AI in Cyber Security

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

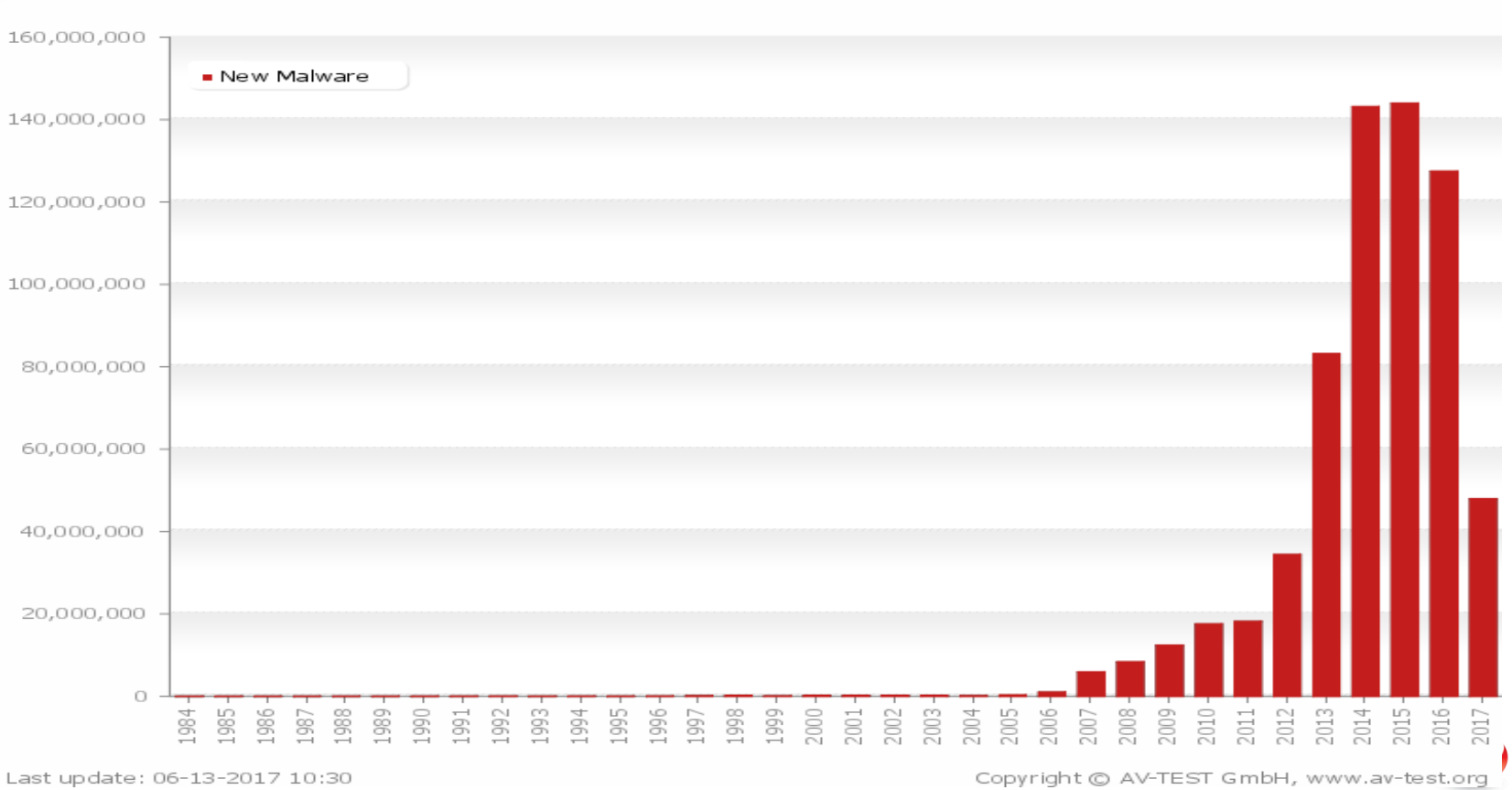
TREND MICRO

Agenda

- AI in Cyber Security
  - Why we need AI
  - AI Application in Trend Micro Services
  - Deep Learning Application
- Adversarial AI in Cyber Security

Why we need AI

Malware Statistic



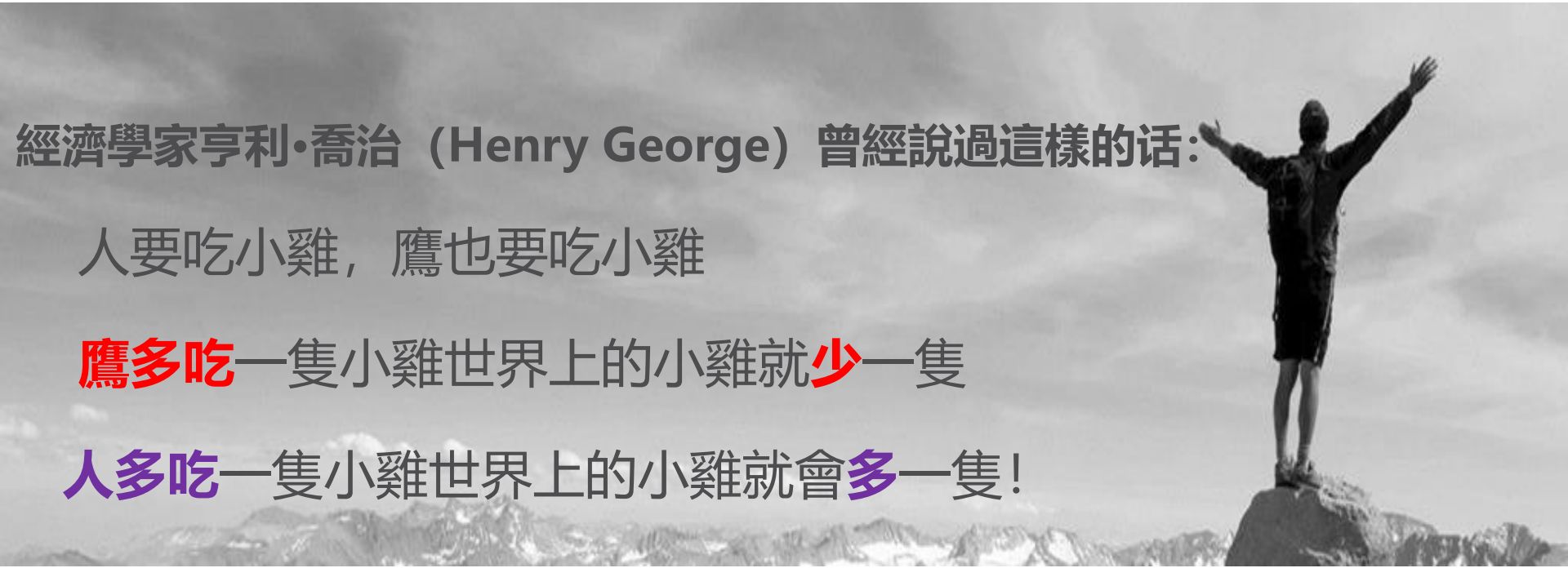
TREND MICRO

Cyber Threat and ML

時代	型態	目的	數量	實例	主要解決方案
初期 ~2006	爆發型	成名	少	Melissa	Signature
中期 2006~2012	潛伏型	資源 資訊	多	Botnet Stuxnet	1-N
現代 2012~	勒索軟體	金錢	極多	WannaCry	ML

TREND MICRO

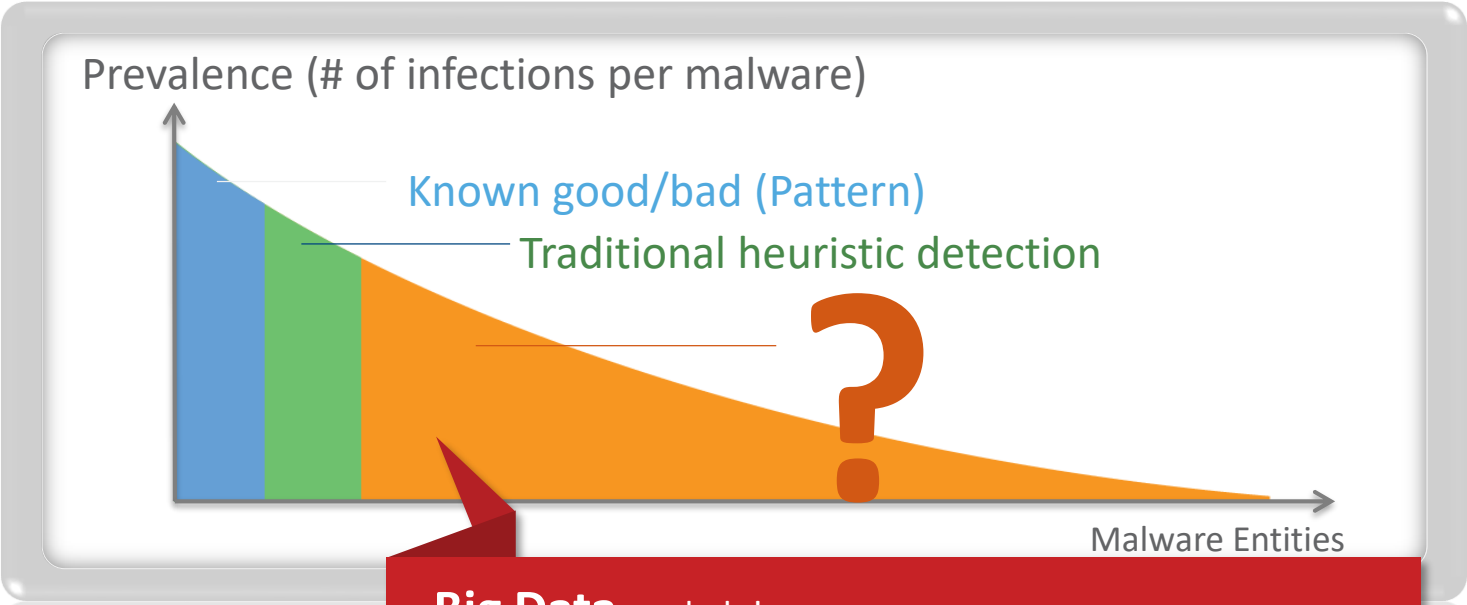
Cyber Threat and ML



Daily Volumes in Trend Micro

- **150M+** sensors worldwide
- **10 TB+** raw logs collected from the world

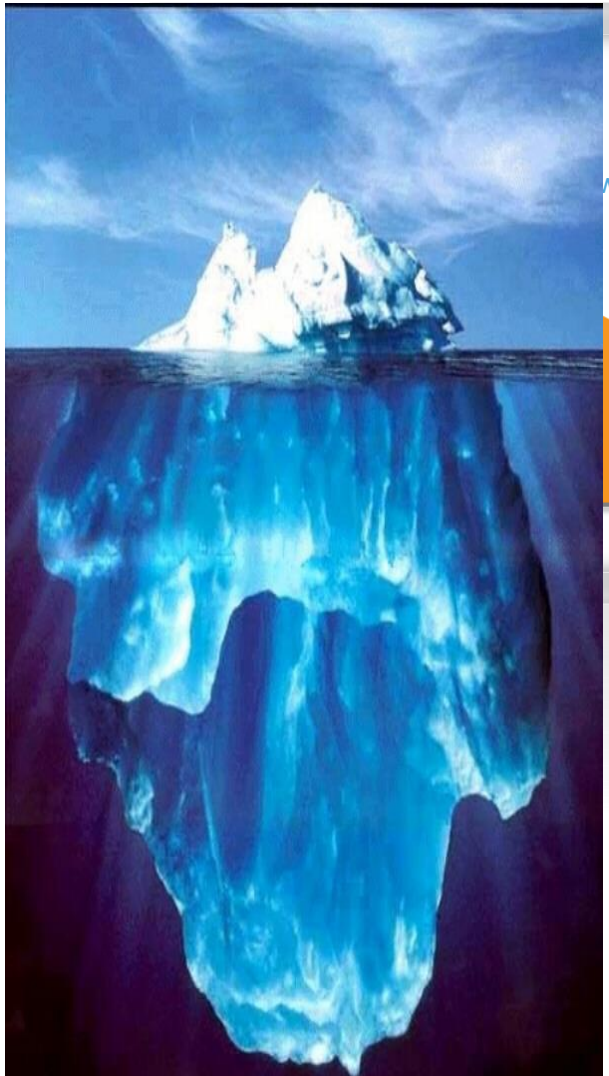
Traditional security solution dilemma – ***Long Tail***



Big Data can help!

15+ billion logs processed per day

- Monitored Files ➔ 717+ million executable files
- Monitored Domains/IP ➔ 707+ million domains



Security in Old Days

wn good/bad

Only Protect What

Traditional heuristic detection

You Can See

~~Intelligence from~~ BigData Analytics (90%)

Next Generation Security

Unleash the Power of Data

Entities

BigData

Data Science

Security Expertise

TREND MICRO

In the begining

- Stage1: drop the e-mail if contains any SPAM WORDS.
  - Hi, you can buy the cheapest iPhone here. ⭕
  - Hi Charles, would you please buy a iPhone and bring back to me? ❌

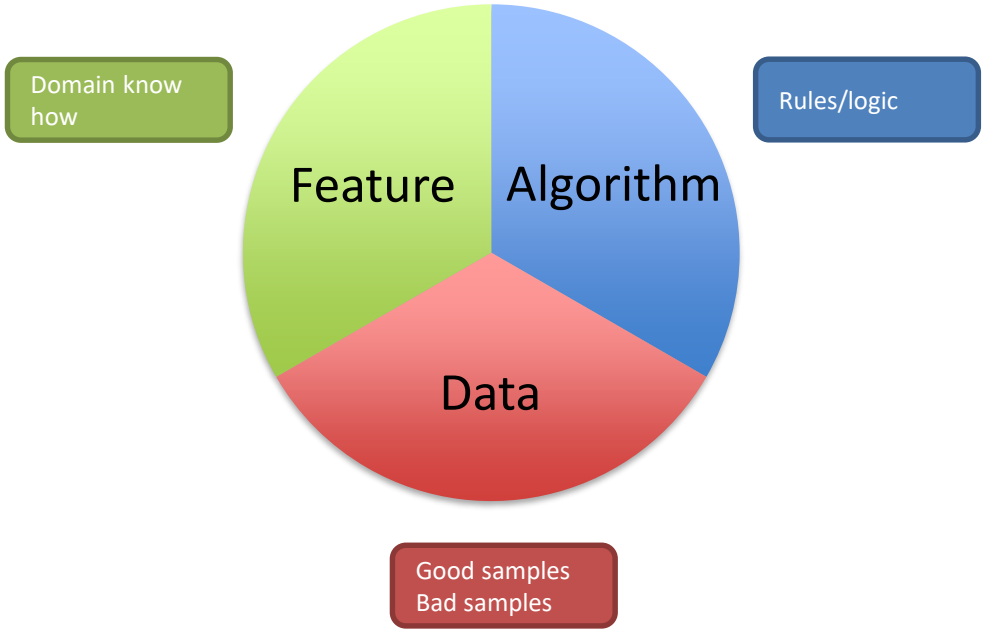
TREND MICRO

In the begining

- Stage2: make more complicated rules
  - if contains any SPAM WORDS.
    - If not contains name.
      - If contains more than 3 SPAM words.
        - » If not contains signature
          - If contains phone-number
            - If not contains URL
              - If….
                - If…..

TREND MICRO

Machine Learning

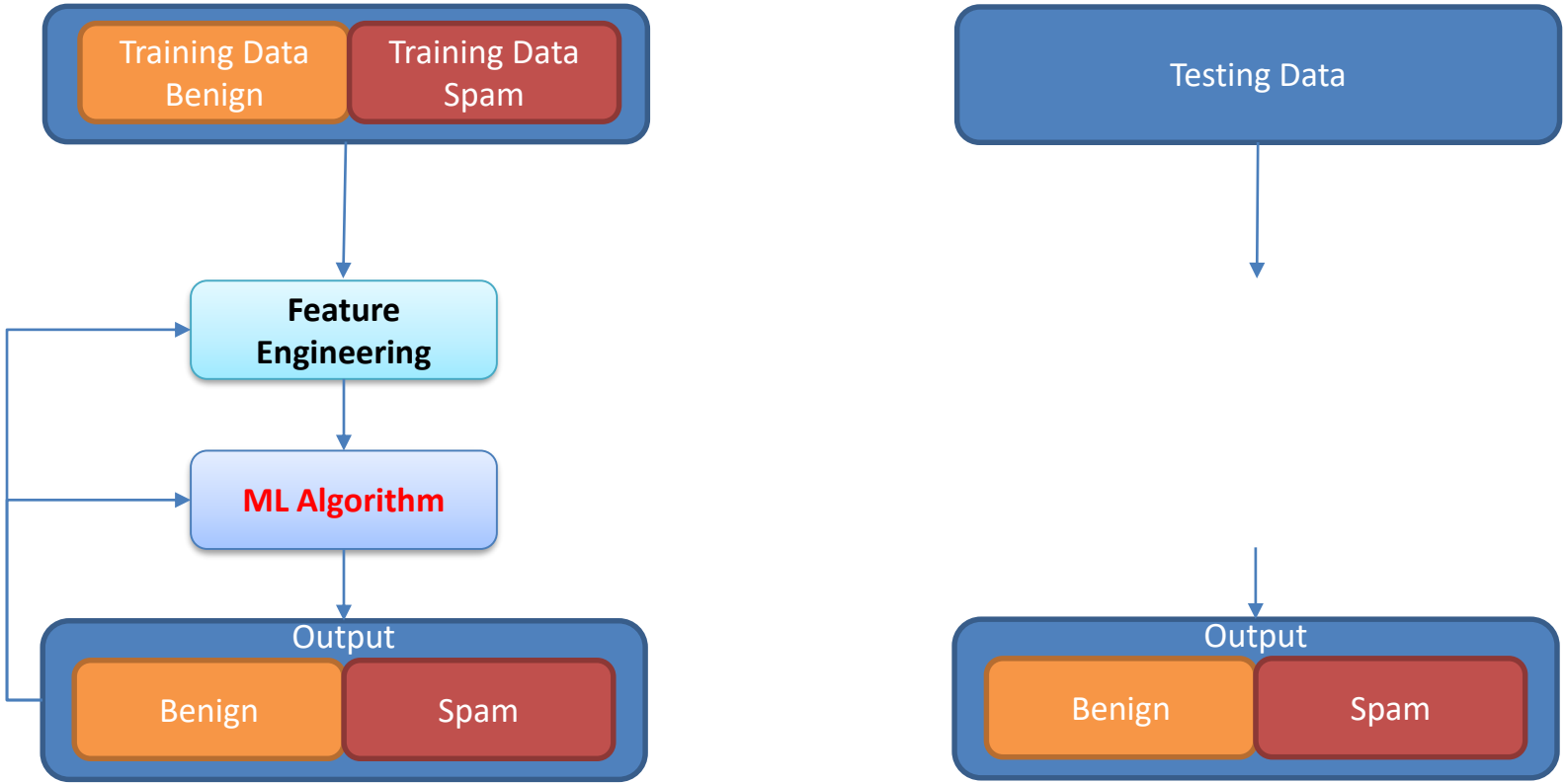


TREND MICRO

Machine Learning

- Goal:
  - A **systematic** way to create **rules/logic** by **data** to get the optimized performance
  - **ML Algorithm**

Machine Learning



Machine Learning

- Example:
  - Training
    - Spam:
      - Hi, you can buy the cheapest iPhone here.
      - ……
    - Benign:
      - Please help to take care of my son, thank you.
      - ……

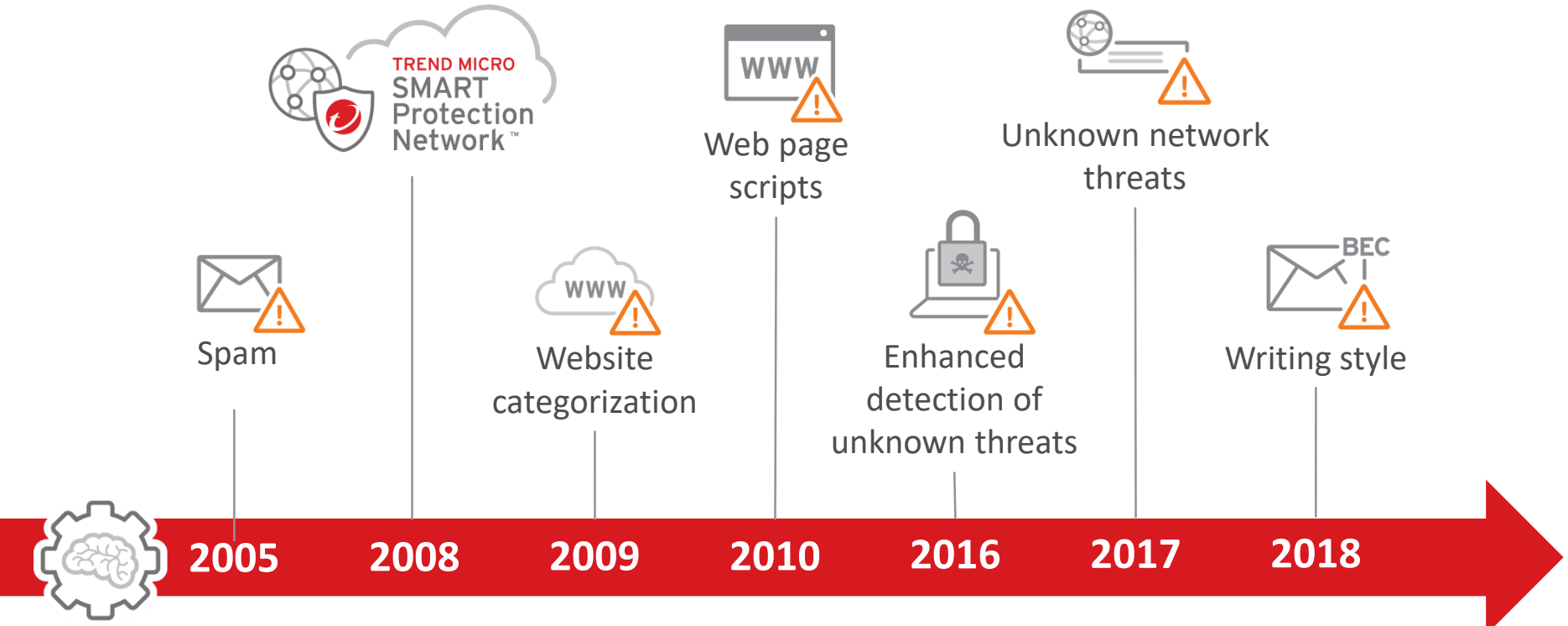
Words	Score
buy	-0.2
cheapest	-0.6
help	0.4
thank you	0.1
…	…

  - Testing
    - Hi Charles, would you please help to buy an iPhone and bring back to me? Thank you.
      - 0.4-0.2+0.1 = **0.3** (Benign)
    - Hi, do you need any help? You can always buy the cheapest stuff here. Thank you.
      - 0.4-0.2-0.6+0.1 = **-0.3** (Spam)

AI Application in Trend Micro Services

Effective AI at Trend Micro since 2005:

40+ applications of AI & machine learning used in Trend Micro products

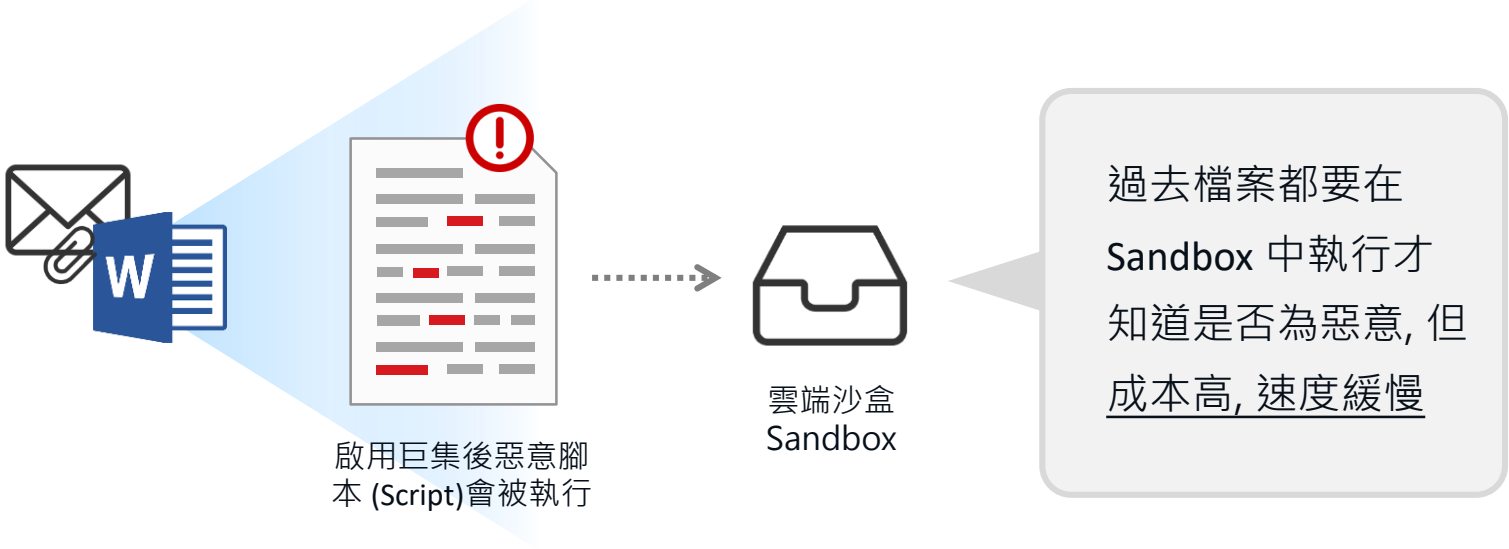


趨勢科技的 AI 經驗

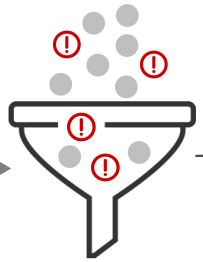
- 惡意巨集偵測 (Cost reduction)
- 垃圾郵件阻擋 (Big data + Model operation)
- 跨世代的威脅防護 X-Gen (Cross generation)
- 變臉詐騙阻擋 (Personal model)

惡意巨集偵測

如何快速的發現夾帶惡意巨集的電子郵件? 和降低雲端沙盒 (sandbox) 處理成本?

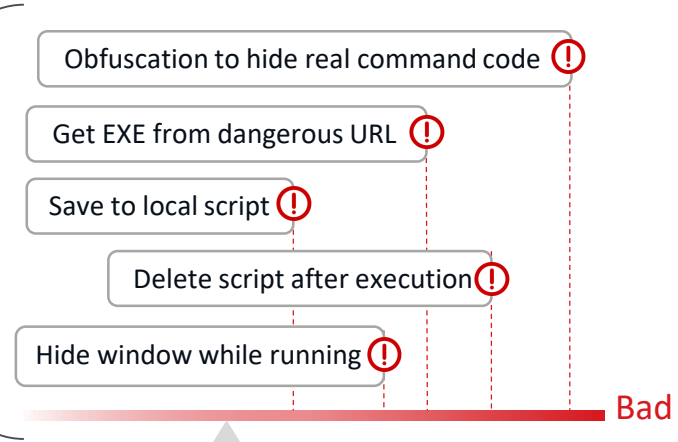


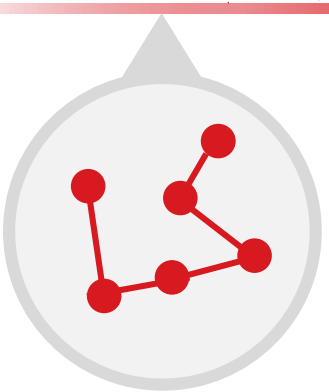
大幅減少進到
Sandbox的檔案數量
使成本降低速度變快



利用機器學習找
出可疑檔案

雲端沙盒
Sandbox





組成 Macro DNA, 透過不同 DNA比對 , 判定是否是惡意巨集

導入人工智慧從小處開始

單純、有明確答案的小工作
解決部份問題
慢慢累積AI元件、ML核心
由AI元件組合成完整解決方案

阻擋垃圾郵件

每天有成千上萬封的垃圾郵件, 如何有效率的過濾郵件? 和降低人力成本?

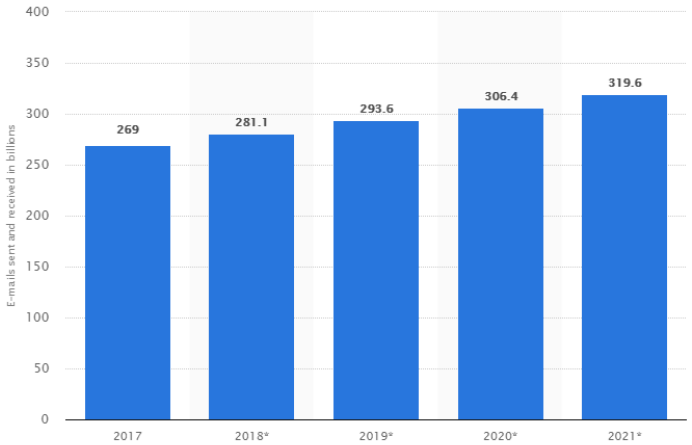
2017全球每天發送和接收的電子郵件總數

2690億

預計到2021年全球每天發送和接收的電子郵件總數

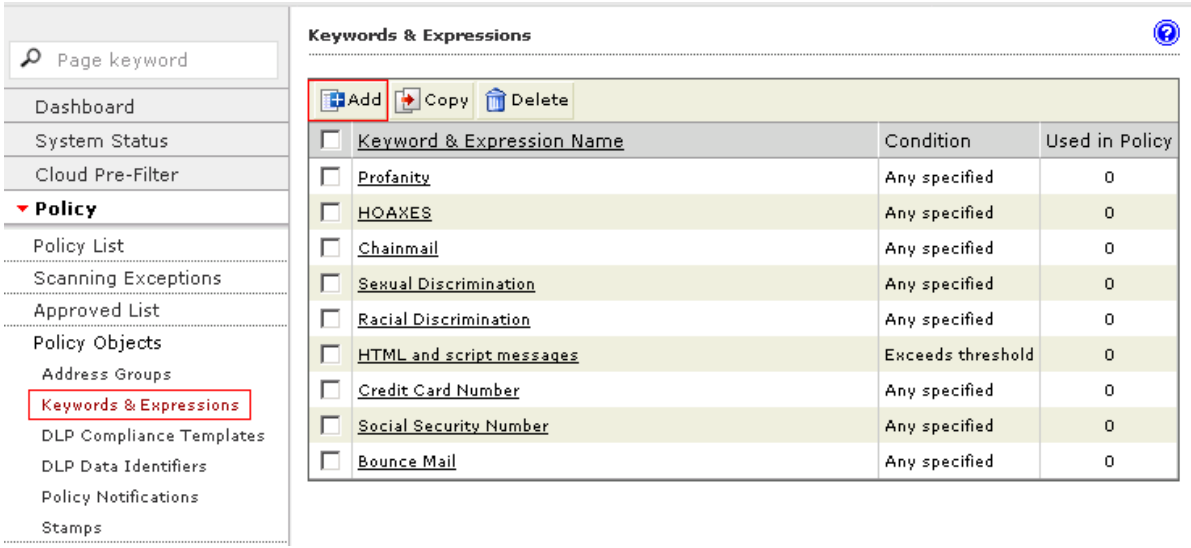
3196億

2017年至2021年全球每天發送和接收的電子郵件數量

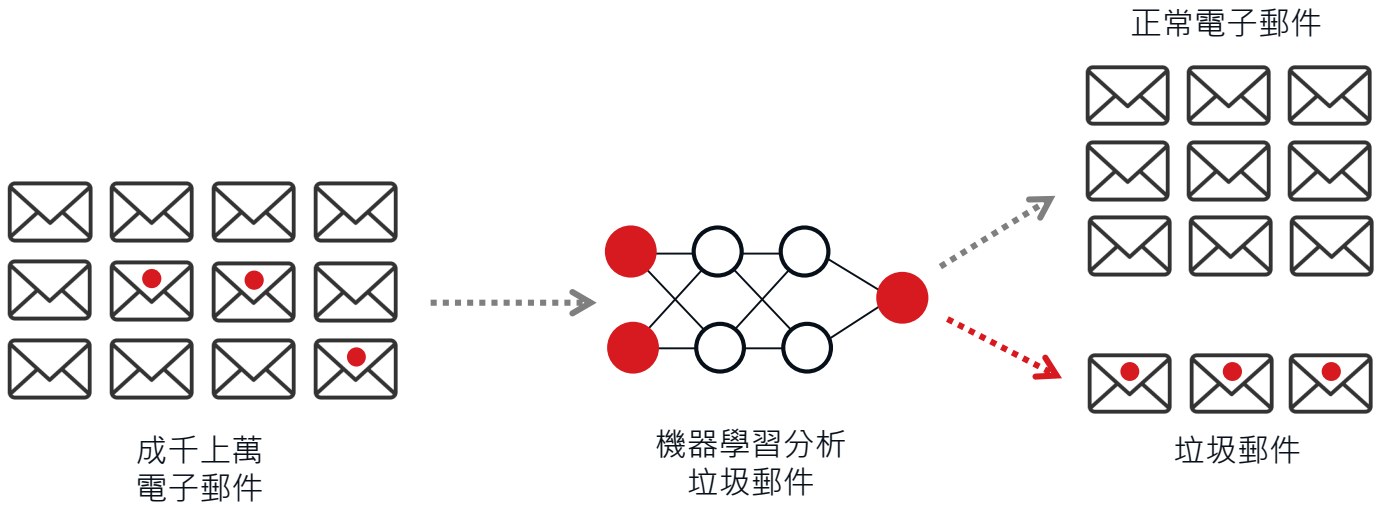


傳統垃圾郵件偵測

設定黑名單
電子郵件地址
垃圾郵件關鍵字
垃圾郵件模板

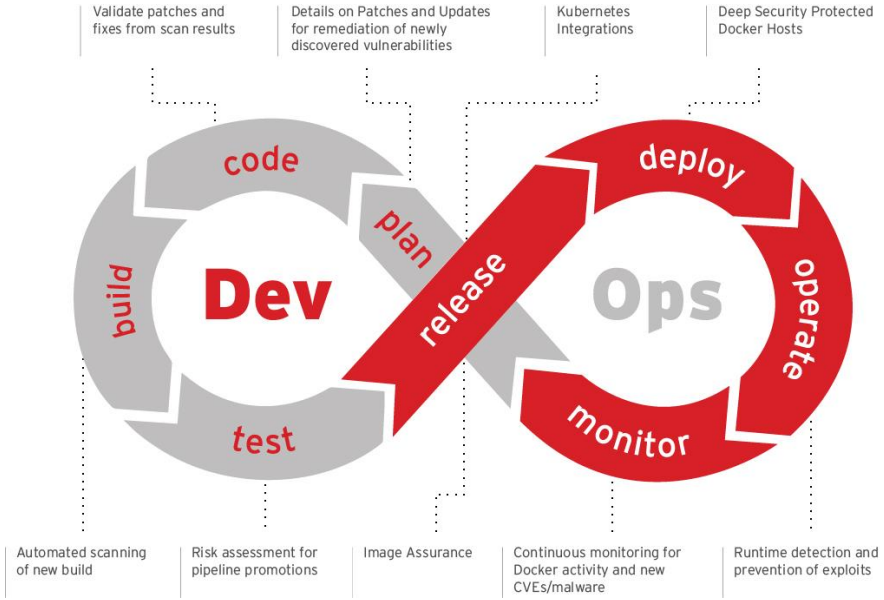


透過 AI 的訓練過濾掉 99% 垃圾郵件
大幅降低人力成本



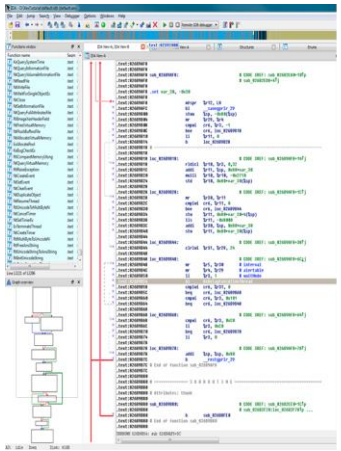
人工智慧/機器學習服務上線
只是開始，不是結束

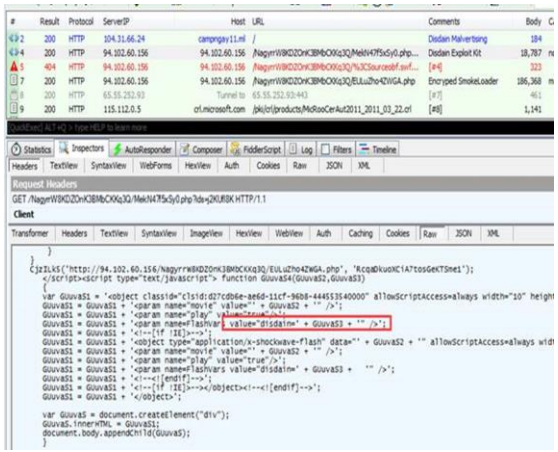
機器學習的日常營運
效能監控
誤判修補
模型更新



傳統攻擊防禦與病毒偵測

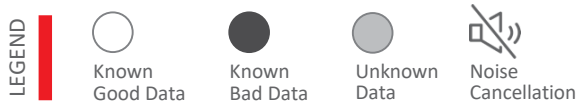
資安專家對病毒樣本和網路封包分析
針對特定攻擊建立特徵 (signature)



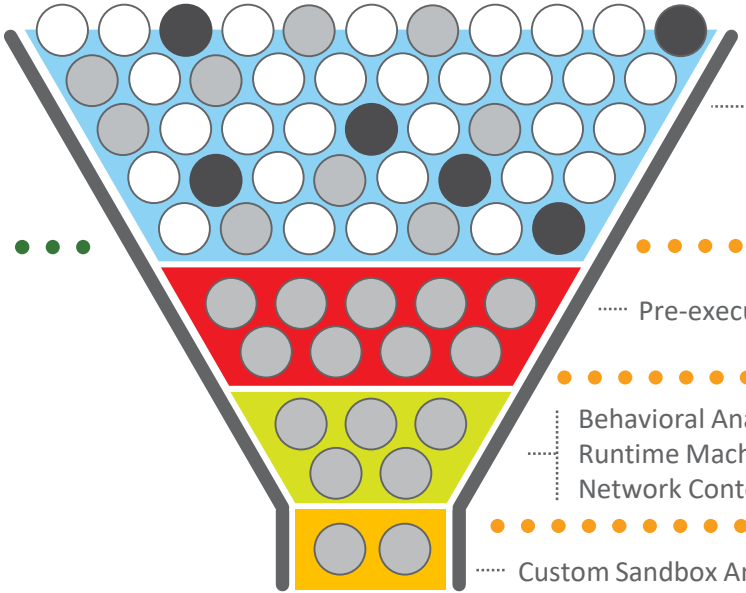


```
rule silent_banker : banker
{
    meta:
        description = "This is just an example"
        thread_level = 3
        in_the_wild = true

    strings:
        $a = {6A 40 68 00 30 00 00 6A 14 8D 91}
        $b = {8D 4D B0 2B C1 83 C0 27 99 6A 4E 59 F7 F9}
        $c = "UVODFRYSIHLNWPEJXQZAKCBGMT"

    condition:
        $a or $b or $c
}
```

跨世代的威脅防護 X-Gen



新舊整合
重層嚇阻



Intrusion Prevention (IPS) & Firewall
Early Zero-Day Protection
Anti-malware & Web Reputation
Application Control
Variant Protection
Integrity Monitoring

Pre-execution Machine Learning

Behavioral Analysis
Runtime Machine Learning
Network Content Correlation

Custom Sandbox Analysis

Investigation & Response

Safe files & actions allowed

Malicious files & actions blocked

導入人工智慧不是砍掉重練
也不是人工智慧的單打獨鬥

將AI與現有系統整合
了解創新初期的不完美
讓AI提早對面真實世界
加速AI開發的流程
發揮1+1>2的效用

變臉詐騙

CEO

2015年2月3日 上午8:09

C

Immediate Wire Transfer

收件人: Chief Financial Office

Please process a wire transfer payment in the amount of $250,000 and code to “admin expenses” by COB today.

Wiring instructions below…

書寫風格異常!!

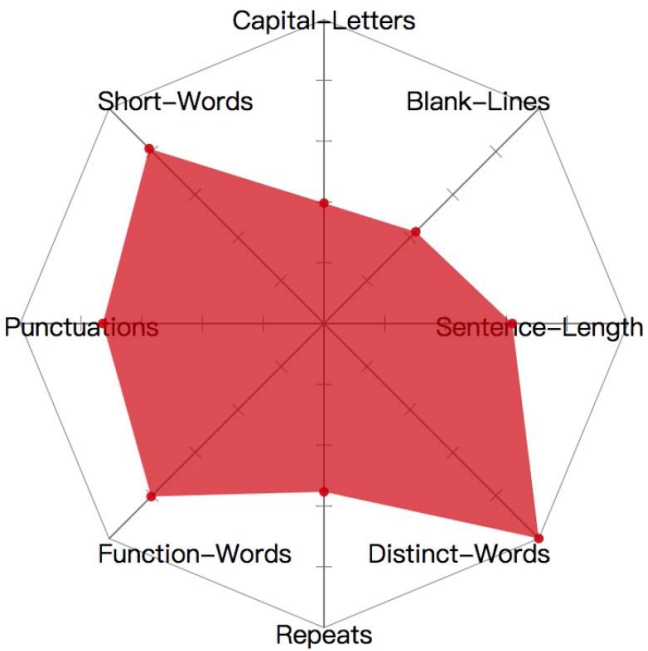
詐騙者偽裝成高階主管要求財務人員匯款

財務人員將款項匯到詐騙者指定的帳戶

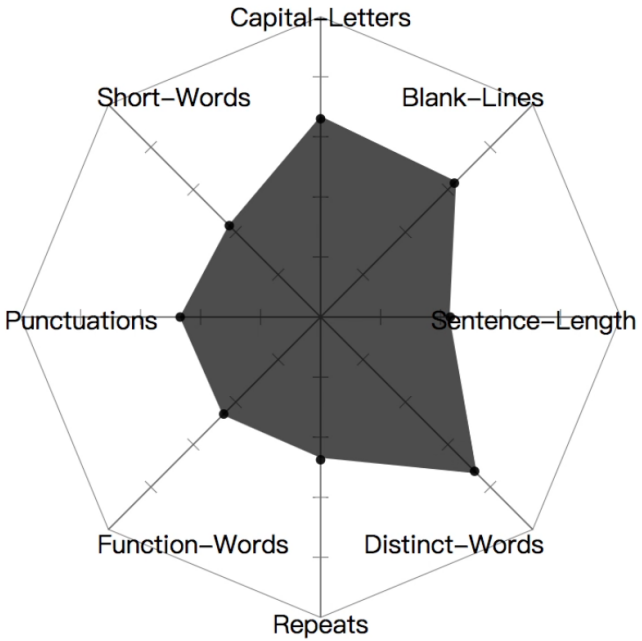
詐騙者成功拿到錢

利用 AI 學習 CEO 書寫風格，判斷信件內容是否為詐騙

CEO



詐騙集團

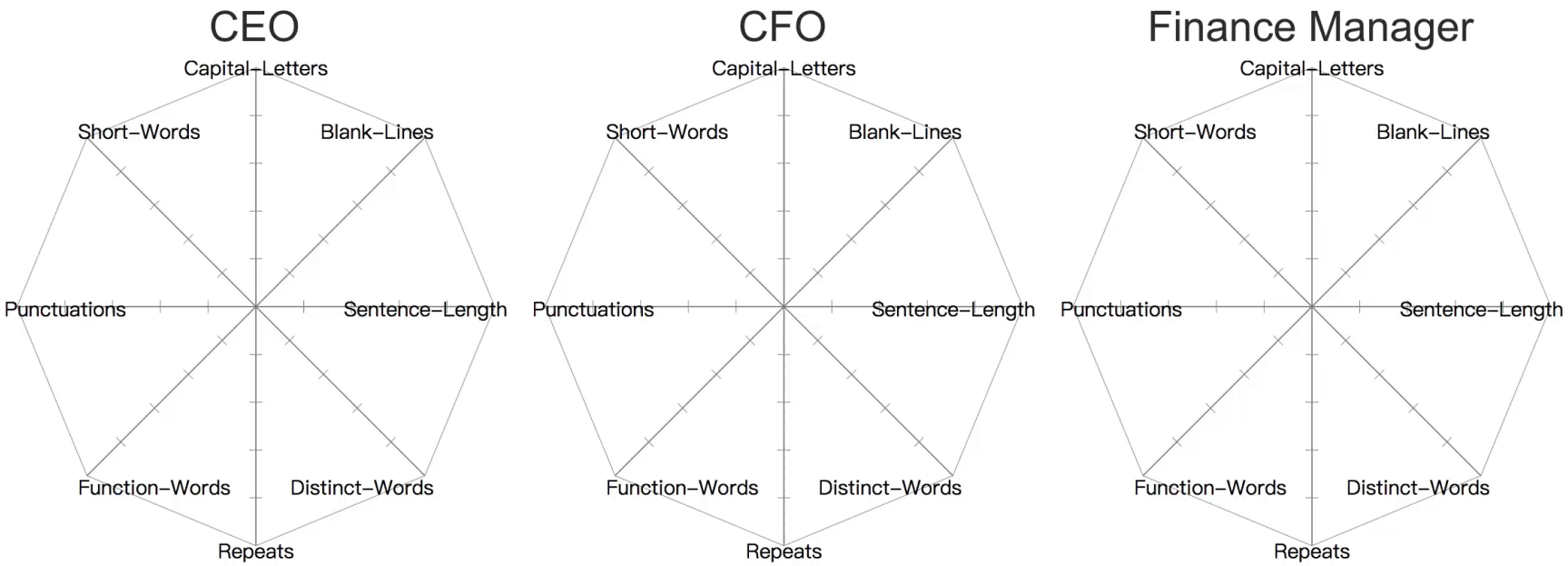


變臉詐騙模型訓練過程

Start training | Stop training | Reset

Mail number : 0

Performance Variance : 0



高度個人化的人工智慧服務

從Global → Personalized
Edge Computing + AI
AI的新挑戰

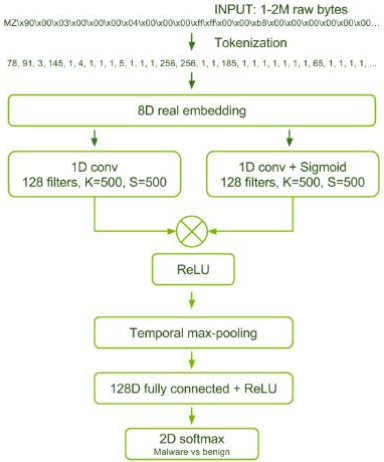
Deep Learning Application

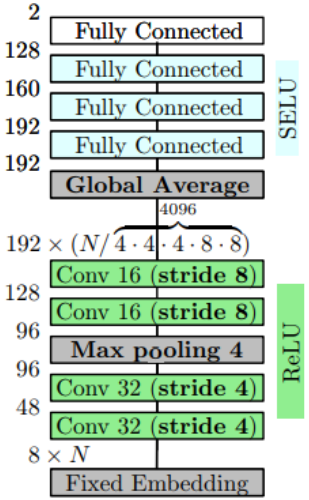
TREND MICRO

Malware Classifier

Static Malware Classification

- NVIDIA
  - Malware Detection by Eating a Whole EXE
- Avast
  - Deep Convolutional Malware Classifiers Can Learn from Raw Executables and Labels Only





Static Malware Classification

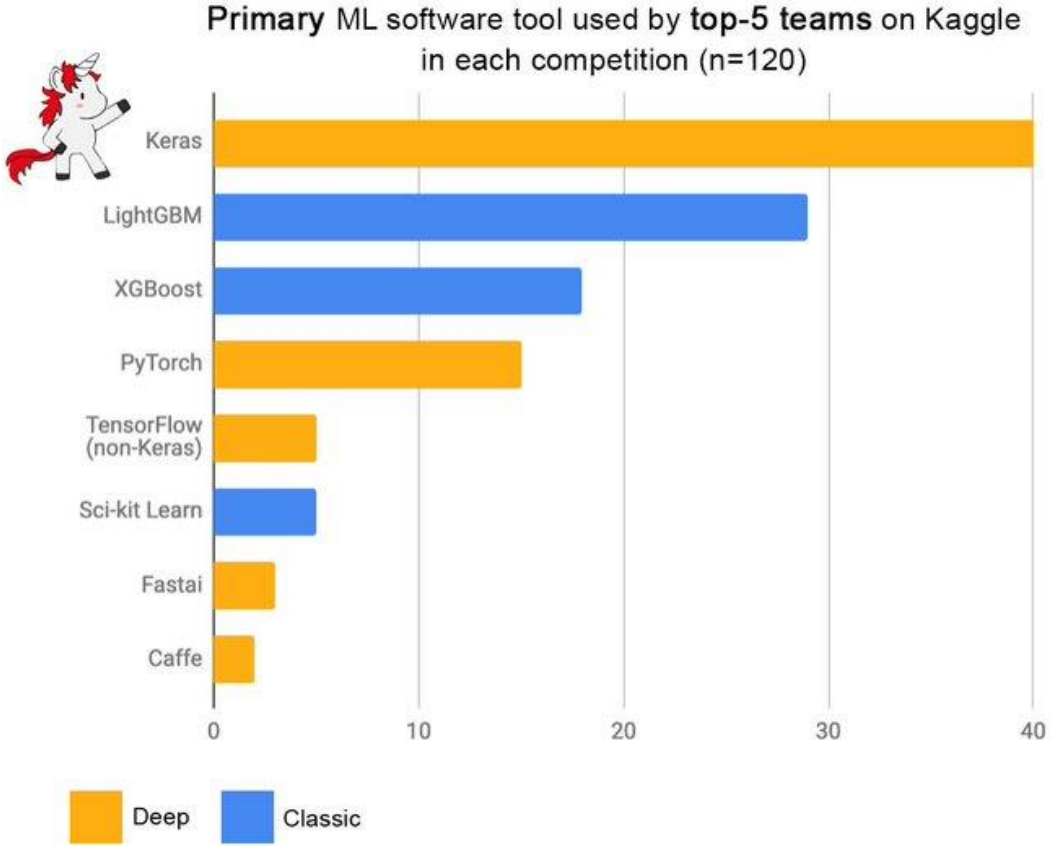
- NVIDIA

	MalConv		MalConv w/o DeCov		Byte n-grams		PE-Header Network	
Test Set	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Group A	*88.1*	**98.5**	83.3	*98.4*	87.0	*98.4*	**90.8**	97.7
Group B	*89.6*	*95.8*	86.6	94.3	**92.5**	**97.9**	83.7	91.4

- Avast

classifier	restricted AUC	cross-entropy	accuracy
MalConv	$66.1\% \pm 0.9$	0.204 ± 0.028	$94.6\% \pm 0.6$
Our convolutional network	$70.4\% \pm 0.5$	0.165 ± 0.020	$96.0\% \pm 0.6$
FNN on handcrafted features	$73.2\% \pm 2.3$	0.151 ± 0.015	$96.2\% \pm 0.3$

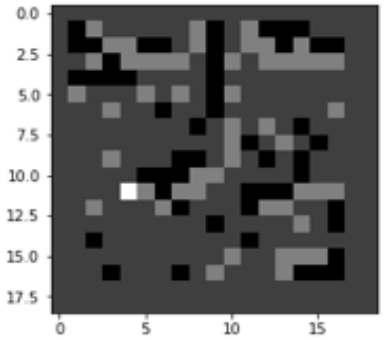
Isn’ t Deep Learning the Best?


Primary ML software tool used by top-5 teams on Kaggle in each competition (n=120)
Keras
LightGBM
XGBoost
PyTorch
TensorFlow (non-Keras)
Sci-kit Learn
Fastai
Caffe
0
10
20
30
40
Deep
Classic

Static Malware Classification

- Avast

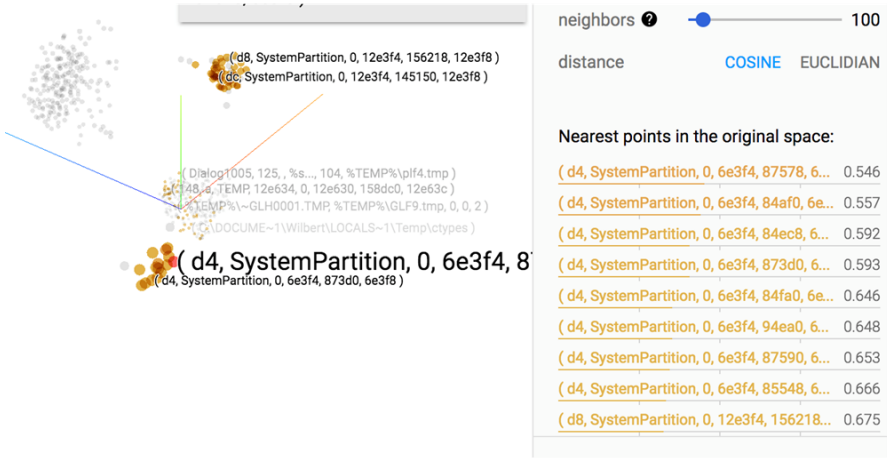
classifier	restricted AUC	cross-entropy	accuracy
MalConv	$66.1\% \pm 0.9$	0.204 ± 0.028	$94.6\% \pm 0.6$
Our convolutional network	$70.4\% \pm 0.5$	0.165 ± 0.020	$96.0\% \pm 0.6$
FNN on handcrafted features	$73.2\% \pm 2.3$	0.151 ± 0.015	$96.2\% \pm 0.3$
FNN on enriched features	$76.1\% \pm 1.0$	0.114 ± 0.006	$97.1\% \pm 0.2$

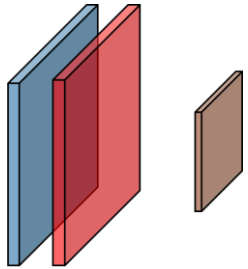


TREND MICRO

Behavior Log

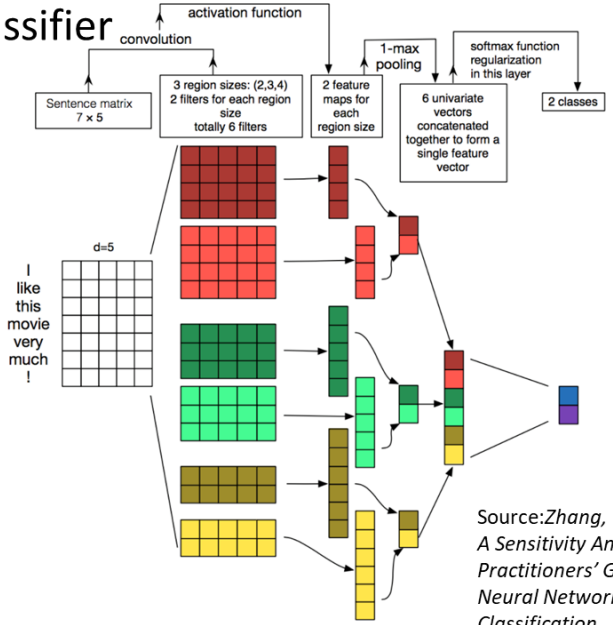
Word embedding





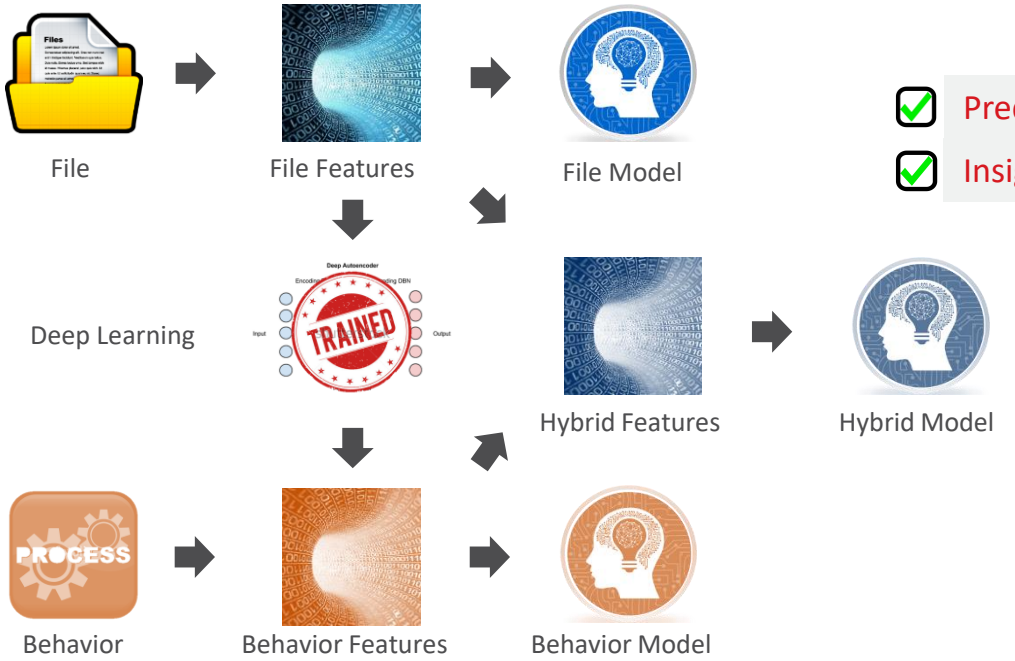
API embedding Args embedding

Classifier



Source:*Zhang, Y., & Wallace, B. (2015). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification.*

Hybrid Model: File + Behavior



- Predictive & Accurate
- Insightful

TREND MICRO

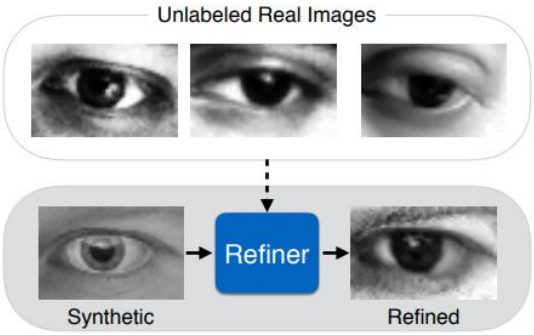
GAN

GAN for Generating Training Data

- Does it make sense?

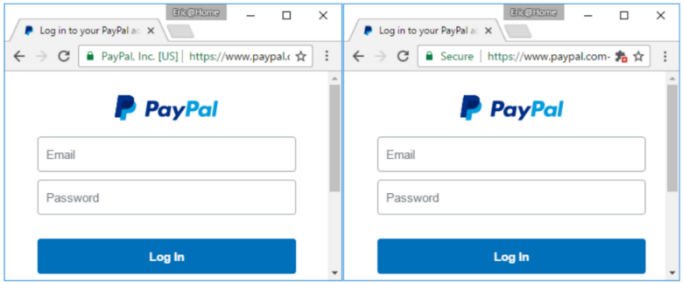
GAN for Generating Training Data

- Learning from Simulated and Unsupervised Images through Adversarial Training

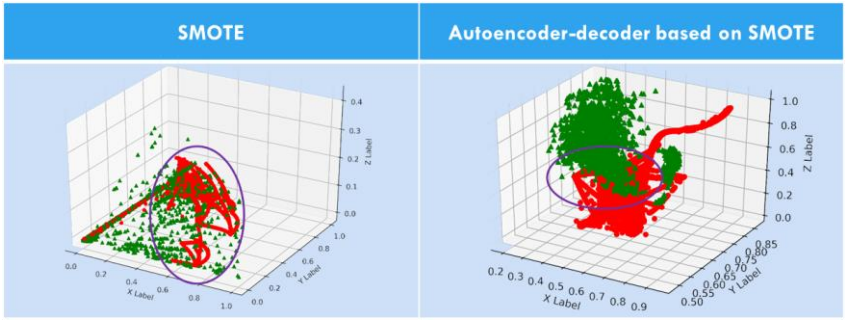


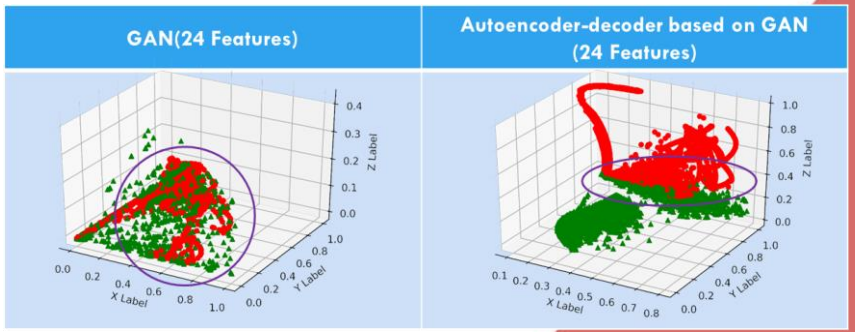
GAN for Generating Training Data

- Intelligent Hybrid Learning Architecture for Cyber-Phishing Attack



- Red point : legal data
- Green point : phishing data



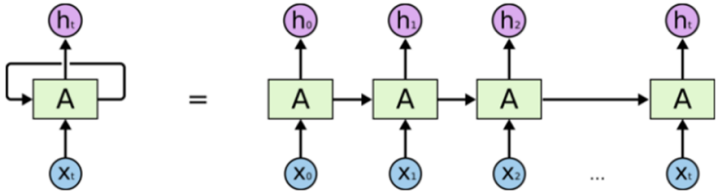


TREND MICRO

Others

URL classification

- Using LSTM to detect malicious URL

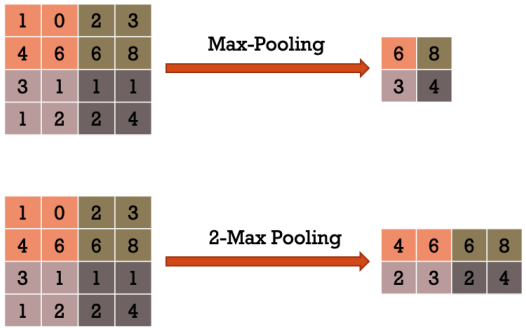


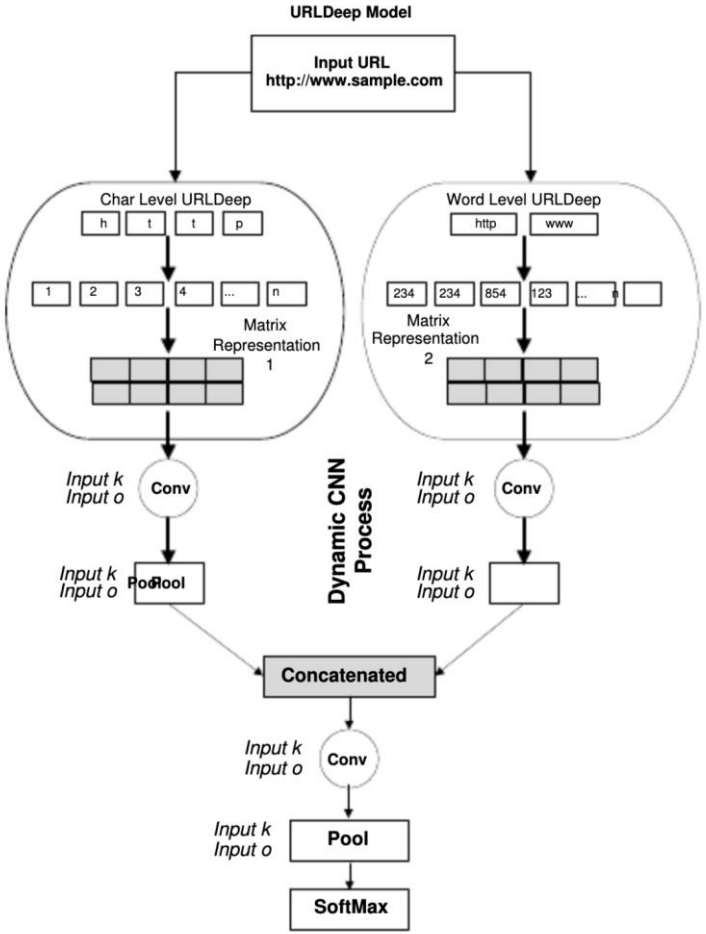

```
htxp://www.tma.tw:80/ltk/106600411.pdf
00000000001111100000001111111111111
```

```
htxps://cdn.fbsbx.com:443/v/t59.2708-21/…/ch14-Fluid.exe
00000000001000000000000000001111111111…1111111111111111
```

URL classification

- URLDEEP
- Quite similar to URLNet, however they use
  - Dynamic amount of convolution layers
  - k-max-pooling schema
- Use dynamic graph to calculate optimal loss value to select best depth of convolution layers.

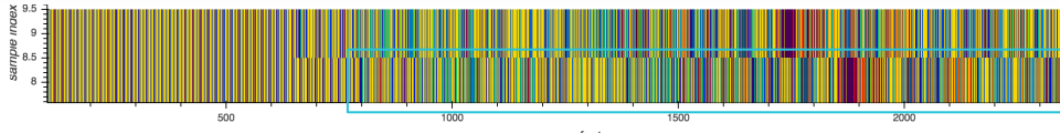


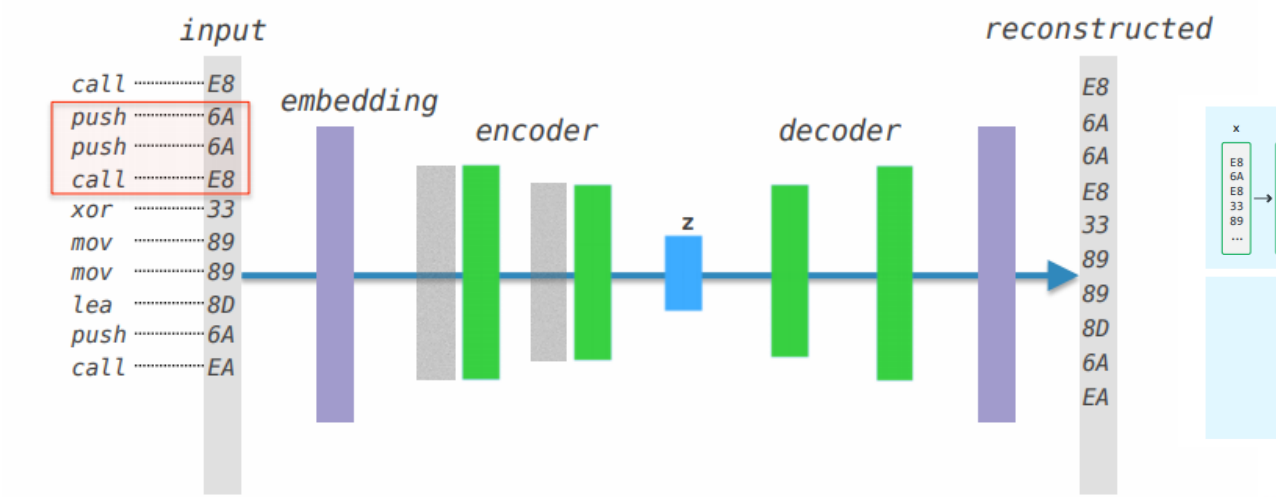


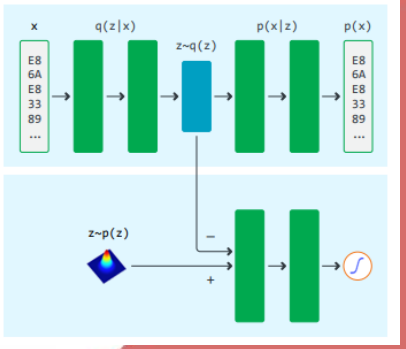
https://www.researchgate.net/publication/331639572_URLDeep_Continuous_Prediction_of_Malicious_URL_with_Dynamic_Deep_Learning_in_Social_Networks

Malware AutoEncoder

	name	id	file	filesize	offset	va	nfuncs	totalfuncsize
312	MAC.OSX.Trojan.FlashBack.AG	db65c02586f7a6555ec86750ca6835a696394df8a73554...	samples/2017-03_base/malware/db65c02586f7a6555...	220784	0	7152	366	41757
284	MAC.OSX.Trojan.FlashBack.F	b99b375c0cbe92c50760240a0eee43d175e2f774474706...	samples/2017-03_base/malware/b99b375c0cbe92c50...	132860	0	6448	280	16402







TREND MICRO

Conclusions

- Deep learning is the best for the data that hard to be represented by features.
  - Image/Voice/NLP
  - There are a lot of progress in the near future.
    - AutoEncoder
    - GAN
    - Attention
    - Explanation

- There is no silver bullet algorithm
  - Try the different algorithms for the problems.

    or
  - Try to convert problems to fit the algorithms.

TREND MICRO

Adversarial AI in Cyber Security

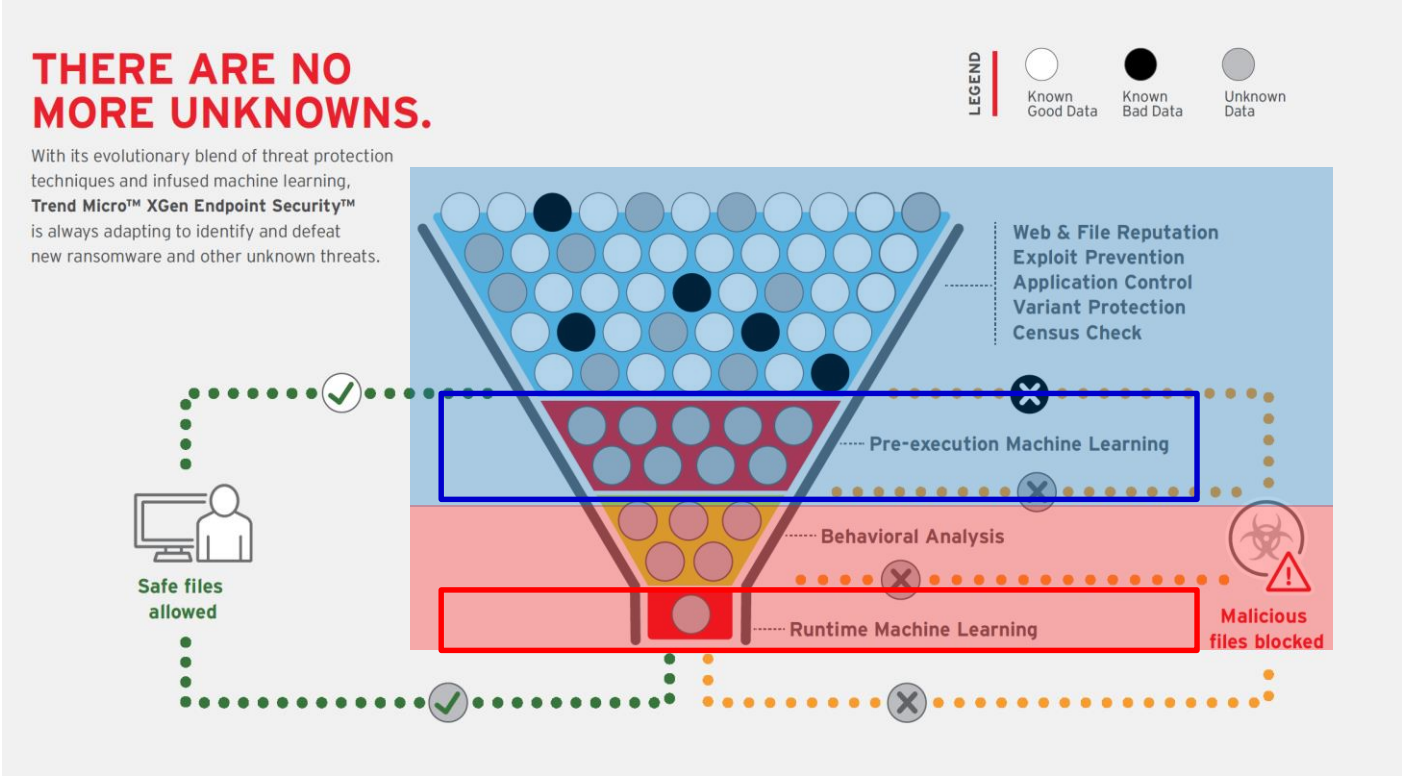
Agenda

- What is Machine Learning ?
- What is Adversarial Machine Learning ?
- Adversarial ML Methodologies
- Possible countermeasures
- Conclusions

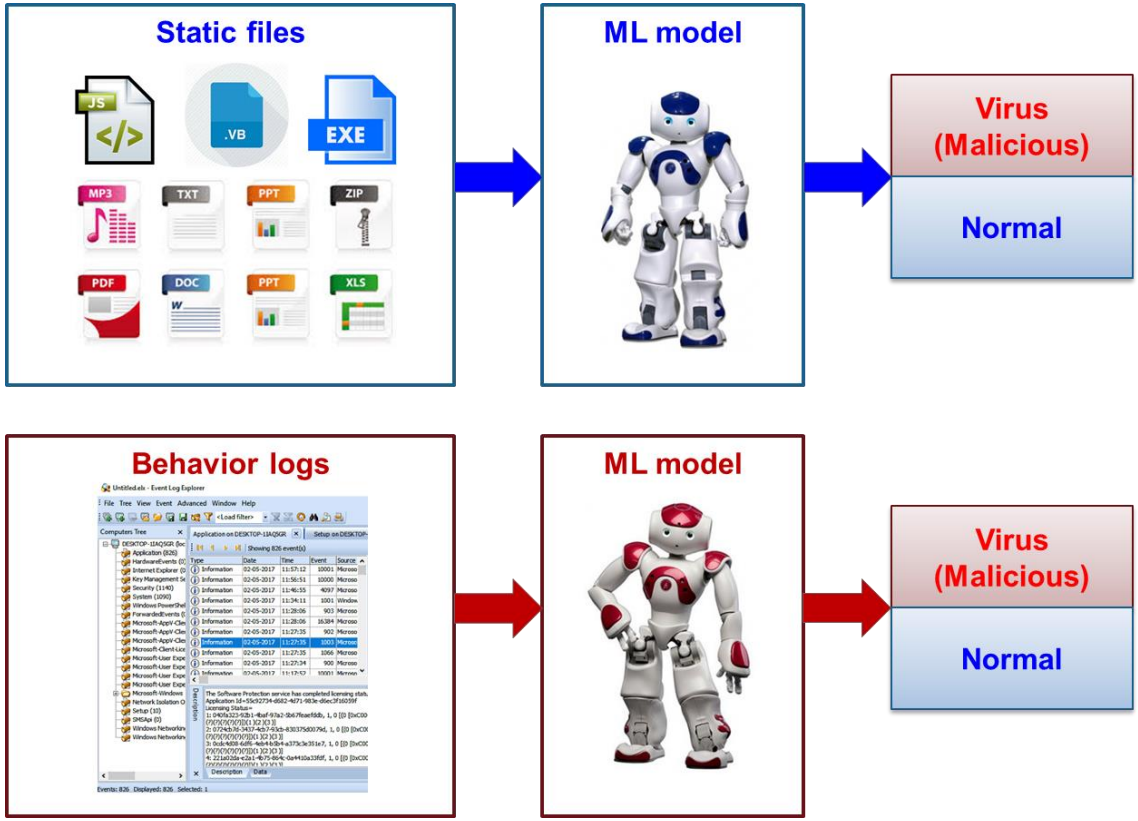
TREND MICRO

Machine Learning
&
Adversarial Machine Learning

XGen ML – Layer protection



What is Machine Learning



TREND MICRO

What is Adversarial Machine Learning

Adversarial machine learning is a technique employed in the field of machine learning which attempts to **fool models** through malicious input.

- Wikipwdia

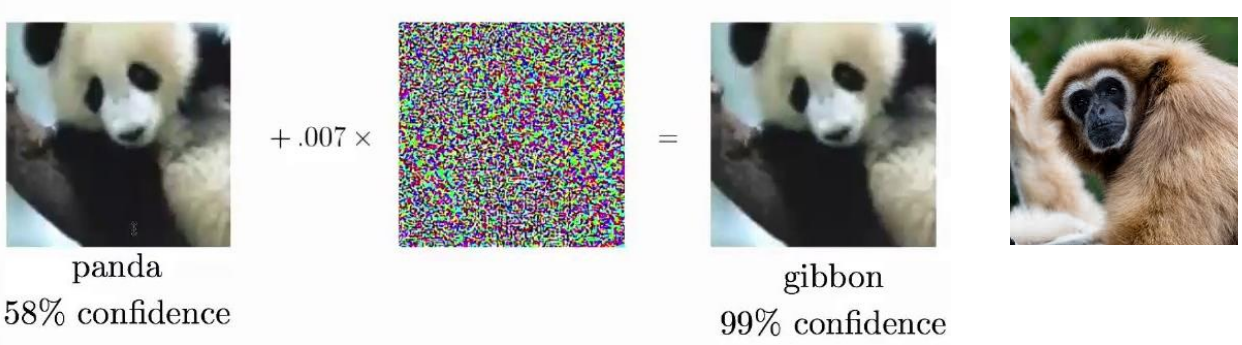
What is Adversarial Machine Learning

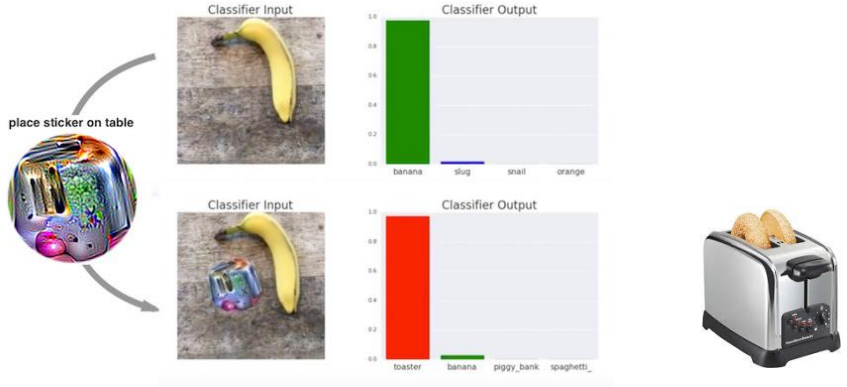
Image Recognition

TREND MICRO

What is Adversarial Machine Learning

Image Recognition

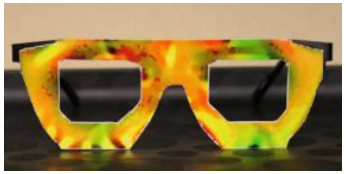




https://arxiv.org/pdf/1412.6572.pdf
https://arxiv.org/pdf/1712.09665.pdf

What is Adversarial Machine Learning

Face Recognition

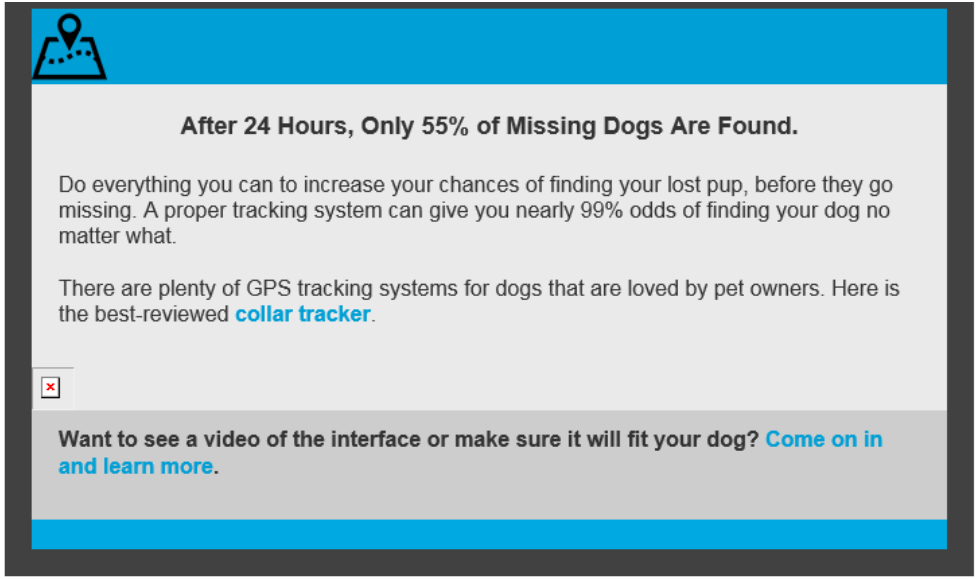


https://www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf

What is Adversarial Machine Learning

Spam Detection

Spam content



salad word

people. no we don't. those are the radical twitter es. i love you same to you bud i love you more <3 i love you both i love you morer you don't love all of us white people that's a bit much since there are great and py people from every creed lol, as sweet as that sounds, but i definitely don't think "most black people hate white people." i appreciate you calling that out. i imagine most of society isn't bigoted and judges on a person by person basis. i could be living under a non-racist rock though. i will never understand how we stereotype each other as racist or thieves or even rapist when there are billions of people across the planet. we've only ever seen a handful of that amount in person. i know. i think it's like a defense mechanism, people don't understand something and just

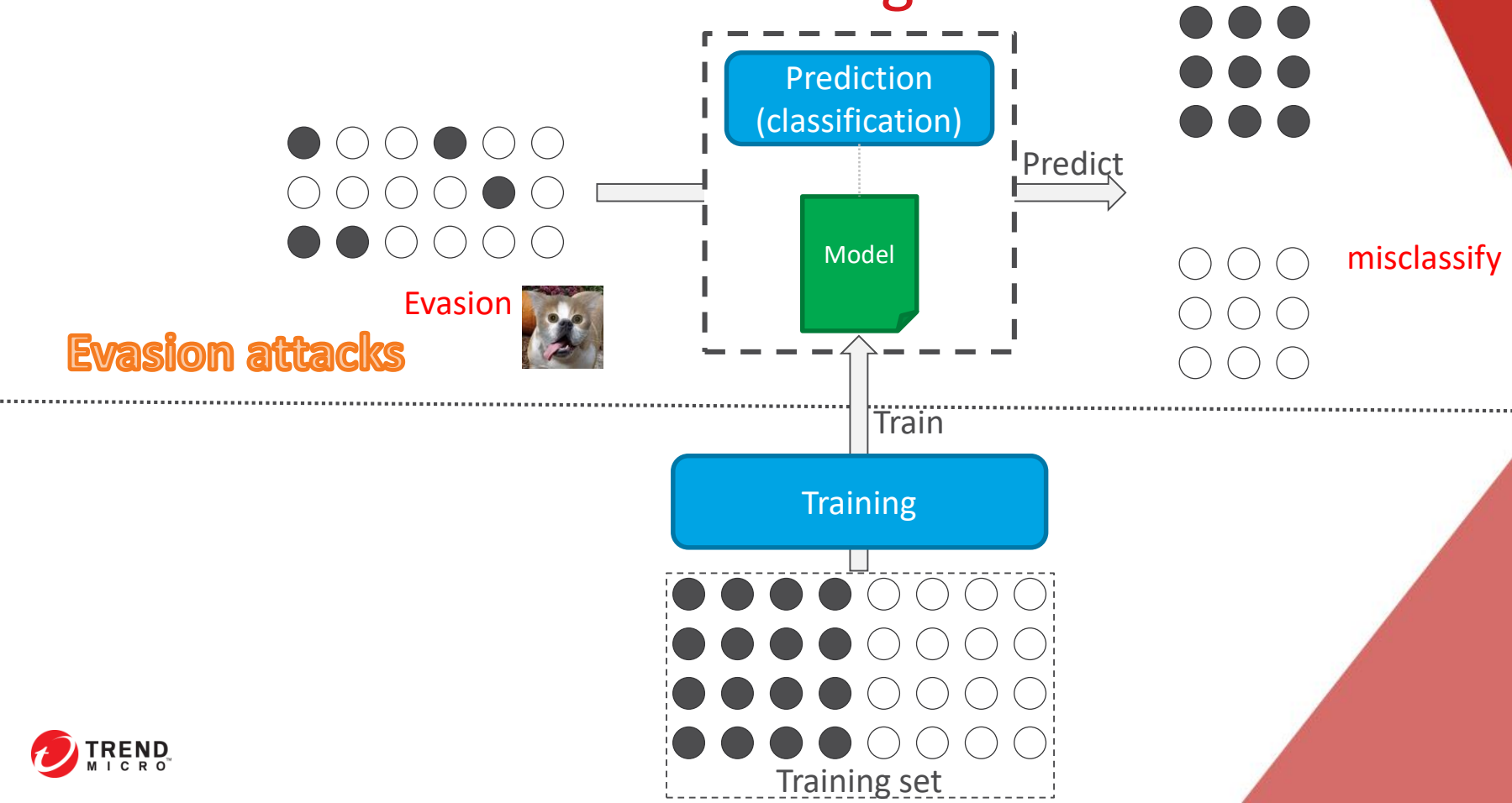
TREND MICRO

Adversarial ML Methodologies

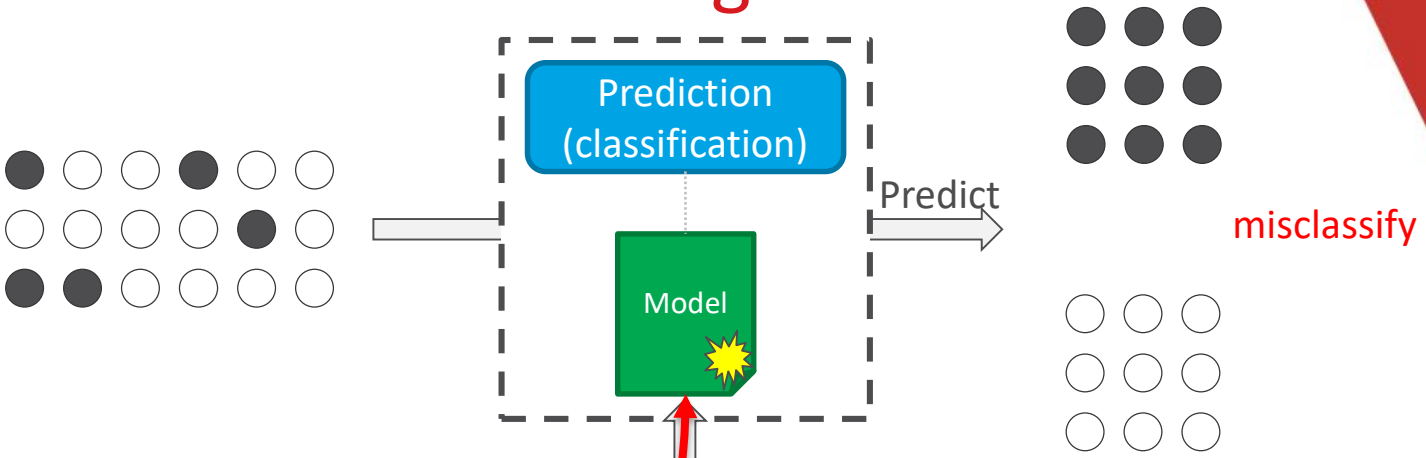
Adversarial ML Methodologies

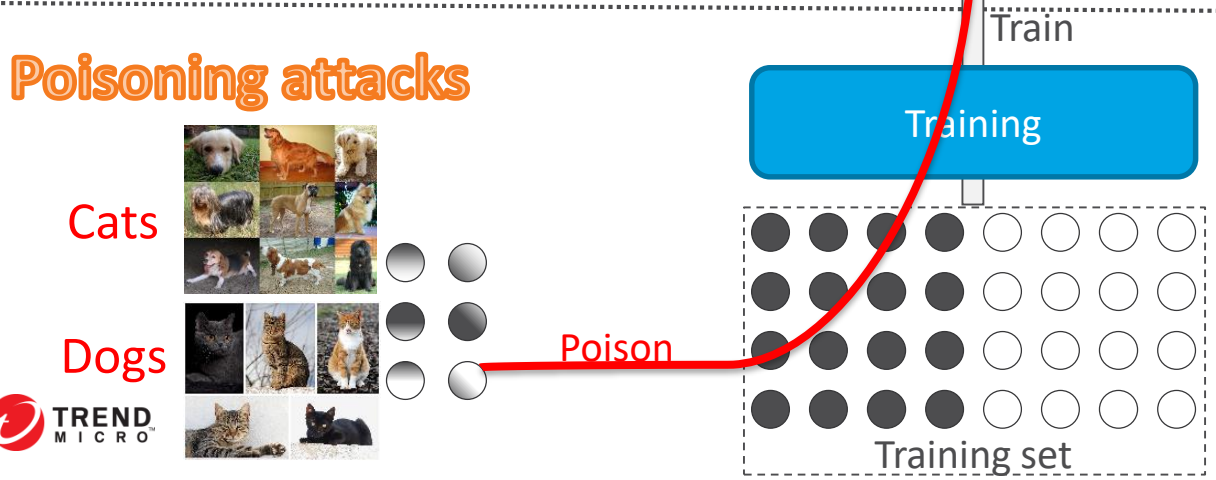
- Evasion Attack
  - Black box
  - White box
    - model stealing
- Poisoning Attack

Adversarial ML Methodologies



Adversarial ML Methodologies

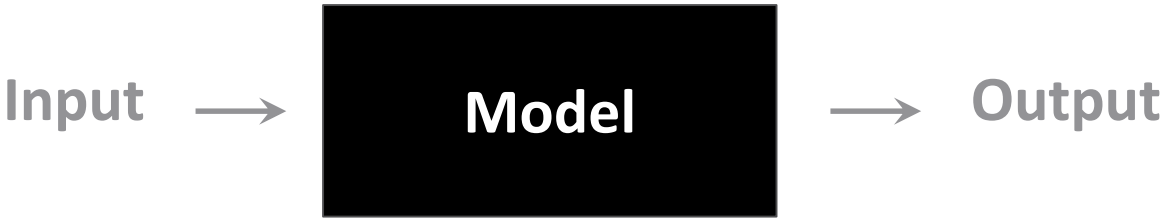




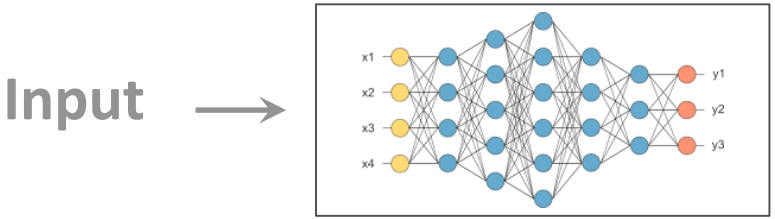
TREND MICRO

Evasion

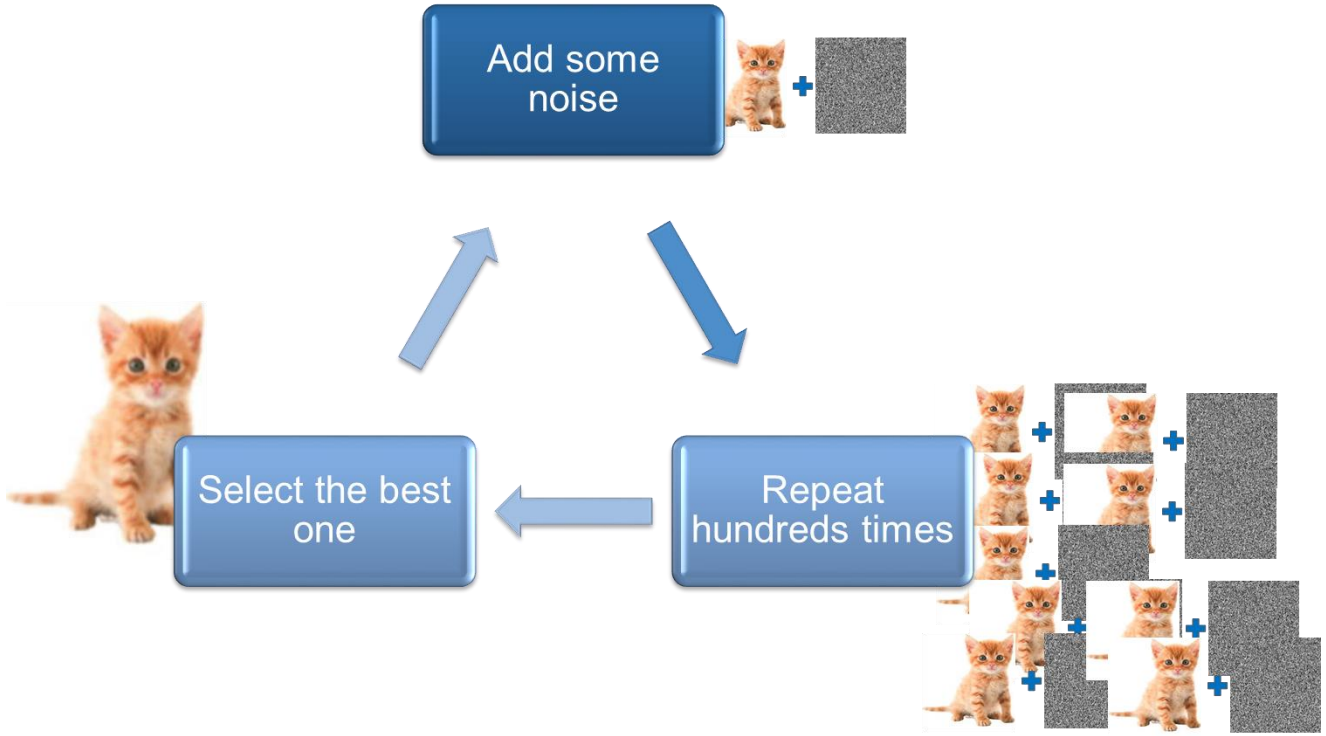
- **Black Box**
  - Hacker can only test model with Input/Output



- **White Box**
  - Hacker knows the detail parameters of the model

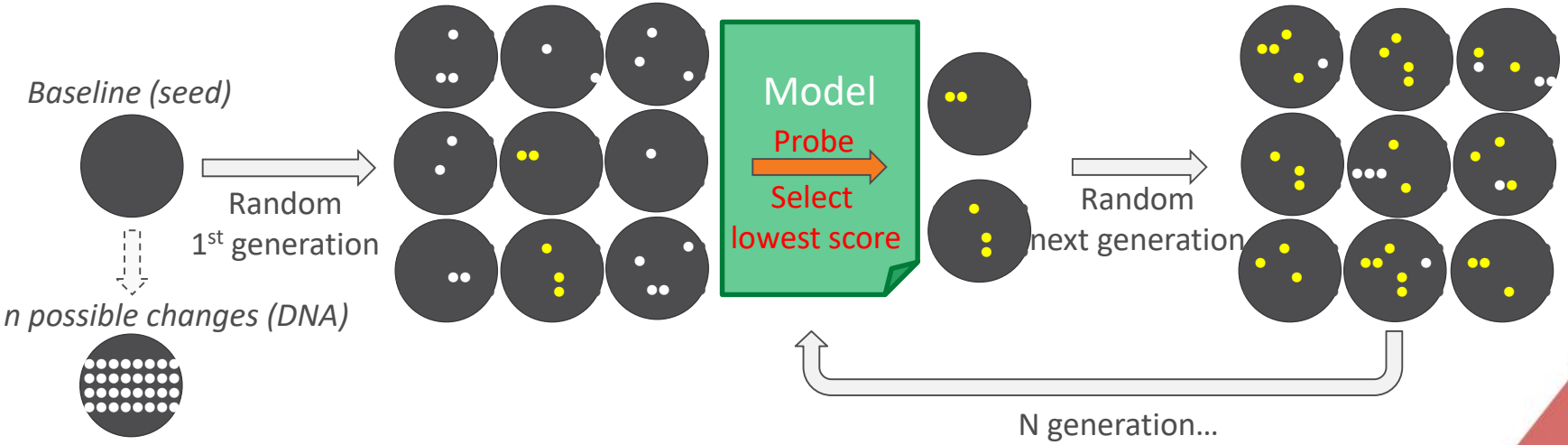


Black Box Evasion: Iterative Random Attack



Evasion successful ratio = 1/1000

Black Box Evasion: Genetic Algorithm



Evasion successful ratio = 1/100

TREND MICRO

Poison Attack

- Online training

Countermeasures

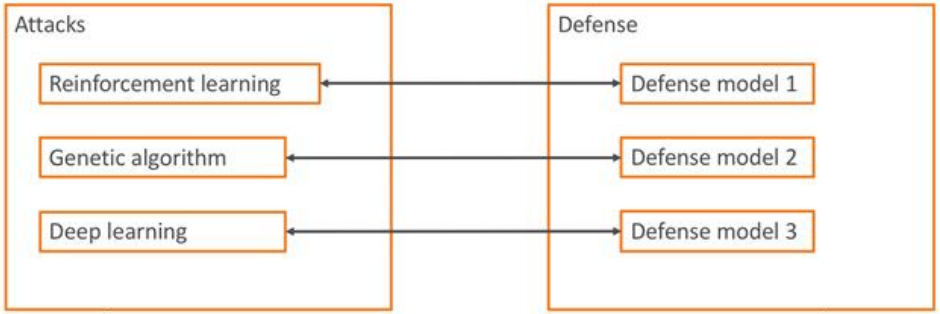
Adversarial ML Countermeasures

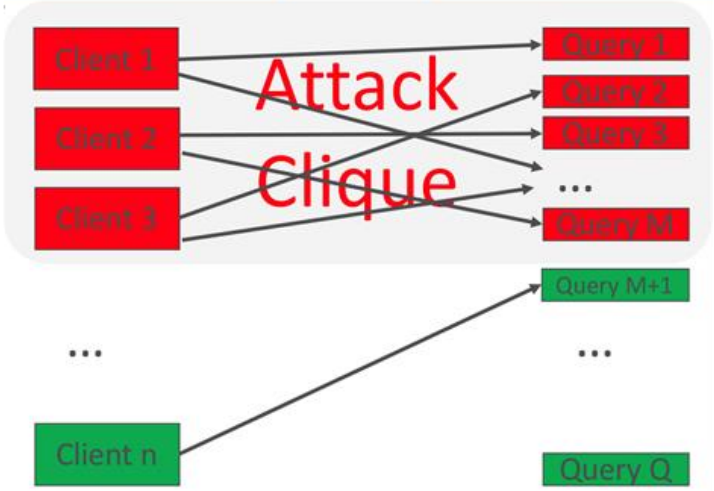
- Evasion Attack - Black box
  - Abuse Protection
  - Model Retrain
    - Reactive
    - Proactive (GAN)
- Evasion Attack - White box
    - Data/feature/model protection
- Poisoning Attack
  - Data/Label quality control

Adversarial ML Countermeasures

- Evasion Attack - Black box
  - Abuse Protection
  - Model Retrain
    - Reactive
    - Proactive (GAN)
- Evasion Attack - White box
  - Data/feature/model protection
- Poisoning Attack
  - Data/Label quality control

Adversarial ML Countermeasures



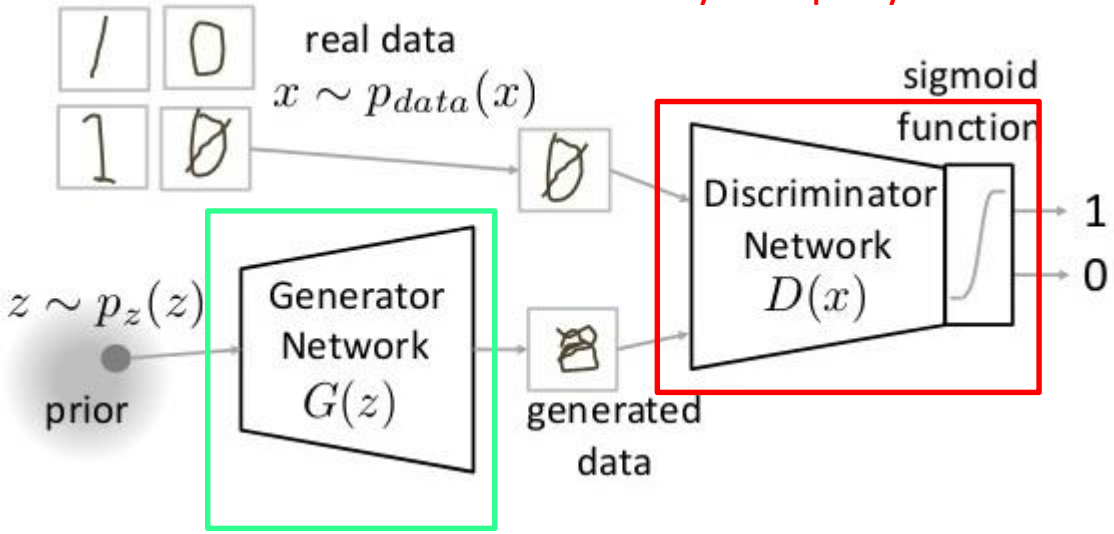


Adversarial ML Countermeasures

- Evasion Attack - Black box
  - Abuse Protection
  - Model Retrain
    - Reactive
    - Proactive
- Evasion Attack - White box
  - Data/feature/model protection
- Poisoning Attack
  - Data/Label quality control

Adversarial ML Countermeasures

Security company model to identify malware

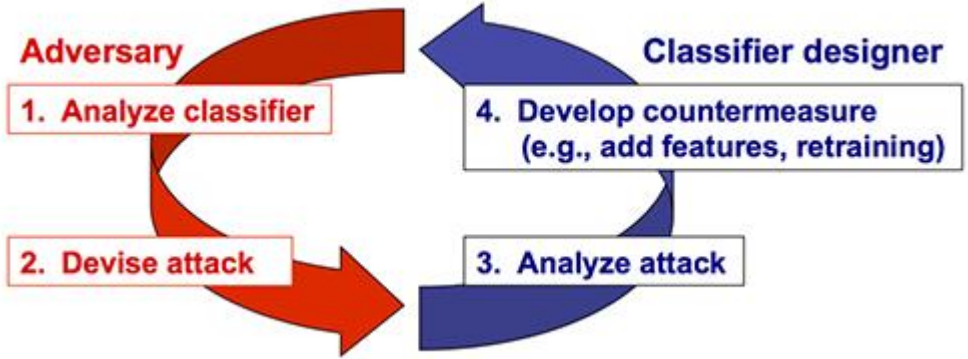


Hacker generate malware to cheat classifier

https://www.analyticsvidhya.com/blog/2017/06/introductory-generative-adversarial-networks-gans/

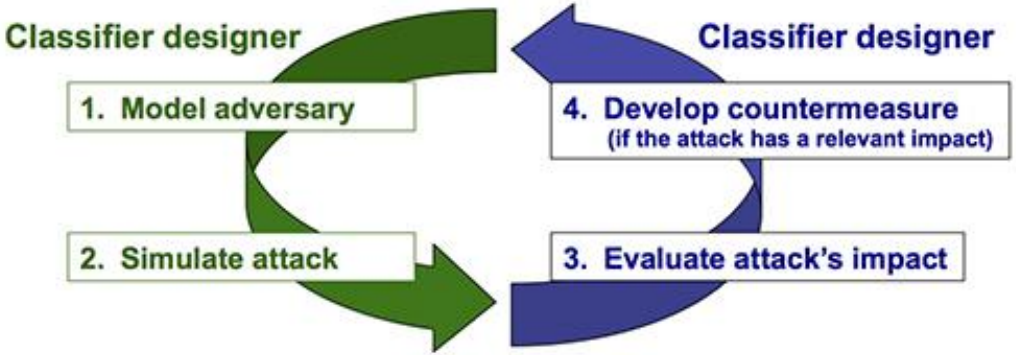
Adversarial ML Countermeasures

Reactive model retrain



Adversarial ML Countermeasures

Proactive model retrain



This method can stop attack A but not stop attack B

Adversarial ML Countermeasures

What if the hair length is an important feature?

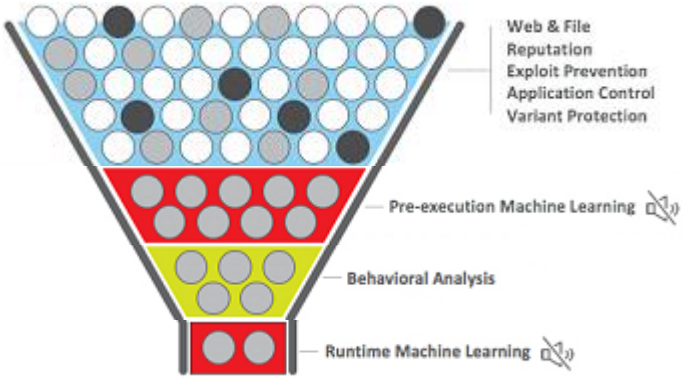
TREND MICRO

Adversarial ML Countermeasures

- Trade off
  - Robustness or Accuracy
  - Proactive or Reactive
  - Fast or Confidence

Adversarial ML Countermeasures

- Trade off
  - Robustness or Accuracy
  - Proactive or Reactive
  - Fast or Confidence



TREND MICRO

Adversarial ML Countermeasures

- Evasion Attack - Black box
  - Abuse Protection
  - Model Retrain
    - Reactive
    - Proactive (GAN)
- Evasion Attack - White box
  - Data/feature/model protection
- Poisoning Attack
  - Data/Label quality control

Adversarial ML Countermeasures

- Evasion Attack - Black box
  - Abuse Protection
  - Model Retrain
    - Reactive
    - Proactive (GAN)
- Evasion Attack - White box
    - Data/feature/model protection
- Poisoning Attack
  - Data/Label quality control

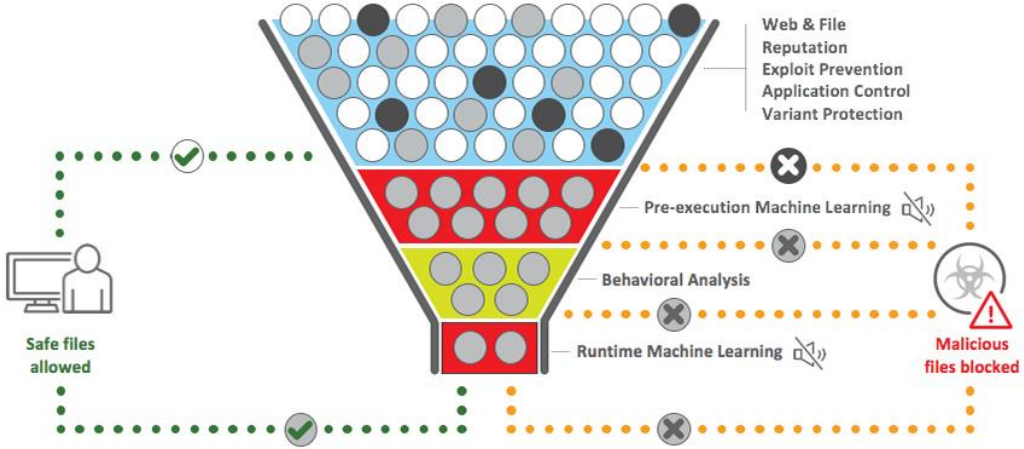
Conclusions

Conclusions

- Almost all models can be cheated
- Find possible vulnerabilities and take the proper actions
- This is an endless battle
  - Pros: Global visibility and excellent operation
  - Cons: 1 FN will cause the damage

Conclusions

- There is no silver bullet for Cyber Security
  - **Dynamic** & **Fast Response** are the key points



TREND MICRO

TREND MICRO

Thank You