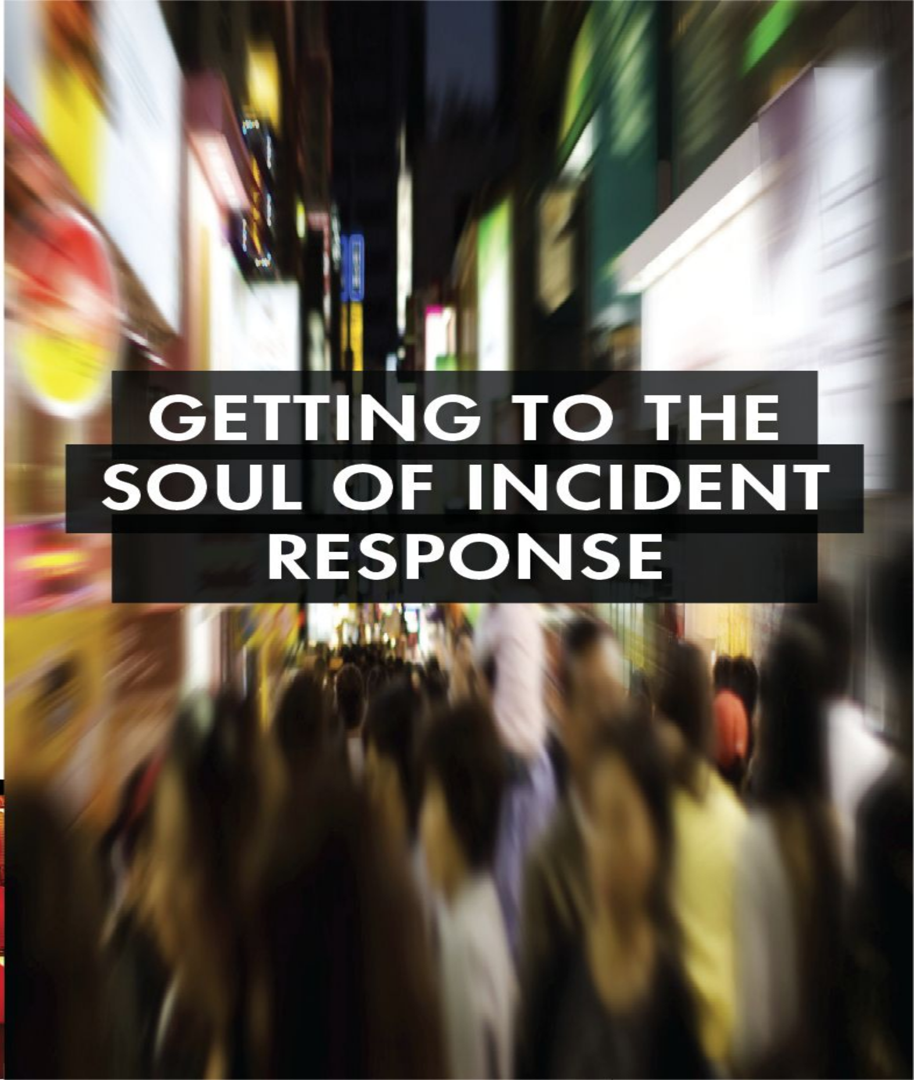
28th ANNUAL FIRST CONFERENCE

SEOUL

JUNE 12 - 17, 2016

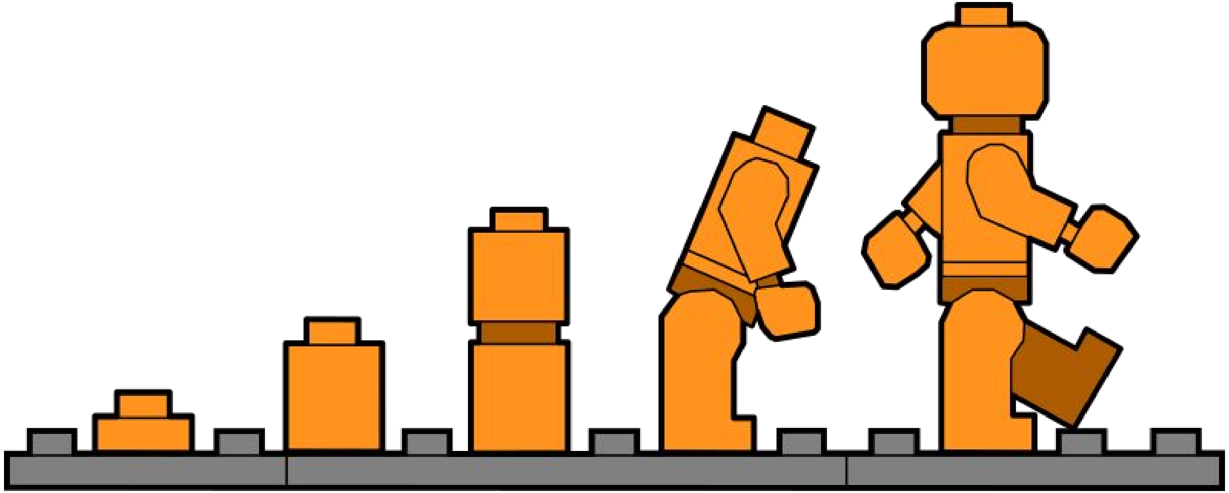


GETTING TO THE SOUL OF INCIDENT RESPONSE

Current AVs only detect

70%

of web shells



web shell classification at scale

About me

Thomas Kastner, MSc.

nimbusec

- Fast detection of hacked websites

SBA Research

- Research center for information security

Definition: web shell

A web shell is a **script** that can be uploaded to a **web server** to enable **remote administration** of the machine. …

A web shell can be written in **any language** that the target web server supports. The most commonly observed web shells are written in languages that are widely supported, such as **PHP** and ASP. Perl, Ruby, Python, and Unix shell scripts are also used. …

US-CERT TA15-314A

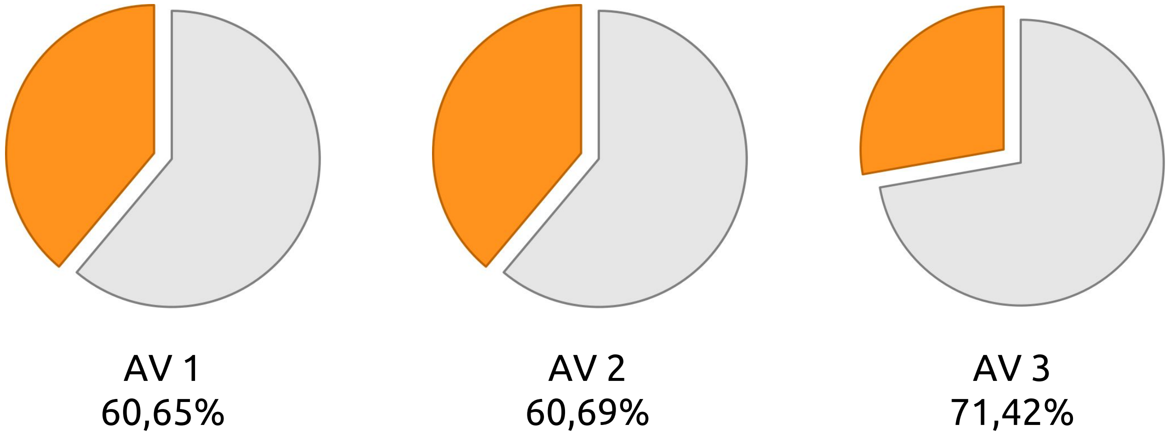
Detection is difficult

Due to the potential **simplicity** and **ease of modification** of web shells, they can be **difficult to detect**. For example, anti-virus products sometimes produce poor results in detecting web shells. …

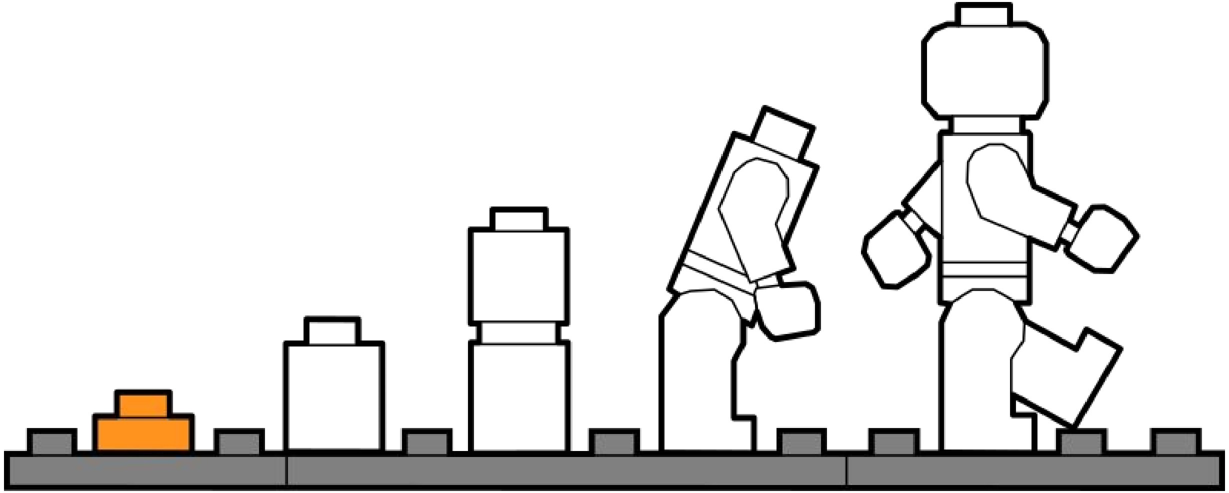
US-CERT TA15-314A

Detection is difficult

Webshell detection rate



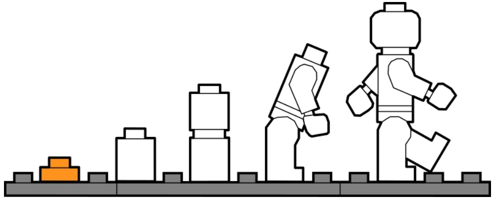
NeoPI & Statistics



NeoPI & Statistics

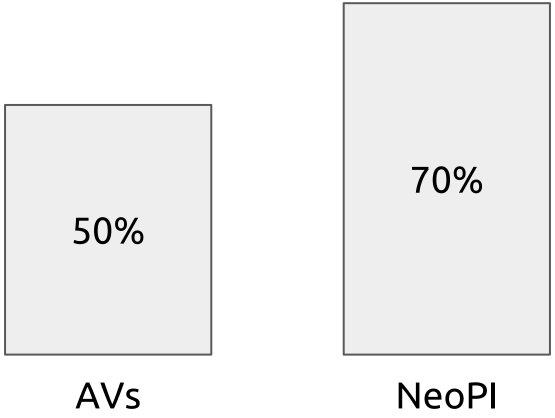
Statistical methods to detect obfuscated content

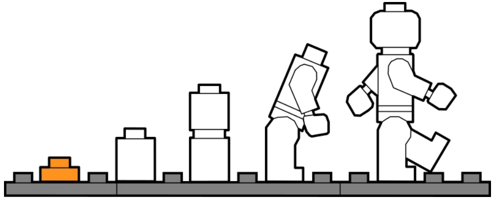
- Index of Coincidence
- Entropy
- Compression
- Longest Word
- Poison words



Real World Data

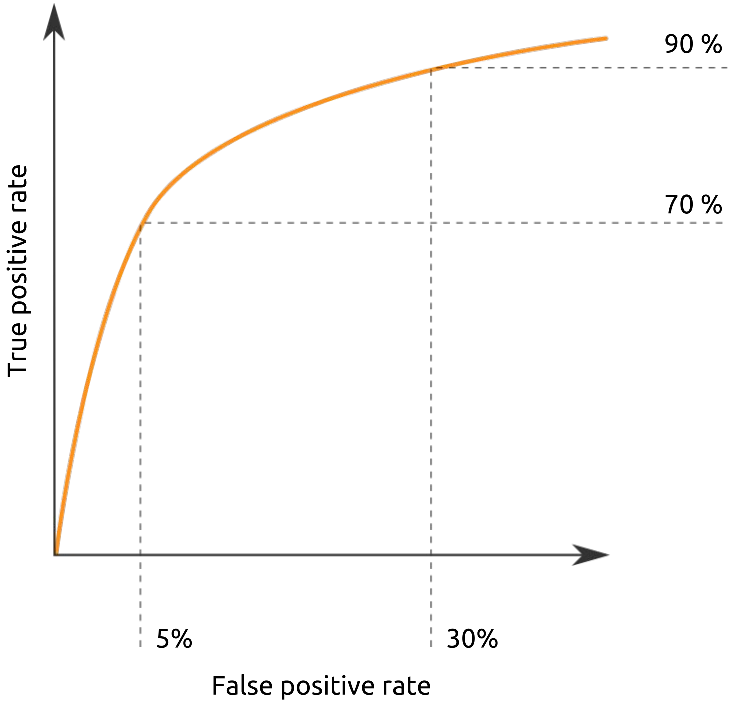
2013

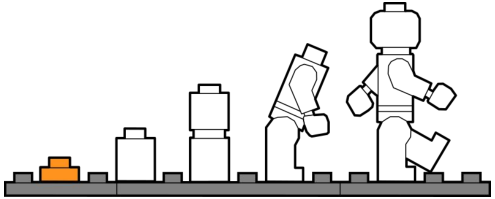




Real World Data

2013



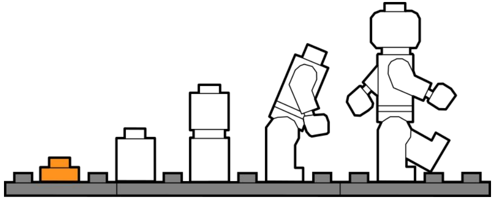


What we learned so far

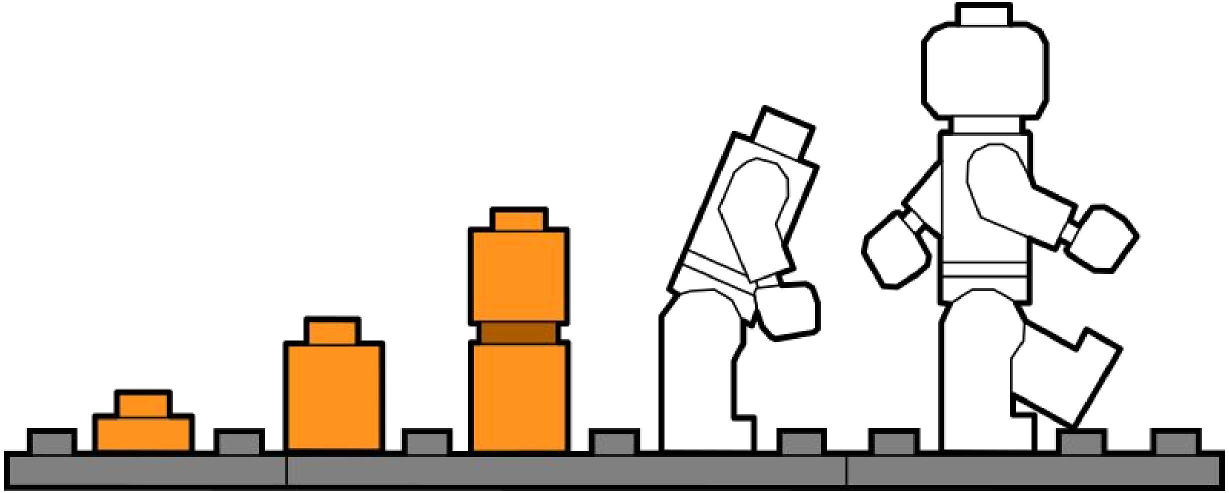
NeoPI was meant for human analysts

Search for thresholds via grid search

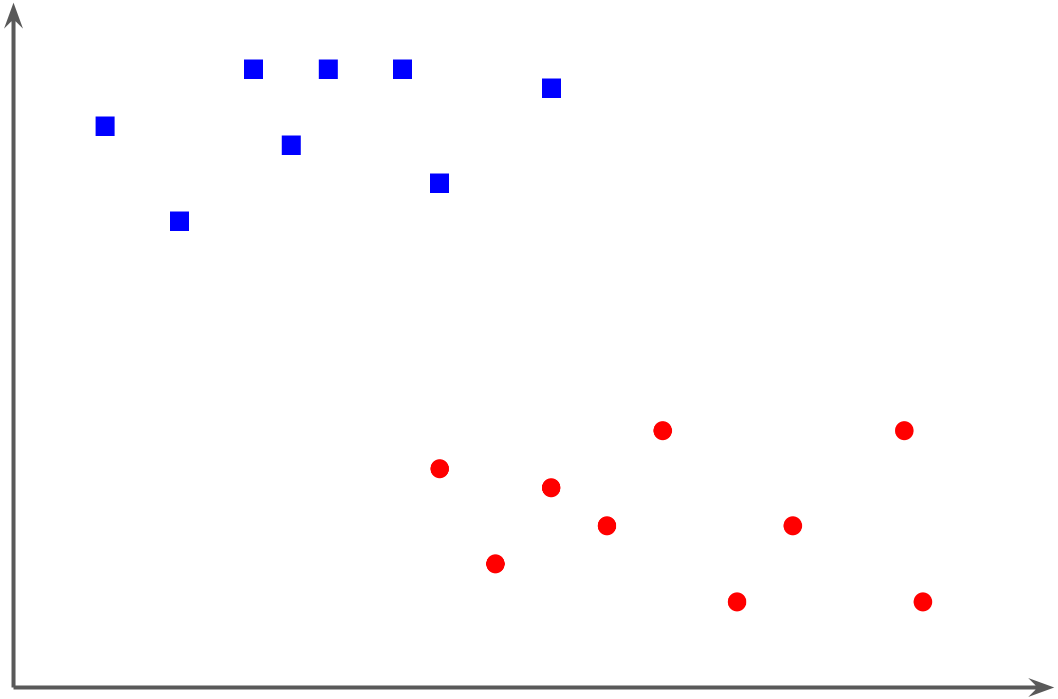
- Marketing called it already machine learning

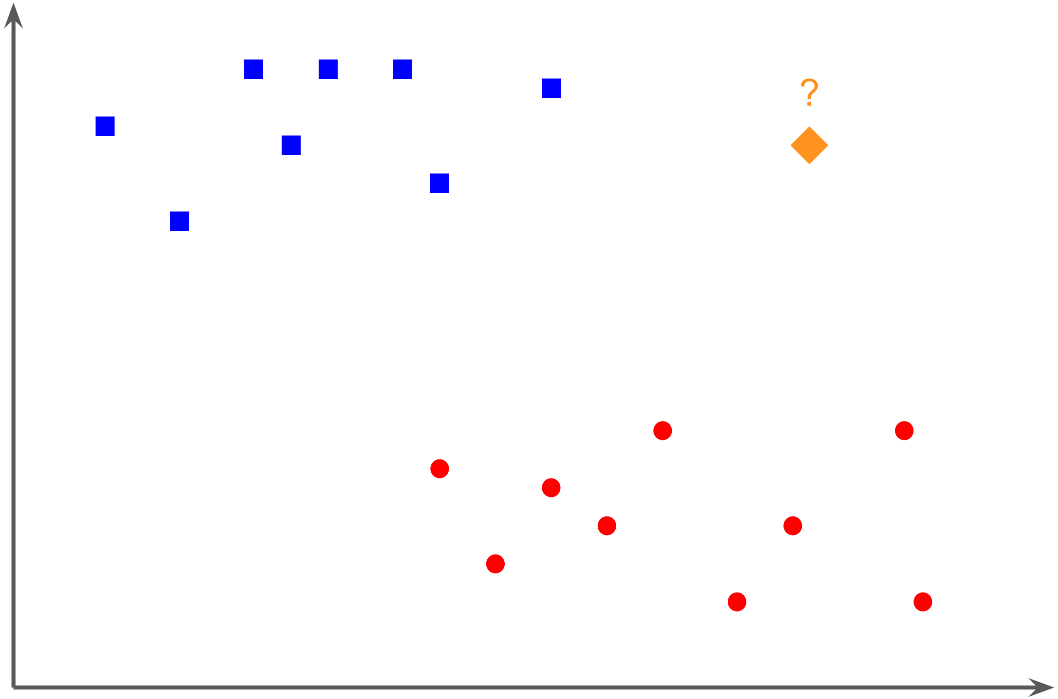


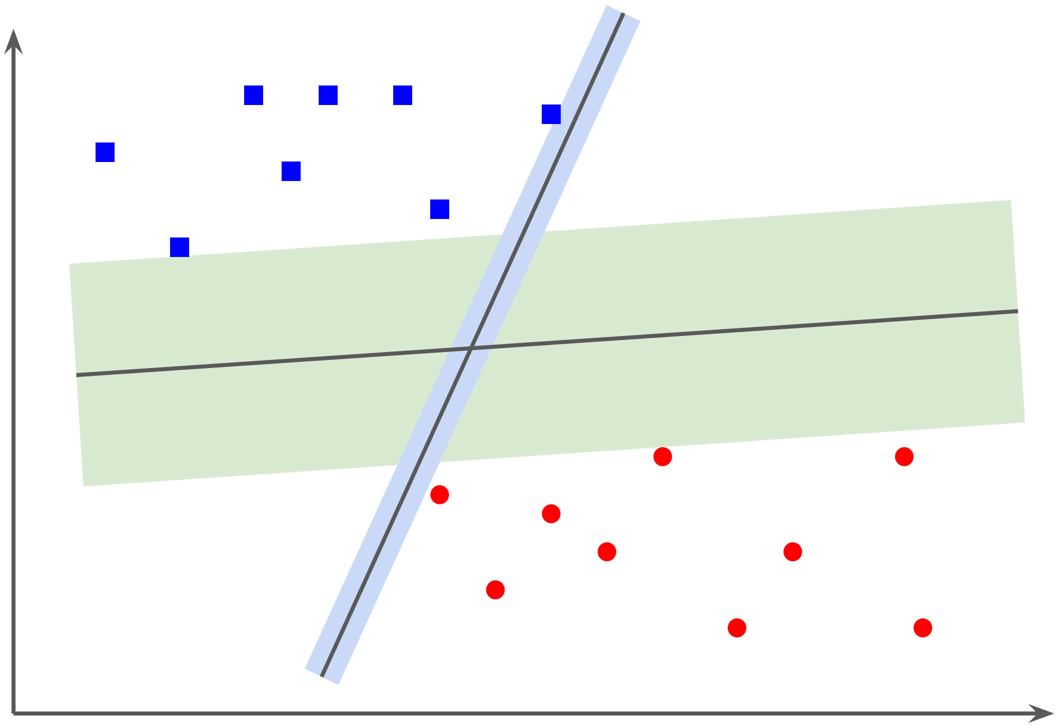
Support Vector Machine

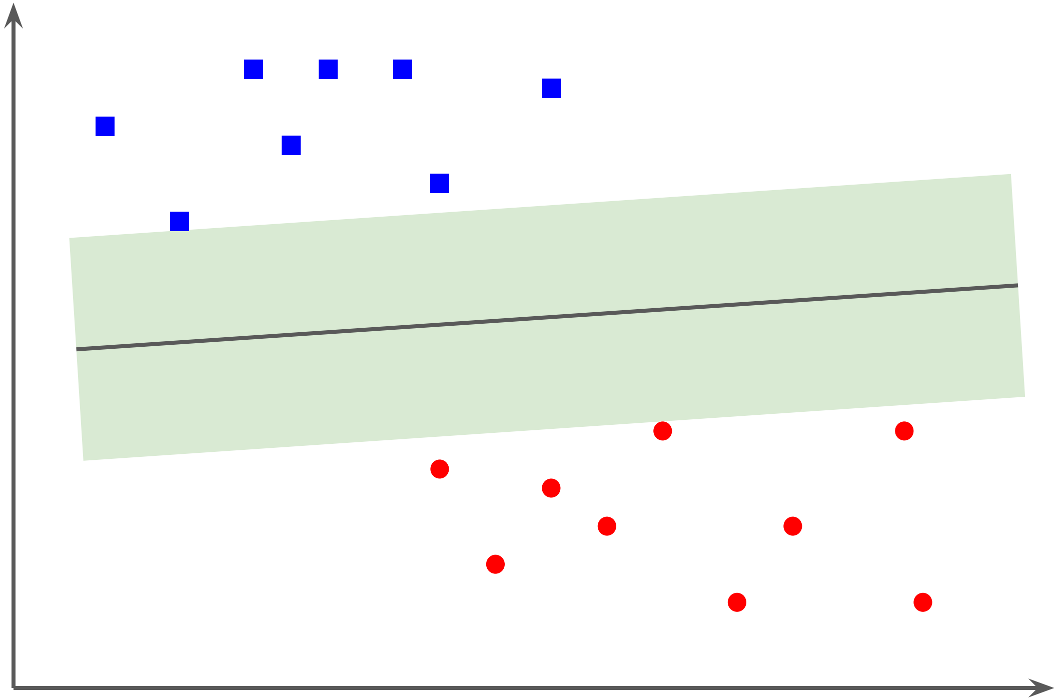


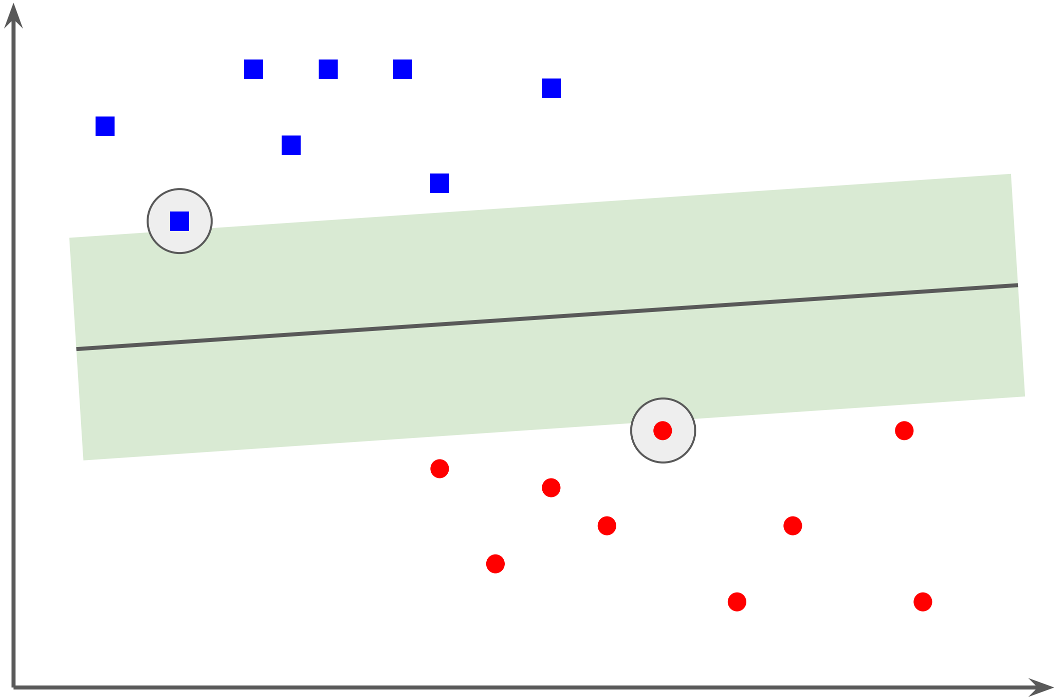
2013 **2014** 2016

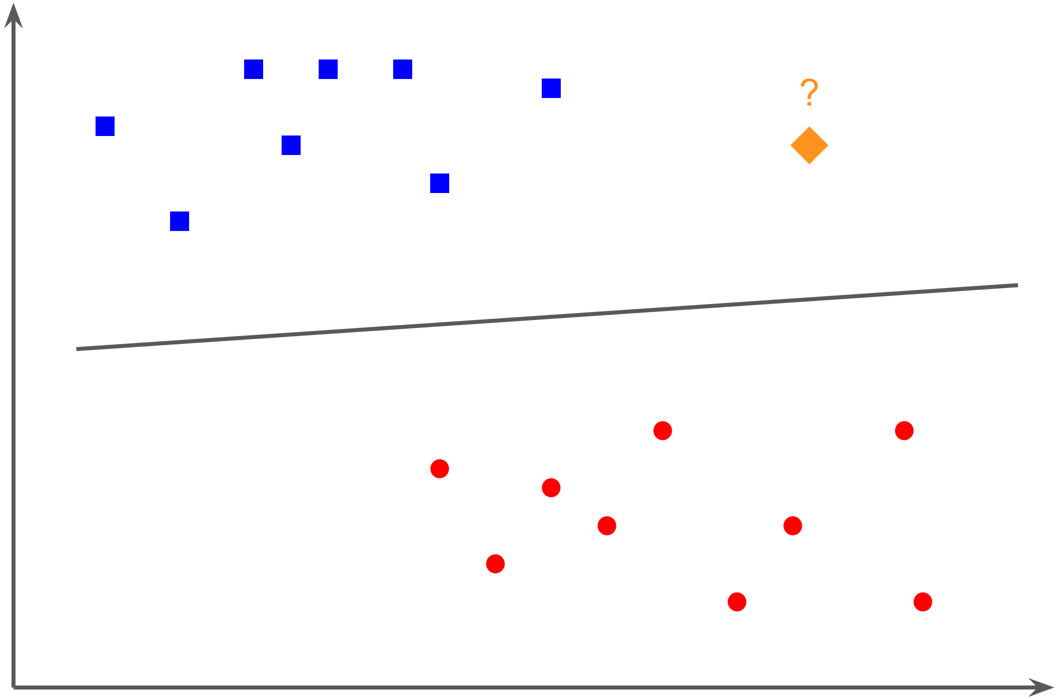


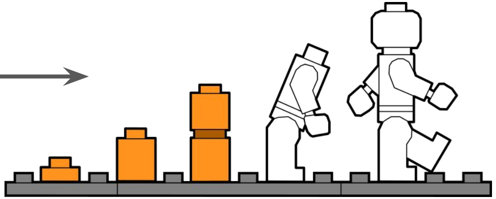




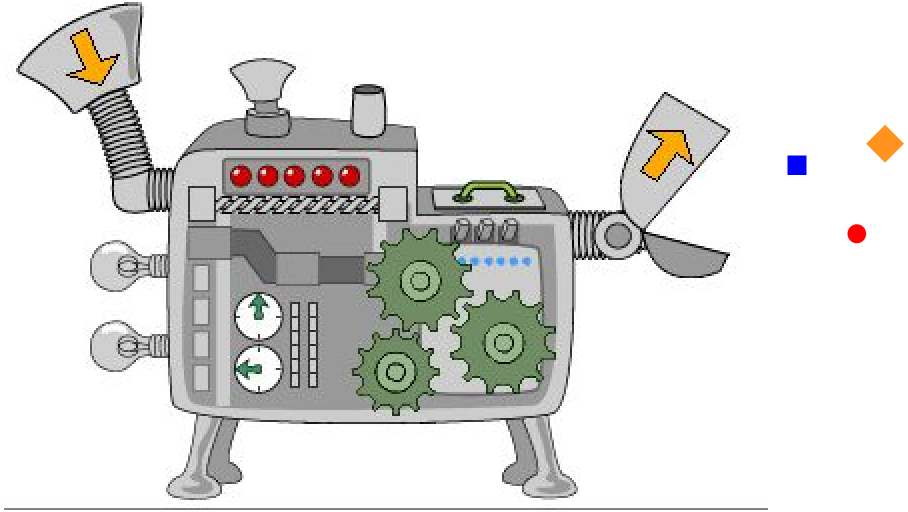


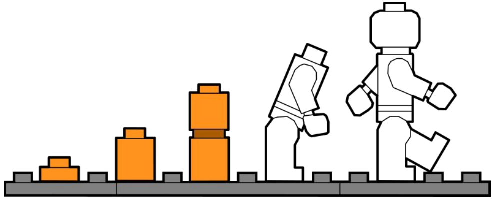


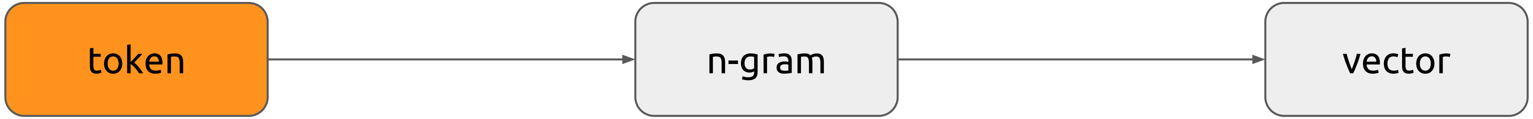




```
<?php
eval(gzip('...
?>
```







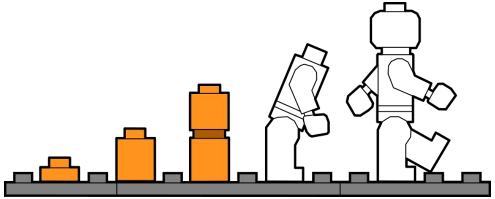
token

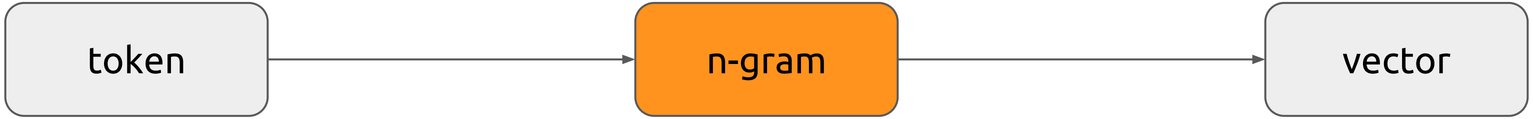
n-gram

vector

```
<?php

eval(base64_decode
(“bWFrZSBuaW1idXNlYyBncmVhdCBhZ
2Fpbg==”));

?>
```

```
function ( function ( string
) ) ;
```





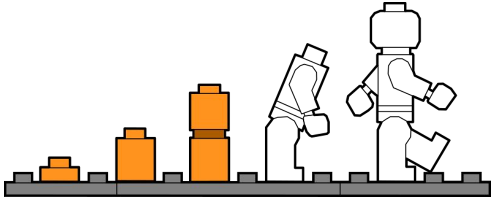
token

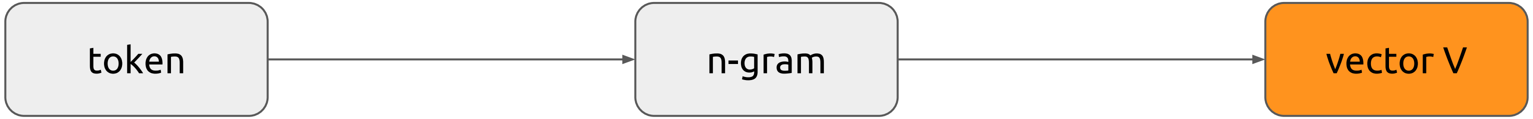
n-gram

vector

```
function ( function ( string
) ) ;
```

```
function ( function

( function (

function ( string

( string )

string ) )

) ) ;
```



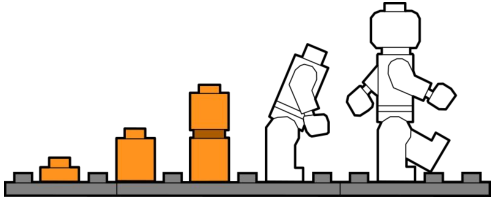


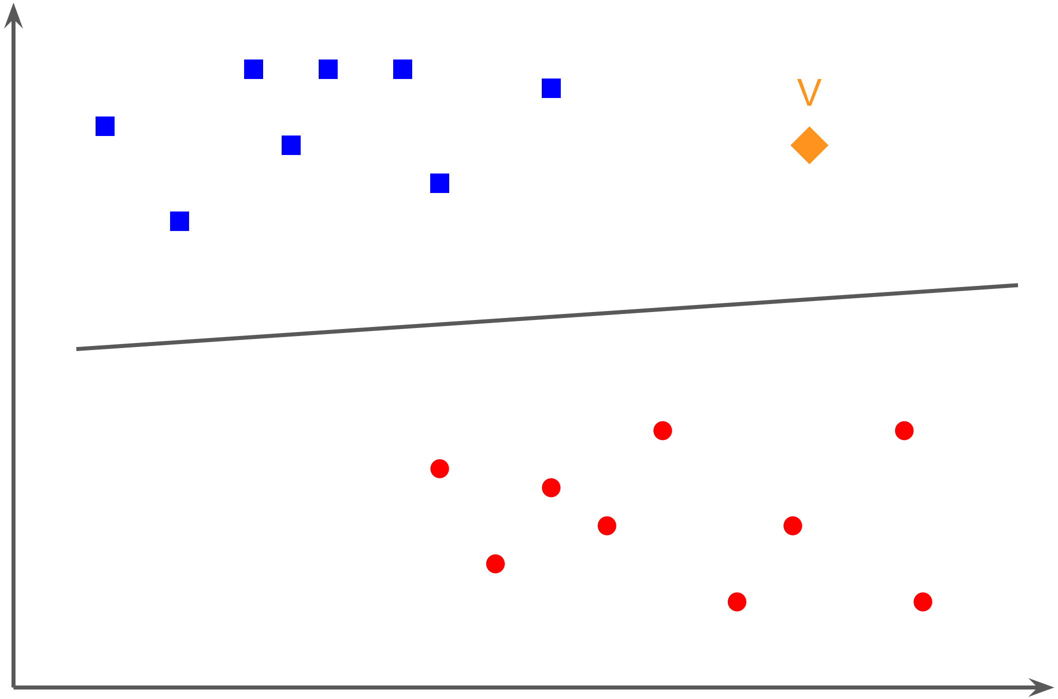
token

n-gram

vector V

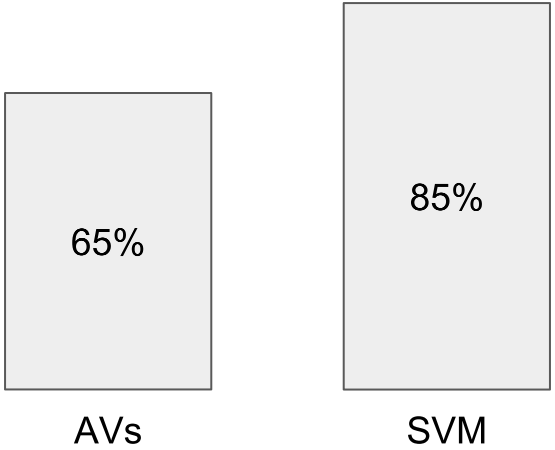
function (function	$131_6 = 55$
(function (	$313_6 = 117$
function (string	$132_6 = 56$
(string)	$324_6 = 124$
string))	$244_6 = 100$
)) ;	$445_6 = 173$

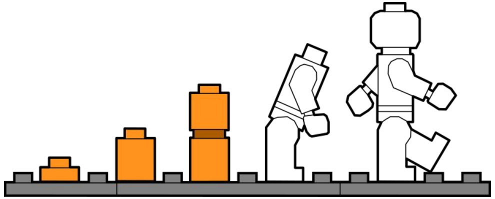



V

Real World Data

2015





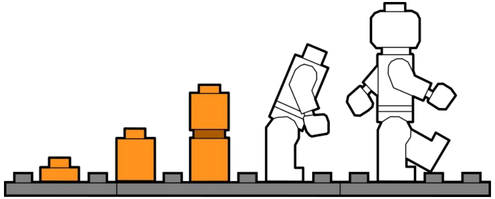
Real World Data

2015

99.9%

Accuracy

5-fold cross-validation



Real World Data

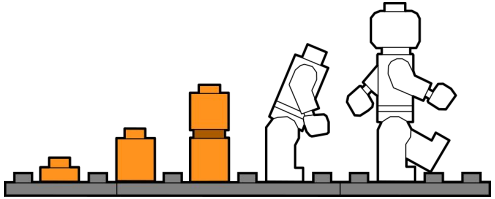
2015

Accuracy paradox

Benign samples:	2,500,000
Malicious samples:	3,000

$$Acc = \frac{TN + TP}{TN + TP + FN + FP}$$

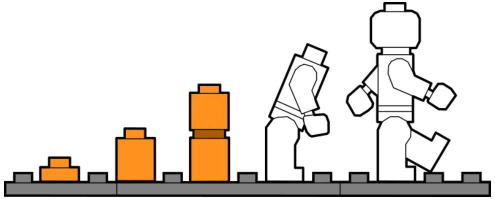
$$Acc = \frac{2500000 + 0}{2500000 + 0 + 3000 + 0} = 99.8\%$$



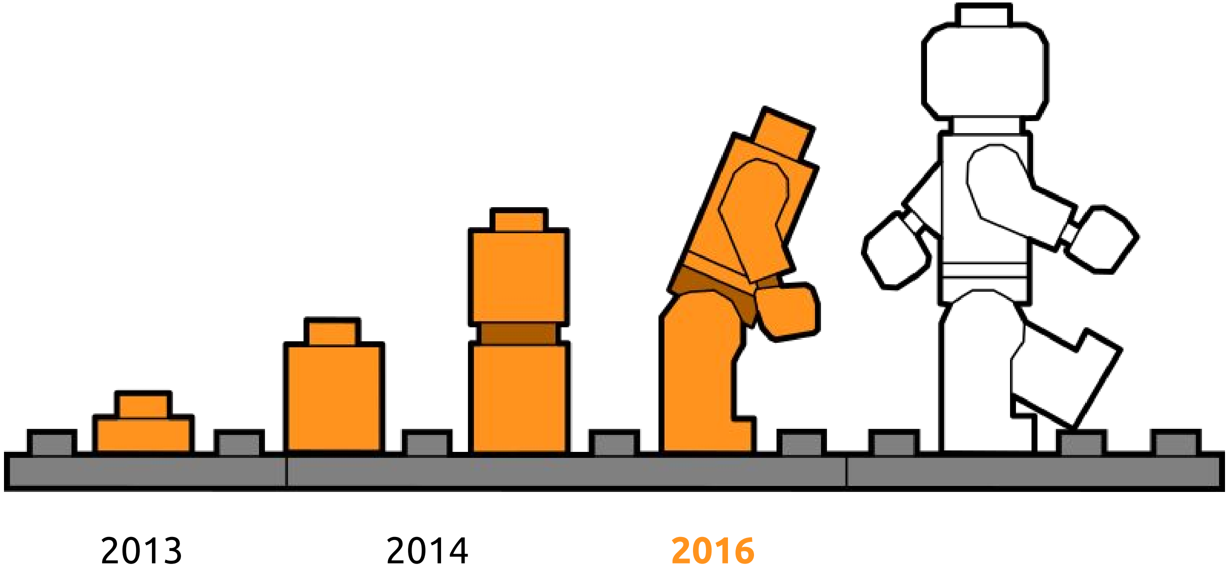
What we learned

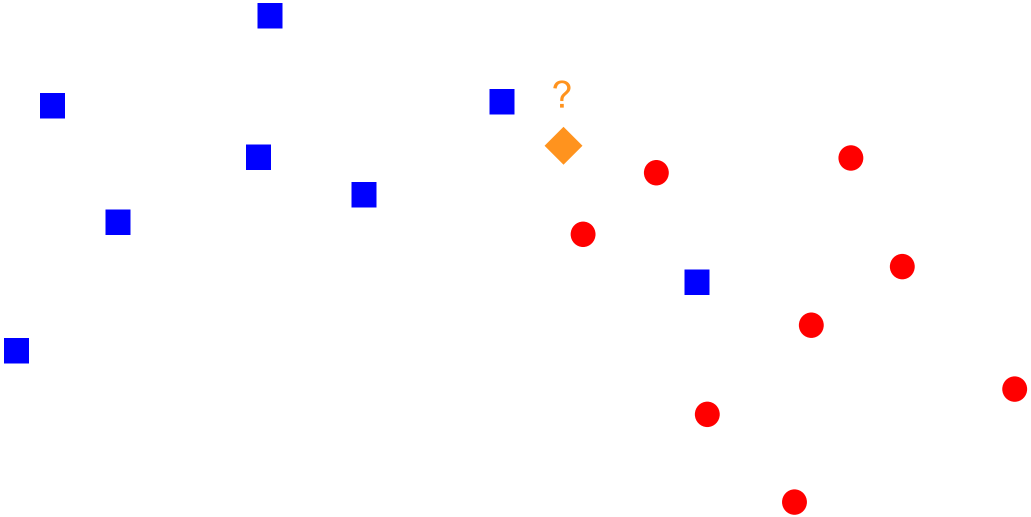
Forget accuracy (and most other metrics)

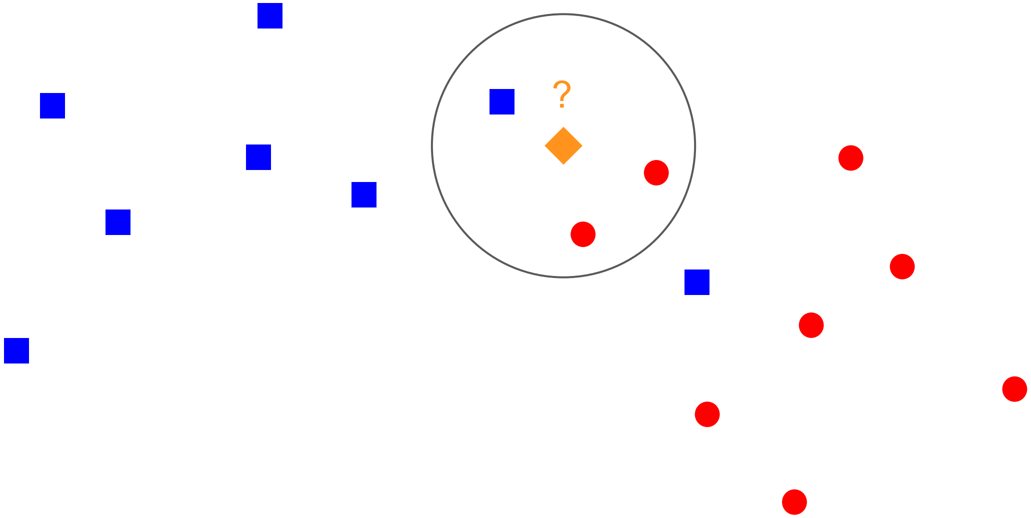
- use TPR and FPR instead



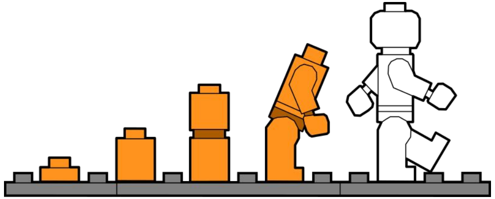
k-Nearest Neighbor

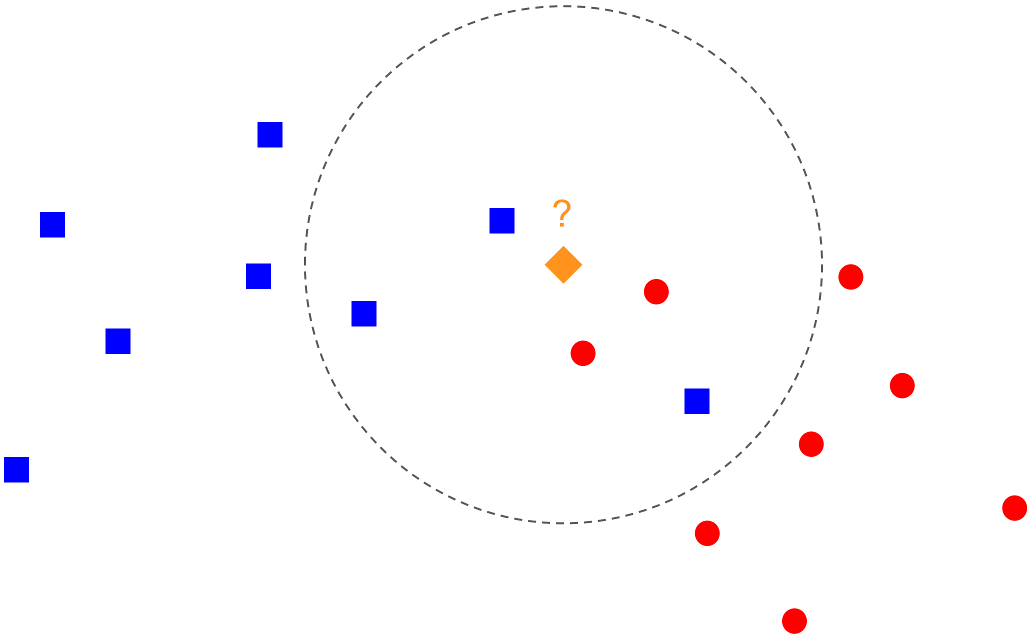




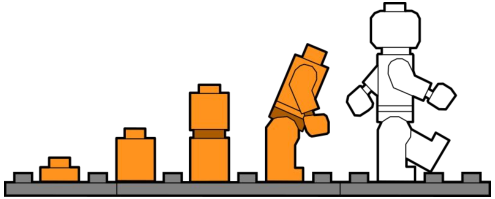


k = 3





k = 5



k-Nearest Neighbors

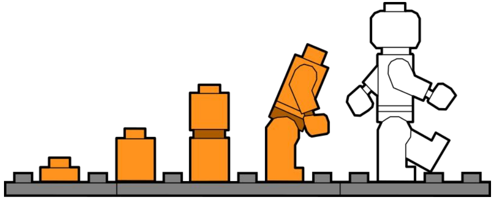
Distance metrics:

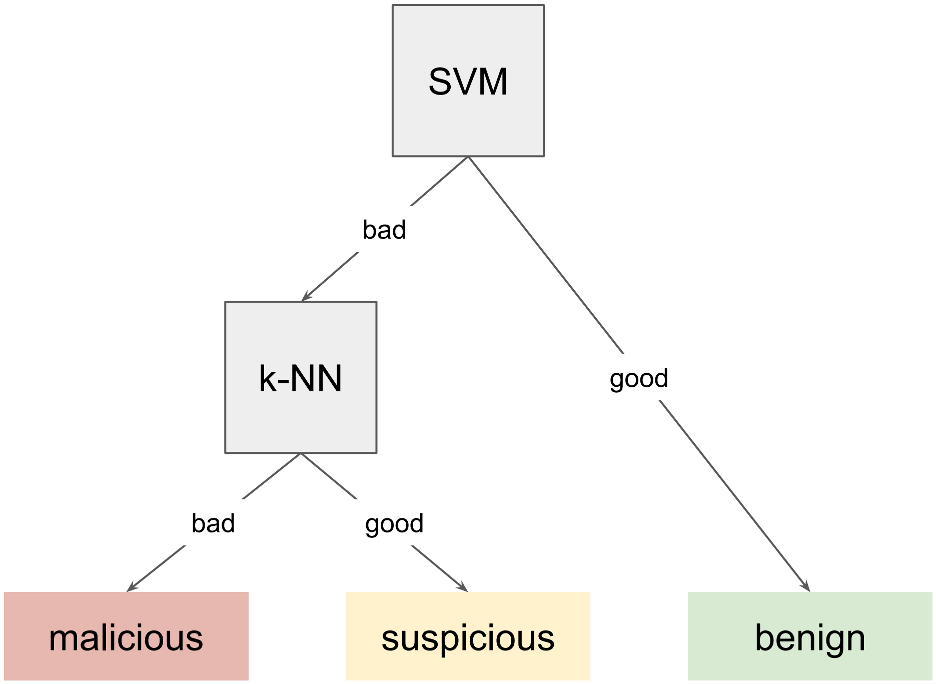
- Euclidean distance
- Hamming distance

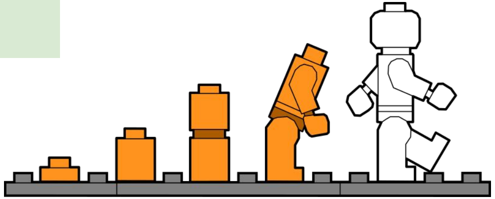
Select *k* by hand or via heuristic

Take distance into account

- weight = 1 / distance




SVM
bad
good
k-NN
bad
good
malicious
suspicious
benign



Benchmark

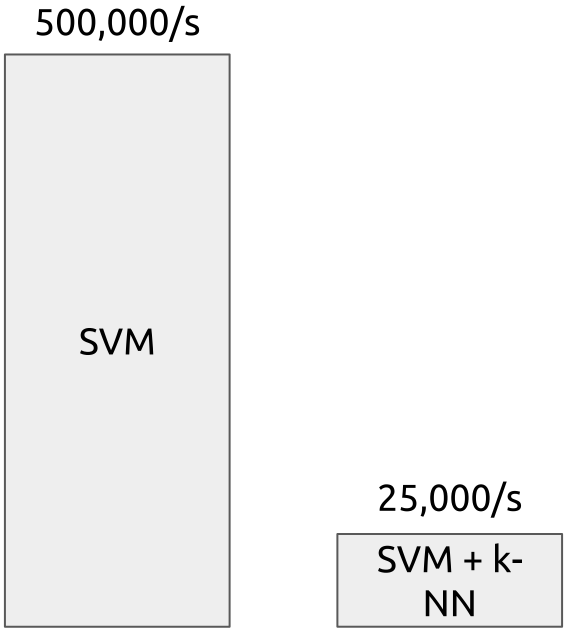
Intel Core i5-3340M CPU @ 2.70GHz

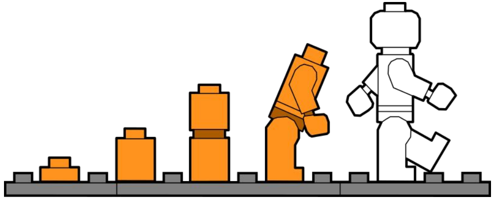
2x8GB DDR3/1600 SODIMM

(SAMSUNG SSD SM841 256GB)

35,000 files

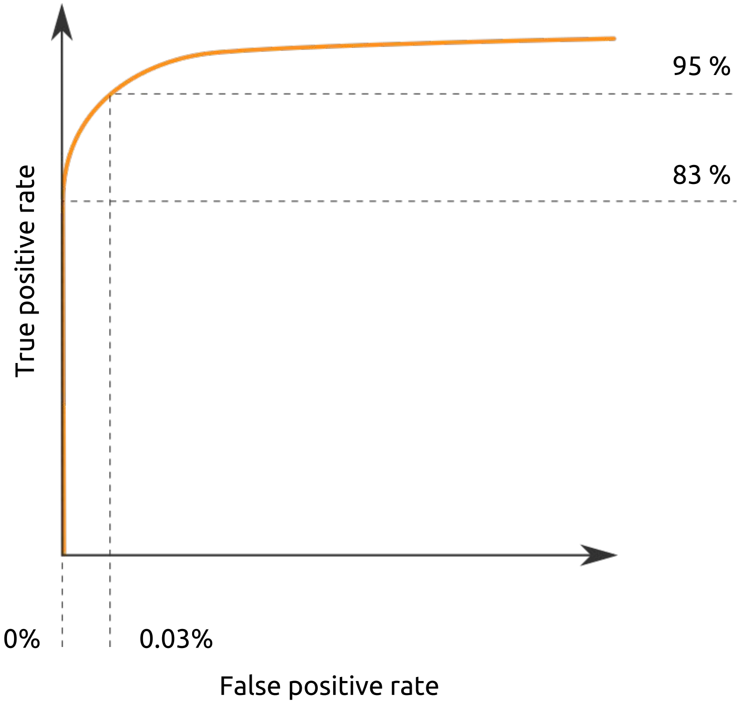
12,000 infected

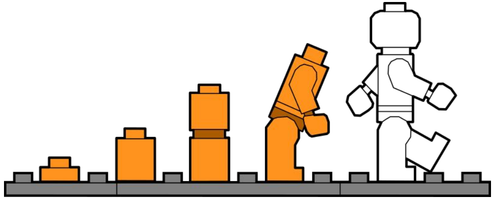




Real World Data

2016



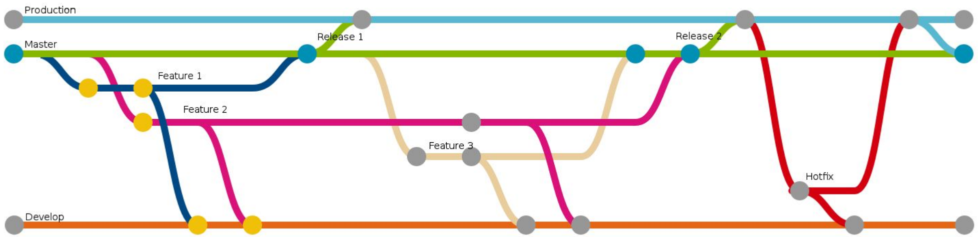


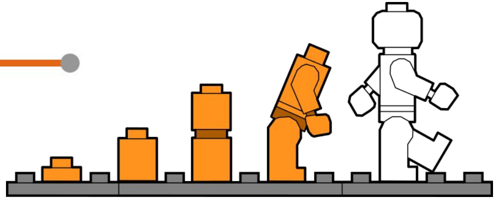
What we learned

16,500,000 classifications / day

Treat training set (web shells) like code

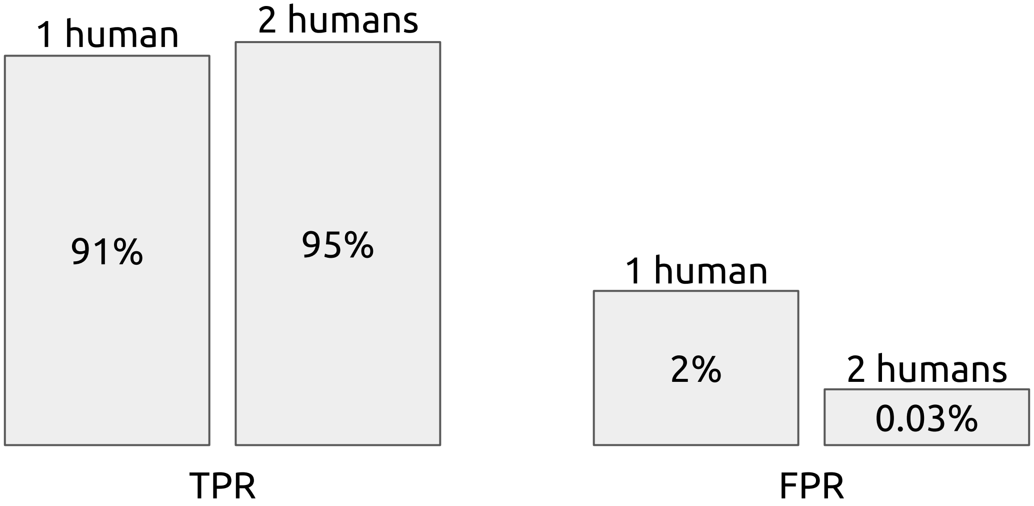
- Version control
- Unit tests

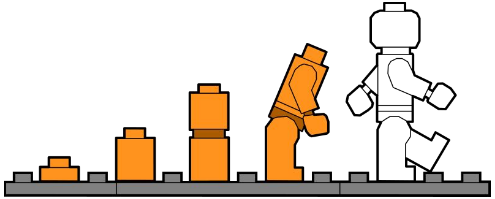




What we learned

Quality of training set

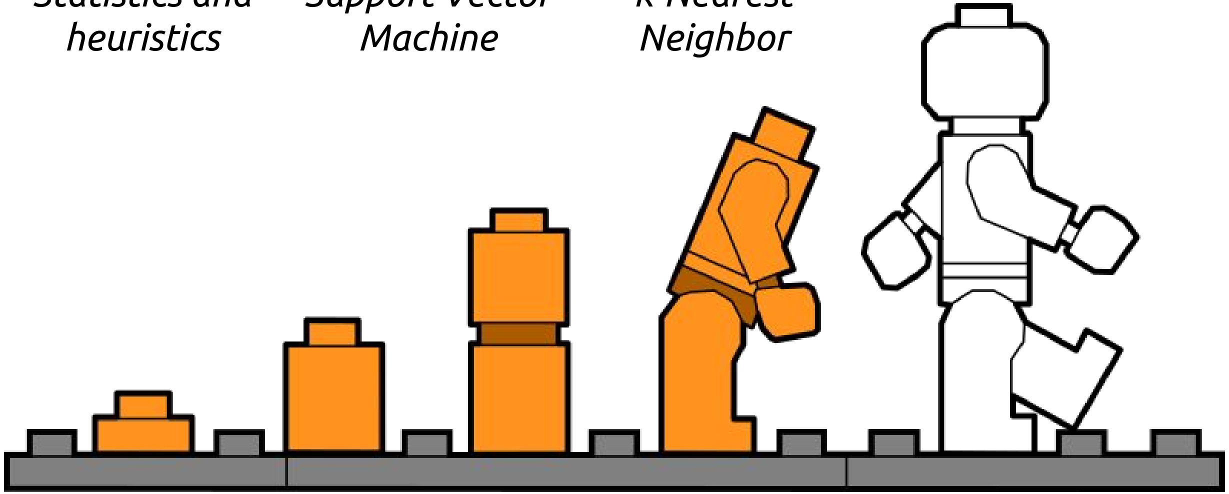




Statistics and heuristics

Support Vector Machine

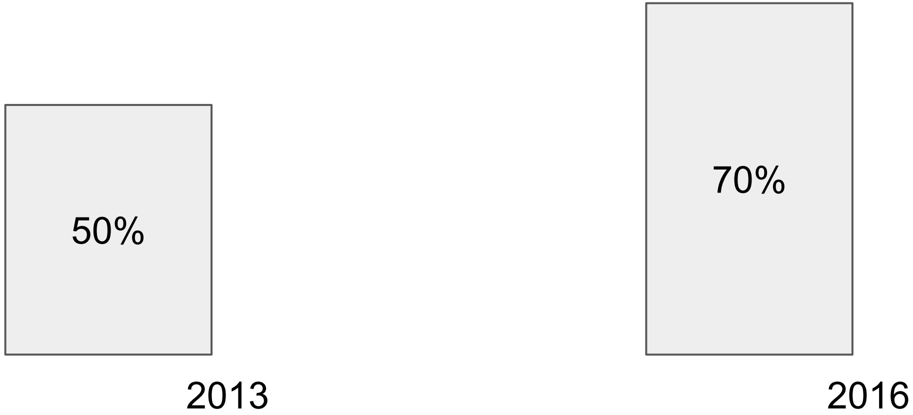
k-Nearest Neighbor

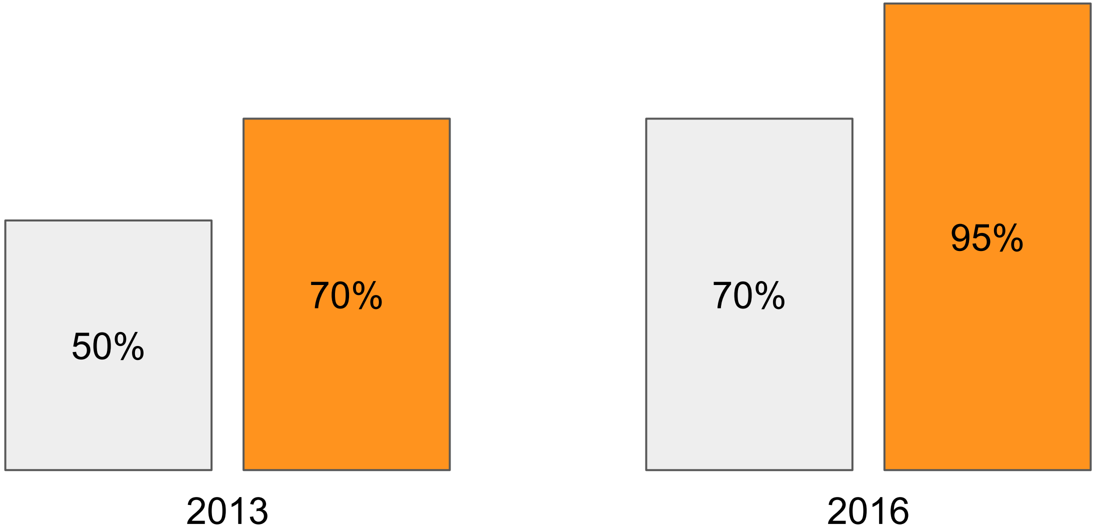


2013

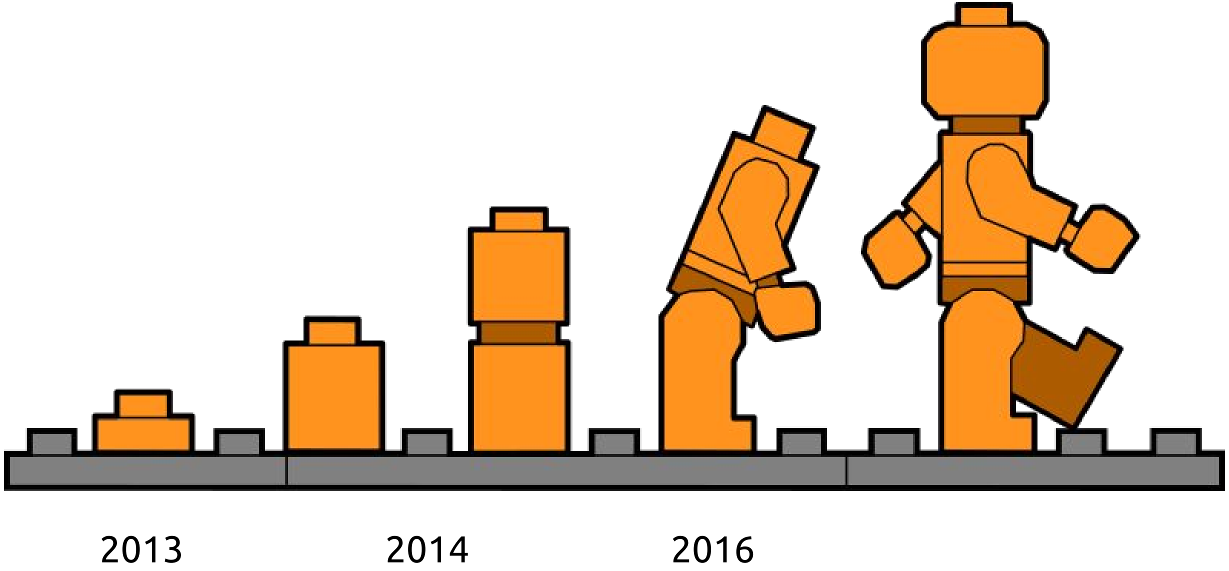
2014

2016


50%
70%
2013
2016


50%
70%
70%
95%
2013
2016

Justin Griner © 2007


2013
2014
2016

Future improvements

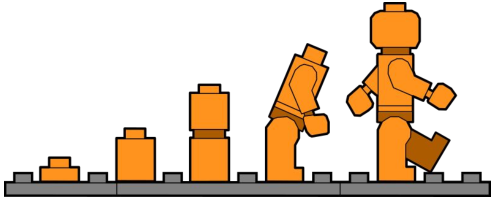
Different strategies for tokenization

New machine learning algorithms

- Deep Learning
- Neural Networks

New frameworks

- Tensor Flow
- DSSTNE



nimbusec
website security monitor

we are currently here

Grossglockner: 3,798 m (12,461 ft)