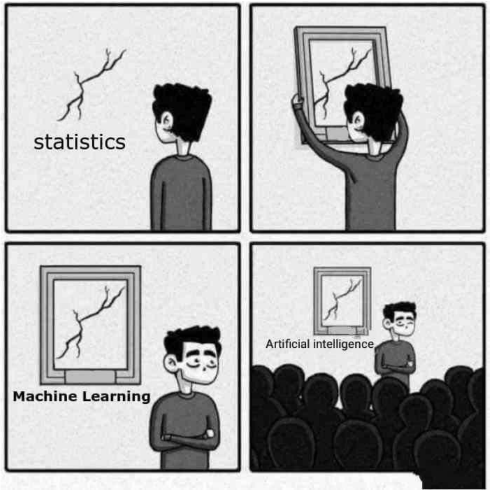
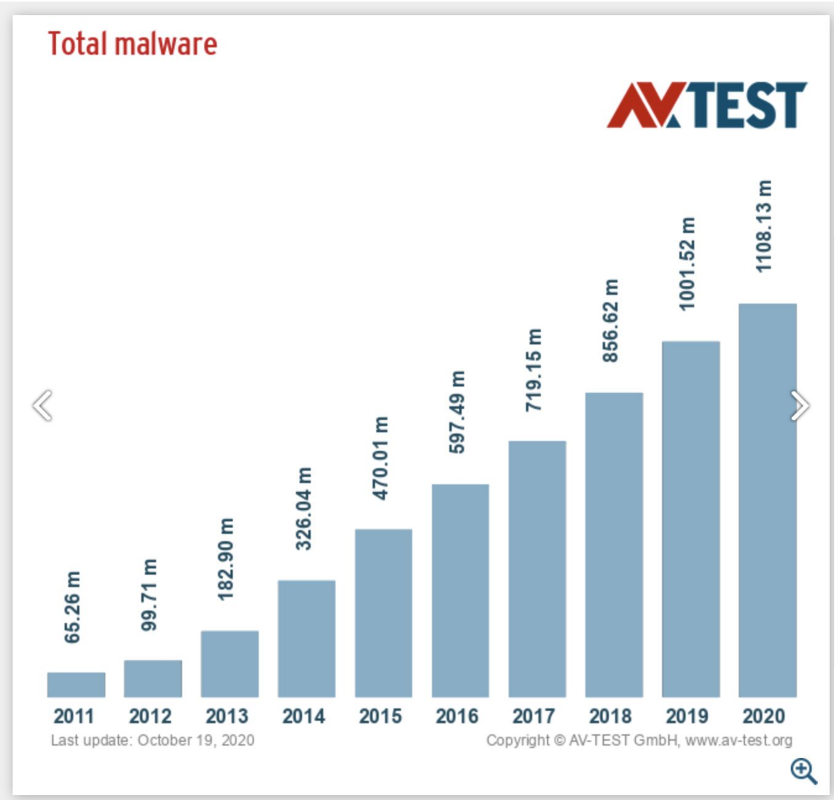
Cybersecurity and Machine Learning



Motivation: **A lot of data!**

- Windows Executables
- Android Applications
- E-mails
- Network Traffic
- Authentication events
- Operation System data such as:
  - System calls
  - Process events
- and more...

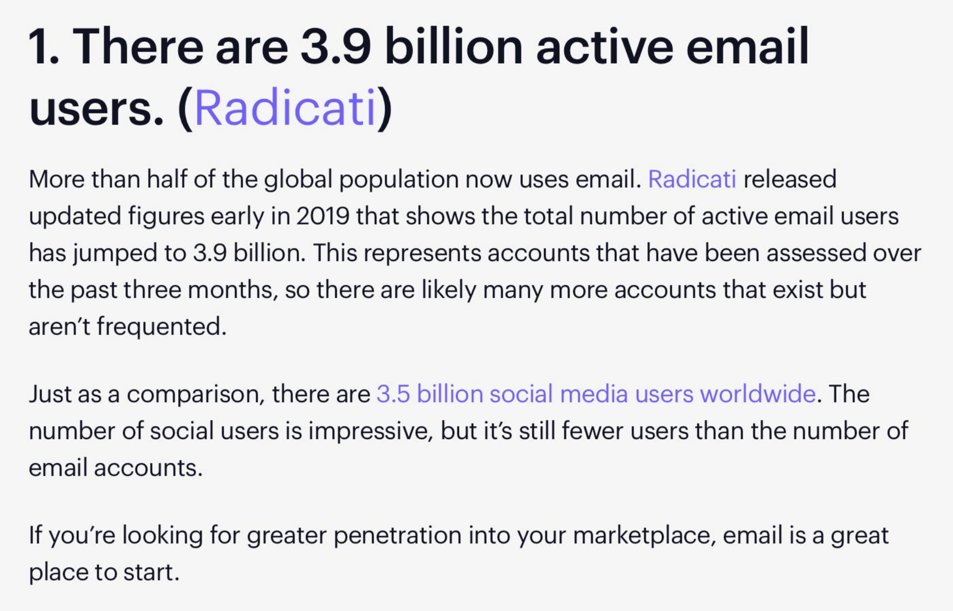


Last update: October 19, 2020

Copyright © AV-TEST GmbH, www.av-test.org

Motivation: **A lot of data!**

- Windows Executables
- Android Applications
- E-mails
- Network Traffic
- Authentication events
- Operation System data such as:
  - System calls
  - Process events
- and more...



1. There are 3.9 billion active email users. (Radicati)

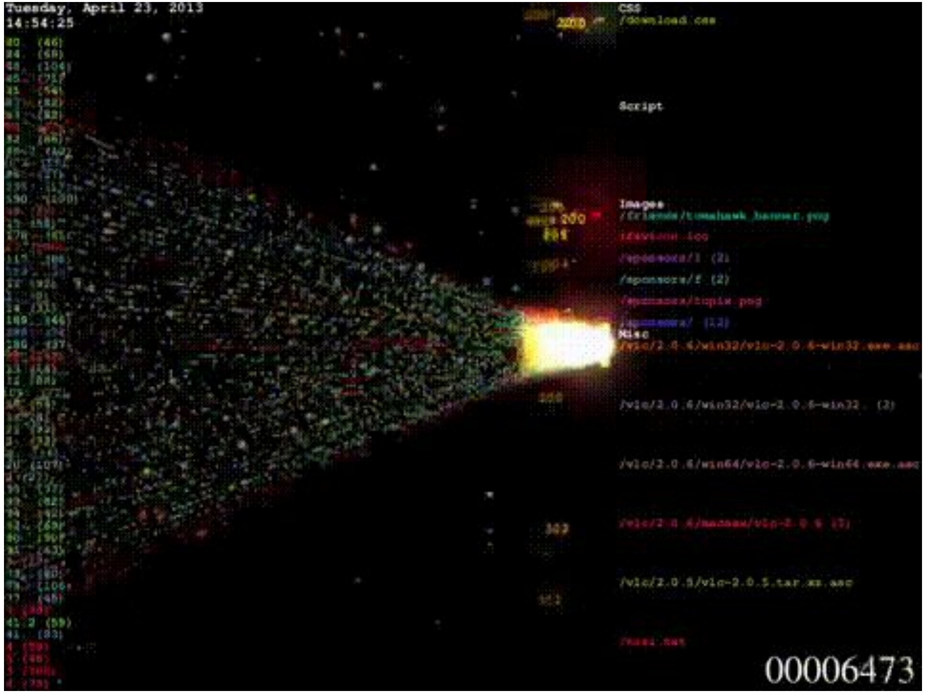
More than half of the global population now uses email. Radicati released updated figures early in 2019 that shows the total number of active email users has jumped to 3.9 billion. This represents accounts that have been assessed over the past three months, so there are likely many more accounts that exist but aren't frequented.

Just as a comparison, there are 3.5 billion social media users worldwide. The number of social users is impressive, but it's still fewer users than the number of email accounts.

If you're looking for greater penetration into your marketplace, email is a great place to start.

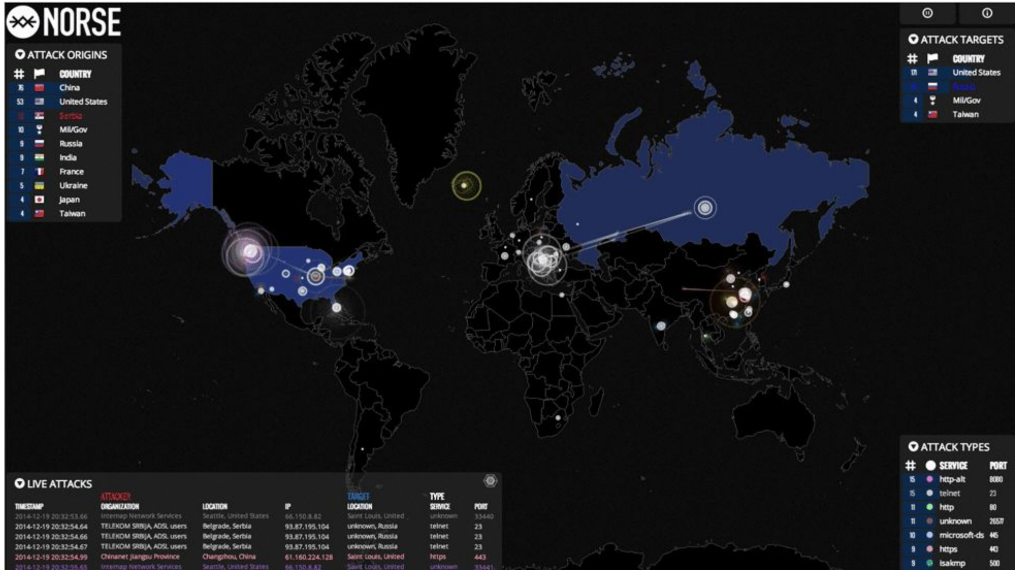
Motivation: **A lot of data!**

- Windows Executables
- Android Applications
- E-mails
- Network Traffic
- Authentication events
- Operation System data such as:
  - System calls
  - Process events
- and more...


Tuesday, April 23, 2013
14:54:25
CSS
/download.css
Script
Images
Misc
/vlc/2.0.6/win32/vlc-2.0.6-win32.exe.asc
/vlc/2.0.6/win32/vlc-2.0.6-win32. (2)
/vlc/2.0.6/win64/vlc-2.0.6-win64.exe.asc
/vlc/2.0.5/vlc-2.0.5.tar.xz.asc
00006473

Motivation: **A lot of data!**

- Windows Executables
- Android Applications
- E-mails
- Network Traffic
- Authentication events
- Operation System data such as:
  - System calls
  - Process events
- and more...



NORSE

ATTACK ORIGINS

#		COUNTRY
76		China
53		United States
[illegible]		Serbia
10		Mil/Gov
9		Russia
9		India
7		France
5		Ukraine
4		Japan
4		Taiwan

ATTACK TARGETS

#		COUNTRY
171		United States
[illegible]		Russia
4		Mil/Gov
4		Taiwan

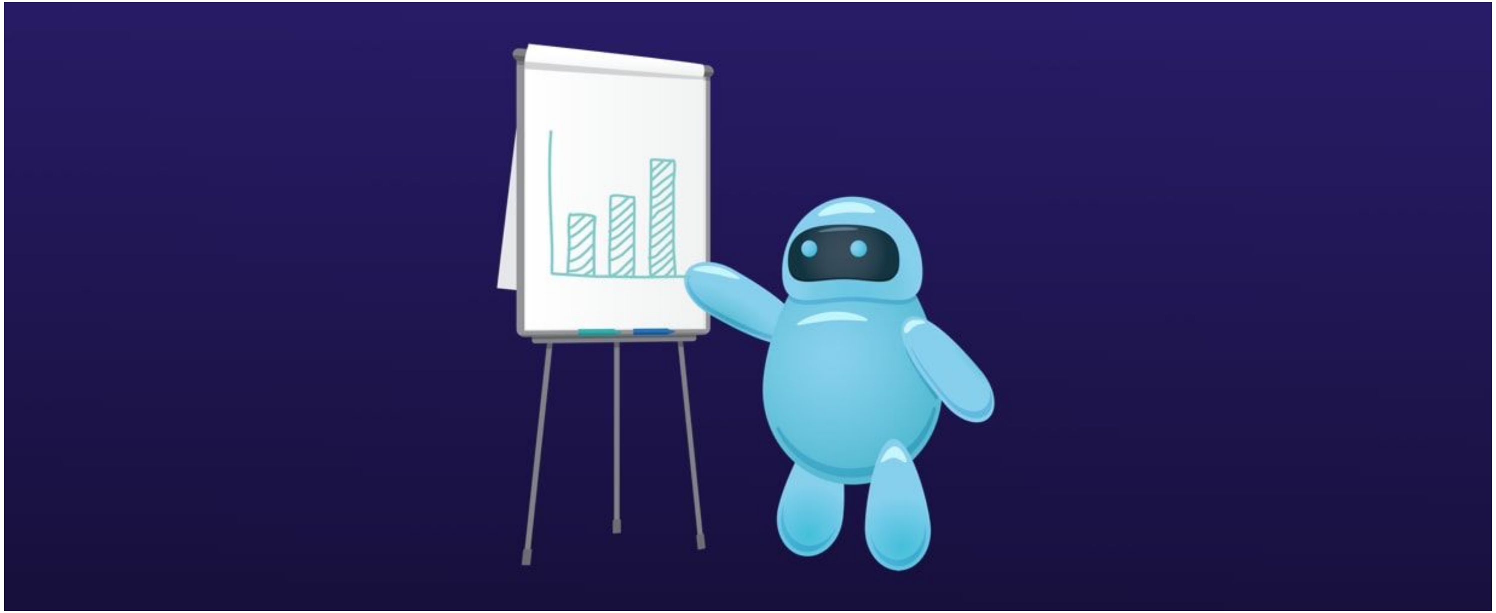
ATTACK TYPES

#		SERVICE	PORT
15	●	http-alt	8080
15	●	telnet	23
11	●	http	80
11	●	unknown	26557
10	●	microsoft-ds	445
9	●	https	443
9	●	isakmp	500

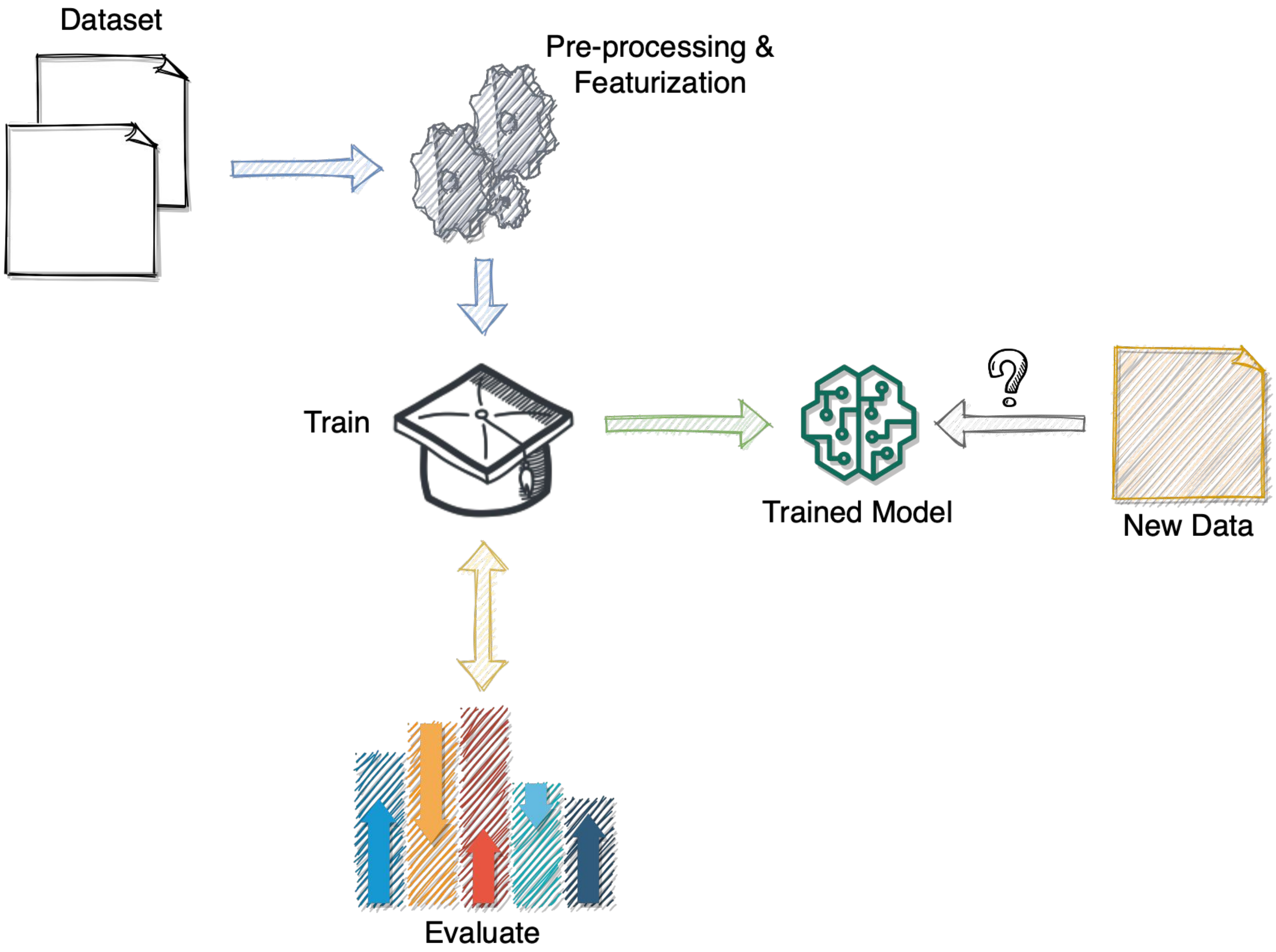
LIVE ATTACKS

	ATTACKER			TARGET	TYPE	
TIMESTAMP	ORGANIZATION	LOCATION	IP	LOCATION	SERVICE	PORT
2014-12-19 20:32:53.66	Internap Network Services	Seattle, United States	66.150.8.82	Saint Louis, United	unknown	33440
2014-12-19 20:32:54.64	TELEKOM SRBIJA, ADSL users	Belgrade, Serbia	93.87.195.104	unknown, Russia	telnet	23
2014-12-19 20:32:54.66	TELEKOM SRBIJA, ADSL users	Belgrade, Serbia	93.87.195.104	unknown, Russia	telnet	23
2014-12-19 20:32:54.67	TELEKOM SRBIJA, ADSL users	Belgrade, Serbia	93.87.195.104	unknown, Russia	telnet	23
2014-12-19 20:32:54.99	Chinanet Jiangsu Province	Changzhou, China	61.160.224.128	Saint Louis, United	https	443
2014-12-19 20:32:55.65	Internap Network Services	Seattle, United States	66.150.8.82	Saint Louis, United	unknown	33441

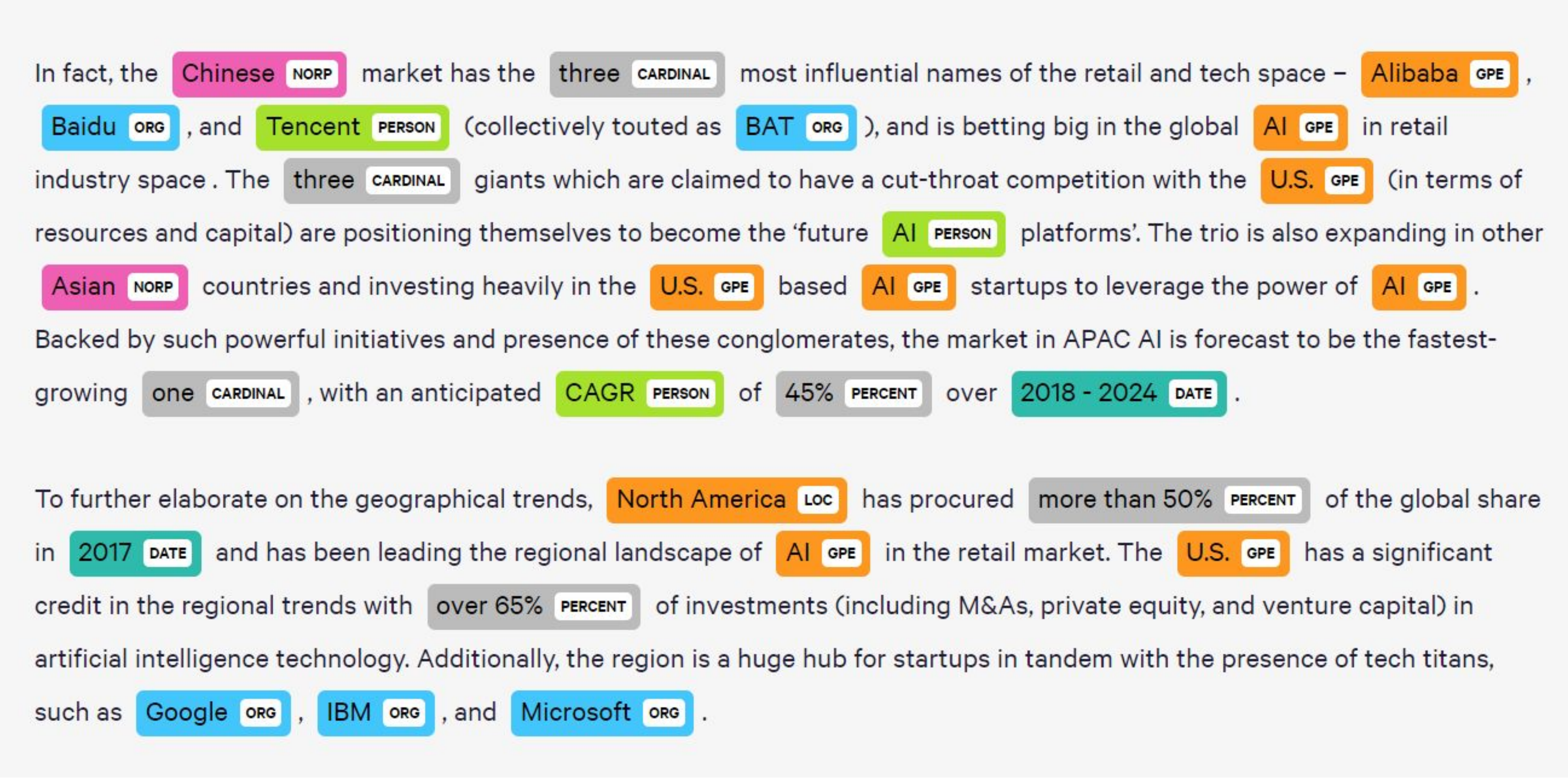
How do you teach it?



[1]


Dataset
Pre-processing &
Featurization
Train
Trained Model
New Data
Evaluate

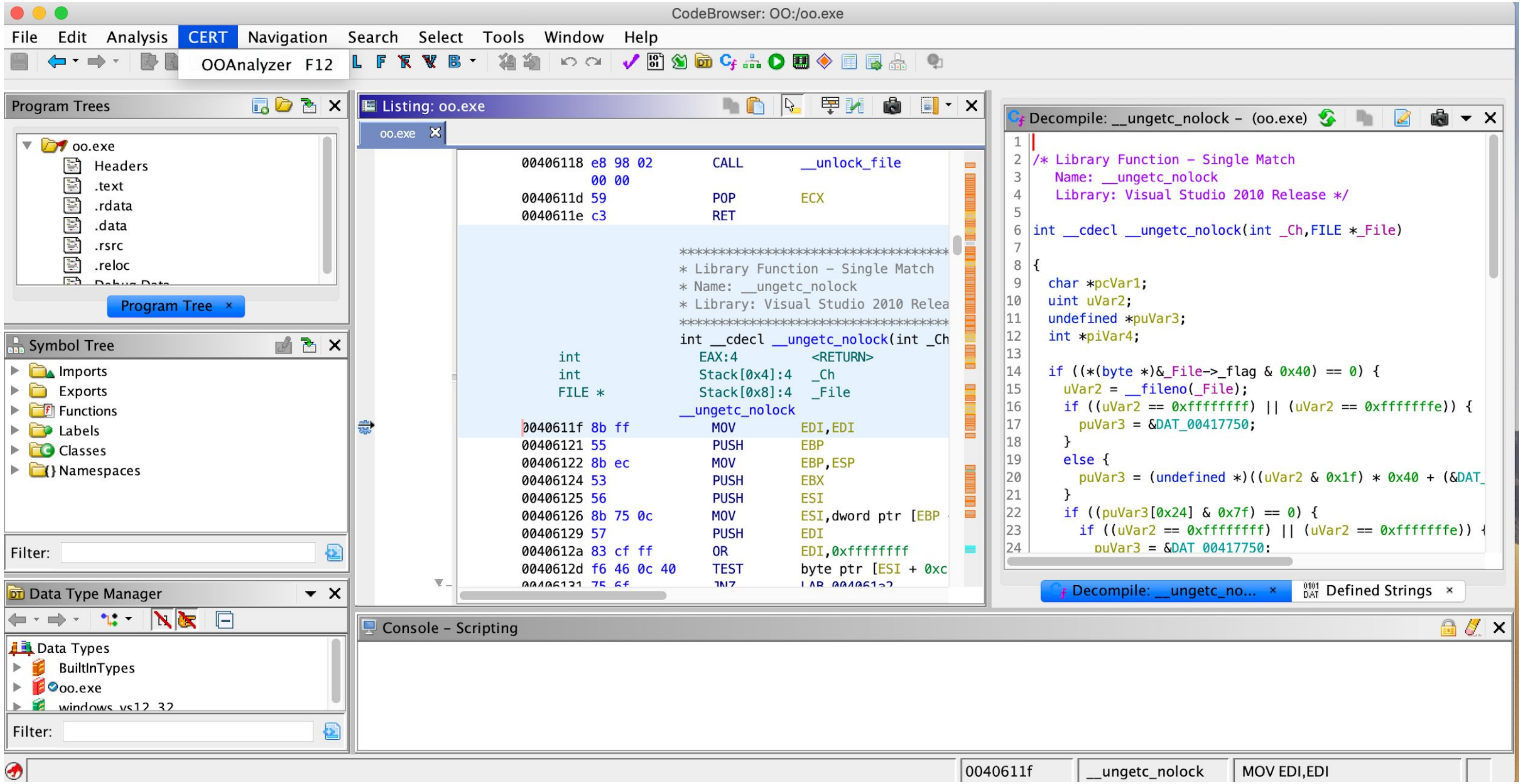
Text



In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE , Baidu ORG , and Tencent PERSON (collectively touted as BAT ORG), and is betting big in the global AI GPE in retail industry space . The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the 'future AI PERSON platforms'. The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL , with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE .

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google ORG , IBM ORG , and Microsoft ORG .

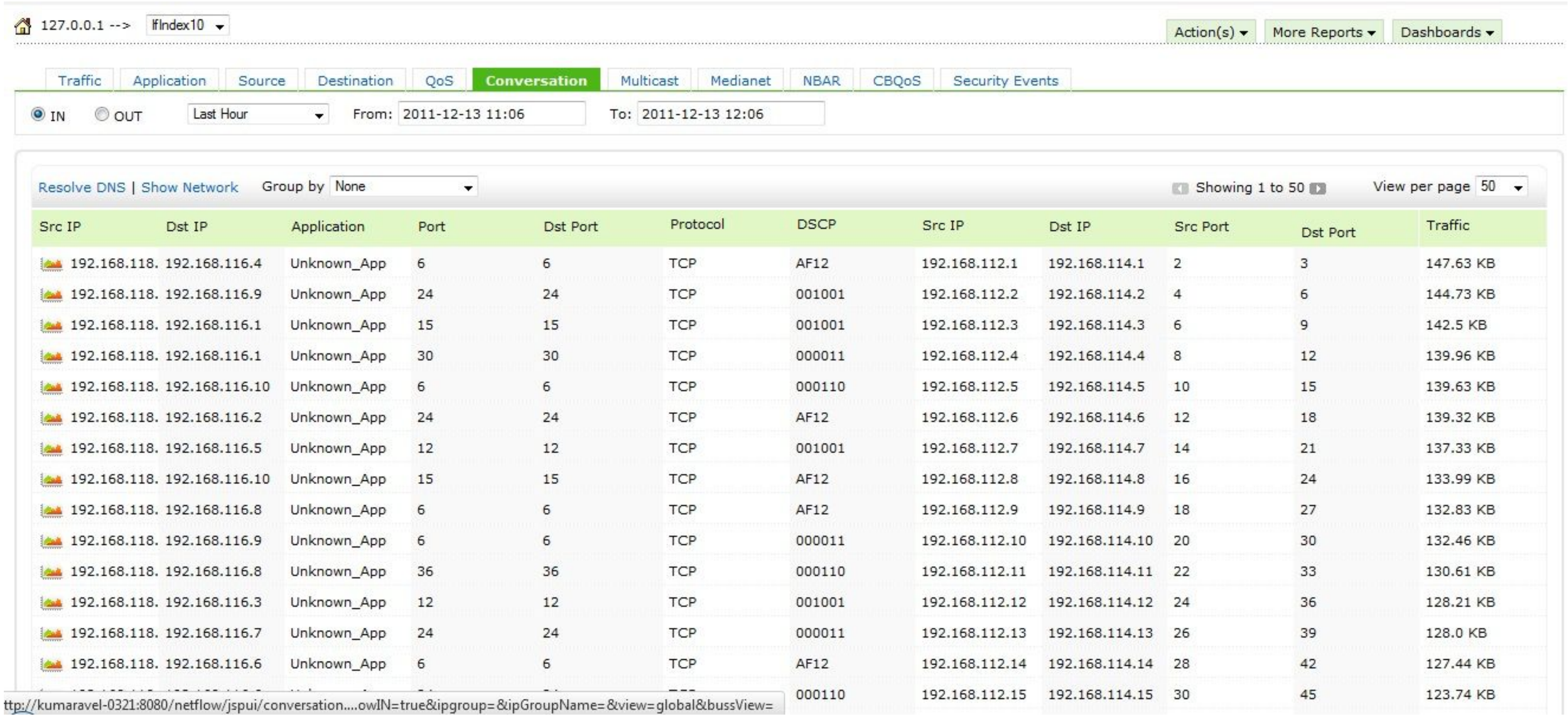
Assembly Instructions



```
        00406118 e8 98 02       CALL       __unlock_file
                 00 00
        0040611d 59             POP        ECX
        0040611e c3             RET

                              ************************************
                              * Library Function - Single Match
                              * Name: __ungetc_nolock
                              * Library: Visual Studio 2010 Relea
                              ************************************
                              int __cdecl __ungetc_nolock(int _Ch
             int                EAX:4          <RETURN>
             int                Stack[0x4]:4   _Ch
             FILE *             Stack[0x8]:4   _File
                              __ungetc_nolock
        0040611f 8b ff          MOV        EDI,EDI
        00406121 55             PUSH       EBP
        00406122 8b ec          MOV        EBP,ESP
        00406124 53             PUSH       EBX
        00406125 56             PUSH       ESI
        00406126 8b 75 0c       MOV        ESI,dword ptr [EBP
        00406129 57             PUSH       EDI
        0040612a 83 cf ff       OR         EDI,0xffffffff
        0040612d f6 46 0c 40    TEST       byte ptr [ESI + 0xc
        00406131 75 6f          JNZ        LAB_004061a2
```

```

/* Library Function - Single Match
   Name: __ungetc_nolock
   Library: Visual Studio 2010 Release */

int __cdecl __ungetc_nolock(int _Ch,FILE *_File)

{
  char *pcVar1;
  uint uVar2;
  undefined *puVar3;
  int *piVar4;

  if ((*(byte *)&_File->_flag & 0x40) == 0) {
    uVar2 = __fileno(_File);
    if ((uVar2 == 0xffffffff) || (uVar2 == 0xfffffffe)) {
      puVar3 = &DAT_00417750;
    }
    else {
      puVar3 = (undefined *)((uVar2 & 0x1f) * 0x40 + (&DAT_
    }
    if ((puVar3[0x24] & 0x7f) == 0) {
      if ((uVar2 == 0xffffffff) || (uVar2 == 0xfffffffe)) {
        puVar3 = &DAT_00417750;
```

Netflow Logs



● IN ○ OUT Last Hour From: 2011-12-13 11:06 To: 2011-12-13 12:06

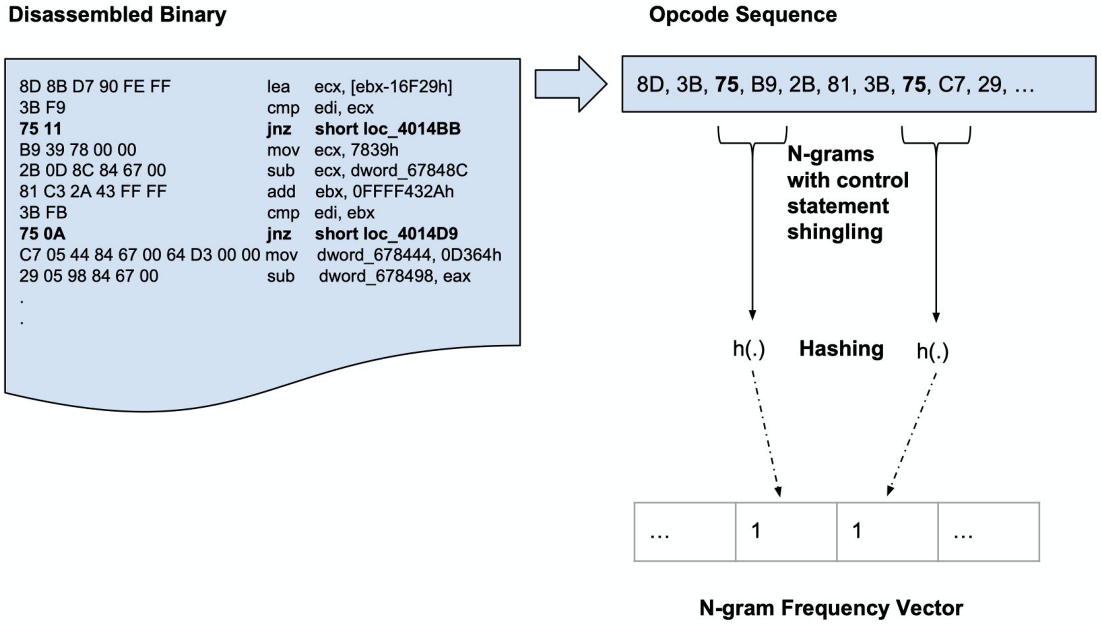
Group by None

View per page 50

Src IP	Dst IP	Application	Port	Dst Port	Protocol	DSCP	Src IP	Dst IP	Src Port	Dst Port	Traffic
192.168.118.	192.168.116.4	Unknown_App	6	6	TCP	AF12	192.168.112.1	192.168.114.1	2	3	147.63 KB
192.168.118.	192.168.116.9	Unknown_App	24	24	TCP	001001	192.168.112.2	192.168.114.2	4	6	144.73 KB
192.168.118.	192.168.116.1	Unknown_App	15	15	TCP	001001	192.168.112.3	192.168.114.3	6	9	142.5 KB
192.168.118.	192.168.116.1	Unknown_App	30	30	TCP	000011	192.168.112.4	192.168.114.4	8	12	139.96 KB
192.168.118.	192.168.116.10	Unknown_App	6	6	TCP	000110	192.168.112.5	192.168.114.5	10	15	139.63 KB
192.168.118.	192.168.116.2	Unknown_App	24	24	TCP	AF12	192.168.112.6	192.168.114.6	12	18	139.32 KB
192.168.118.	192.168.116.5	Unknown_App	12	12	TCP	001001	192.168.112.7	192.168.114.7	14	21	137.33 KB
192.168.118.	192.168.116.10	Unknown_App	15	15	TCP	AF12	192.168.112.8	192.168.114.8	16	24	133.99 KB
192.168.118.	192.168.116.8	Unknown_App	6	6	TCP	AF12	192.168.112.9	192.168.114.9	18	27	132.83 KB
192.168.118.	192.168.116.9	Unknown_App	6	6	TCP	000011	192.168.112.10	192.168.114.10	20	30	132.46 KB
192.168.118.	192.168.116.8	Unknown_App	36	36	TCP	000110	192.168.112.11	192.168.114.11	22	33	130.61 KB
192.168.118.	192.168.116.3	Unknown_App	12	12	TCP	001001	192.168.112.12	192.168.114.12	24	36	128.21 KB
192.168.118.	192.168.116.7	Unknown_App	24	24	TCP	000011	192.168.112.13	192.168.114.13	26	39	128.0 KB
192.168.118.	192.168.116.6	Unknown_App	6	6	TCP	AF12	192.168.112.14	192.168.114.14	28	42	127.44 KB
						000110	192.168.112.15	192.168.114.15	30	45	123.74 KB

ttp://kumaravel-0321:8080/netflow/jspui/conversation....owIN=true&ipgroup=&ipGroupName=&view=global&bussView=

Must speak in their language!



```
8D 8B D7 90 FE FF             lea   ecx, [ebx-16F29h]
3B F9                         cmp   edi, ecx
75 11                         jnz   short loc_4014BB
B9 39 78 00 00                mov   ecx, 7839h
2B 0D 8C 84 67 00             sub   ecx, dword_67848C
81 C3 2A 43 FF FF             add   ebx, 0FFFF432Ah
3B FB                         cmp   edi, ebx
75 0A                         jnz   short loc_4014D9
C7 05 44 84 67 00 64 D3 00 00 mov   dword_678444, 0D364h
29 05 98 84 67 00             sub   dword_678498, eax
.
.
```

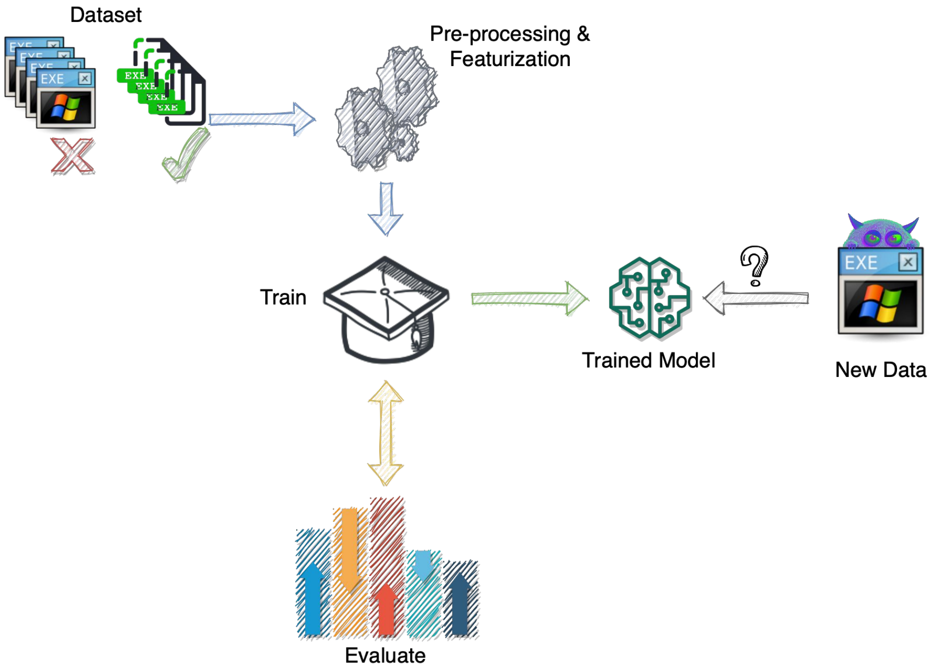
Opcode Sequence
8D, 3B, **75**, B9, 2B, 81, 3B, **75**, C7, 29, …
N-grams with control statement shingling
h(.)
Hashing
h(.)
…
1
1
…
N-gram Frequency Vector

Fig. 2. Extracting opcode n-grams and hashing to reduce dimensionality

Hassen, Mehadi & Carvalho, Marco & Chan, Philip. (2017). Malware classification using static analysis based features. 1-7. 10.1109/SSCI.2017.8285426.

What can we do?

Malware Classification



Dataset

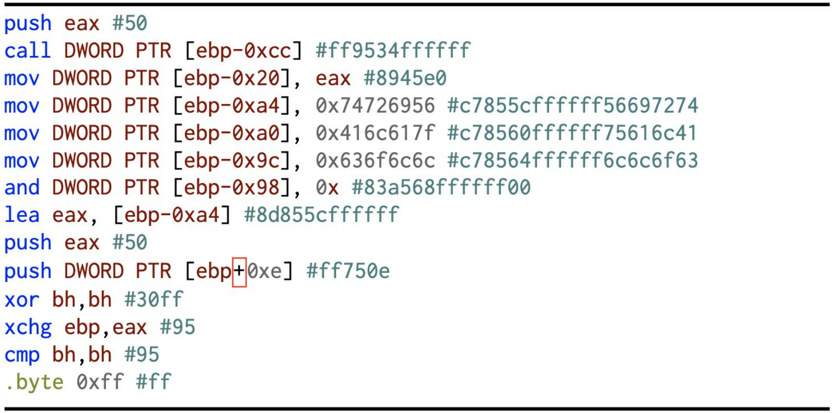
Pre-processing & Featurization

Train

Trained Model

New Data

Evaluate



```
push eax #50
call DWORD PTR [ebp-0xcc] #ff9534ffffff
mov DWORD PTR [ebp-0x20], eax #8945e0
mov DWORD PTR [ebp-0xa4], 0x74726956 #c7855cffffff56697274
mov DWORD PTR [ebp-0xa0], 0x416c617f #c78560ffffff75616c41
mov DWORD PTR [ebp-0x9c], 0x636f6c6c #c78564ffffff6c6c6f63
and DWORD PTR [ebp-0x98], 0x #83a568ffffff00
lea eax, [ebp-0xa4] #8d855cffffff
push eax #50
push DWORD PTR [ebp+0xe] #ff750e
xor bh,bh #30ff
xchg ebp,eax #95
cmp bh,bh #95
.byte 0xff #ff
```

Figure 3: Example of a disassembled 64-gram feature found in the EMBER dataset. The hex values of the raw bytes are shown in comments for each line of assembly.

Raff, E., Fleming, W., Zak, R., Anderson, H., Finlayson, B., Nicholas, C., & McLean, M. (2019). KiloGrams: Very Large N-Grams for Malware Classification. ArXiv, abs/1908.00200.

Malware Clustering

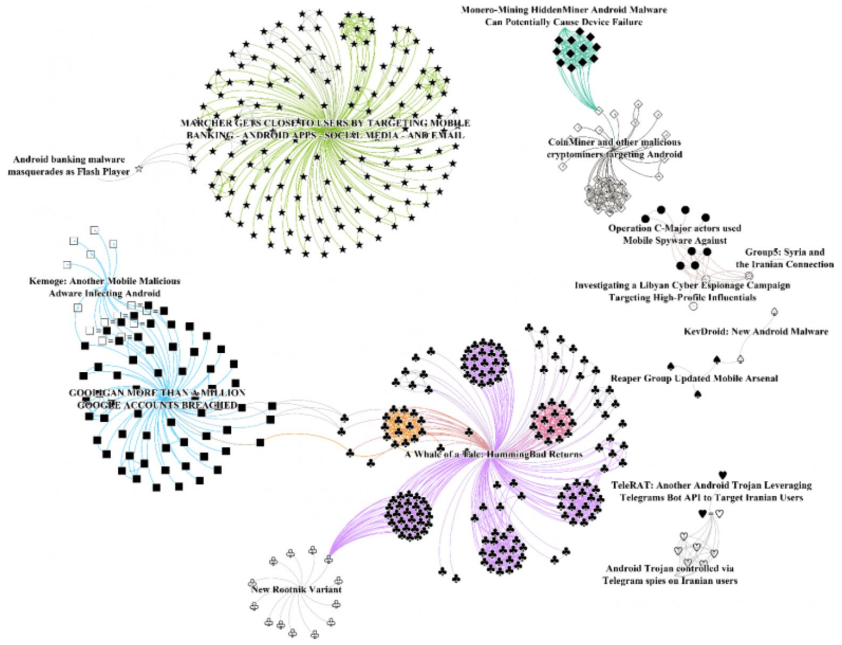
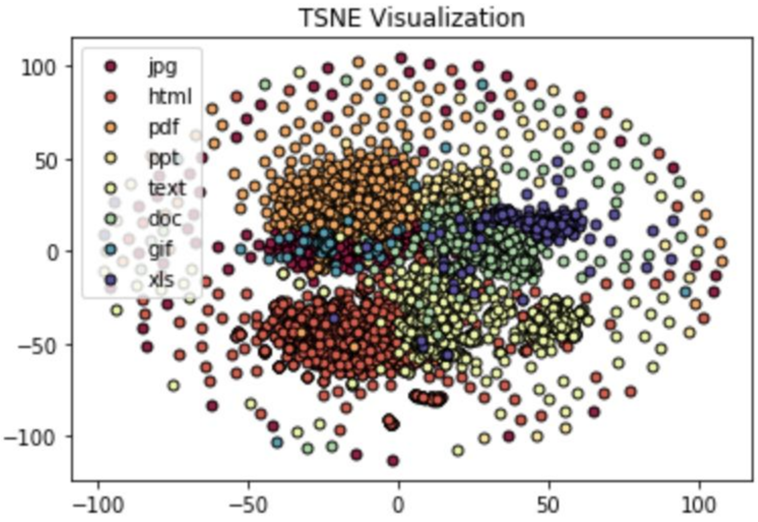
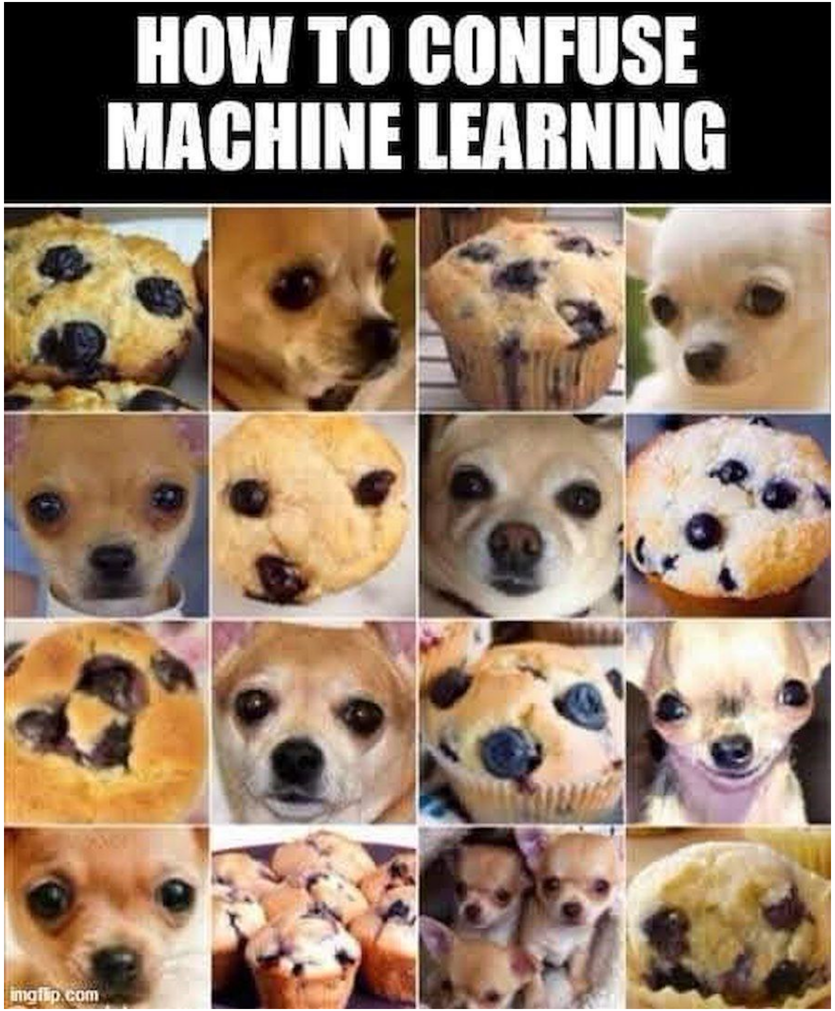

Monero-Mining HiddenMiner Android Malware Can Potentially Cause Device Failure
MARCHER GETS CLOSE TO USERS BY TARGETING MOBILE BANKING - ANDROID APPS - SOCIAL MEDIA - AND EMAIL
Android banking malware masquerades as Flash Player
CoinMiner and other malicious cryptominers targeting Android
Operation C-Major actors used Mobile Spyware Against
Group5: Syria and the Iranian Connection
Investigating a Libyan Cyber Espionage Campaign Targeting High-Profile Influentials
Kemoge: Another Mobile Malicious Adware Infecting Android
KevDroid: New Android Malware
GOOLIGAN MORE THAN A MILLION GOOGLE ACCOUNTS BREACHED
Reaper Group Updated Mobile Arsenal
A Whale of a Tale: HummingBad Returns
TeleRAT: Another Android Trojan Leveraging Telegrams Bot API to Target Iranian Users
Android Trojan controlled via Telegram spies on Iranian users
New Rootnik Variant

Figure 17: Relationship of 15 AlienVault OTX reports.

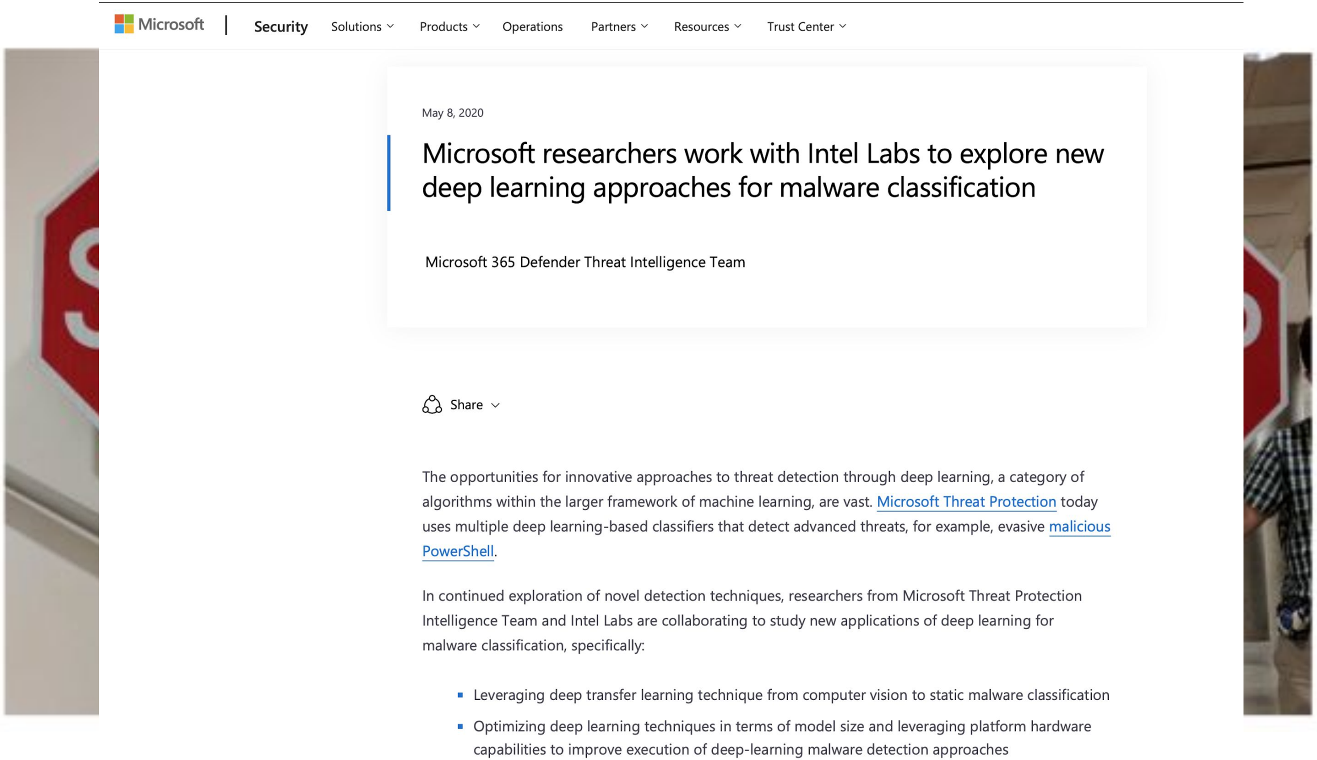
Lee et al. Dexofuzzy: Android malware similarity clustering method using opcode sequence.



Raff, E., & Nicholas, C. (2017). An Alternative to NCD for Large Sequences, Lempel-Ziv Jaccard Distance. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.


HOW TO CONFUSE
MACHINE LEARNING
imgflip.com

Adversarial Machine Learning



May 8, 2020

Microsoft researchers work with Intel Labs to explore new deep learning approaches for malware classification

Microsoft 365 Defender Threat Intelligence Team

The opportunities for innovative approaches to threat detection through deep learning, a category of algorithms within the larger framework of machine learning, are vast. Microsoft Threat Protection today uses multiple deep learning-based classifiers that detect advanced threats, for example, evasive malicious PowerShell.

In continued exploration of novel detection techniques, researchers from Microsoft Threat Protection Intelligence Team and Intel Labs are collaborating to study new applications of deep learning for malware classification, specifically:

- Leveraging deep transfer learning technique from computer vision to static malware classification
- Optimizing deep learning techniques in terms of model size and leveraging platform hardware capabilities to improve execution of deep-learning malware detection approaches

Goodfellow, I. (2020, October 05). Attacking Machine Learning with Adversarial Examples. Retrieved October 21, 2020, from https://openai.com/blog/adversarial-example-research/

Statistical User Behaviour Analysis

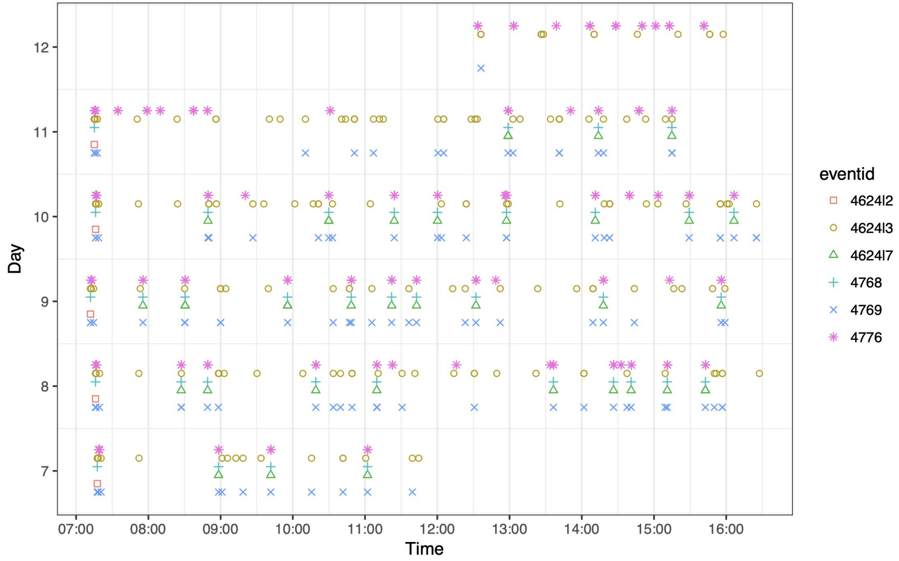


Figure 9: Event times for User205265. 4624l2 corresponds to *EventID* 4624 - *LogonType* 2.

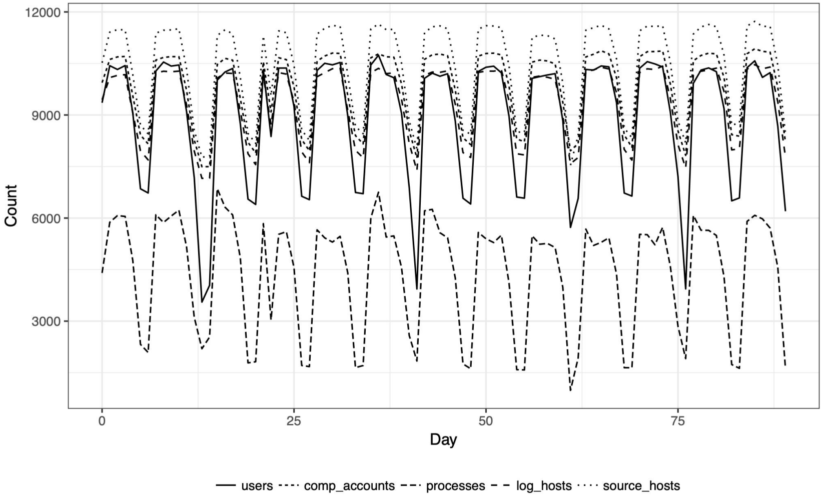
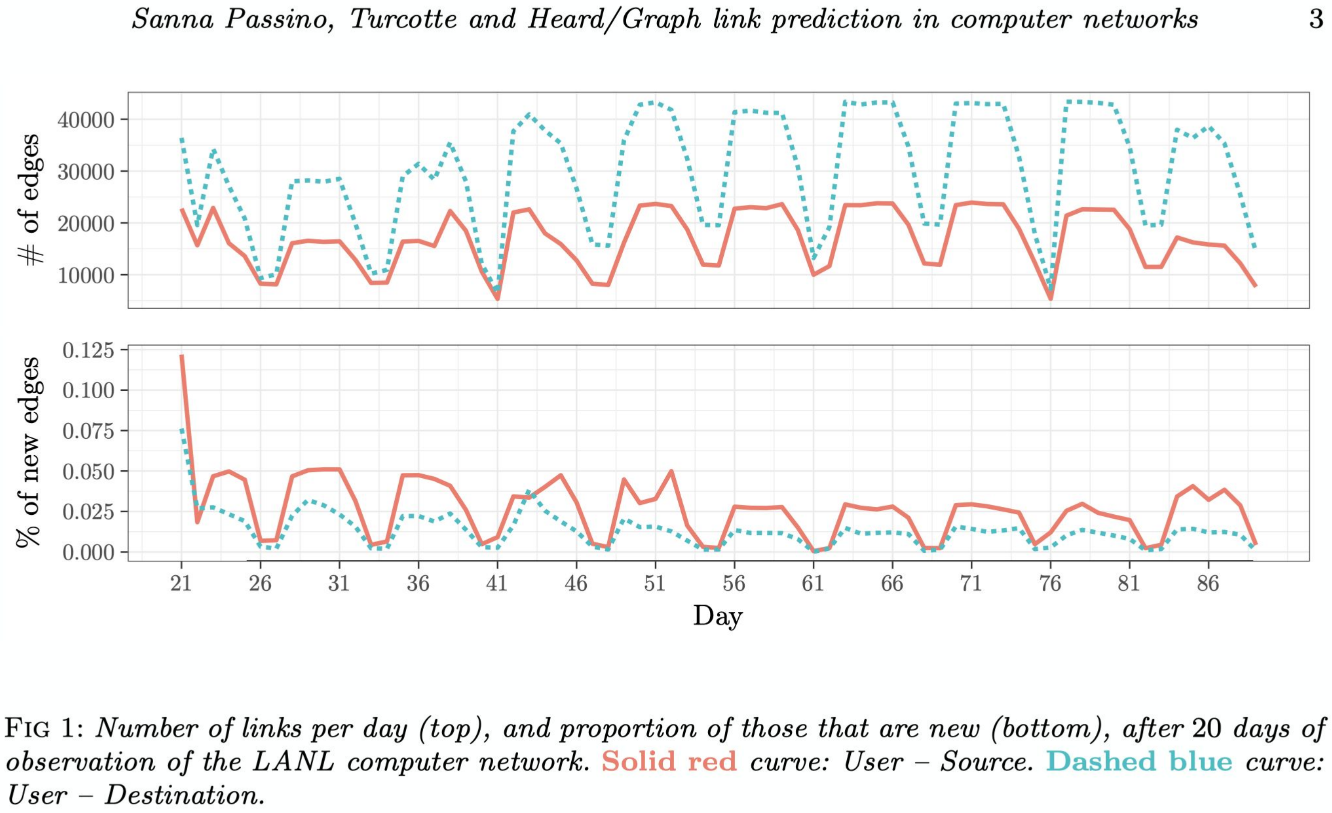


Figure 8: Daily count of the fields in Figure 7.

Turcotte, M.J., Kent, A., & Hash, C. (2017). Unified Host and Network Data Set. ArXiv, abs/1708.07518.

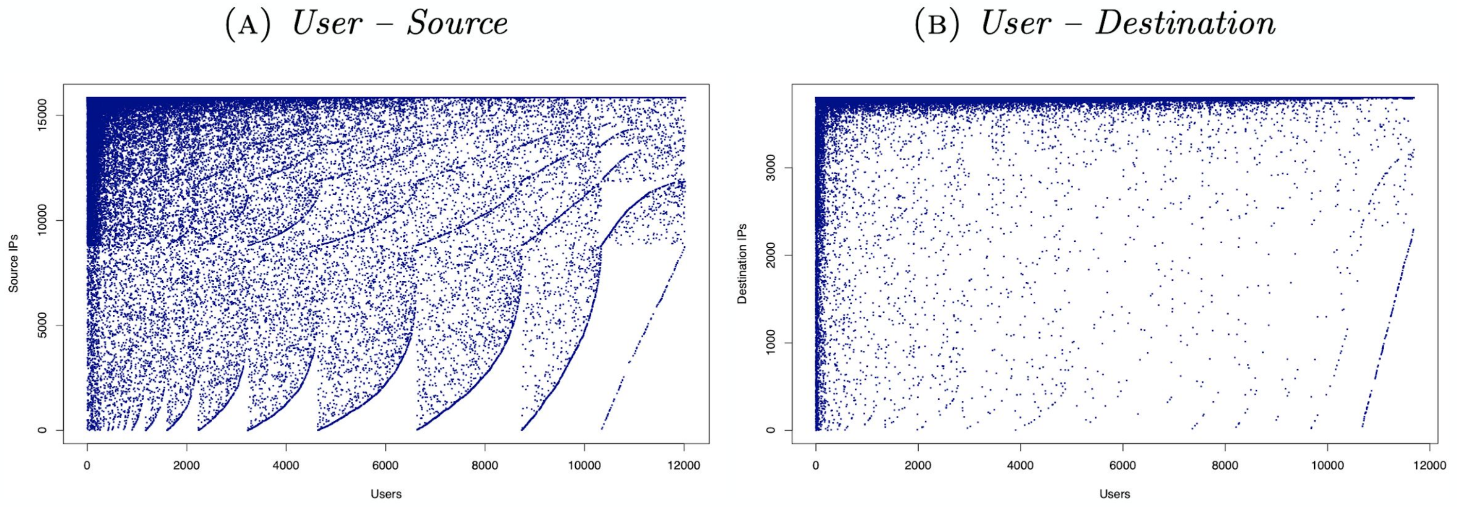


Sanna Passino, Turcotte and Heard/Graph link prediction in computer networks 3

FIG 1: *Number of links per day (top), and proportion of those that are new (bottom), after* 20 *days of observation of the LANL computer network.* **Solid red** *curve: User – Source.* **Dashed blue** *curve: User – Destination.*

Turcotte, M.J., Kent, A., & Hash, C. (2017). Unified Host and Network Data Set. ArXiv, abs/1708.07518.

Anomaly Detection: Link Prediction



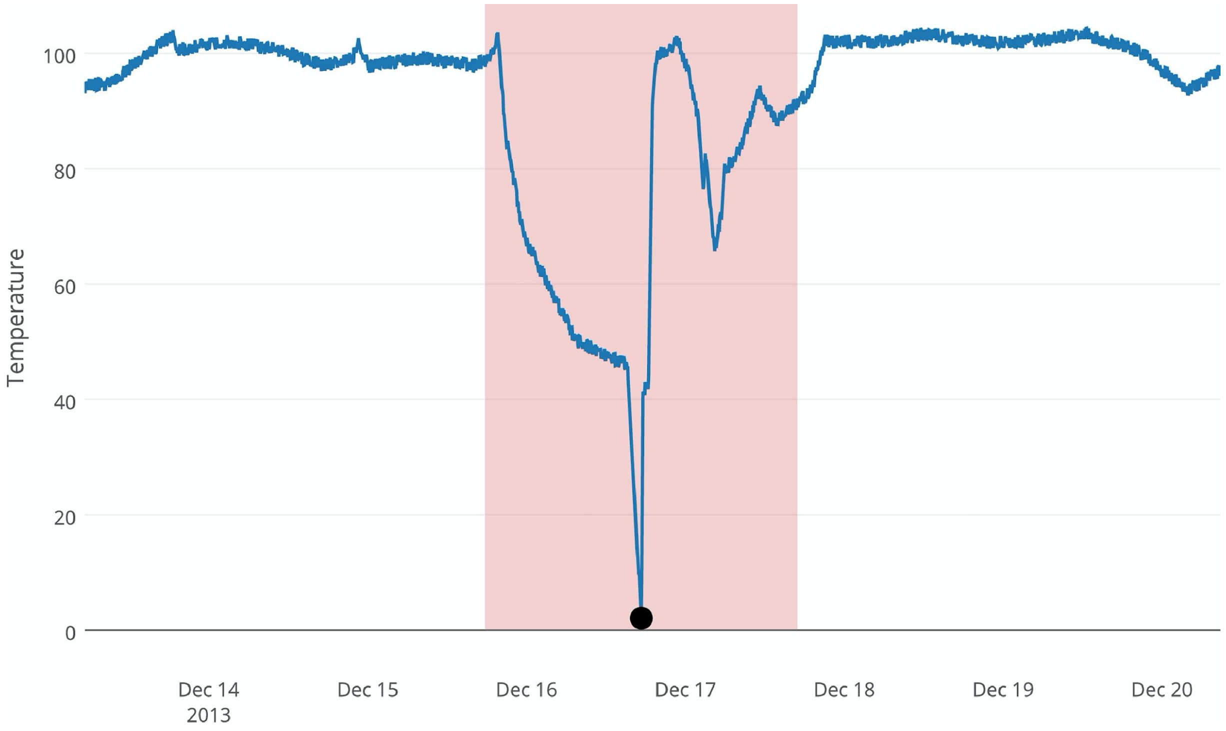
(A) *User – Source*

(B) *User – Destination*

FIG 2: *Training set adjacency matrices for the two graphs (spy-plot). Nodes are sorted by in-degree and out-degree.*

Passino, F.S., Turcotte, M.J., & Heard, N. (2020). Graph link prediction in computer networks using Poisson matrix factorisation. ArXiv, abs/2001.09456.

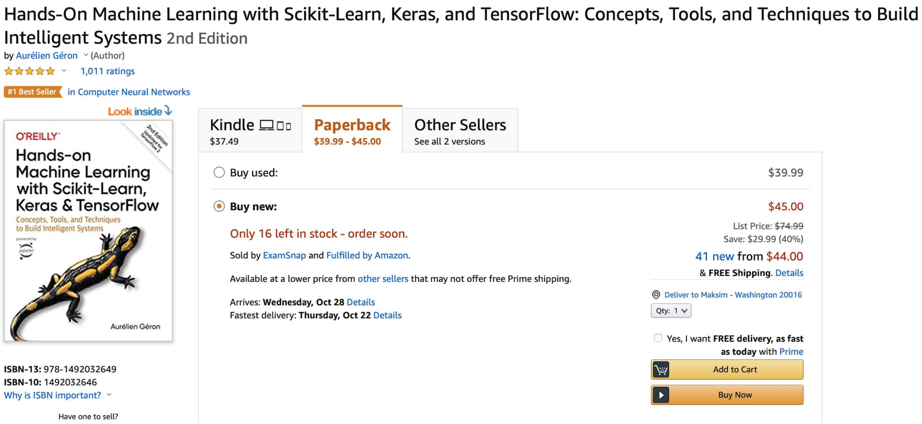
Anomaly Detection: Time Series



Ahmad, Subutai & Lavin, Alexander & Purdy, Scott & Agha, Zuha. (2017). Unsupervised real-time anomaly detection for streaming data. Neurocomputing. 10.1016/j.neucom.2017.04.070.

MACHINE LEARNING
MACHINE LEARNING EVERYWHERE
makeameme.org

Woa!! Where do I sign-up?



Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 2nd Edition

by Aurélien Géron (Author)

★★★★★ 1,011 ratings

#1 Best Seller in Computer Neural Networks

Look inside

O'REILLY
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
2nd Edition
Aurélien Géron

ISBN-13: 978-1492032649
ISBN-10: 1492032646
Why is ISBN important?

Have one to sell?

Kindle	Paperback	Other Sellers
$37.49	$39.99 - $45.00	See all 2 versions

○ Buy used: $39.99

● Buy new: $45.00

Only 16 left in stock - order soon.

List Price: $74.99
Save: $29.99 (40%)

Sold by ExamSnap and Fulfilled by Amazon.

41 new from $44.00
& FREE Shipping. Details

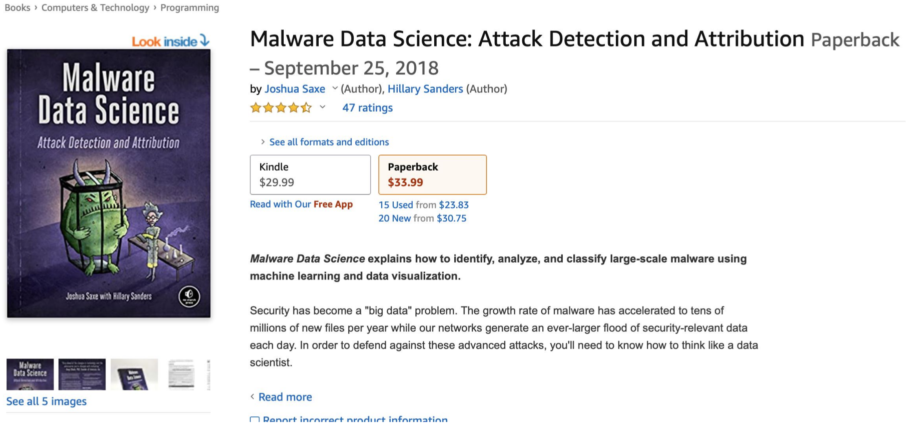
Available at a lower price from other sellers that may not offer free Prime shipping.

Arrives: Wednesday, Oct 28 Details
Fastest delivery: Thursday, Oct 22 Details

Deliver to Maksim - Washington 20016

Qty: 1

□ Yes, I want FREE delivery, as fast as today with Prime



Books › Computers & Technology › Programming

Malware Data Science: Attack Detection and Attribution Paperback – September 25, 2018

by Joshua Saxe (Author), Hillary Sanders (Author)

★★★★☆ 47 ratings

Look inside

Malware Data Science
Attack Detection and Attribution
Joshua Saxe with Hillary Sanders

See all formats and editions

Kindle	Paperback
$29.99	$33.99
Read with Our Free App	15 Used from $23.83 20 New from $30.75

Malware Data Science **explains how to identify, analyze, and classify large-scale malware using machine learning and data visualization.**

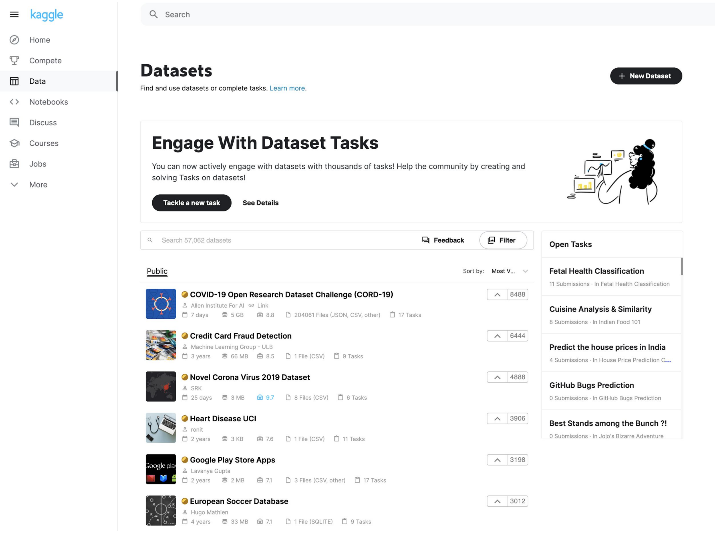
Security has become a "big data" problem. The growth rate of malware has accelerated to tens of millions of new files per year while our networks generate an ever-larger flood of security-relevant data each day. In order to defend against these advanced attacks, you'll need to know how to think like a data scientist.

Read more

See all 5 images

□ Report incorrect product information

- Books
- Malware Research Group
- Internships
- Projects on your free time



Datasets

Find and use datasets or complete tasks. Learn more.

Engage With Dataset Tasks

You can now actively engage with datasets with thousands of tasks! Help the community by creating and solving Tasks on datasets!

COVID-19 Open Research Dataset Challenge (CORD-19) — 8488
Allen Institute For AI · Link
7 days · 5 GB · 8.8 · 204061 Files (JSON, CSV, other) · 17 Tasks

Credit Card Fraud Detection — 6444
Machine Learning Group - ULB
3 years · 66 MB · 8.5 · 1 File (CSV) · 9 Tasks

Novel Corona Virus 2019 Dataset — 4888
SRK
25 days · 3 MB · 9.7 · 8 Files (CSV) · 6 Tasks

Heart Disease UCI — 3906
ronit
2 years · 3 KB · 7.6 · 1 File (CSV) · 11 Tasks

Google Play Store Apps — 3198
Lavanya Gupta
2 years · 2 MB · 7.1 · 3 Files (CSV, other) · 17 Tasks

European Soccer Database — 3012
Hugo Mathien
4 years · 33 MB · 7.1 · 1 File (SQLITE) · 9 Tasks

Open Tasks

Fetal Health Classification
11 Submissions · In Fetal Health Classification

Cuisine Analysis & Similarity
8 Submissions · In Indian Food 101

Predict the house prices in India
4 Submissions · In House Price Prediction C...

GitHub Bugs Prediction
0 Submissions · In GitHub Bugs Prediction

Best Stands among the Bunch ?!
0 Submissions · In Jojo's Bizarre Adventure

References

[1] Sarver, C. (n.d.). Introduction to Regression and Classification in Machine Learning. Retrieved from https://www.springboard.com/blog/introduction-regression-classification-machine-learning/

[2] Email Usage Statistics in 2019. (n.d.). Retrieved October 21, 2020, from https://www.campaignmonitor.com/blog/email-marketing/2019/07/email-usage-statistics-in-2019/

[3] Malware Statistics & Trends Report: AV-TEST. (n.d.). Retrieved October 21, 2020, from https://www.av-test.org/en/statistics/malware/

[4] Microsoft researchers work with Intel Labs to explore new deep learning approaches for malware classification. (2020, May 08). Retrieved October 21, 2020, from https://www.microsoft.com/security/blog/2020/05/08/microsoft-researchers-work-with-intel-labs-to-explore-new-deep-learning-approaches-for-malware-classification/

[5] Goodfellow, I. (2020, October 05). Attacking Machine Learning with Adversarial Examples. Retrieved October 21, 2020, from https://openai.com/blog/adversarial-example-research/

[6] Dexofuzzy: Android malware similarity clustering method using opcode sequence. (n.d.). Retrieved October 21, 2020, from https://www.virusbulletin.com/virusbulletin/2019/11/dexofuzzy-android-malware-similarity-clustering-method-using-opcode-sequence/

[7] Raff, E., & Nicholas, C. (2017). An Alternative to NCD for Large Sequences, Lempel-Ziv Jaccard Distance. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

References

[8] Turcotte, M.J., Kent, A., & Hash, C. (2017). Unified Host and Network Data Set. ArXiv, abs/1708.07518.

[9] Passino, F.S., Turcotte, M.J., & Heard, N. (2020). Graph link prediction in computer networks using Poisson matrix factorisation. ArXiv, abs/2001.09456.

[10] Raff, E., Fleming, W., Zak, R., Anderson, H., Finlayson, B., Nicholas, C., & McLean, M. (2019). KiloGrams: Very Large N-Grams for Malware Classification. ArXiv, abs/1908.00200.

[11] Ahmad, Subutai & Lavin, Alexander & Purdy, Scott & Agha, Zuha. (2017). Unsupervised real-time anomaly detection for streaming data. Neurocomputing. 10.1016/j.neucom.2017.04.070.

[12] Hassen, Mehadi & Carvalho, Marco & Chan, Philip. (2017). Malware classification using static analysis based features. 1-7. 10.1109/SSCI.2017.8285426.