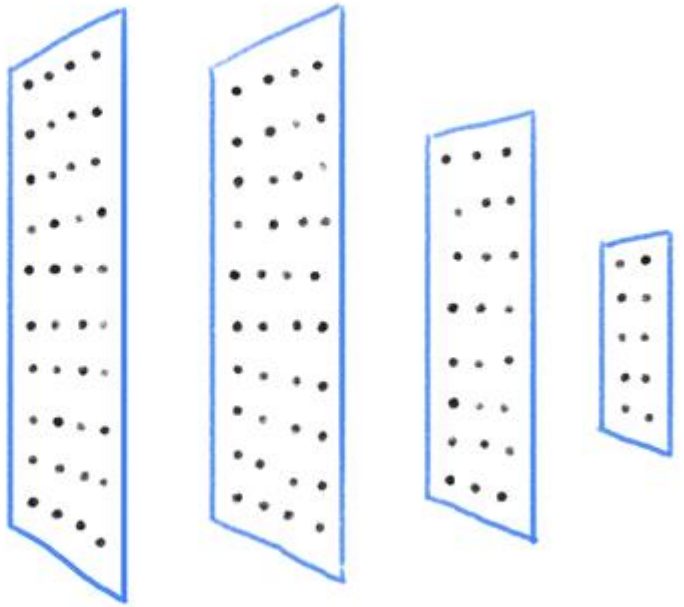

Machine Learning Types
Supervised Learning
Unsupervised Learning
Semi-supervised Learning
Reinforcement Learning
Continuous Target Variable
Categorical Target Variable
Target variable not available
Categorical Target Variable
Categorical Target Variable
Target variable not available
Regression
Classification
Clustering
Association
Classification
Clustering
Classification
Control
Housing Price Prediction
Medical Imaging
Customer Segmentation
Market Basket Analysis
Text Classification
Lane-finding on GPS data
Optimized Marketing
Driverless Cars

Components of Machine Learning Pipeline

Let’s go through most of them with a study on Classification of Malwares

Supervised Learning: Classification

Malware Classification



Malware: Malicious Software

Problem: How traditional anti virus systems work, and if machine learning could be help full.

Traditional antiviruses works on:

1. Signature-based detection
2. Heuristic-based detection
3. Behavior based detection
4. Sandbox detection
5. Data mining techniques

Malware Classification

Classify an application as malware or not based on behavior i.e. to train computer to learn boundary between behavior of a normal application as compared to a malware

Step 1: Define your problem and see if you can gather data + Domain Knowledge
定义您的问题，看看您是否可以收集数据

Problems:

1. Missing Items
2. Incorrect Items, specifically labels
3. Skewness
4. Low Volume
5. Outdated data

Data Source for demo: https://github.com/Te-k/malware-classification

ALWAYS REMEMBER:
Garbage IN Garbage OUT

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

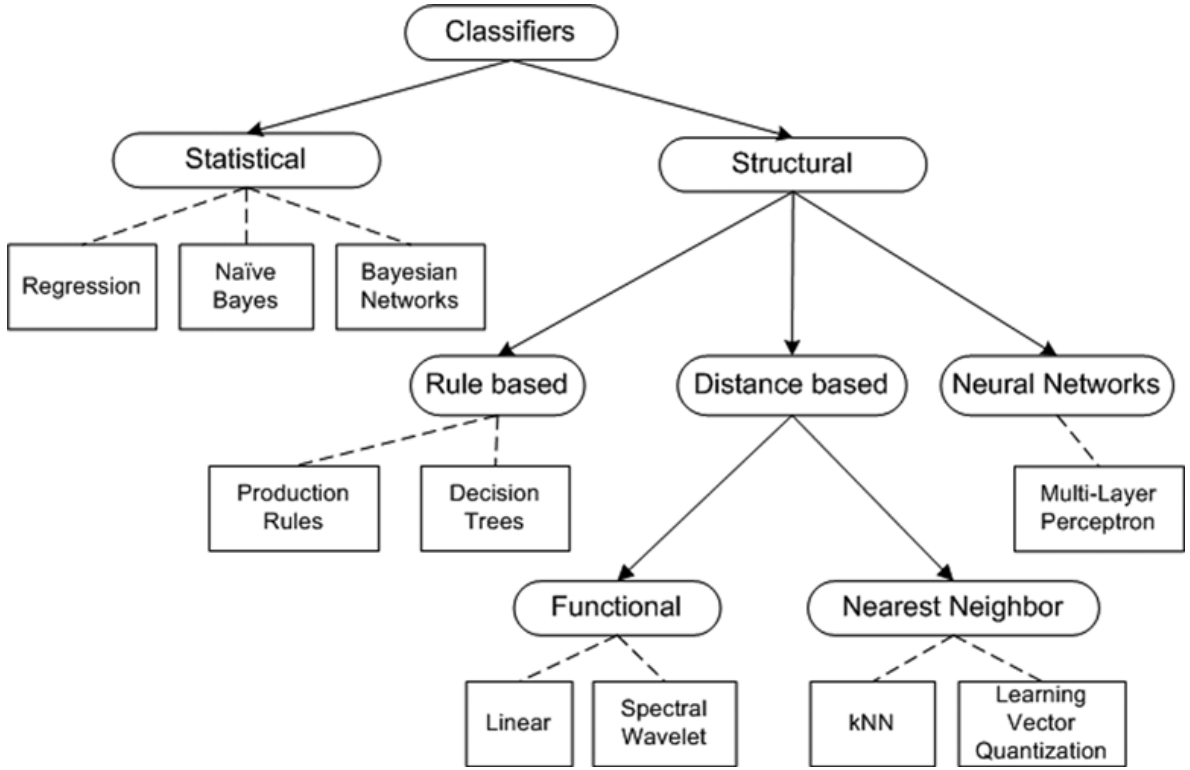
kaggle GitHub

Step 2: Feature Engineering

- Feature Extraction
- Feature Addition
- Feature Selection
  - Manual
  - Automatic

Step 3: Choice of Algorithm

There are wide range of algorithms from which we can choose based on whether we are trying to do prediction, classification or clustering. We can also choose between linear and non-linear algorithms. Naive Bayes, Support Vector Machines, Decision Trees, k-Means Clustering are some common algorithms used.



scikits learn machine learning in Python

dmlc XGBoost

Microsoft CNTK

TensorFlow

K Keras

Step 4: Training

- In this step we tune our algorithm based on the data we already have. This data is called training set as it is used to train our algorithm. This is the part where our machine or software learn and improve with experience.

- **Test Train Split**
- We divide our data (randomly) to testing and training datasets to be evaluate the capabilities of our models with unknown datasets.

scikits learn machine learning in Python
dmlc XGBoost
Microsoft CNTK
TensorFlow
K Keras

Step 5: Choice of Metrics / Evaluation Criteria

- Accuracy
- False Positive Rate (FPR)
- False Negative Rate (FNR)
- Precision
- Recall
- f1-measure
- & More…

scikits learn machine learning in Python

dmlc XGBoost

Microsoft CNTK

TensorFlow

K Keras

Step 6: Testing

Lastly, we test how our machine learning algorithm performs on an unseen set of test cases. One way to do this, is to partition the data into training and testing set. The training set is used in step 4 while the test set is then used in this step. Techniques such as cross-validation and leave-one-out can be used to deal with scenarios where we do not have enough data.

Another interesting example
另一个有趣的例子

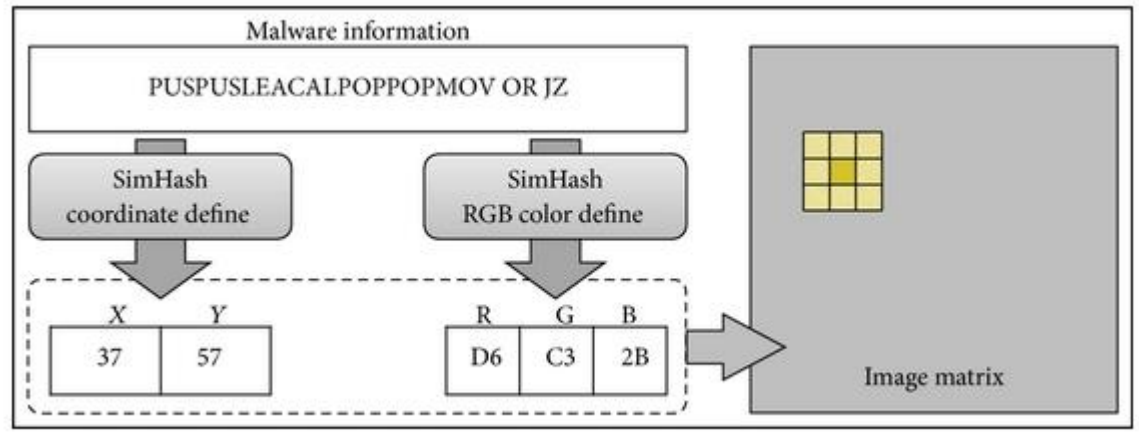
Another interesting way to do malware classification has by converting malwares to images and applying machine learning / deep learning techniques on top of them;

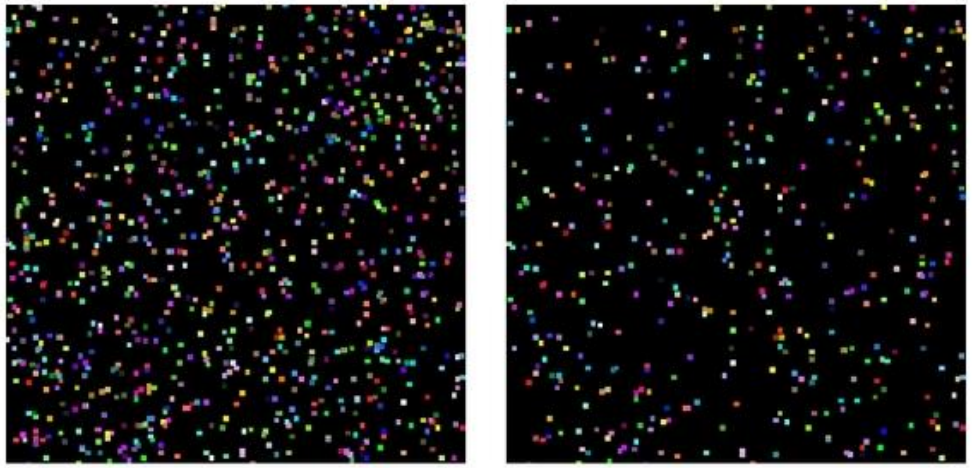
The proposed method generates RGB-colored pixels on image matrices using the opcode sequences extracted from malware samples and calculates the similarities for the image matrices.

Reference: Malware Analysis Using Visualized Image Matrices

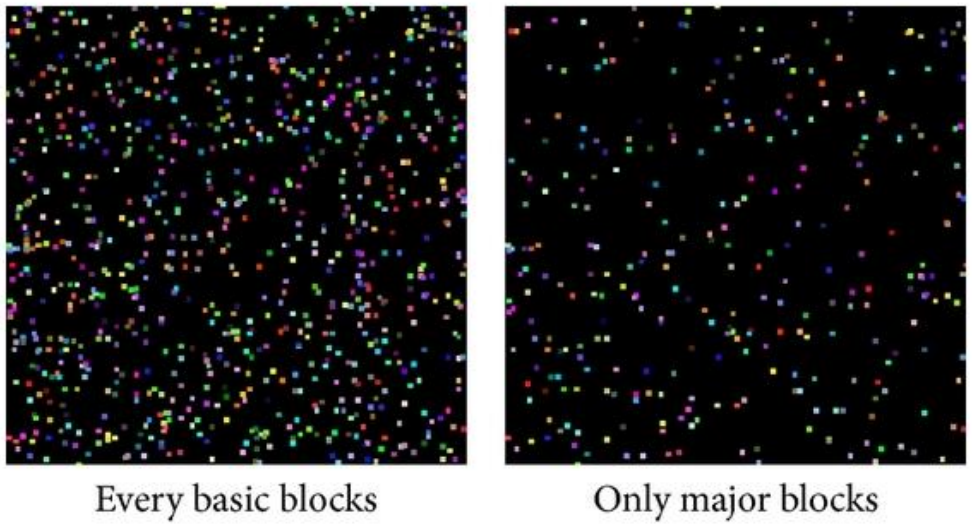
https://www.hindawi.com/journals/tswj/2014/132713/

```
0042A68B  Main   JBE   SHORT Exploit_.0042A69C
0042A68D  Main   MOV   AL, BYTE PTR DS: [EDX]
0042A68F  Main   INC   EDX
0042A690  Main   MOV   BYTE PTR DS: [EDI], AL
0042A692  Main   INC   EDI
0042A693  Main   DEC   ECX
0042A694  Main   JNZ   SHORT Exploit_.0042A68D
0042A68D  Main   MOV   AL, BYTE PTR DS: [EDX]
0042A68F  Main   INC   EDX
0042A690  Main   MOV   BYTE PTR DS:[EDI],AL
0042A692  Main   INC   EDI
0042A693  Main   DEC   ECX
0042A694  Main   JNZ   SHORT Exploit_.0042A68D
0042A68D  Main   MOV   AL, BYTE PTR DS: [EDX]
0042A68F  Main   INC   EDX
0042A690  Main   MOV   BYTE PTR DS: [EDI], AL
0042A692  Main   INC   EDI
0042A693  Main   DEC   ECX
0042A694  Main   JNZ   SHORT Exploit_.0042A68D
0042A68D  Main   MOV   AL,BYTE PTR DS: [EDX]
0042A68F  Main   INC   EDX
0042A690  Main   MOV   BYTE PTR DS: [EDI], AL
0042A692  Main   INC   EDI
0042A693  Main   DEC   ECX
0042A694  Main   JNZ   SHORT Exploit_.0042A68D
0042A696  Main   JMP   Exploit_.0042A5FE
0042A5FE  Main   ADD   EBX, EBX
```

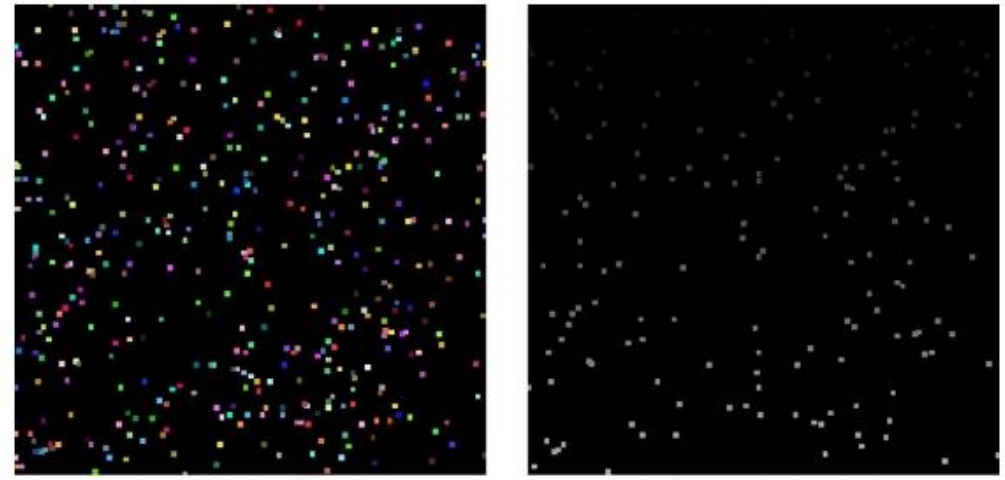




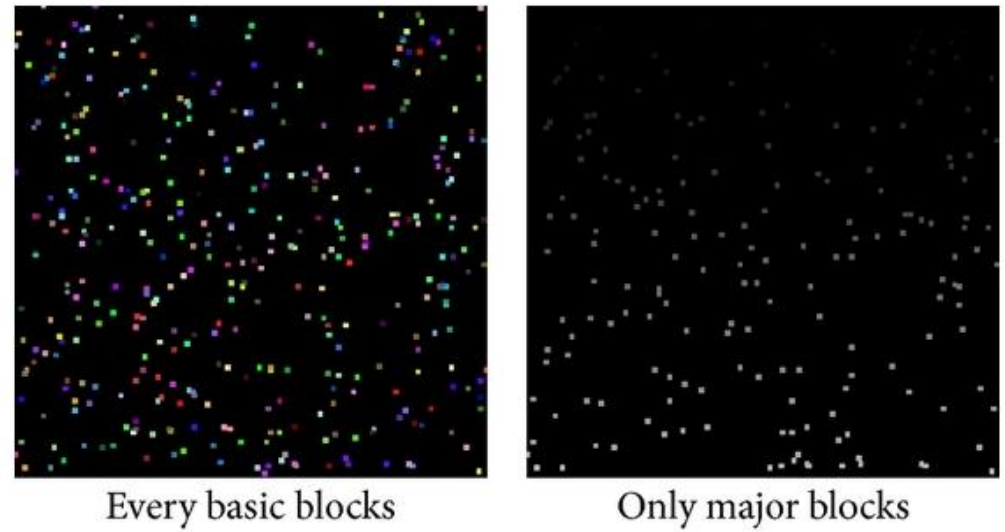
(a) Email-Worm.Win32.Klez.a



(b) Trojan-DDos.Win32.Boxed.a



(c) Trojan-Downloader.Win32.Lemmy.e



(d) Virus.Win32.HLLP.Zepp.a

So is malware detection being done using machine learning as of now?

Kaspersky: Machine Learning for Malware Detection

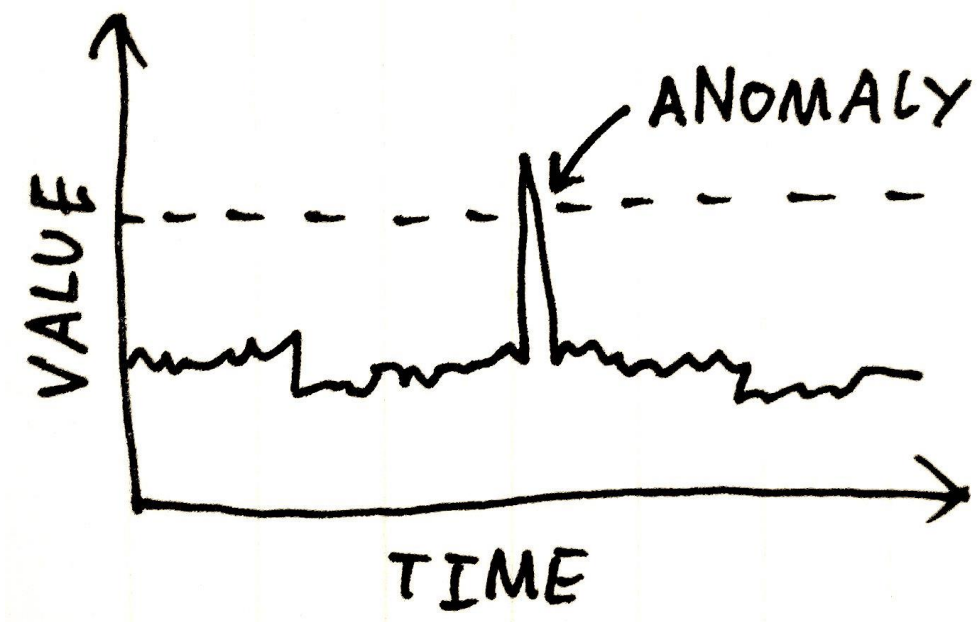
Key points they mention are:

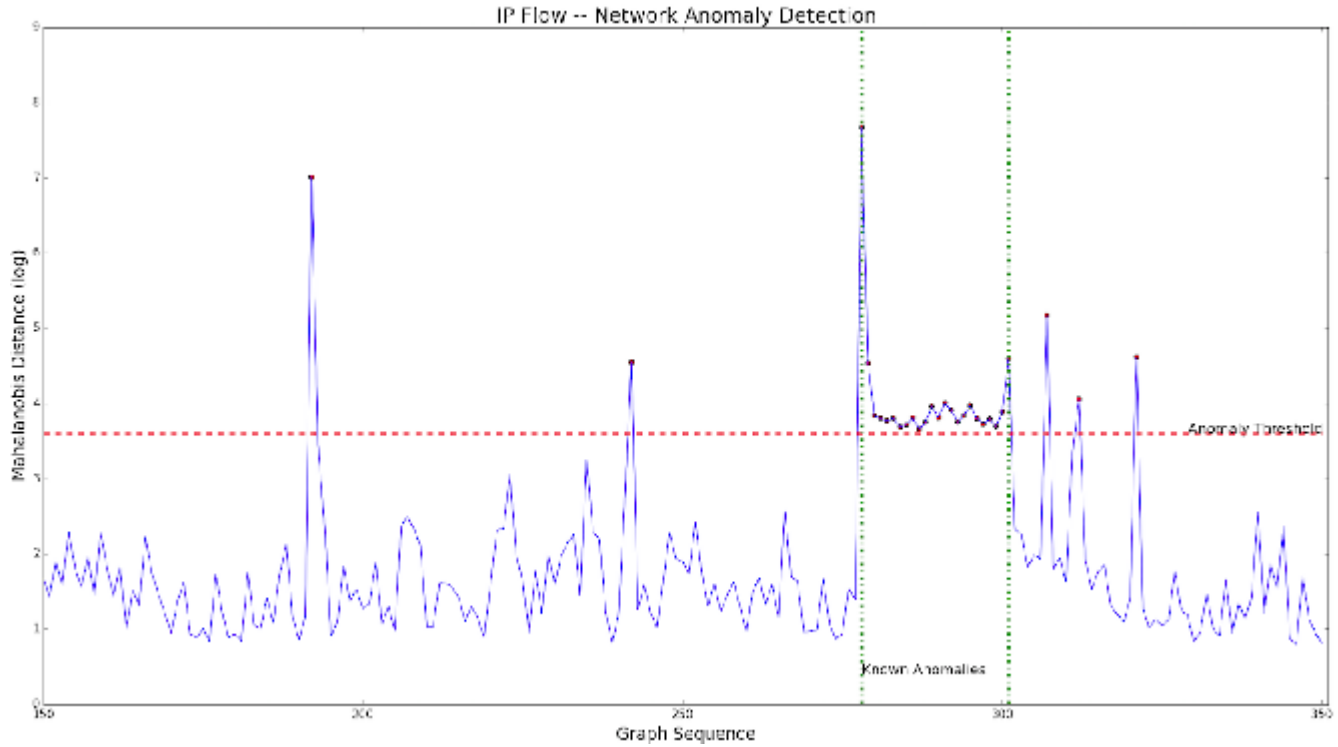
- **Have the right data**.
- Know theoretical machine learning and how to apply it to cybersecurity.
- Know user practical needs and be an expert at implementing machine learning into products
- Earn a sufficient user base and use the power of **feedback loop and crowdsourcing.**
- **Keep detection methods in multi-layered synergy.**

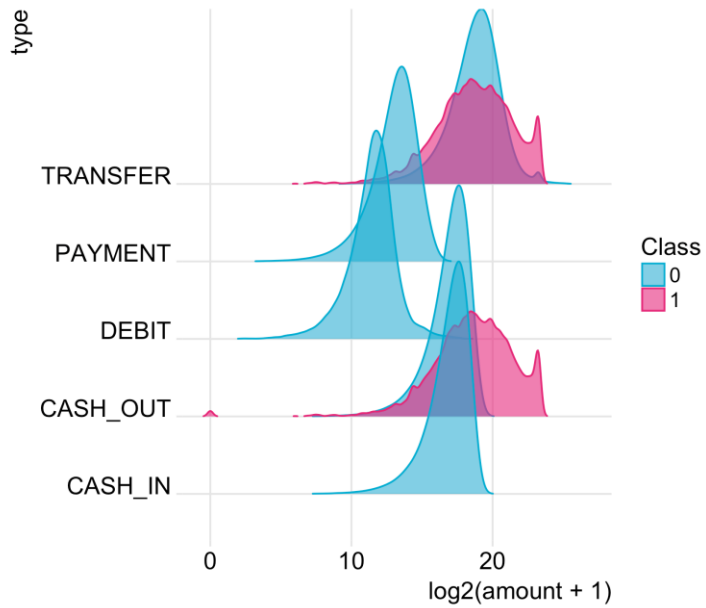
Link to whitepaper

Unsupervised Learning

Anomaly Detection

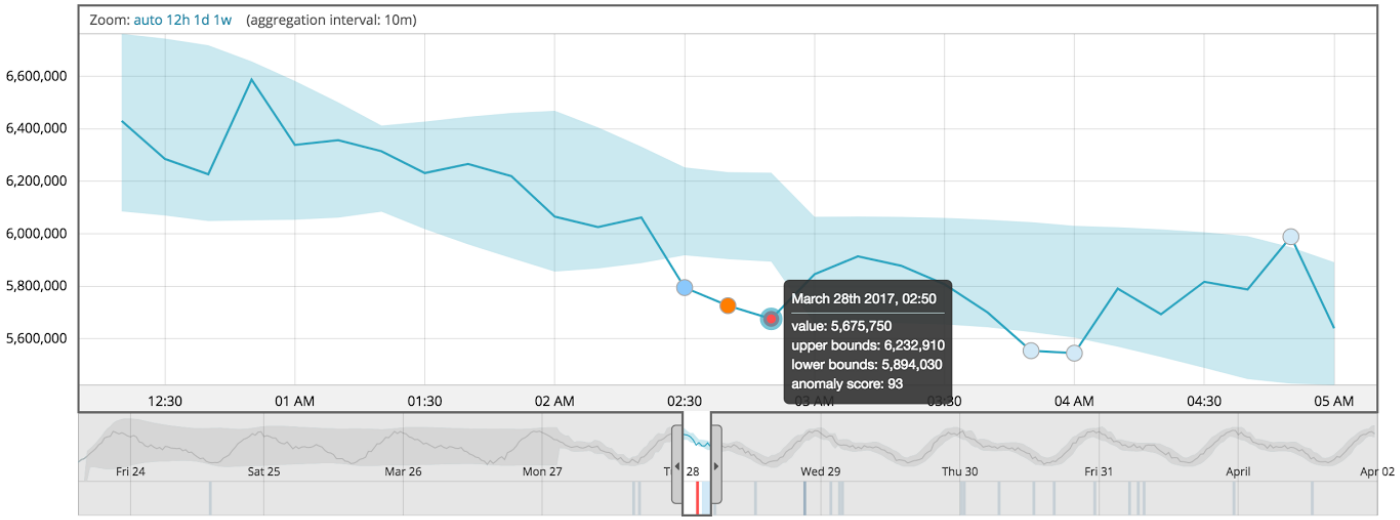






RISK LEVEL	DETECTION TYPE	RISK EVENT TYPE	RISK EVENTS CLOSED	LAST UPDATED (UTC)
High	Offline	Users with leaked credentials	44 of 45	12/7/2016 1:04 AM
Medium	Real-time	Sign-ins from anonymous IP addresses	76 of 78	1/17/2017 2:44 PM
Medium	Offline	Impossible travels to atypical locations	11 of 14	1/17/2017 2:44 PM
Medium	Real-time	Sign-in from unfamiliar location	0 of 1	11/15/2016 7:18 PM
Low	Offline	Sign-ins from infected devices	76 of 78	1/17/2017 2:44 PM

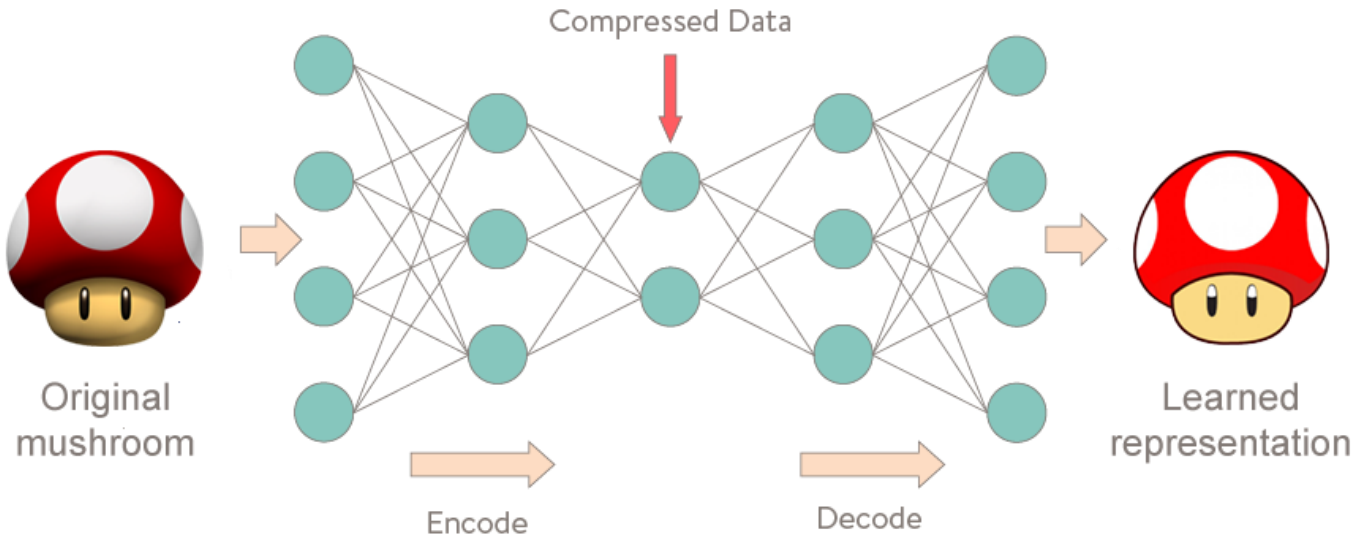
Anomaly Detection

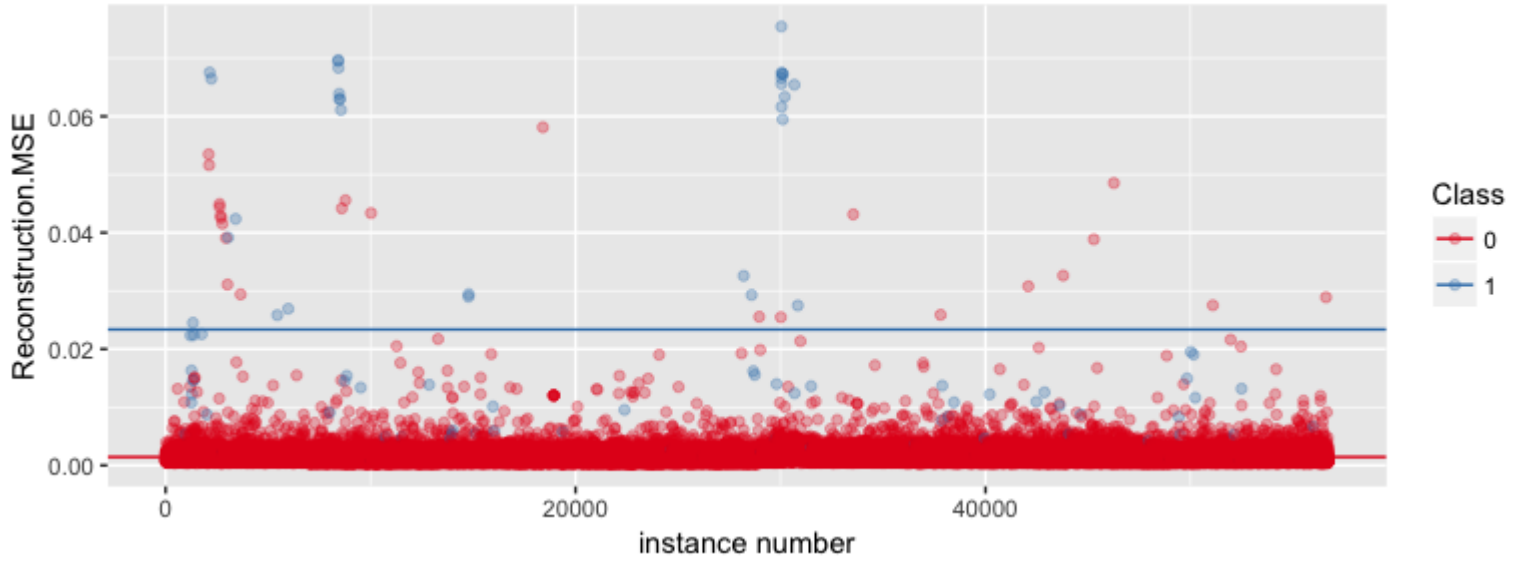


- Statistical techniques: Mean,
- Supervised Algorithms: KNN,
- Unsupervised Algorithms: SC Factor
- Deep Learning Models: LSTM, Auto Encoders

Twitter Anomaly Detection | scikit-learn | Facebook Prophet | LinkedIn Luminol

Auto Encoders





Use cases in other areas
其他领域的用例

Supervised Learning

Classification

Malware Detection / Classification

Spam detection

Phishing Detection

Regression

Risk Scoring

User Behavior Analysis and Fraud Detection

Unsupervised Learning

Clustering

Forensic analysis

Anomaly Detection

Network Traffic Analysis

Fraud Detection

Recommendations

Remediation Action Recommendations

In incident response

Pattern Detection, Correlation and NLP

Log Correlation

Noise Reduction

Why now?

为什么现在?

1. Volume of data (数据量)

Data has posed perhaps the single greatest challenge in cybersecurity over the past decade. For a human, or even a large team of humans, the amount of data produced daily on a global scale is unimaginable. For every minute in 2017 there were:

Data Produced in 2017

455,000 New Tweets

510,000 Comments and 293,000 Status Updates on Facebook

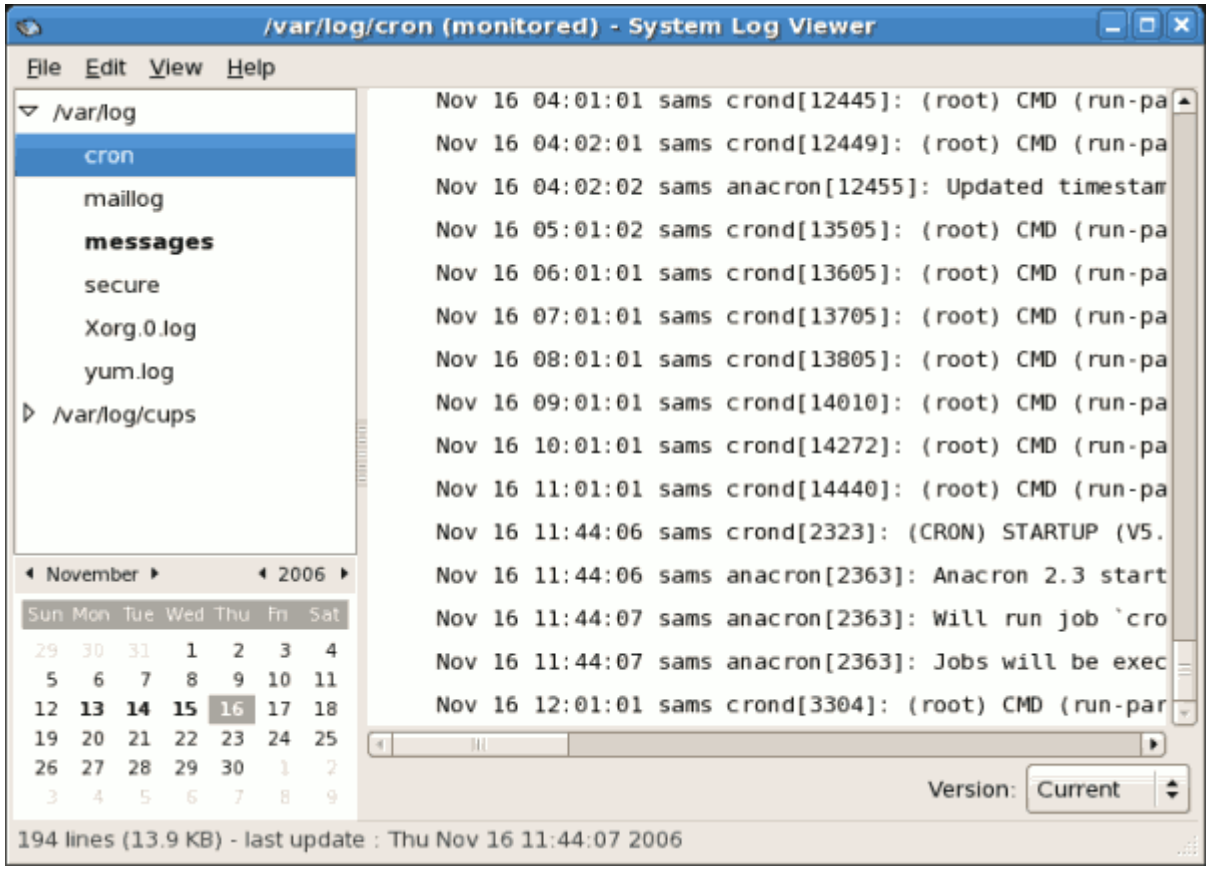
400 Hours of Video Uploaded to YouTube

3,607,080 Google Searches

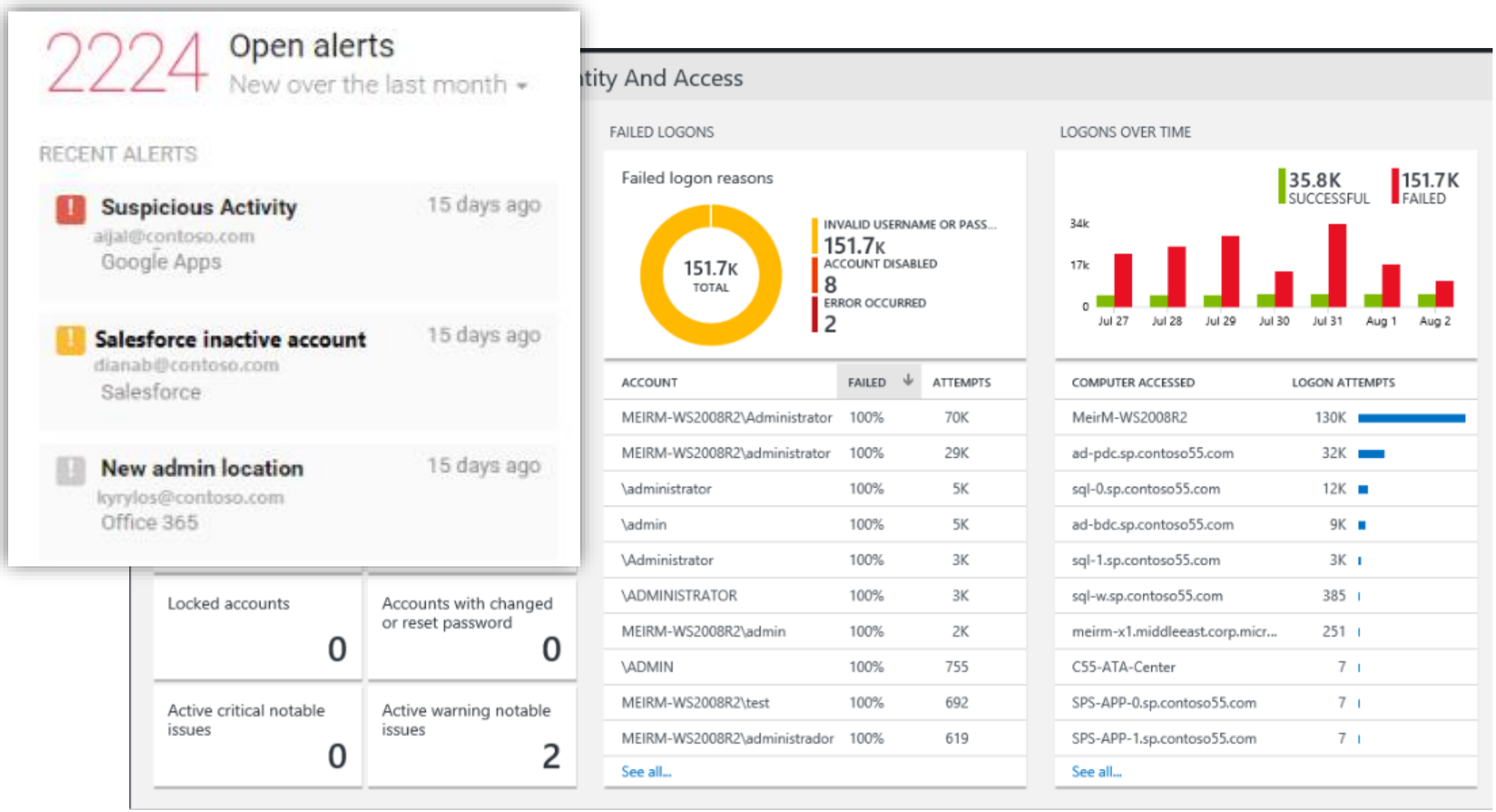
Over 186,000,000 Emails Sent

Source: Domo Infographic

Recorded Future

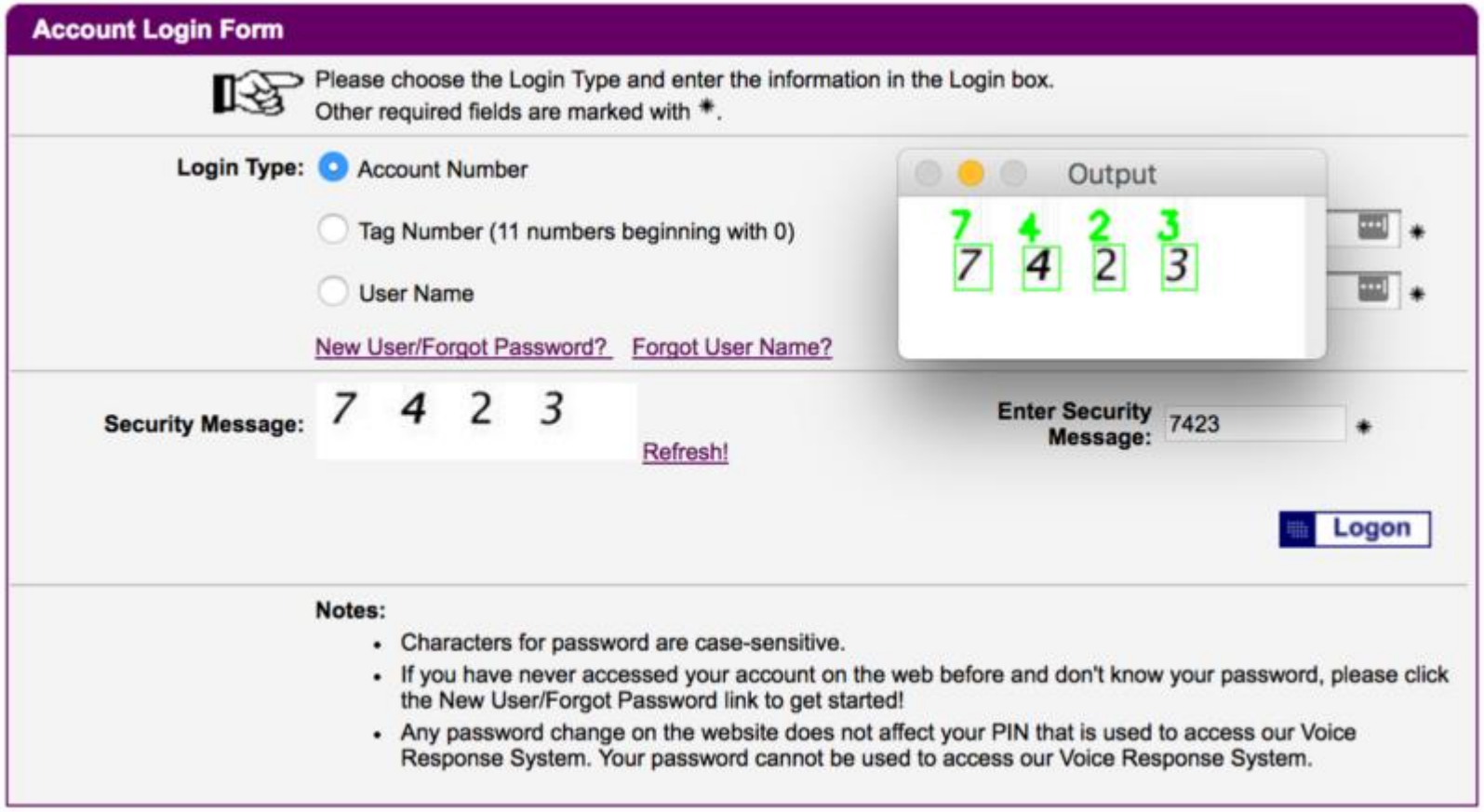

/var/log/cron (monitored) - System Log Viewer
File Edit View Help
/var/log
cron
maillog
messages
secure
Xorg.0.log
yum.log
/var/log/cups
November
2006
Sun Mon Tue Wed Thu Fri Sat
29 30 31 1 2 3 4
5 6 7 8 9 10 11
12 13 14 15 16 17 18
19 20 21 22 23 24 25
26 27 28 29 30 1 2
3 4 5 6 7 8 9
Nov 16 04:01:01 sams crond[12445]: (root) CMD (run-pa
Nov 16 04:02:01 sams crond[12449]: (root) CMD (run-pa
Nov 16 04:02:02 sams anacron[12455]: Updated timestam
Nov 16 05:01:02 sams crond[13505]: (root) CMD (run-pa
Nov 16 06:01:01 sams crond[13605]: (root) CMD (run-pa
Nov 16 07:01:01 sams crond[13705]: (root) CMD (run-pa
Nov 16 08:01:01 sams crond[13805]: (root) CMD (run-pa
Nov 16 09:01:01 sams crond[14010]: (root) CMD (run-pa
Nov 16 10:01:01 sams crond[14272]: (root) CMD (run-pa
Nov 16 11:01:01 sams crond[14440]: (root) CMD (run-pa
Nov 16 11:44:06 sams crond[2323]: (CRON) STARTUP (V5.
Nov 16 11:44:06 sams anacron[2363]: Anacron 2.3 start
Nov 16 11:44:07 sams anacron[2363]: Will run job `cro
Nov 16 11:44:07 sams anacron[2363]: Jobs will be exec
Nov 16 12:01:01 sams crond[3304]: (root) CMD (run-par
Version: Current
194 lines (13.9 KB) - last update : Thu Nov 16 11:44:07 2006


4031/VT.5 SVT 17
3836/VT.6 SVT 45
31029/VT.4 SVT 18
4133/VT.5 SVT 13
3631/VT.6 SVT 29
3756/VT.6 SVT 4
31034/VT.5 SVT 12


2224
Open alerts
New over the last month
RECENT ALERTS
Suspicious Activity
15 days ago
aijal@contoso.com
Google Apps
Salesforce inactive account
15 days ago
dianab@contoso.com
Salesforce
New admin location
15 days ago
kyrylos@contoso.com
Office 365
tity And Access
FAILED LOGONS
Failed logon reasons
151.7K
TOTAL
INVALID USERNAME OR PASS...
151.7K
ACCOUNT DISABLED
8
ERROR OCCURRED
2
LOGONS OVER TIME
35.8K
SUCCESSFUL
151.7K
FAILED
34k
17k
0
Jul 27
Jul 28
Jul 29
Jul 30
Jul 31
Aug 1
Aug 2
ACCOUNT
FAILED
ATTEMPTS
MEIRM-WS2008R2\Administrator 100% 70K
MEIRM-WS2008R2\administrator 100% 29K
\administrator 100% 5K
\admin 100% 5K
\Administrator 100% 3K
\ADMINISTRATOR 100% 3K
MEIRM-WS2008R2\admin 100% 2K
\ADMIN 100% 755
MEIRM-WS2008R2\test 100% 692
MEIRM-WS2008R2\administrador 100% 619
See all...
COMPUTER ACCESSED
LOGON ATTEMPTS
MeirM-WS2008R2 130K
ad-pdc.sp.contoso55.com 32K
sql-0.sp.contoso55.com 12K
ad-bdc.sp.contoso55.com 9K
sql-1.sp.contoso55.com 3K
sql-w.sp.contoso55.com 385
meirm-x1.middleeast.corp.micr... 251
C55-ATA-Center 7
SPS-APP-0.sp.contoso55.com 7
SPS-APP-1.sp.contoso55.com 7
See all...
Locked accounts
0
Accounts with changed or reset password
0
Active critical notable issues
0
Active warning notable issues
2

2. To focus on what’s important

3. Attacks are getting more sophisticated
攻击越来越复杂

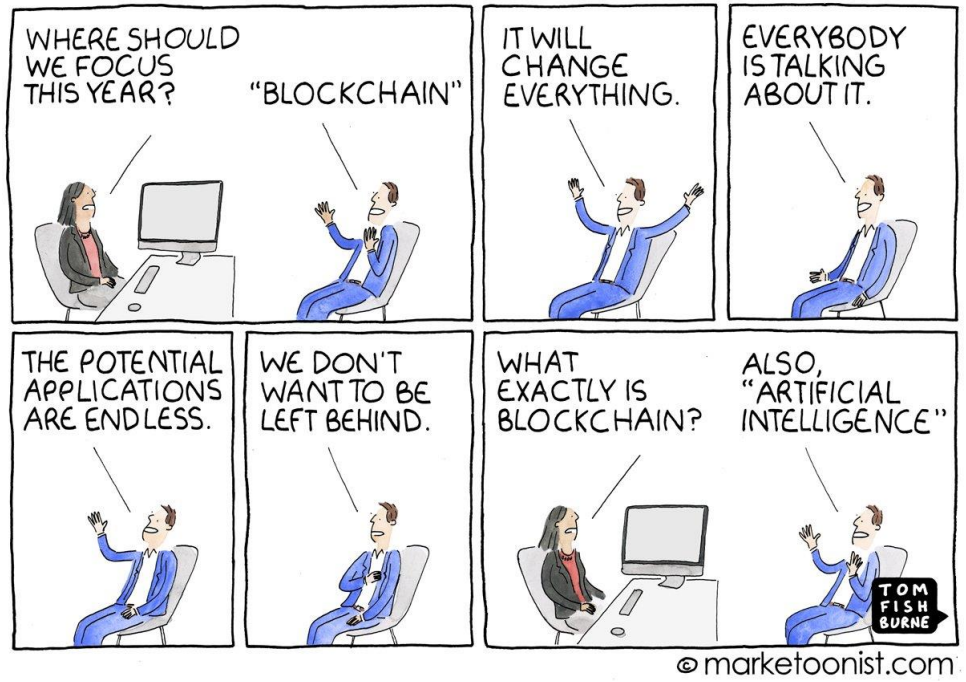


Breaking captcha using deep convolutional networks

4. Solve set of problems like we solved for SPAMS

5. Vendor Management - New vendors coming up every other day

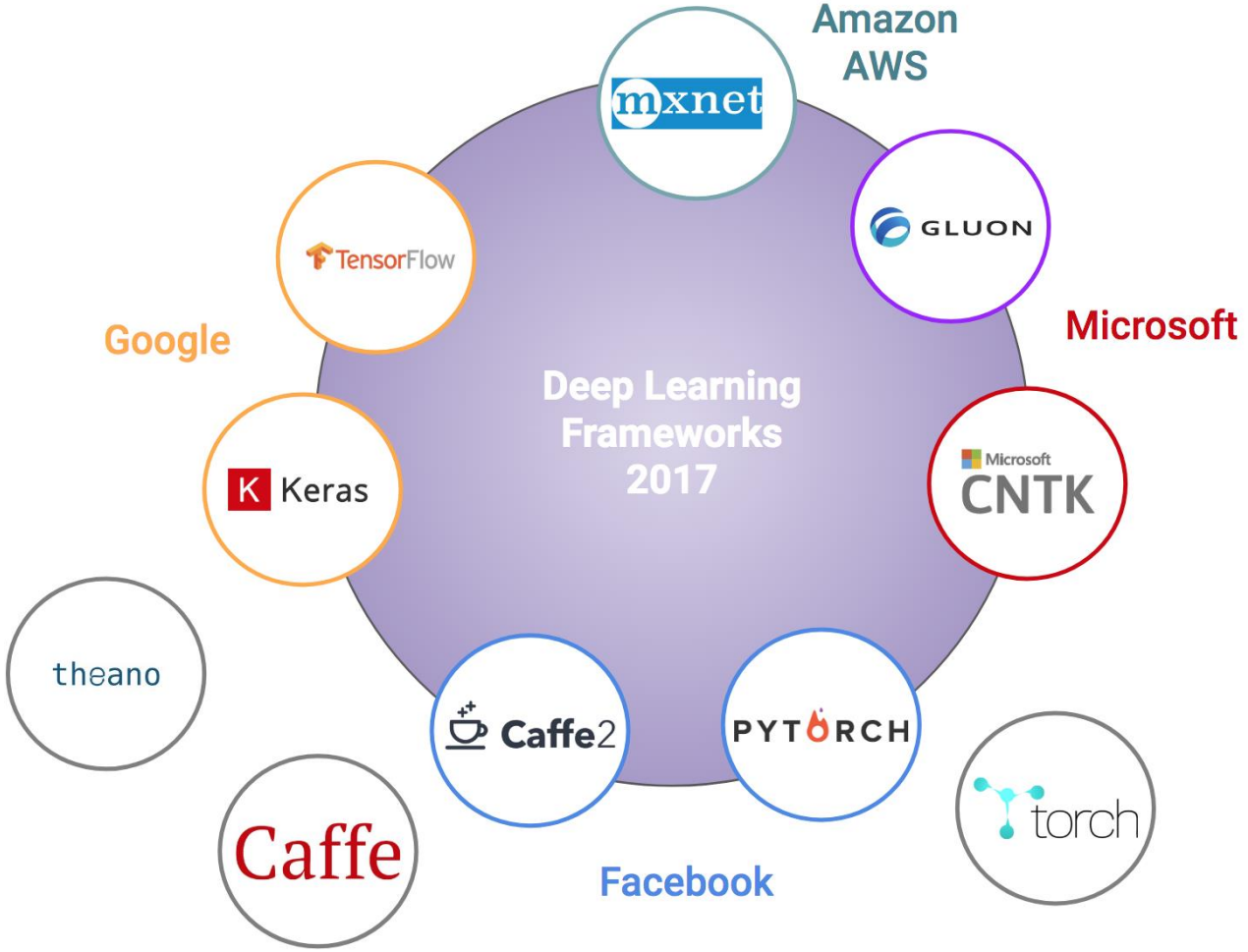
- You need to brace yourself and know what the technology has to offer before evaluating what they offer.
- AI/ML is no longer just a buzz word. It has strong capabilities. But it's a tool at the end.



And how?

As individual or an Enterprise
作为个人或企业

- Online Courses Online
- Online Challenges and Open Source Tools to try out stuff and proof of concepts
- Using power of cloud to do things at scale
- There is no lack of content out there on this topic.



kagg

- Outside the Closed World: On Using Machine Learning for Network Intrusion Detection
- Anomalous Payload-Based Network Intrusion Detection
- Malicious PDF detection using metadata and structural features
- Adversarial support vector machine learning
- Exploiting machine learning to subvert your spam filter
- CAMP – Content Agnostic Malware Protection
- Notos – Building a Dynamic Reputation System for DNS
- Kopis – Detecting malware domains at the upper dns hierarchy
- Pleiades – From Throw-away Traffic To Bots – Detecting The Rise Of DGA-based Malware
- EXPOSURE – Finding Malicious Domains Using Passive DNS Analysis
- Polonium – Tera-Scale Graph Mining for Malware Detection
- Nazca – Detecting Malware Distribution in Large-Scale Networks
- PAYL – Anomalous Payload-based Network Intrusion Detection
- Anagram – A Content Anomaly Detector Resistant to Mimicry Attacks
- Applications of Machine Learning in Cyber Security
- Data Mining для построения систем обнаружения сетевых атак (RUS)
- Выбор технологий Data Mining для систем обнаружения вторжений в корпоративную сеть (RUS)
- Нейросетевой подход к иерархическому представлению компьютерной сети в задачах информационной безопасности (RUS)
- Методы интеллектуального анализа данных и обнаружение вторжений (RUS)
- Dimension Reduction in Network Attacks Detection Systems
- Rise of the machines: Machine Learning & its cyber security applications
- Machine Learning in Cyber Security: Age of the Centaurs
- Automatically Evading Classifiers A Case Study on PDF Malware Classifiers
- Weaponizing Data Science for Social Engineering—Automated E2E Spear Phishing on Twitter
- Machine Learning: A Threat-Hunting Reality Check
- Neural Network-based Graph Embedding for Cross-Platform Binary Code Similarity Detection
- Practical Secure Aggregation for Privacy-Preserving Machine Learning
- DeepLog: Anomaly Detection and Diagnosis from System Logs through Deep Learning
- eXpose: A Character-Level Convolutional Neural Network with Embeddings For Detecting Malicious URLs, File Paths and

Research related to Machine Learning
And cyber security.

https://github.com/jivoi/awesome-ml-for-cybersecurity

MACHINE LEARNING EXPERTISE TO THING BEYOND STANDARD TOOLKITS.

DATA ACROSS THE STACK
HOST (EVENT LOGS, SYS LOGS, AV LOGS)
NETWORK LOGS
SERVICE & APPLICATION LOGS

SECURE AND SCALABLE PLATFORM

EYES ON GLASS

TESTING WITH REAL ATTACKS

Security ML requirements

01

Create , Share and Validate Open Data Repositories.

02

Involve in crowdsourced generation of labelled data.

03

Initiate research in this area and collaborate.

04

Brace ourselves for next generation of attack and defence.

Open Source Communities

Takeaways

Takeaways

- ML/DL are here, embrace the change: the correct applicability of ML can enhance defensive practices.
- There is a lot of possibilities in InfoSec for these techniques.
- Machine Learning / Deep Learning / AI – they are tools. It’s a tool you have to know how to apply in order for it to reveal true insight. And while it’s not the only tool we need to use but it’s bound to get more powerful with time. We need to mix in experience. We have to work with experts to capture their knowledge for the algorithms to reveal actual security insights or issues.

Thanks
谢谢

Appendix

- Visual introduction to machine learning - http://www.r2d3.us/visual-intro-to-machine-learning-part-1/
- Microsoft Malware Challenge on Kaggle - https://www.kaggle.com/c/malware-classification
- Malware Detection and Classification Using Machine Learning on Microsoft Malware Classification challenge - https://github.com/dchad/malware-detection
- Collection of deep learning research papers - https://medium.com/@jason_trost/collection-of-deep-learning-cyber-security-research-papers-e1f856f71042
- Security data science papers - http://www.covert.io/security-datascience-papers/

All Code and references available at

https://github.com/mebjas/owasp.tw.0718