Practical AI

Malware Classifier from scratch

Charles Chang 張佳彥

TREND MICRO™

Securing Your Journey to the Cloud

WHO AM I

- Join Trend Micro on 2009
  - Infra Developer
  - Threat Researcher
  - Machine Learning Researcher
- Join XGen ML project on 2015
- Now leading the Machine Learning Research/Operation team of XGen

TREND MICRO

Agenda

- XGen Malware Classifier Introduction
- How to Start
- What We Have Done
- What's Next
- Highlight

TREND MICRO

Securing Your Journey to the Cloud

XGen Malware Classifier

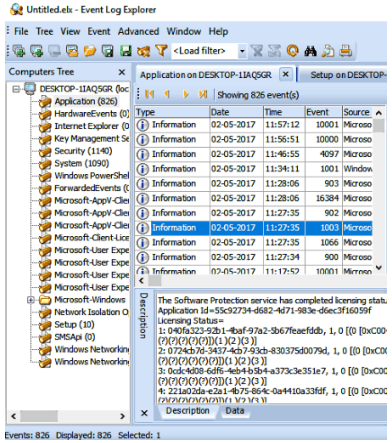
Static files

ML model

Virus (Malicious)

Normal

Behavior logs



ML model

Virus (Malicious)

Normal

TREND MICRO

Trade-off

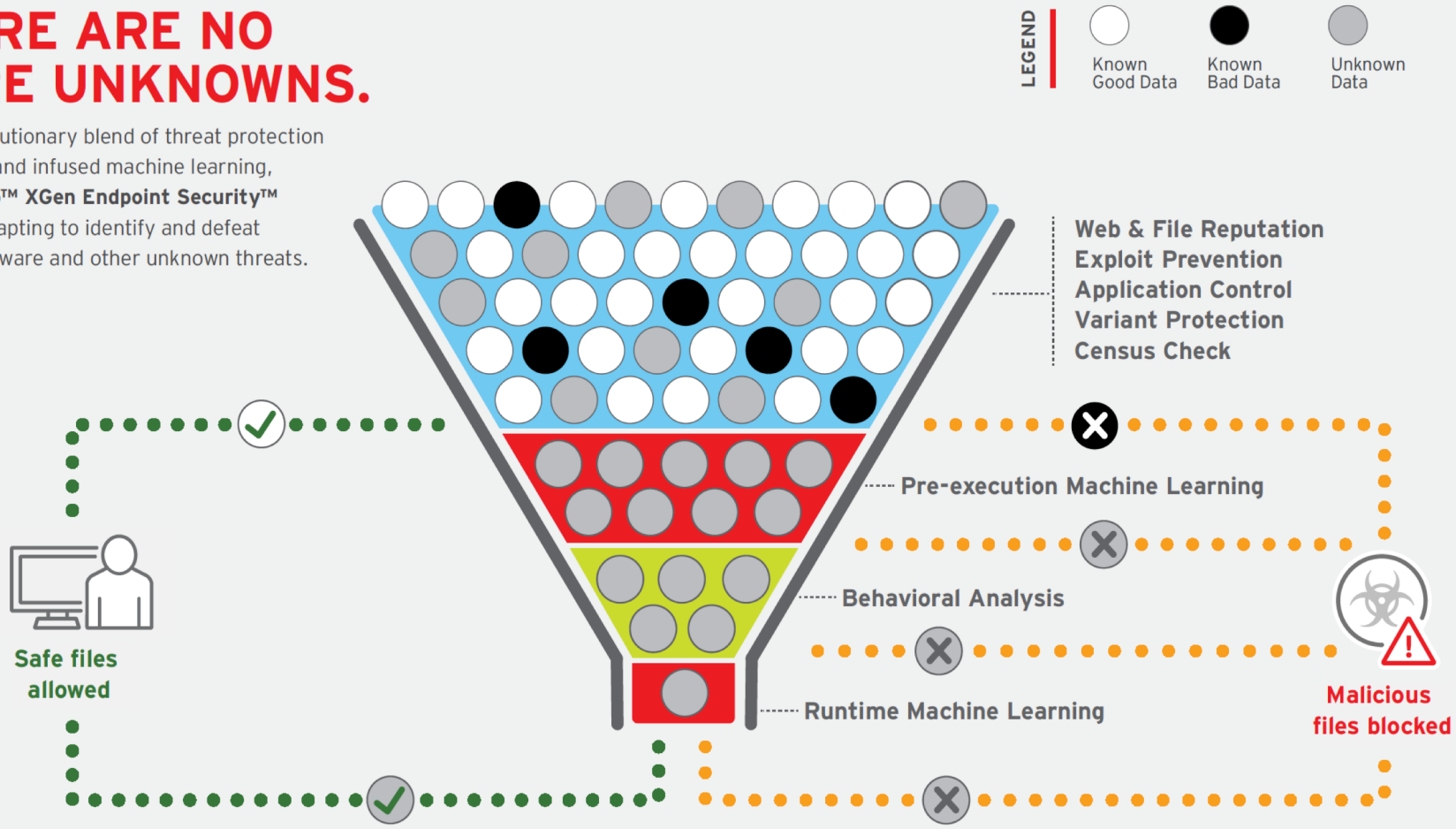
- File machine learning:
  - Pros: fast, low cost
  - Cons: need fast model rolling, obfuscation problem

- Behavior machine learning:
  - Pros: malicious evidence as the feature
  - Cons: Expensive, detect after bad things happened

XGen ML – Layer protection

THERE ARE NO MORE UNKNOWNS.

With its evolutionary blend of threat protection techniques and infused machine learning, **Trend Micro™ XGen Endpoint Security™** is always adapting to identify and defeat new ransomware and other unknown threats.



TREND MICRO

Securing Your Journey to the Cloud

How to Start

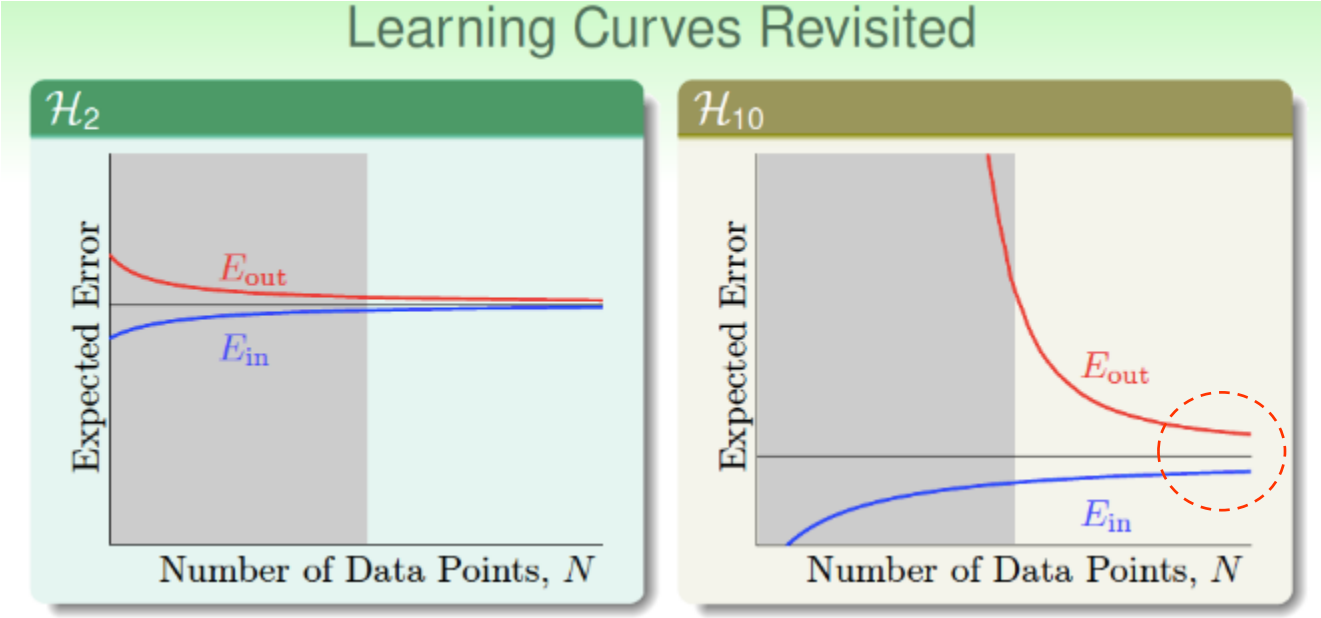
What is your problem ?

- Invariant V.S. Variant
  - Invariant
    - 手寫辨識
    - 語音辨識
    - 圍棋

> 750 Million NTD

  - Variant
    - 金融預測
      - Daily update
    - 病毒預測
      - Monthly update

The best solution



https://zh-tw.coursera.org/learn/ntumlone-mathematicalfoundations

A **complex** model with as many **reliable data**(feature/label) as possible

Where is your data ?

- Rules of Machine Learning:
  - Rule #1: Don’t be afraid to launch a product without machine learning.
    - Machine learning is cool, but it requires data.
    - If machine learning is not absolutely required for your product, don't use it until you have data.

- I have some data, maybe I can try it……..
  - Rule #2: First, design and implement metrics.
  - Rule #4: Keep the first model simple and get the infrastructure right.

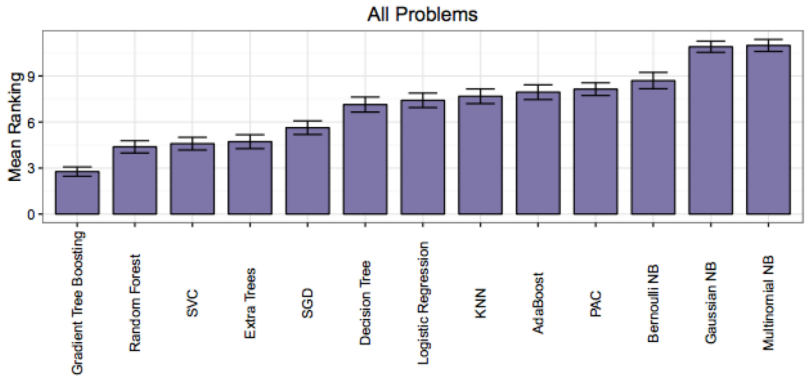
https://developers.google.com/machine-learning/rules-of-ml/

What should the first model be

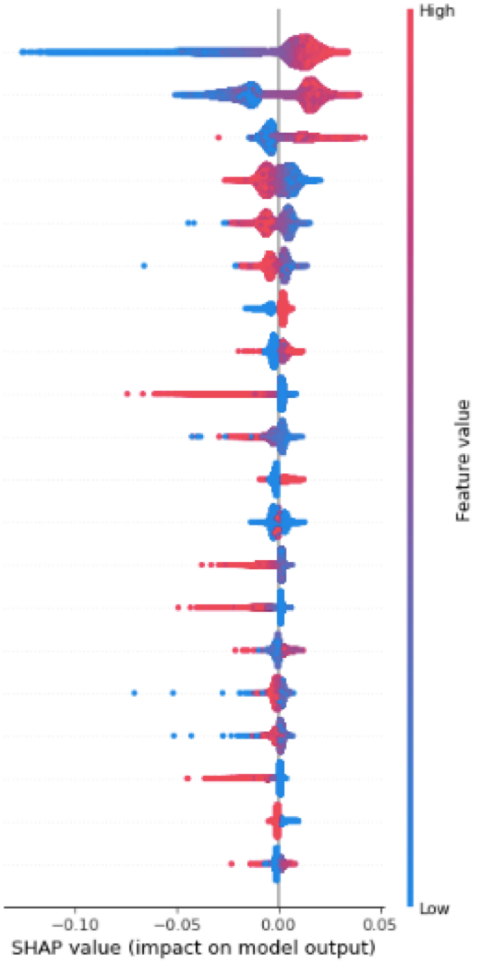
- Features
  - Rule #7: Turn heuristics into features, or handle them externally.
  - Statistic features
- Algorithm
  - Linear/logistic regression ?
  - Requirements
    - No need to do complex feature pre processing
    - No need to do feature selection
      - Can evaluate the feature importance
    - No need to tune parameters hardly
    - Can evaluate the problem complexity
    - Fast training speed
    - Stable accuracy for building the baseline

https://developers.google.com/machine-learning/rules-of-ml/

Algorithms in Bioinformatics Problems



Algorithm	Parameters	Datasets Covered
GradientBoostingClassifier	loss="deviance" learning_rate=0.1 n_estimators=500 max_depth=3 max_features="log2"	51
RandomForestClassifier	n_estimators=500 max_features=0.25 criterion="entropy"	19
SVC	C=0.01 gamma=0.1 kernel="poly" degree=3 coef0=10.0	16
ExtraTreesClassifier	n_estimators=1000 max_features="log2" criterion="entropy"	12
LogisticRegression	C=1.5 penalty="l1" fit_intercept=True	8



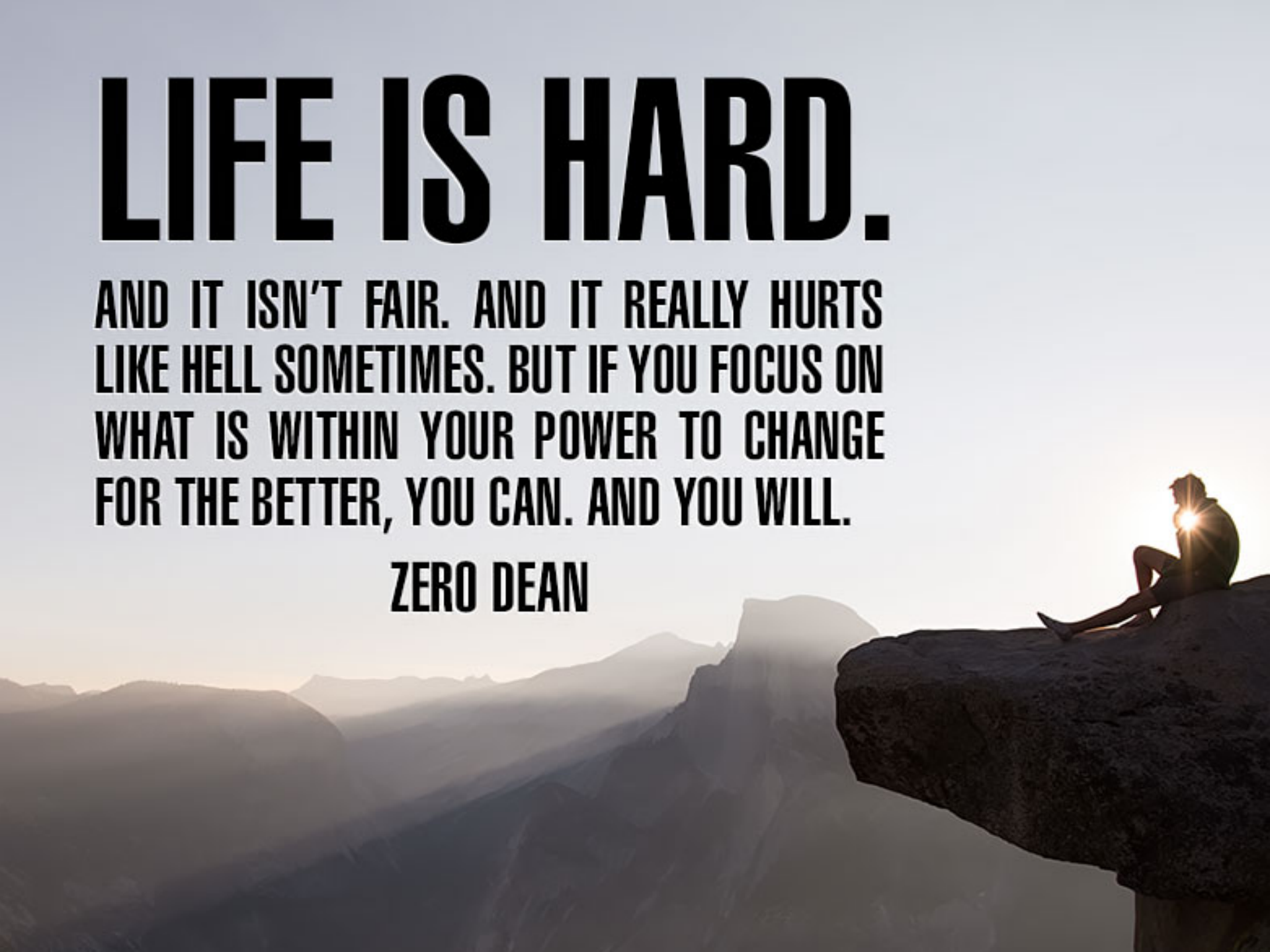
The first model result- NG

- Why?
  - Data(Problem)? Feature? Algorithm?

KKBOX Data Game - 17.06

#	△pub	Team Name	Kernel	Team Members	Score	Entries	Last
1	—	kst			0.28810	36	1y
2	—	Ricky Chou			0.28796	56	1y
3	—	tjw			0.28627	10	1y
4	—	Hey, dude! There's no prizes b...			0.28334	45	1y
5	▲1	simonlin			0.28103	31	1y

- Always check your training error first
  - Upper Bound
- **Change the way you ask/think.**


LIFE IS HARD.
AND IT ISN'T FAIR. AND IT REALLY HURTS LIKE HELL SOMETIMES. BUT IF YOU FOCUS ON WHAT IS WITHIN YOUR POWER TO CHANGE FOR THE BETTER, YOU CAN. AND YOU WILL.
ZERO DEAN

The first model result- Good

1st Round

Malicious	Test	Training
Count	15K	84K

Good	Test	Training
Count	20K	80K

Precision	Recall
0.9873	0.9878

- How can we move from the first model to production?
  - Don't trust your lab number at all
  - Do the simulation or pre-production

More test

Malicious	Test	Training
Count	90K	84K

Good	Test	Training
Count	77K	80K

Precision	Recall
0.8140	0.9530

- Try to improve it

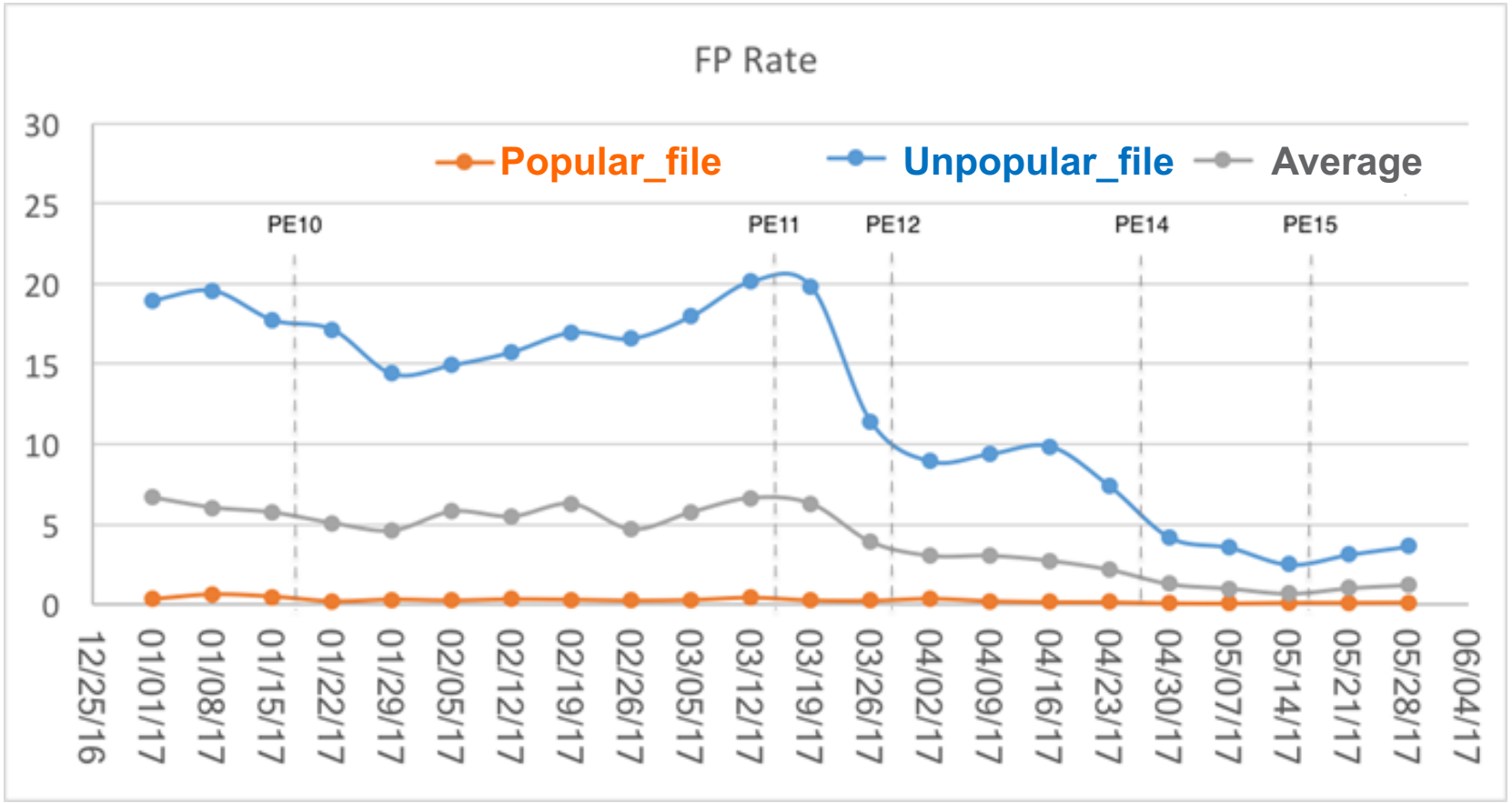
More training

Malicious	Test	Training
Count	90K	374K

Good	Test	Training
Count	77K	362K

Precision	Recall
0.9926	**0.9829**

FP Trending


FP Rate
Popular_file
Unpopular_file
Average
PE10
PE11
PE12
PE14
PE15
30
25
20
15
10
5
0
12/25/16
01/01/17
01/08/17
01/15/17
01/22/17
01/29/17
02/05/17
02/12/17
02/19/17
02/26/17
03/05/17
03/12/17
03/19/17
03/26/17
04/02/17
04/09/17
04/16/17
04/23/17
04/30/17
05/07/17
05/14/17
05/21/17
05/28/17
06/04/17

TREND MICRO

Securing Your Journey to the Cloud

What We Have Done

Build the infrastructure and data pipeline

- The data pipeline can process 1 year log and 20M samples within 3 hours. (With label……..)
- A new model can be deployed to production with full test within 2 hours for serving >10M queries every day.

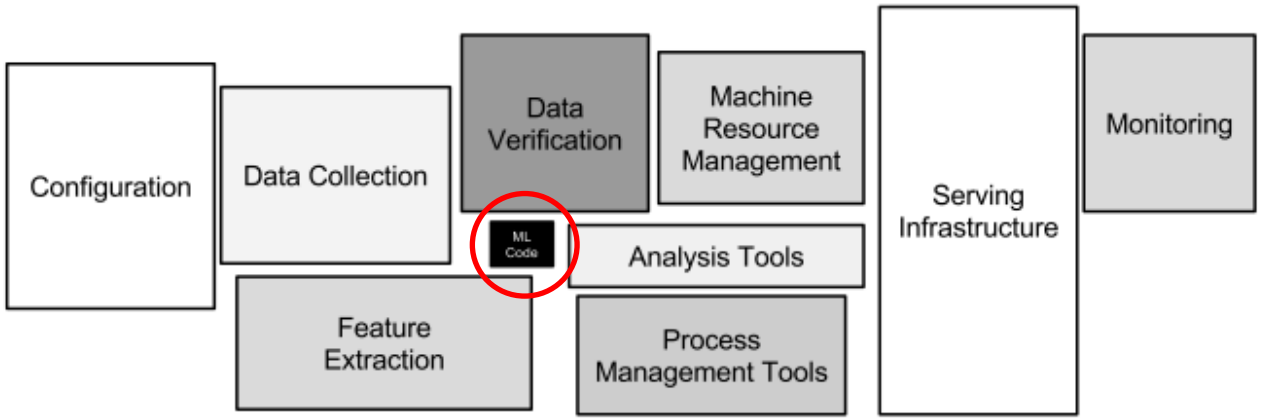
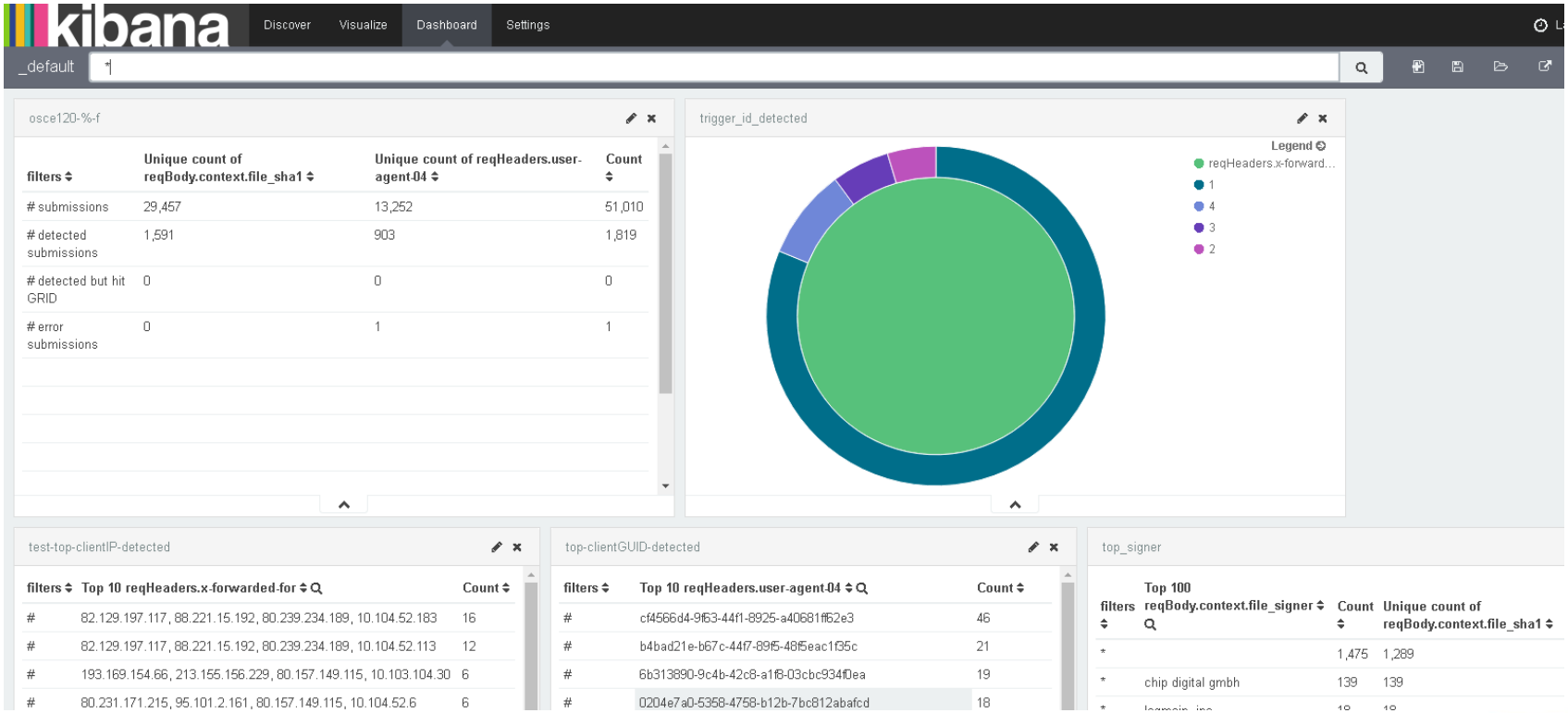


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

Build monitoring system

- Operation excellence
  - Know/Fix the problem before your customer tell you.



Enhance feature and algorithm

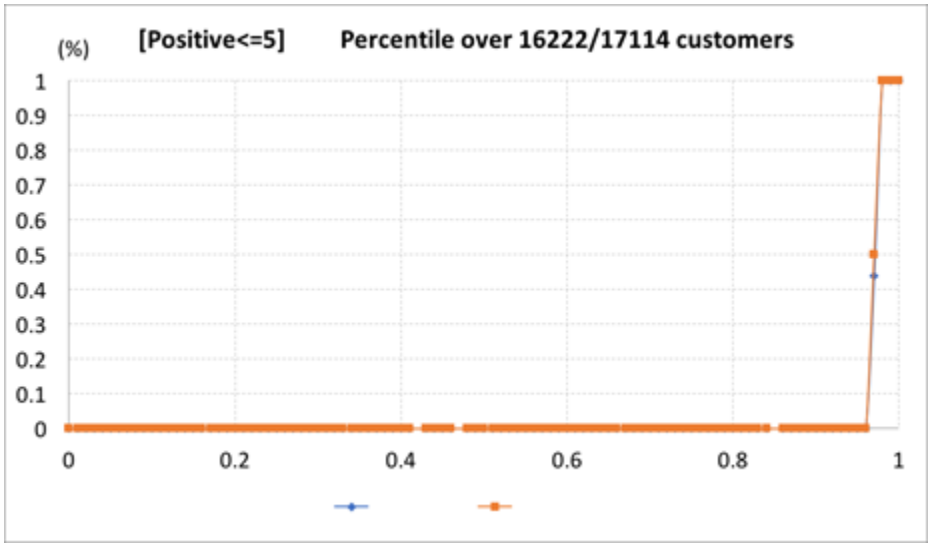
- Rule #7: Turn heuristics into features, or handle them externally.
  - Really? It depends……..
- Look Kaggle, it helps.

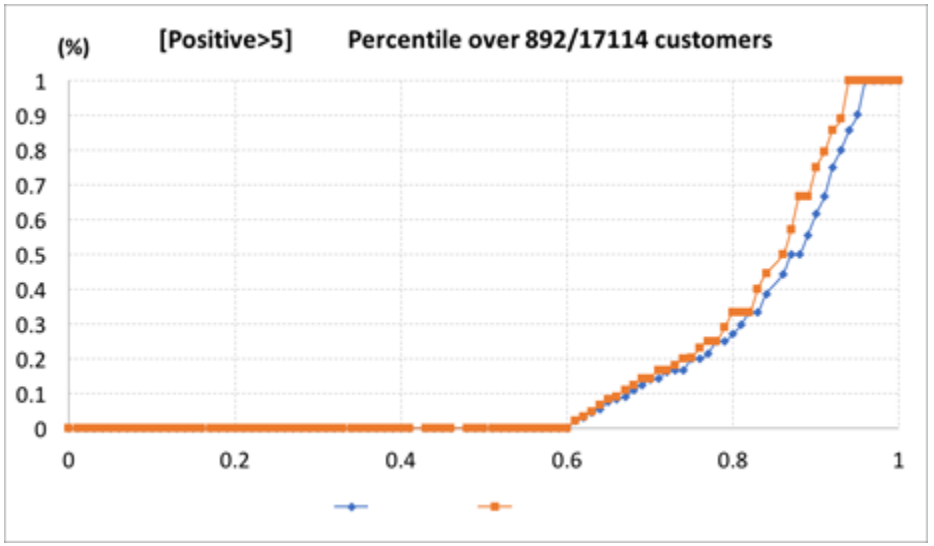
- It may not make sense to evaluate feature and algorithm if your data distribution is not stable.

https://developers.google.com/machine-learning/rules-of-ml/

Know more about the numbers

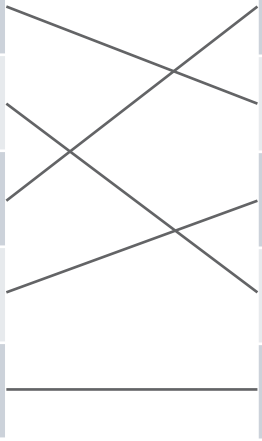
- Potential unhappy customer
  - False discovery rate >= 20%, is 4.8%





Know more about the numbers

Where	MN ratio	Total count
Web Download	241%	11633
E-mail Attach	1.8%	858118
USB Autorun	0.7%	105796
Special Program	0.3%	189778
Others	0.04%	1096194



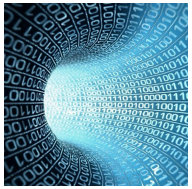
TREND MICRO
Securing Your Journey to the Cloud

From ML to AI

Spark Tsao 曹文光

Hybrid Model: File + Behavior

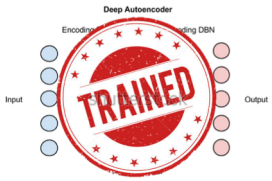
File



File Features

File Model

Deep Learning



Hybrid Features

Hybrid Model

Behavior

Behavior Features

Behavior Model

- [x] **Predictive & Accurate**
- [x] **Insightful**

Highlight

Highlight

- ML is the silver bullet.
  - No, it needs to be used in the correct situation.
- We can do ML solution with only some data.
  - Not recommend, but you can do POC and get to know how to enhance.
- Data is the king.
  - For supervised learning, label quality is more important.
- I need many ML engineers to do a solution.
  - No, you need more infra or data engineers.
  - Data pipeline quality is the upper bound of your ML solution.
- ML skills are the top priority of a ML engineer.
  - Get the intelligences from the data and tell a good story for making actions.

We are not ML engineers

We are the engineers who solve the problems with data

Thank You