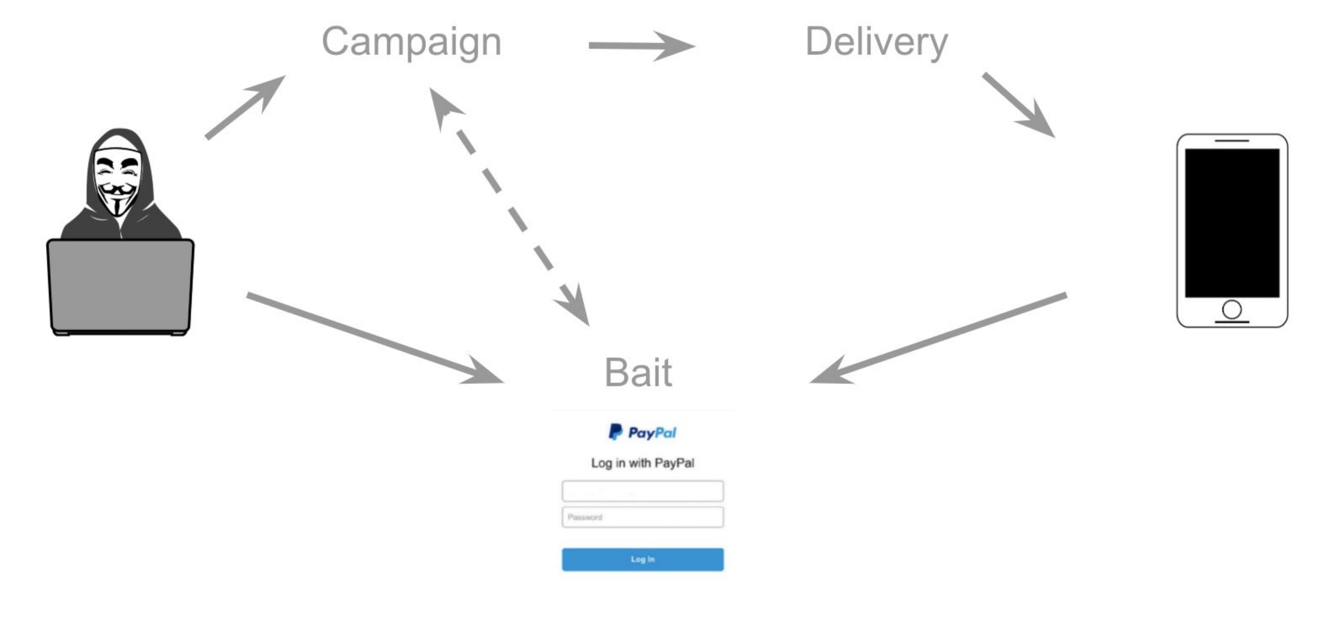
0-day Security Detections at Scale

Zdenek Letko

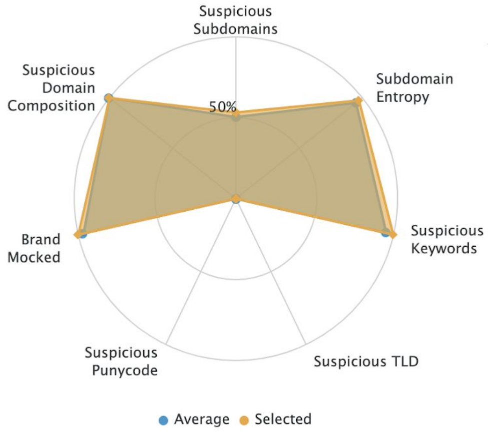
Motivation



Campaign
Delivery
Bait
PayPal
Log in with PayPal
Password
Log In

Motivation

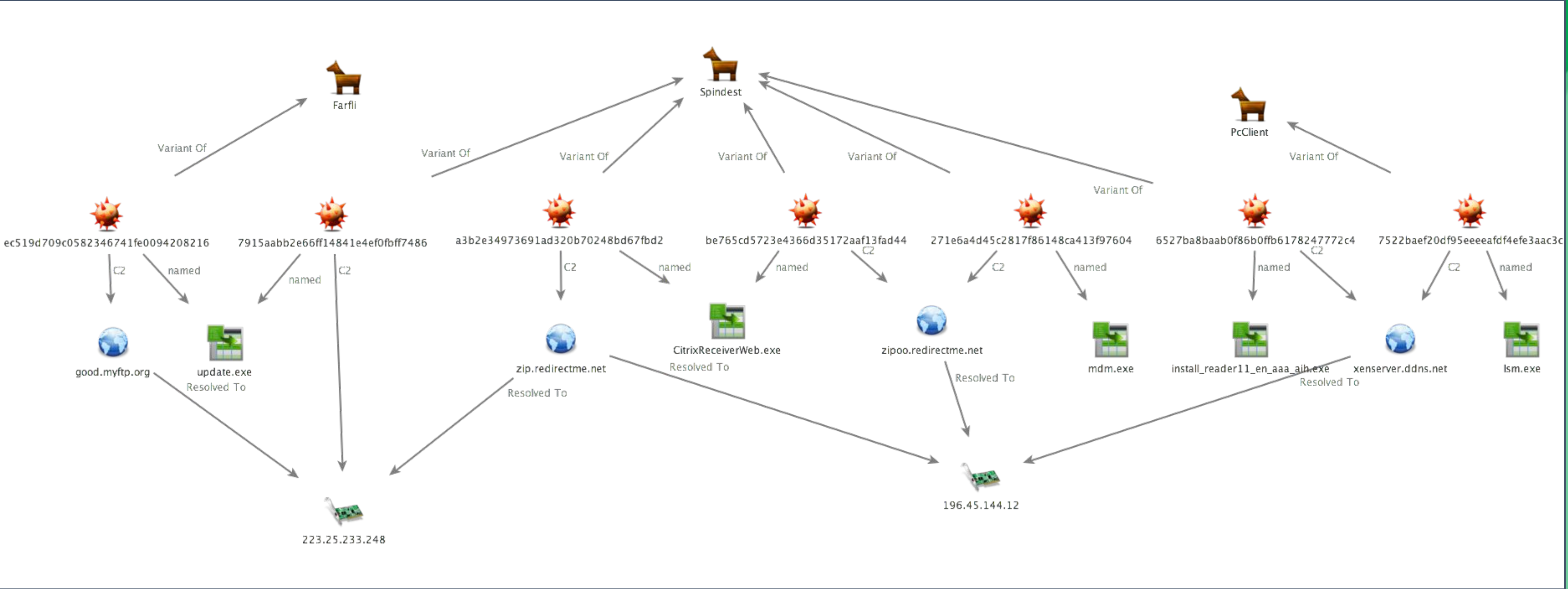
ZERO-DAY PHISHING



Timestamp	URL	Phishing Score
11:32 28 Oct 2019	1572258739062.billing.facebook.com-user----account-----overview-----new-----terms8eb493ef.4061151572247010552.1528986948509.snappli.co.uk	83%
08:15 28 Oct 2019	1572246900960.billing.facebook.com-user----account-----overview-----new-----terms8eb493ef.4061151572247010552.1528986948509.snappli.co.uk	86%
03:18 27 Oct 2019	1572142714769.billing.facebook.com-user----account-----overview-----new-----terms8eb493ef.4061151572247010552.1528986948509.snappli.co.uk	86%
04:09 26 Oct 2019	1572055763891.billing.facebook.com-user----account-----overview-----new-----terms8eb493ef.4061151572247010552.1528986948509.snappli.co.uk	82%
09:32 25 Oct 2019	1572031958595.billing.facebook.com-user----account-----overview-----new-----terms8eb493ef.4061151572247010552.1528986948509.snappli.co.uk	86%
06:23 25 Oct 2019	1572020625896.billing.facebook.com-user----account-----overview-----new-----terms8eb493ef.4061151572247010552.1528986948509.snappli.co.uk	83%
03:29 25 Oct 2019	1572010171199.billing.facebook.com-user----account-----overview-----new-----terms8eb493ef.4061151572247010552.1528986948509.snappli.co.uk	86%

< 1 2 3 4 5 ••• 15 > 7

Motivation

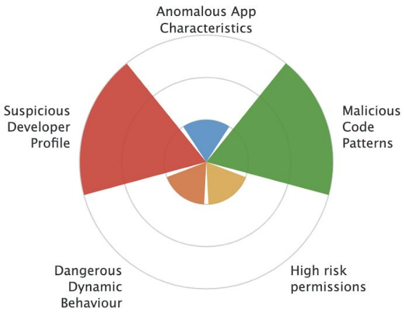


Source: researchcenter.paloaltonetworks.com

Motivation

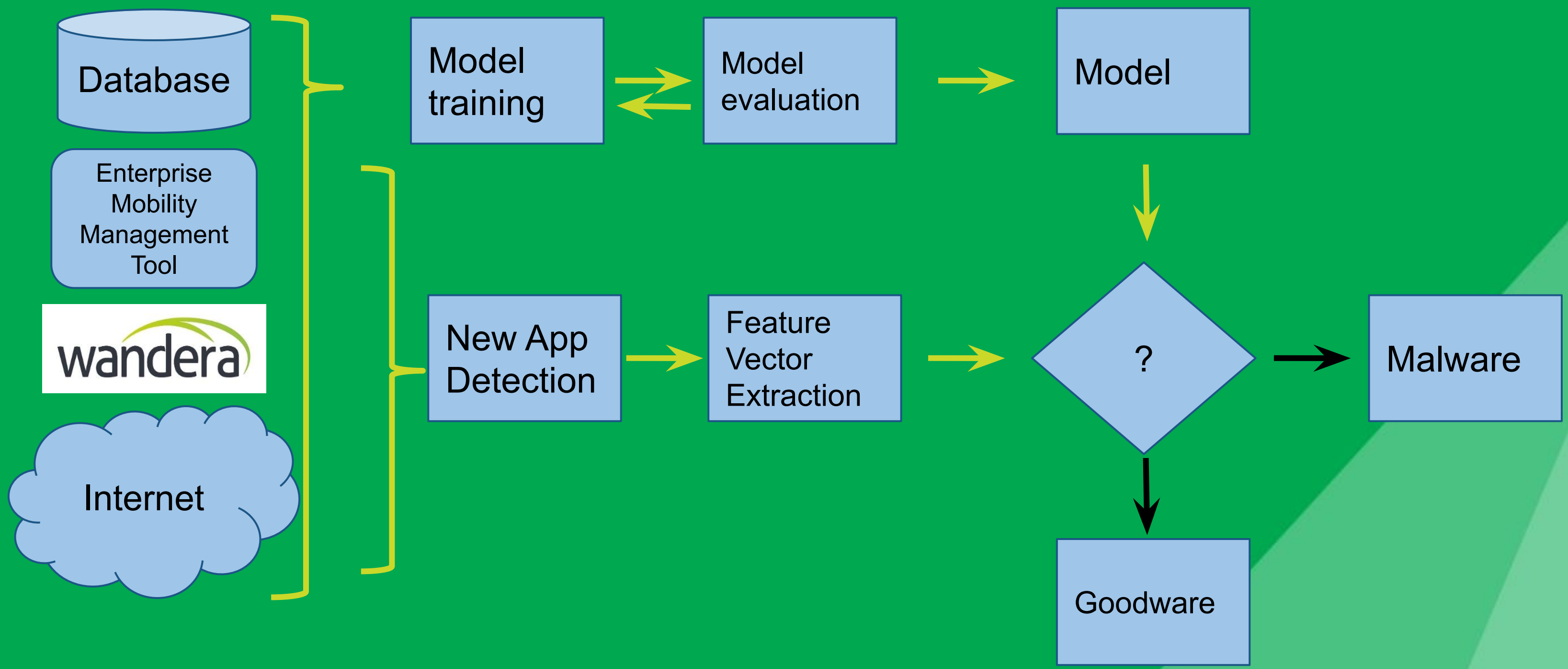
RISKY APPS

欢乐捕鱼 - 天天斗地主 69%



App	Risk Score
欢乐捕鱼 - 天天斗地主	69%
SuperX Phone Cleaner	69%
Totally TABS 2019 : Accurate Battle Simulator	60%
Master of Spin - Daily Reward Link	54%
Call On Duty Mobile Free Games: Offline Game	52%
Manga Geek - Free Manga Reader App	71%
QR Scanner Pro - Scan All QR Codes & Barcodes	62%
Video Editor Pro - Music, Vlog, Effect, Filter	62%

Detection Process

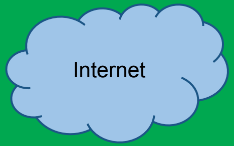


Agenda

Malware

ML-based Detection

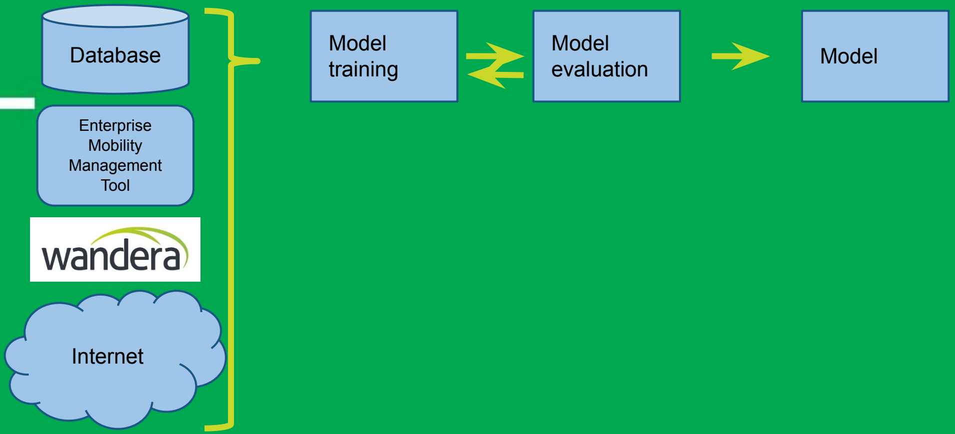
ML-based Detection as a Service



Malware

Types of Mobile Malware

- Ransomware
- Spyware
- Adware
- Trojan
- Rooting
- SMS-fraud
- Cryptojacking
- Banker


Database
Enterprise Mobility Management Tool
wandera
Internet
Model training
Model evaluation
Model

Training

Possible Data Sources

- Static analysis
  - App code (API calls, strings, domains, obfuscation, patterns, …)
  - App packaging (author, source, certificates, permissions, …)

- Dynamic analysis
  - Communication destinations IP/domain
  - Communication security / content / patterns
  - Device behaviour (battery drain, …)

- Manual inspection by threat ops team -- **possible trusted labels for training data**

- Internet databases and providers (free or paid = various quality)
  - IP/Domain blacklists
  - App analyses results, researches reports, … -- **possible labels for the training data**

Feature Vector Encoding

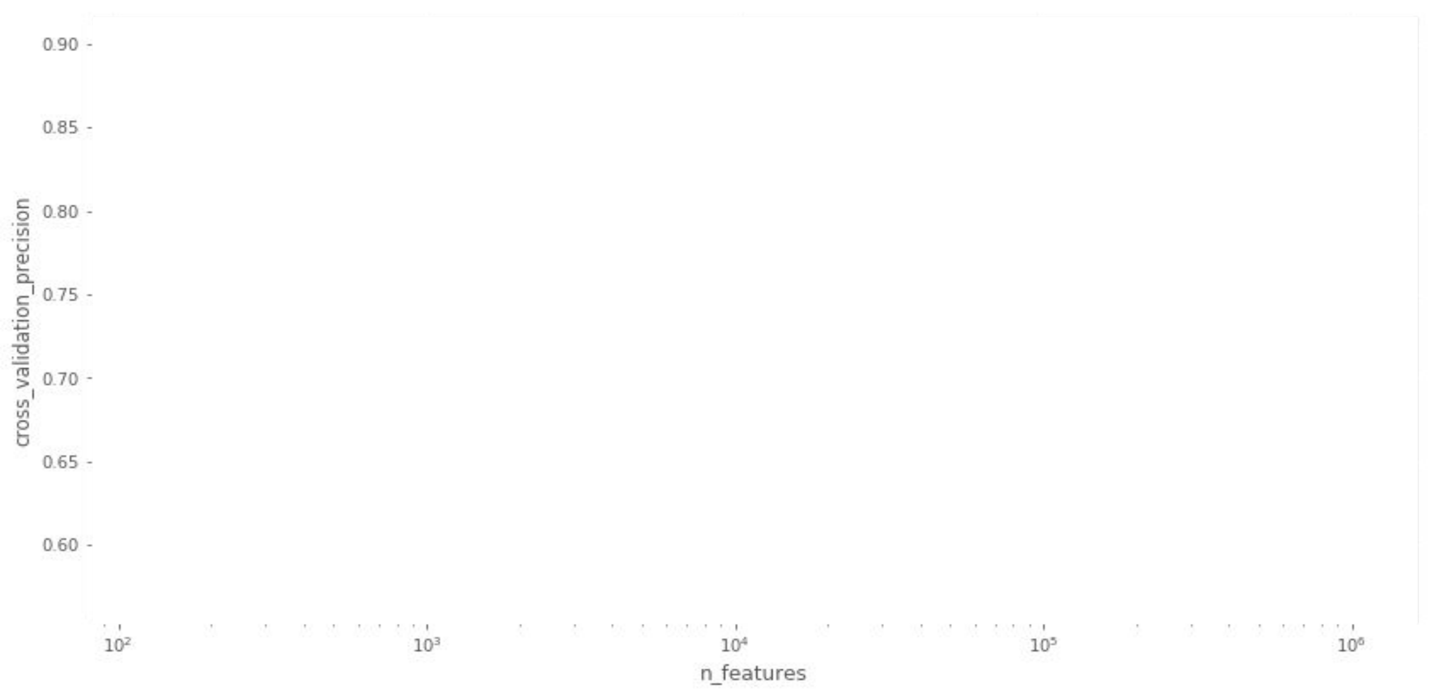
- Unique app identifiers -- SHA / MD5 hashes
- Feature vector
  - Very sparse binary vector (over million of elements and growing)
- Categorical features
- Sparse feature domain

Q. Shi, J. Petterson, G. Dror, J. Langford, A. Smola, and S. Vishwanathan:
Hash kernels for structured data. Journal of Machine Learning Research, 10(Nov):2615–2637, 2009

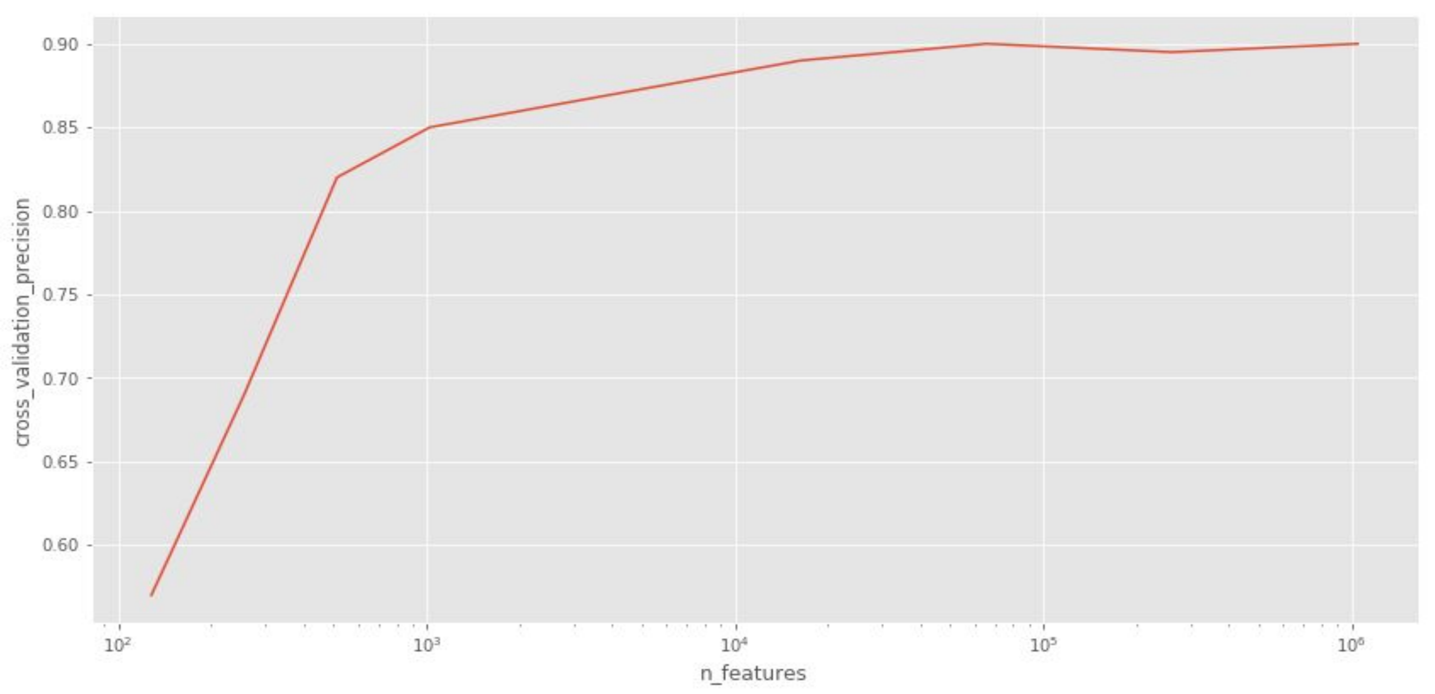
Feature Hashing

- Effective representation and encoding -- Feature Hashing
  - Handles increasing size of the binary vector
  - Fixed size of final feature vector
- Pros
  - No dictionary (mapping)
  - Preserves sparsity
- Cons
  - No inverse mapping
  - Hash collisions

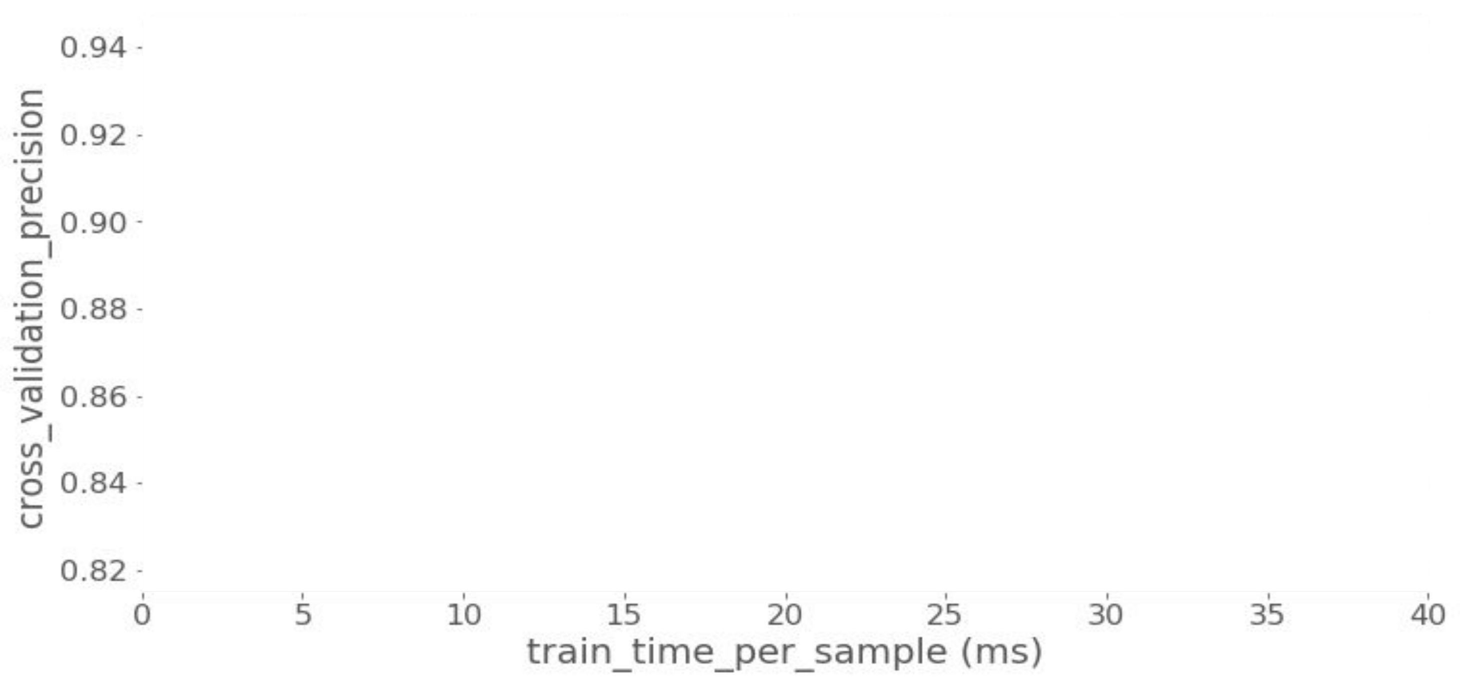
Hashing function - Number of Features after Hashing



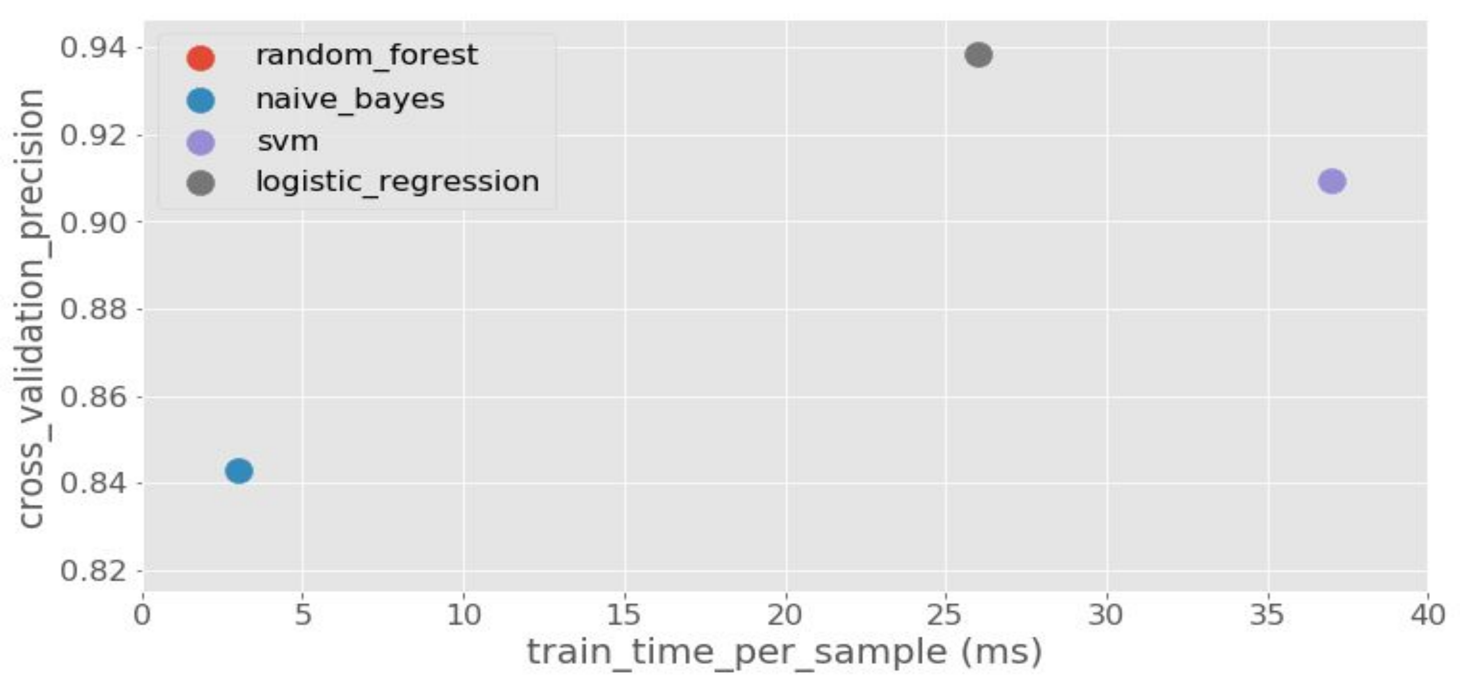
Hashing function - Number of Features after Hashing



Classification Algorithms

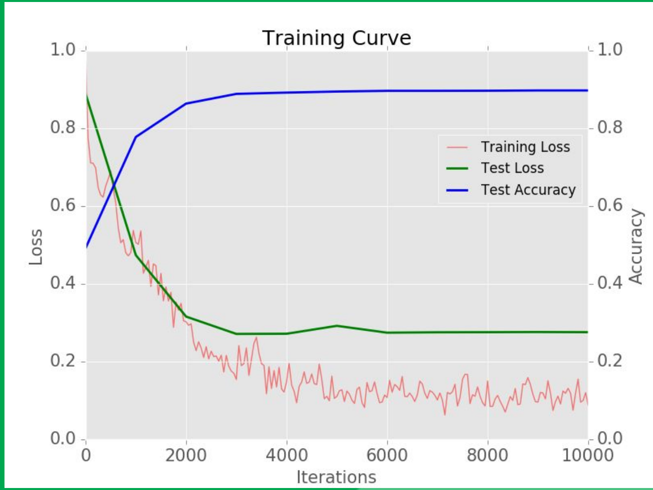


Classification Algorithms



Model Training Loop

- Supervised learning
  - Balanced training set
- Logistic regression
  - Limited memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) algorithm
- Training Process & Termination
  - No improvement in recent iterations
    Training loss stabilised
  - Overfitting
    Test loss increases
  - Number of iteration



Trained Model Evaluation -- QA

- Cross validation -- accuracy, precision, recall, ...
- Impact estimation -- validation on all previously classified samples
- Top 50 -- manually crafted sample set

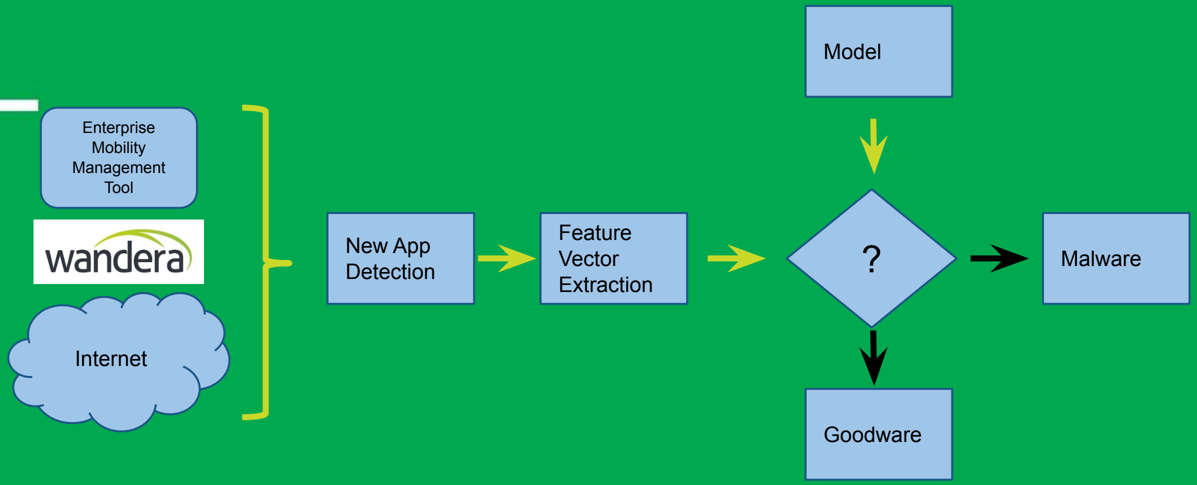
		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

Source: wikipedia

Online vs. Offline Learning

Online learning is still challenging in our environment

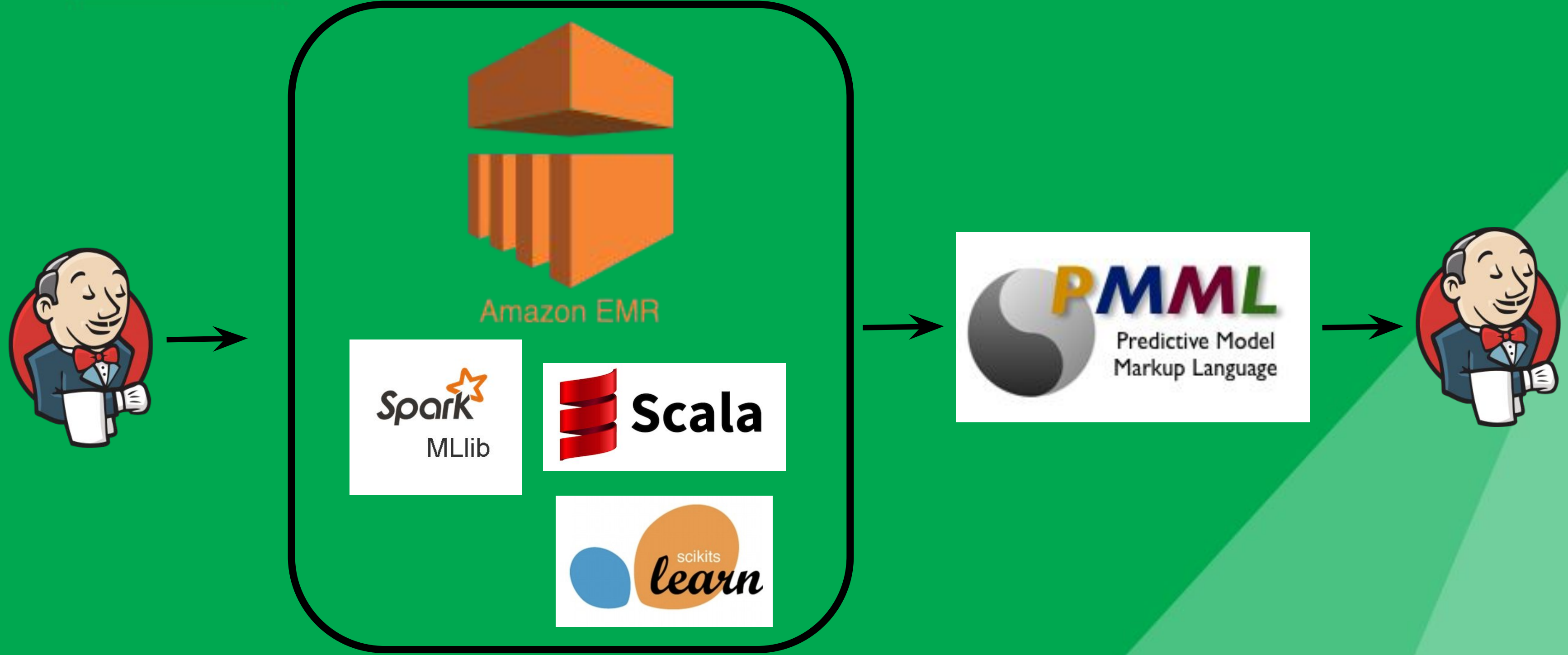
Offline learning with possibly high frequency of training



Malware Detection as a Service

Clear Separation of ML & Production

Technical Background - Full Automation



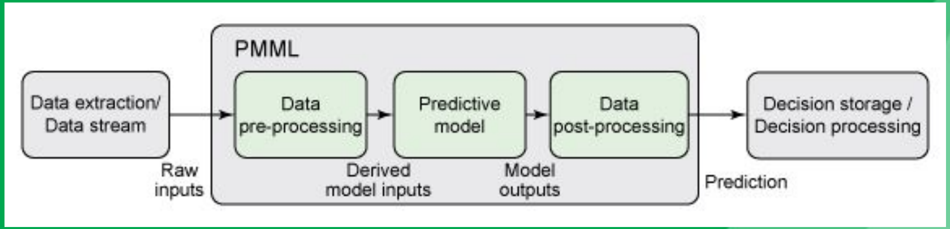
PMML

- Header
- Data Dictionary
- Mining Schema
- Data Transformations
- Model
- Targets
- Output
- Model Explanation
- Model verification

Pre-processing

Post-processing

http://dmg.org/pmml/



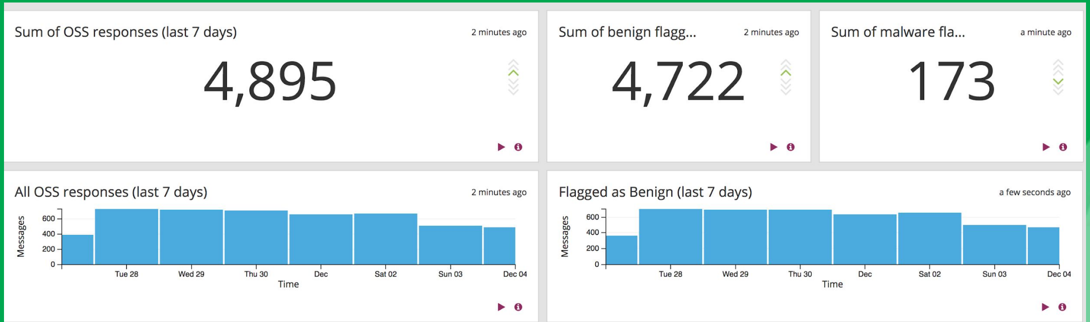

```
<NeuralLayer numberOfNeurons="2">
    <Neuron id="3" bias="-3.1808306946637">
        <Con from="0" weight="0.119477686963504" />
        <Con from="1" weight="-1.97301278112877" />
        <Con from="2" weight="3.04381251760906" />
    </Neuron>
    <Neuron id="4" bias="0.743161353729323">
        <Con from="0" weight="-0.49411146396721" />
        <Con from="1" weight="2.18588757615864" />
        <Con from="2" weight="-2.01213331163562" />
    </Neuron>
</NeuralLayer>
```

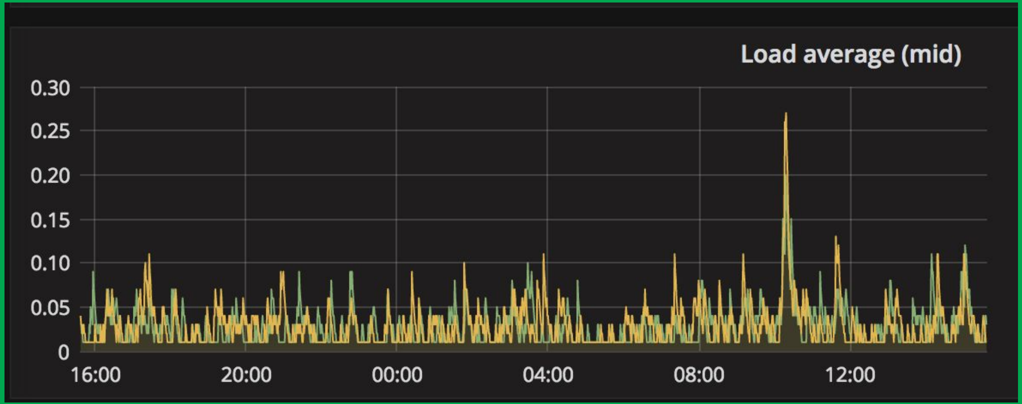
Scoring Service

- Model agnostic micro-service written in Java
  - Loads PMML (XML) and executes the model for given inputs
- “Low” resources requirements
- Super fast for small models
- Work distribution (data science vs. engineering)

Monitoring

- Metrics
- Logs
- Is it enough for DS?





Lesson Learnt: Overfitting Investigation 1/2

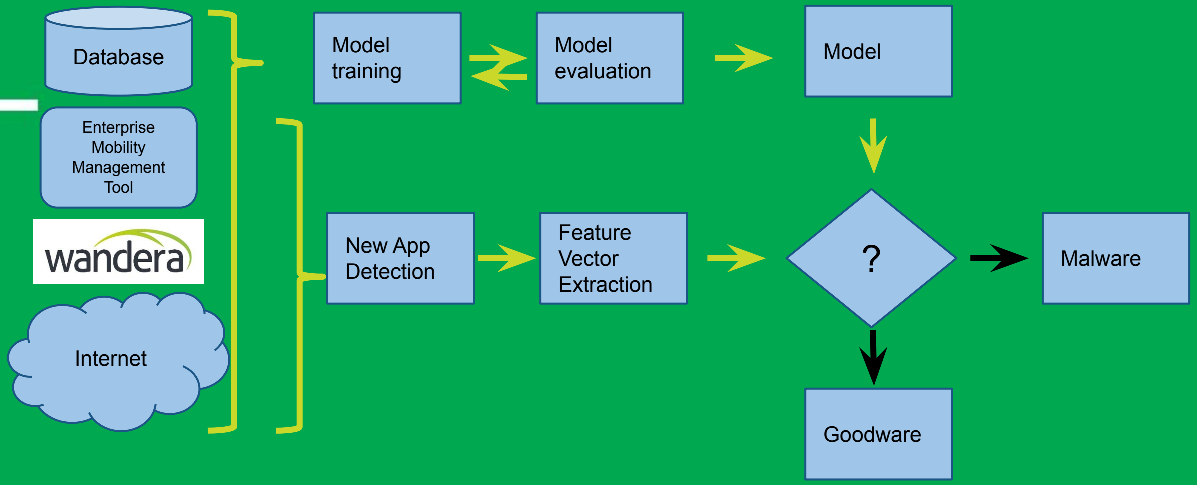
- Problem: Discrepancy between cross validation results and production results (accuracy, precision, recall)
- Contributing factors
  - Duplicates in training data (different app identifiers)
  - Size of feature space (hashing setting or model)
  - Data distribution 50/50 (goodware/malware)

Lesson Learnt: Overfitting Investigation 2/2

- Actions
  - Remove duplicates after hashing
  - Increased regularisation parameter (gradient descent parameter of logistic regression)
  - Data distribution reflects production

Other Lessons Learnt

- Hashing is super useful
- Overfitting / Underfitting -- Check accuracy, precision, recall, …
- Data can become really huge
- Check Openscoring vs. Spark results -- Look for implementation bugs
- Automate everything and document your decisions
- Data scientists and threat ops like to see context in logs and they like to query those data (marketing and users love stories and visualisations)


Database
Enterprise Mobility Management Tool
wandera
Internet
Model training
Model evaluation
Model
New App Detection
Feature Vector Extraction
?
Malware
Goodware

Thank You

Wandera Ecosystem