



머신러닝 기반의 보안데이터 분석 연구

2019.06.21(금)

국민대 소프트웨어융합대학
부교수 윤명근

발표 순서

- 보안 빅데이터
- 학습 모델
- 악성코드 분석 사례
- 보안관제 데이터 분석 사례

발표자 소개

•약력

- 2010~현재 : 국민대학교 소프트웨어융합대학 부교수
- 2004~2008: University of Florida, 컴퓨터공학 박사 (세부전공: 네트워크 보안)
- 1998~2010: 금융결제원 전자금융연구소, 금융ISAC, IT기획부

•연구 분야

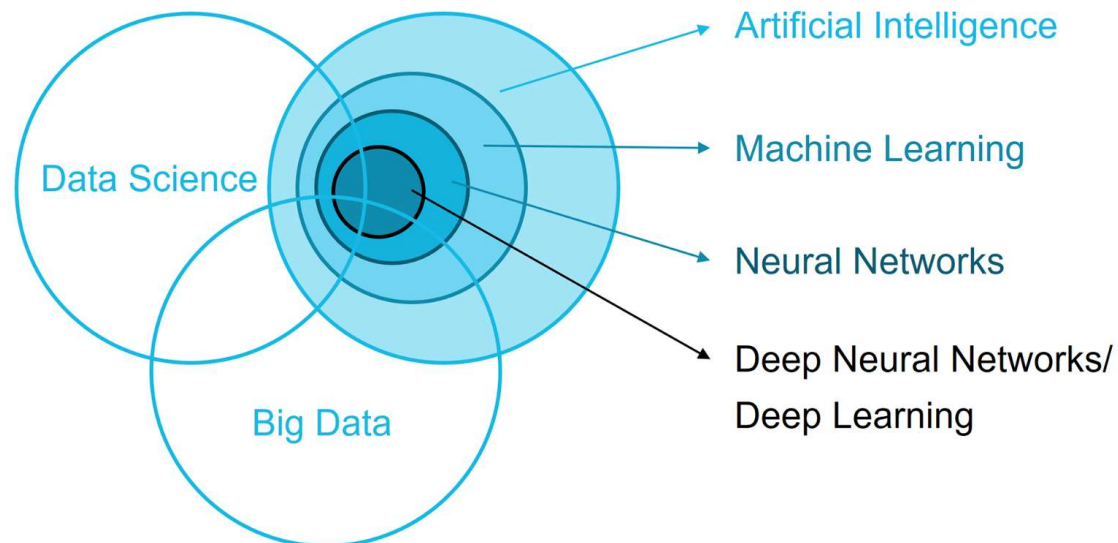
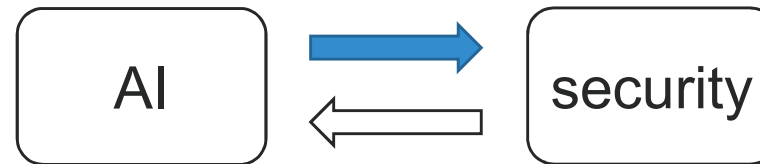
- 인공지능 보안, 보안 빅데이터 분석, 사이버 보안, 핀테크 등

•주요 수상

- 2018년 한국정보보호학회 & 한국인터넷진흥원 정보보호R&D 데이터챌린지, AI기반 악성코드 탐지트랙 대학(원) 1위, AI기반 안드로이드 악성앱 탐지 1위 (지도교수)
- 2018년 금융보안원 논문공모전 최우수상
- 2018년 한국정보과학회 한국소프트웨어종합학술대회 정보보안및고신뢰컴퓨팅 부문 최우수논문상
- 2017년 금융보안원 논문공모전 최우수상
- 2017년 사이버보안 논문공모전 최우수상
- 2017년 한국정보보호학회 & 한국인터넷진흥원 정보보호R&D 데이터챌린지 AI기반 악성코드 탐지트랙 대학(원) 2위

보안 빅데이터

- 보안을 위한 인공지능 기술
 - 보안 빅데이터 분석을 위한 머신러닝 응용 기술

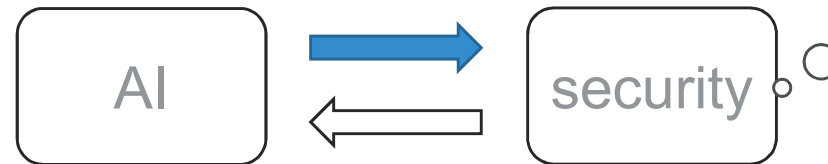


VENN diagram of AI, Big Data and Data Science Fraunhofer FOKUS

<https://towardsdatascience.com/role-of-data-science-in-artificial-intelligence-950efedd2579>

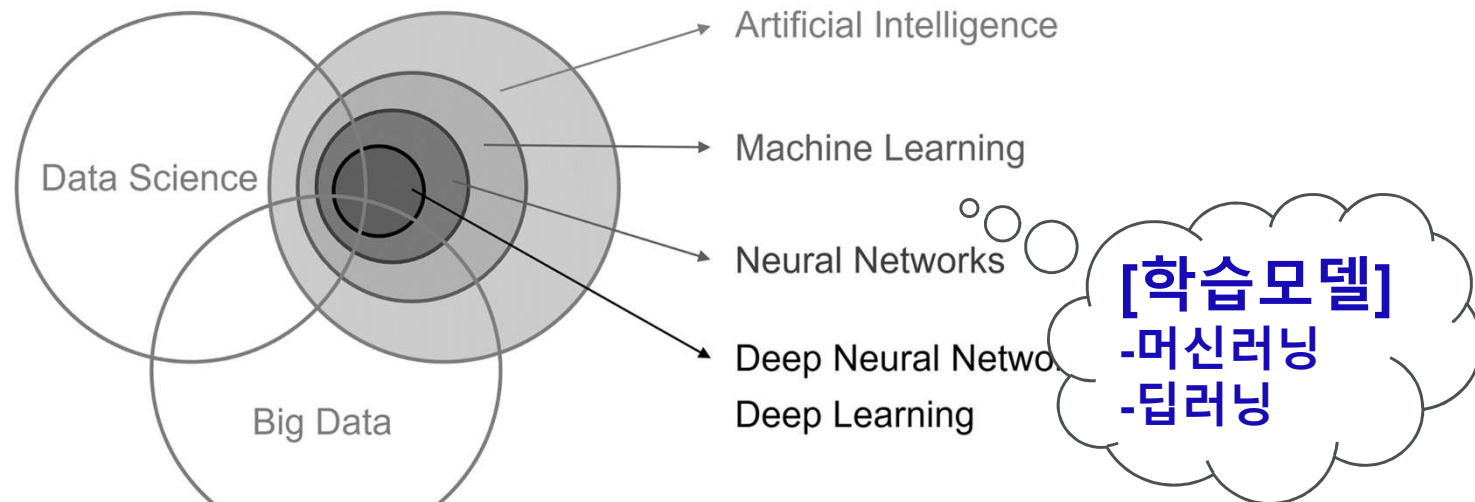
보안 빅데이터

- 보안을 위한 인공지능 기술
 - 보안 빅데이터 분석을 위한 머신러닝 응용 기술



[데이터]
-악성코드
-보안관제

기계학습 기반의 보안 분야 빅데이터 분석에 관한 연구

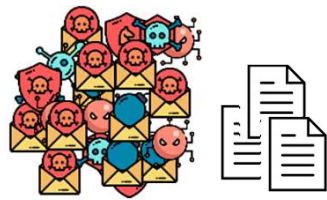


VENN diagram of AI, Big Data and Data Science Fraunhofer FOKUS

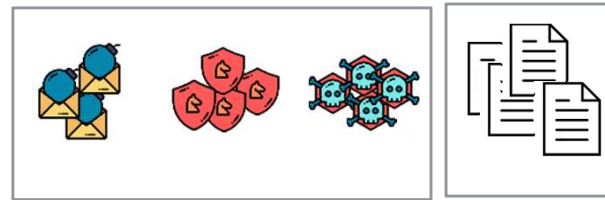
<https://towardsdatascience.com/role-of-data-science-in-artificial-intelligence-950efedd2579>

보안 빅데이터

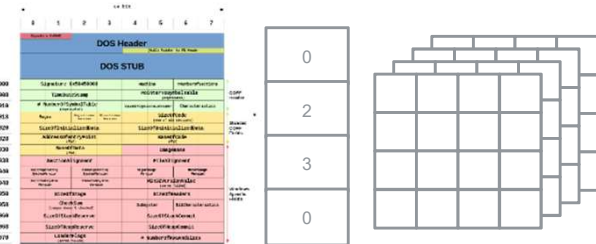
머신러닝 기반 보안데이터 분석 과정



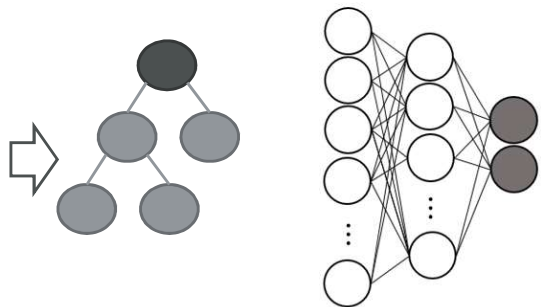
[data collecting]



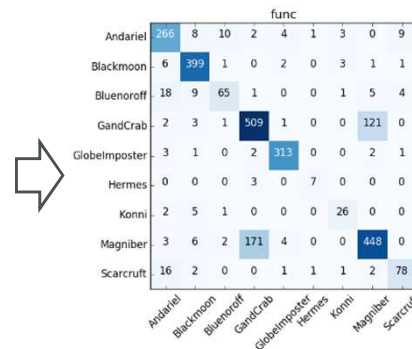
[data preprocessing]



[feature engineering]



[model development]



[test]

보안 빅데이터

- 빅 데이터

- 크기 (Volume), 속도 (Velocity), 다양성 (variety)

- 보안 빅 데이터

- 악성코드 (Malware)

- ✓ 실행 파일

- 윈도우 **PE(Portable Executable)**

- ✓ 문서형 악성코드

- **PDF**, MS Office, HWP

- ✓ URL, 스크립트,...

- 보안관계 로그

- ✓ 침입탐지/차단시스템 이벤트 로그

- ✓ 침입차단시스템 로그

- ✓ 네트워크 플로우 정보

- ✓ 서버, EDR(Endpoint Detection & Response), IoT 디바이스 로그

보안 빅데이터

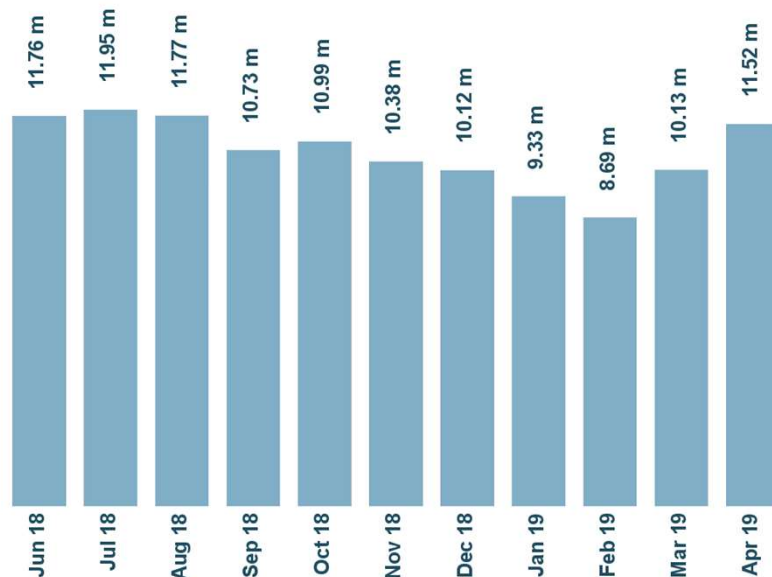
•악성코드

■실행 파일

✓윈도우 PE(Portable Executable)

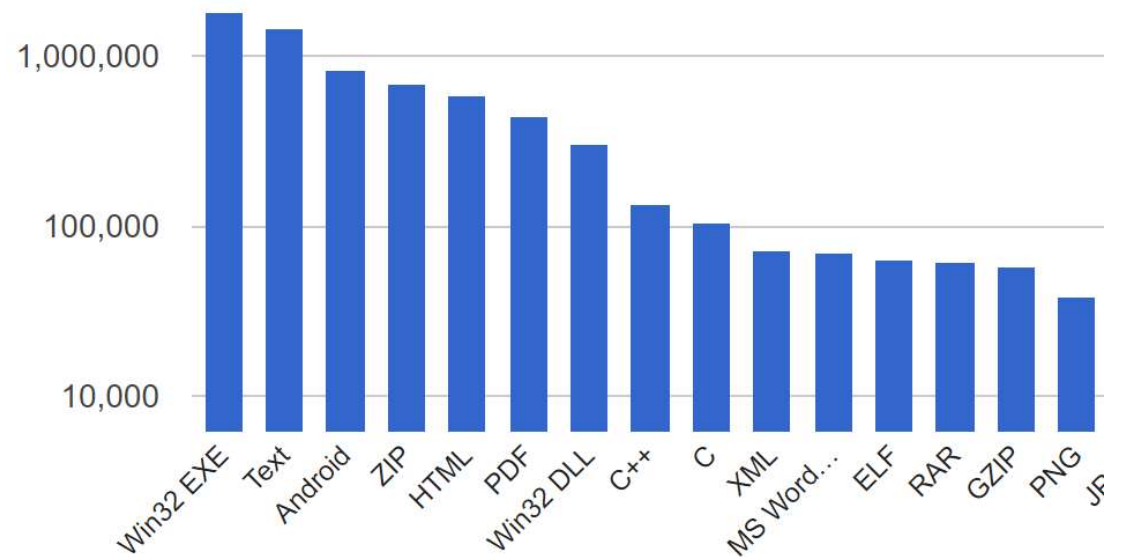
■문서형 악성코드

✓PDF, MS Office, HWP



매달 발견된 새로운 악성코드의 개수.

2018.6~2019.4 (출처: AVTEST)

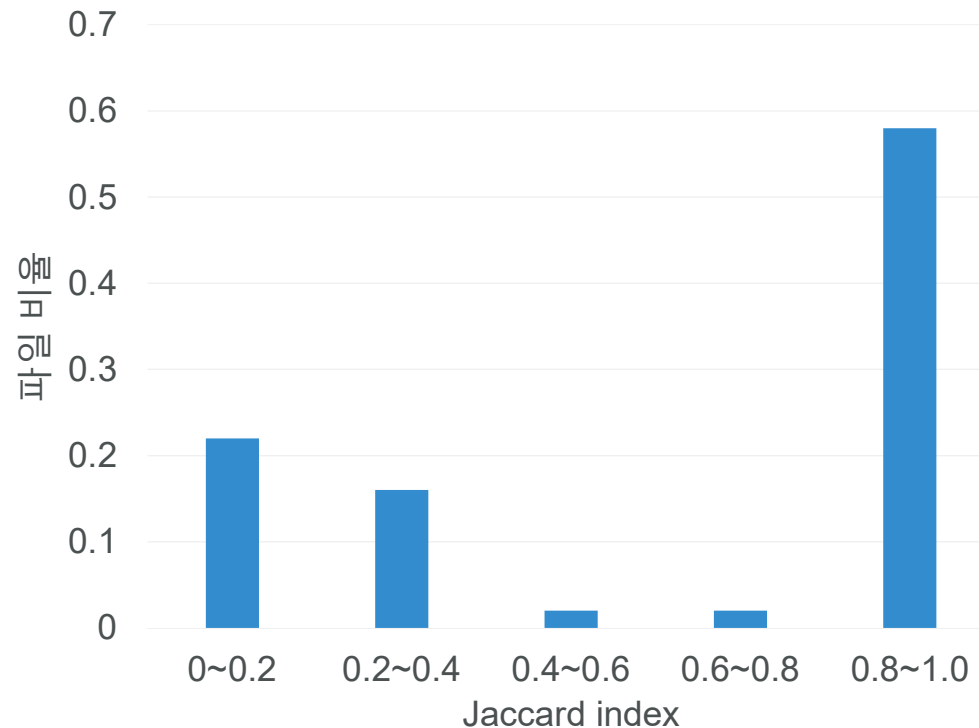


바이러스토탈 업로드 파일 수 (2019.6.3~10)

보안 빅데이터

- 악성코드

- Elasticsearch 인덱싱 & 검색
- 인덱싱: 악성코드 하루 분량 (약 5만개)
- 검색: 다음날 수집된 악성코드
 - ✓ Byte-stream 4-gram



보안 빅데이터

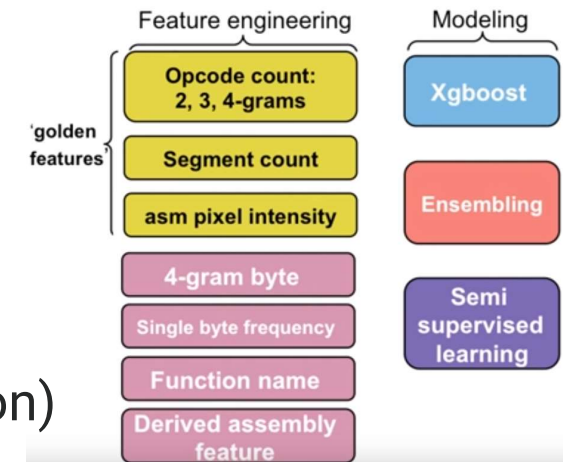
- 악성코드 수집
 - Virusshare (무료)
 - Virussign (유료)
 - Virustotal (academic request)
 - 데이터분석 경진대회 데이터 셋
- 악성코드 분석 서비스
 - Virustotal
 - malwares.com (국내)
 - Hybrid analysis
 - Intezer

보안 빅데이터

•악성코드 분석 경진 대회

■MS2015

- ✓캐글 대회
- ✓악성코드 20,000개
- ✓Disassembled code (by IDA Pro)
- ✓Bytecode (non-executable)
- ✓다중 분류 (9개 라벨, multi-class classification)
- ✓보안 분야 비전문가 우승 (정확도 99.83%)



<http://blog.kaggle.com/2015/05/26/microsoft-malware-winners-interview-1st-place-no-to-overfitting/>

■MS2019

- ✓캐글 대회
- ✓PC 정보 82개 피쳐 → 악성코드 감염 여부 예측

Microsoft Malware Prediction, 2019, <https://www.kaggle.com/c/microsoft-malware-prediction>

보안 빅데이터

•악성코드 분석 경진 대회

■정보보호 R&D 데이터 챌린지 (2017, 2018)

- ✓Windows executable files (PE 32bits)
- ✓악성코드+정상파일 40,000개 (dataset 1,2,3,4), 악성:정상=7:3
- ✓이진 분류 (binary classification)
- ✓일반인 우승: XGBoost + 다양한 피쳐 (정확도 96.83%)
- ✓대학팀 우승: 딥러닝 + 앙상블 (정확도 96.10%)

정보보호 R&D 데이터챌린지 2018_본선대회

2018-12-01 / 15:30:53

악성코드 탐지_일반			악성코드 탐지_대학(원)			안드로이드 악성앱 탐지			차량주행 데이터기반 도난 탐지		
순위	팀명	탐지율	순위	팀명	탐지율	순위	팀명	탐지율	순위	팀명	분류정확도
1		96.83%	1		96.1%	1		97.53%	1		49.98%
2		94.48%	2		95.84%	2		95.85%	2		42.45%
3		93.41%	3		95.46%	3		94.33%	3		36.19%
4		91.51%	4		94.67%				4		35.58%

보안 빅데이터

• 악성코드 피처 추출

■ PE (Portable Executable)

✓ 헤더 정보

• 크기, 섹션, 임포트,...

✓ 실행 코드 (어셈블리 코드)

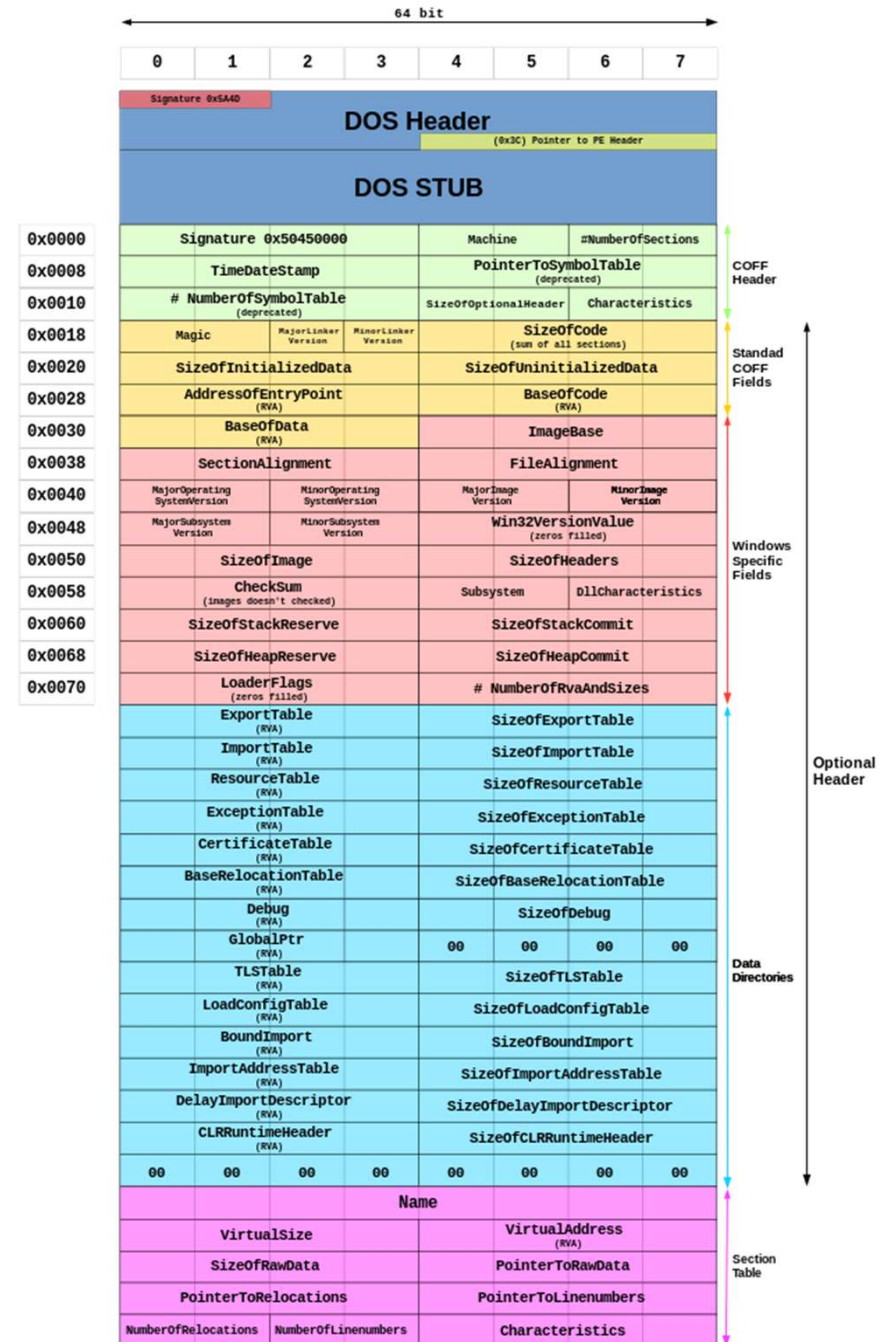
• Mnemonics, opcode,...

✓ strings

✓ 실행 정보

• API 호출 시퀀스, 생성 파일, 레지스트리,...

✓ 정적 정보 vs 동적 정보



보안 빅데이터

• 악성코드 피처 추출

■ 문서형 악성코드

- ✓ PDF, Msoffice, hwp,...
- ✓ 구조 정보, 태그 정보, 실행 정보,...

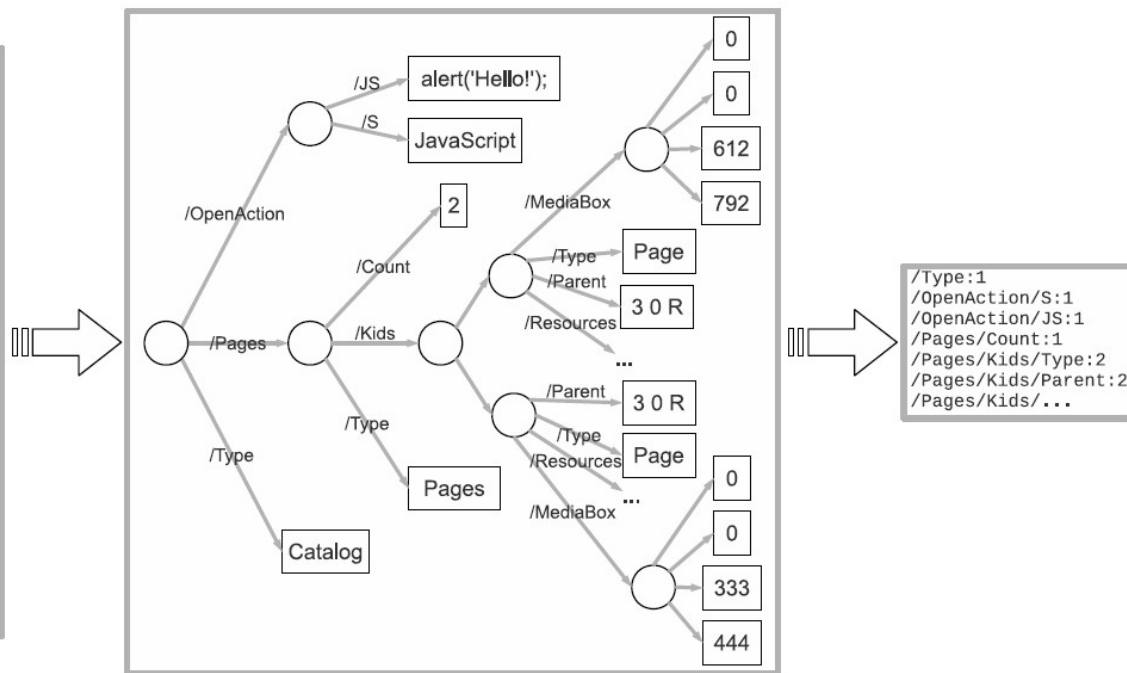
```

1 0 obj <<
  /Type /Catalog
  /OpenAction <<
    /S /JavaScript
    /JS (alert('Hello!'))
  >>
  /Pages 3 0 R
>> endobj

3 0 obj <<
  /Type /Pages
  /Kids [ 22 0 R 23 0 R ]
  /Count 2
>> endobj

22 0 obj <<
  /Type /Page
  /Parent 3 0 R
  /MediaBox [ 0 0 612 792 ]
  /Resources ...
>> endobj

23 0 obj <<
  /Type /Page
  /Parent 3 0 R
  /MediaBox [ 0 0 333 444 ]
  /Resources ...
>> endobj
  
```



N. Srndic and P. Laskov, "Detection of Malicious PDF Files Based on Hierarchical Document Structure," *NDSS 2013*

보안 빅데이터

•악성 코드 피처 가공

■ Disassembled code → 시퀀스 → 테이블 (term frequency)

```

push    ebp
mov     ebp, esp
push    ecx
push    ebx
push    esi
push    edi
xor     eax, eax
mov     [ebp+var_4], eax
mov     ebx, 1
mov     esi, 2
mov     edi, 3
push    esi
push    ebx
call    sub_401150
add     esp, 8
push    edi
push    esi
push    ebx
call    sub_401164
add     esp, 0Ch
push    [ebp+var_4]
push    edi
push    esi
push    ebx
call    sub_401189
add     esp, 10h
push    edi
push    esi
push    ebx
call    sub_401185
add     esp, 0Ch
  
```



```

push
mov
push
push
push
push
xor
mov
mov
mov
mov
push
push
push
call
add
push
push
push
call
add
push
push
push
call
add
push
push
push
call
add
  
```



	word	freq.
1	push	3403
2	mov	2134
3	add	1234
4	call	856
...		
n	xor	13

보안 빅데이터

•악성 코드 피처 가공

■ Disassembled code → 시퀀스 → 테이블 (feature hashing)

```

push    ebp
mov     ebp, esp
push    ecx
push    ebx
push    esi
push    edi
xor     eax, eax
mov     [ebp+var_4], eax
mov     ebx, 1
mov     esi, 2
mov     edi, 3
push    esi
push    ebx
call    sub_401150
add     esp, 8
push    edi
push    esi
push    ebx
call    sub_401164
add     esp, 0Ch
push    [ebp+var_4]
push    edi
push    esi
push    ebx
call    sub_401189
add     esp, 10h
push    edi
push    esi
push    ebx
call    sub_401185
add     esp, 0Ch
  
```



```

push
mov
push
push
push
push
xor
mov
mov
mov
mov
push
push
push
call
add
push
push
push
call
add
push
push
push
push
call
add
push
push
push
push
call
add
  
```



H("pushmovpush")



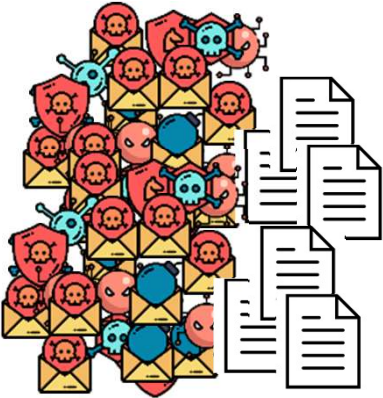
343
5626
34+1
0
856
...
344

G("pushmovpush")

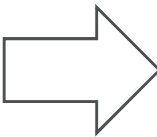
3-gram frequency table

보안 빅데이터

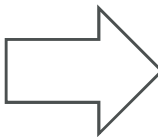
•보안 빅데이터 → 고정 크기 입력 벡터 생성 → 학습



[보안 데이터]



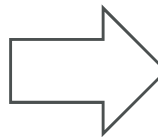
[정적/동적분석]



정보 손실 발생!

343
5626
34
0
856
...
344

[고정 크기 벡터]



[학습 모델]

E2E(end-to-end) deep learning

보안 빅데이터

•분류가 되는 이유?

■유사 파일 증가

■악성코드

✓변종 악성코드 $A \rightarrow A'$

- 정적 분석 피쳐 유사
- 난독화(예: packer) 적용 시 동적 분석 피쳐 유사
 - 쿠크샌드박스 리포트 비교만으로 탐지 가능

✓X. UGARTE-PEDRERO, M. GRAZIANO, "A Close Look at a Daily Dataset of Malware Samples," ACM Transactions on Privacy and Security, Vol. 22, No. 1, Article 6, January, 2019

■정상코드

✓Localized performance optimization (예: .NET framework)

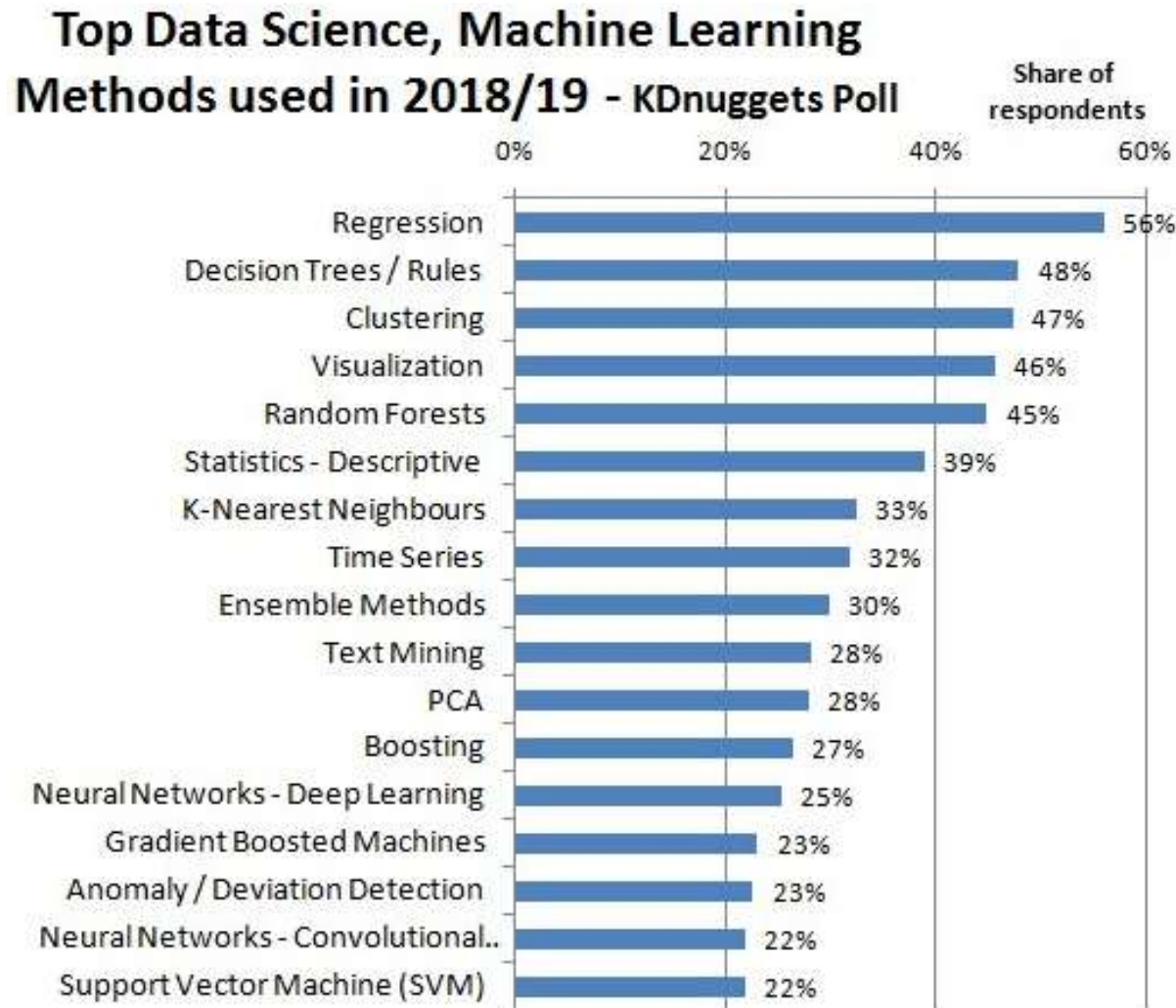
✓Product serial numbers, license keys

✓B. Li, K. Roundy, C. Gates, Y. Vorobeychik, "Large-Scale Identification of Malicious Singleton Files," CODASPY'17

■유사도 분석 연구 활용

✓Fuzzy hash, ssdeep, TLSH, imphash,...

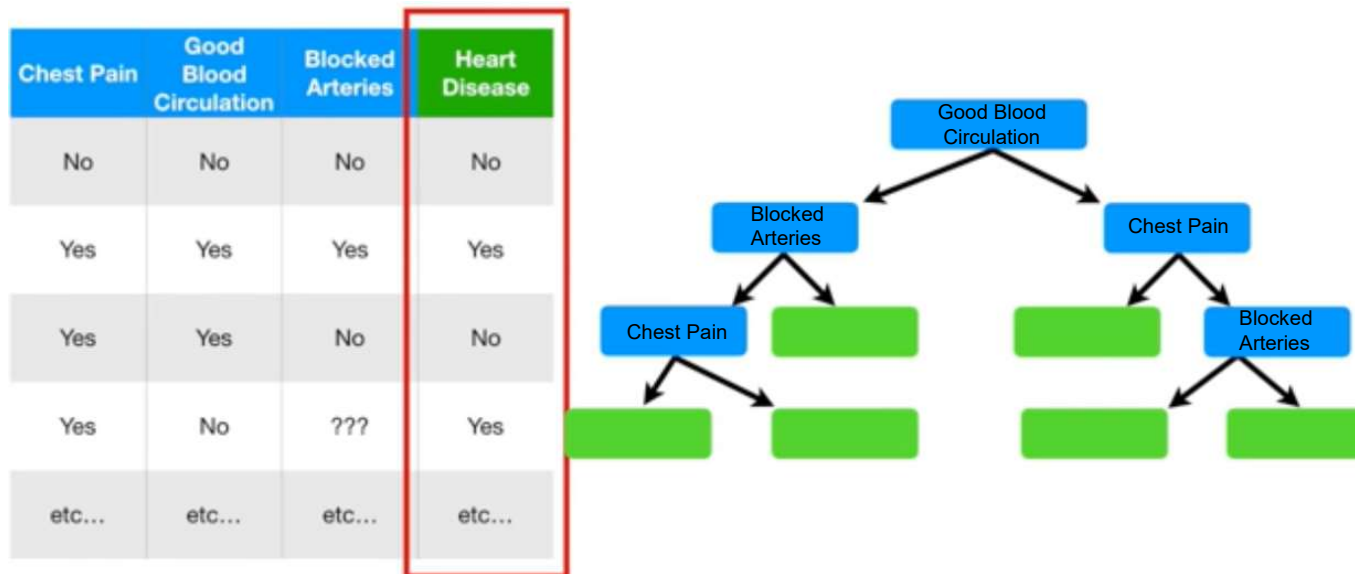
학습 모델



<https://www.kdnuggets.com/2019/04/top-data-science-machine-learning-methods-2018-2019.html>

학습 모델

- 분류 (classification)
 - Decision tree



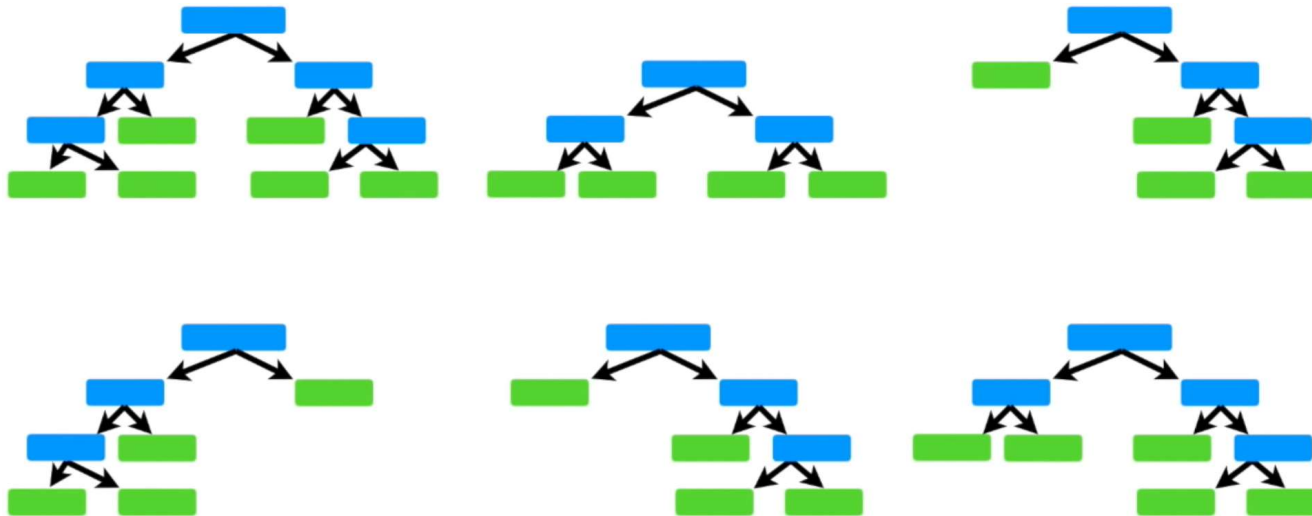
StatQuest: Decision Trees, <https://www.youtube.com/watch?v=7VeUPuFGJHk>

low bias, high variance

학습 모델

- 분류 (classification)
 - Random Forest → bootstrap aggregating (bagging)

Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No
Yes	Yes	Yes
Yes	No	No



StatQuest: Random Forests Part 1 - Building, Using and Evaluating https://www.youtube.com/watch?v=J4Wdy0Wc_xQ

Sampling with replacement → low variance

학습 모델

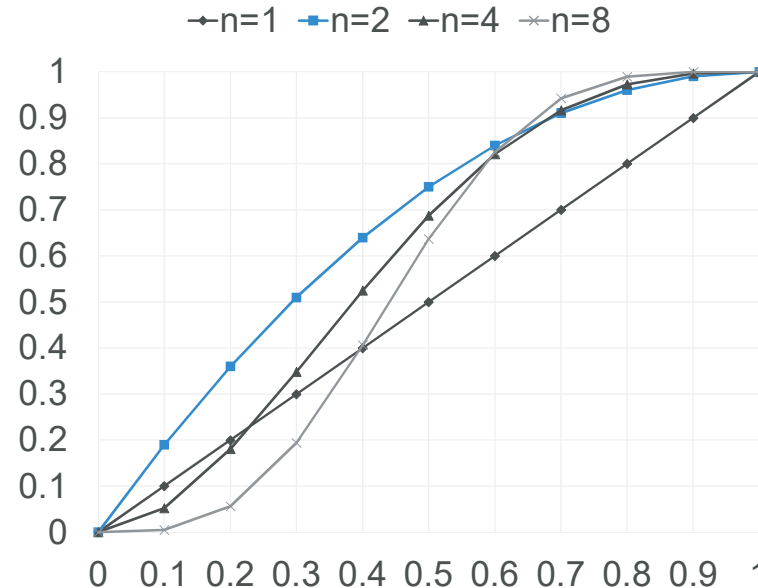
•분류 (classification)

■ **Random Forest** → 일종의 **ensemble**

■ p =단일 트리(모델) 맞춤 확률

■ 단순 투표 방식, 앙상블 모델 맞춤 확률 = $\sum_{r=n/2}^n \binom{n}{r} p^r (1-p)^{n-r}$

앙상블 정확도

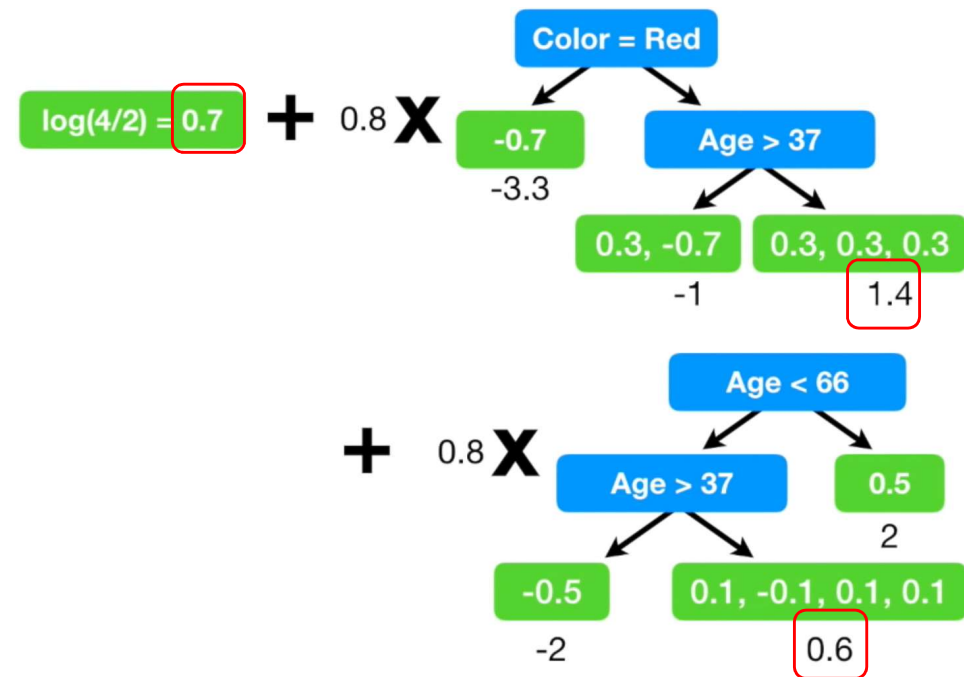


학습 모델

- 분류 (classification)
 - Gradient Boost, AdaBoost, XGBoost, LightGBM
 - Example: 3 trees

Likes Popcorn	Age	Favorite Color	Loves Troll 2
Yes	12	Blue	Yes
Yes	87	Green	Yes
No	44	Blue	No
Yes	19	Red	No
No	32	Green	Yes
No	14	Blue	Yes

Likes Popcorn	Age	Favorite Color	Loves Troll 2
Yes	25	Green	???



$$0.7 + (0.8 \times 1.4) + (0.8 \times 0.6) = 2.3, \frac{e^{2.3}}{1 + e^{2.3}} = 0.9$$

Gradient Boost Part 1: Regression Main Ideas <https://www.youtube.com/watch?v=3CC4N4z3GJc>

학습 모델

- Gradient Boosting vs Random Forest

- “*With excellent performance on all eight metrics, **calibrated boosted trees were the best learning algorithm overall. Random forests are close second.***”

- ✓ R. Caruana, A. Niculescu-Mizil, “An Empirical Comparison of Supervised Learning Algorithms Using Different Performance Metrics,” ICML'06

- *boosted decision trees perform exceptionally well when dimensionality is low. In this study **boosted trees are the method of choice for up to about 4000 dimensions. Above that, random forests have the best overall performance.***

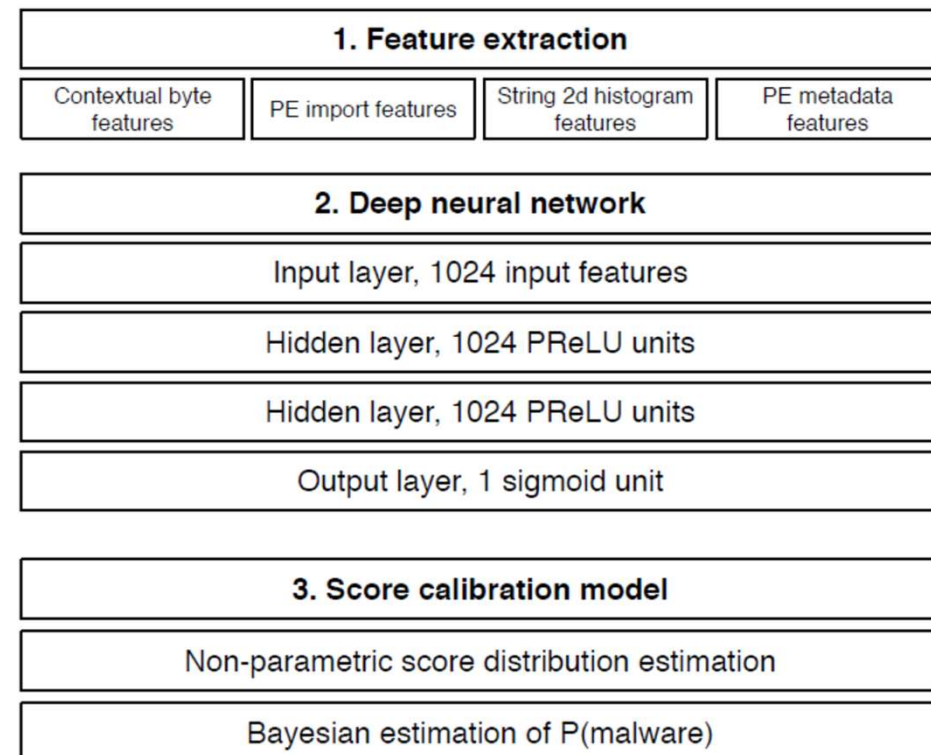
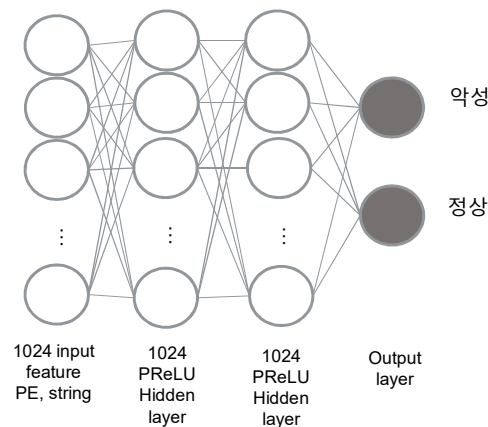
- ✓ R. Caruana, N. Karampatziakis, A. Yessenalina, “An empirical evaluation of supervised learning in high dimensions,” ICML'08

- <http://fastml.com/what-is-better-gradient-boosted-trees-or-random-forest/>

학습 모델

• 딥러닝 모델

- Deep Neural Network Based Malware Detection Using Two Dimensional Binary Program Features, IEEE Malware'15
- Deep-learning, ANN, byte/entropy histogram, PE Import features
- 약 400,000 malware samples
- 4-fold cross-validation

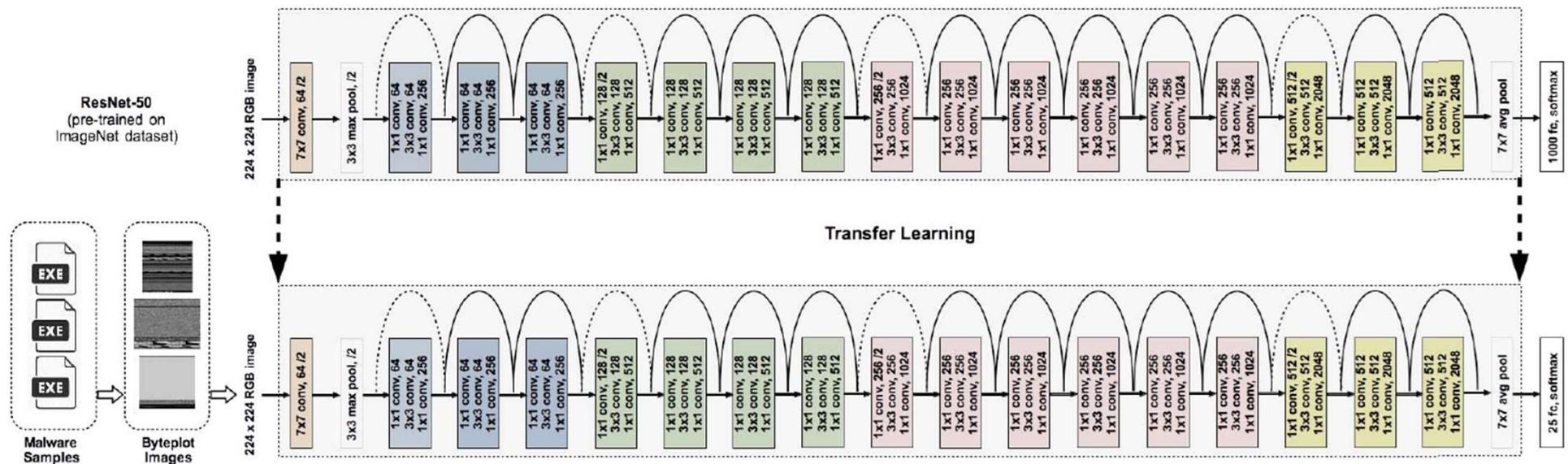


학습 모델

• 딥러닝 모델

- E. Rezende, "Malicious Software Classification using Transfer Learning of ResNet-50 Deep Neural Network", 2017 16th IEEE ICMLA, 2017

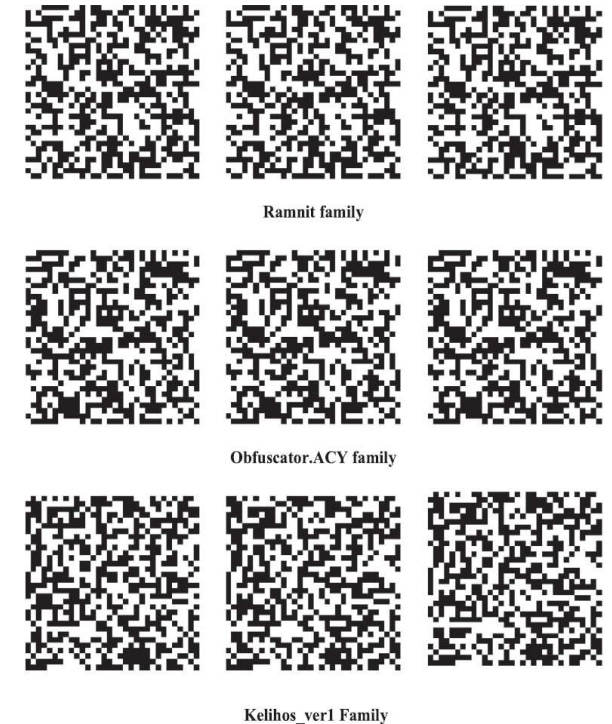
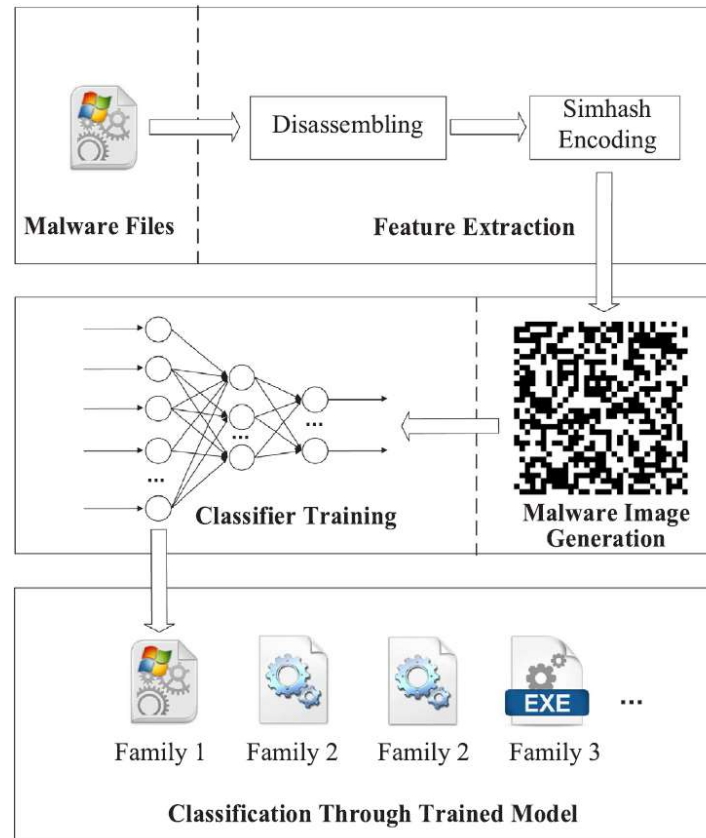
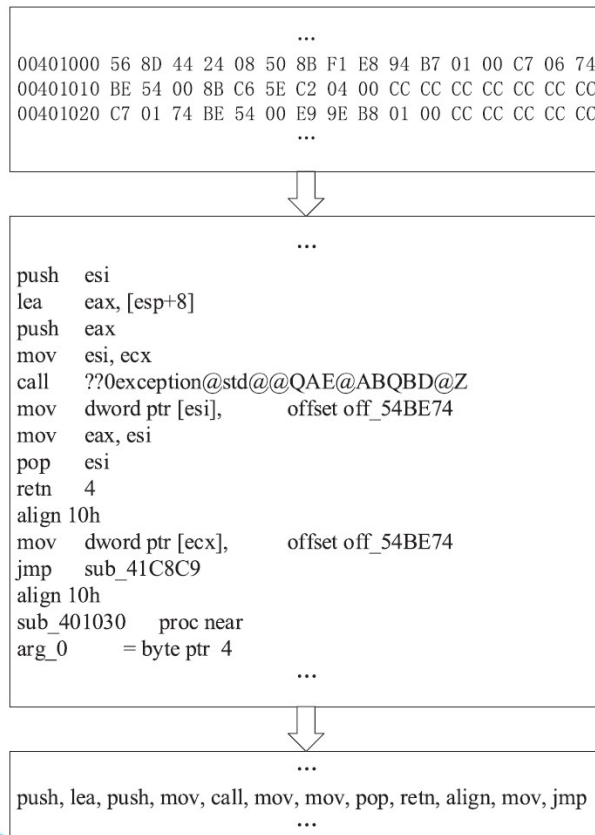
✓ 25개 패밀리 분류, 10-cv 평균 정확도: 0.9862



학습 모델

• 딥러닝 + simHash

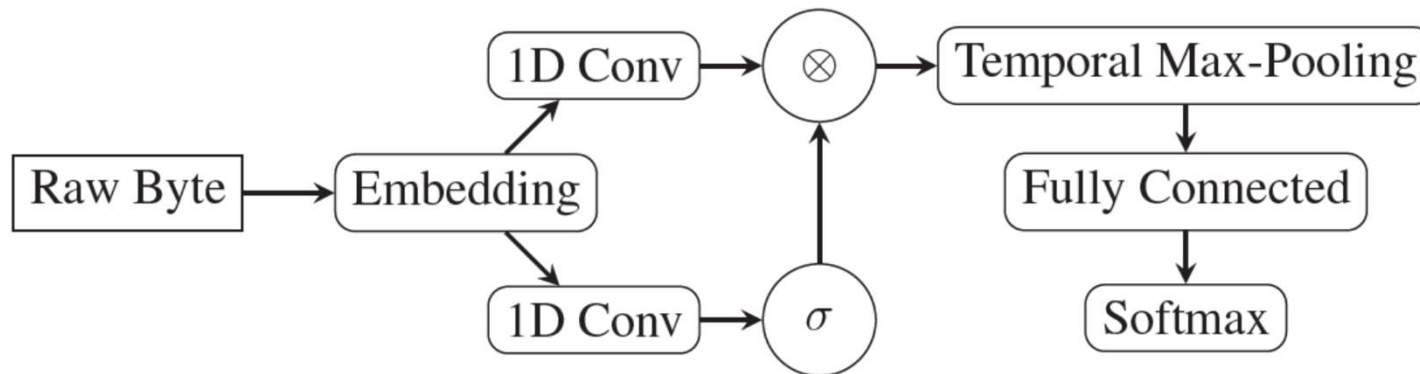
- Malware identification using visualization images and deep learning, Elsevier Computers and Security, May 2018
- 악성코드 이미지 변환 + simHash + CNN



학습 모델

• 딥러닝 모델

- E. Raff, J. Barker, J. Sylvester, R. Brandon, B. Catanzaro, C. Nicholas, "Malware Detection by Eating a Whole EXE", AAAI Workshop on AI for Cyber Security, 2018
 - ✓ 2MB from a file
 - ✓ 8 GPUs of a DGX-1 in 16.75 hours per epoch, for 10 epochs, and using all available GPU memory. Training on the larger 2 million set took one month.
- E2E 딥러닝 주장



학습 모델

- Deep learning > Other machine learning algo.

- image classification
- natural language processing
- speech recognition

<https://www.kdnuggets.com/2016/04/deep-learning-vs-svm-random-forest.html>

- Other machine learning algo. > Deep learning

- Tabular
- Tuning deep learning models are challenging!
 - ✓ Fill missing values
 - ✓ Convert categorical data into numerical data
 - ✓ Scale features
 - ✓ Architecture, layers, activation functions, learning rate, optimizers, batch size, ...



https://www.huffpost.com/entry/understanding-trump-the-hammer-the-nail-and-his_b_5a473d7fe4b06cd2bd03dff1

- 보안데이터 분석 + 딥러닝 적용, 현재 완성도 낮으나 잠재력 큼!

- **Tabular** 표현 가능한 피처를 딥러닝 적용?
- 가변 **Operand** 실행 코드를 딥러닝 적용?

악성코드 분석 사례

- KISA2018 dataset (1, 2번 학습 → 3번 테스트)
 - Not cross-validation

- 피쳐 추출 + XGBoost

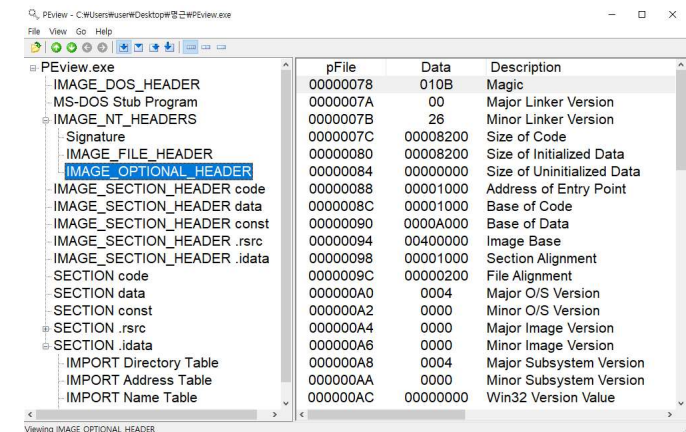
- PE HEADER (7개 피쳐) → 90.0%
- OPTAIONAL HEADER(30개 피쳐) → 94.1%

```
import pefile
with open('.\PEview.exe', 'rb') as f:
    file = f.read()
pe = pefile.PE(data = file)
print(pe.OPTIONAL_HEADER)
```

<https://github.com/erocarrera/pefile/blob/master/pefile.py>

- 피쳐 추출 + Random Forest

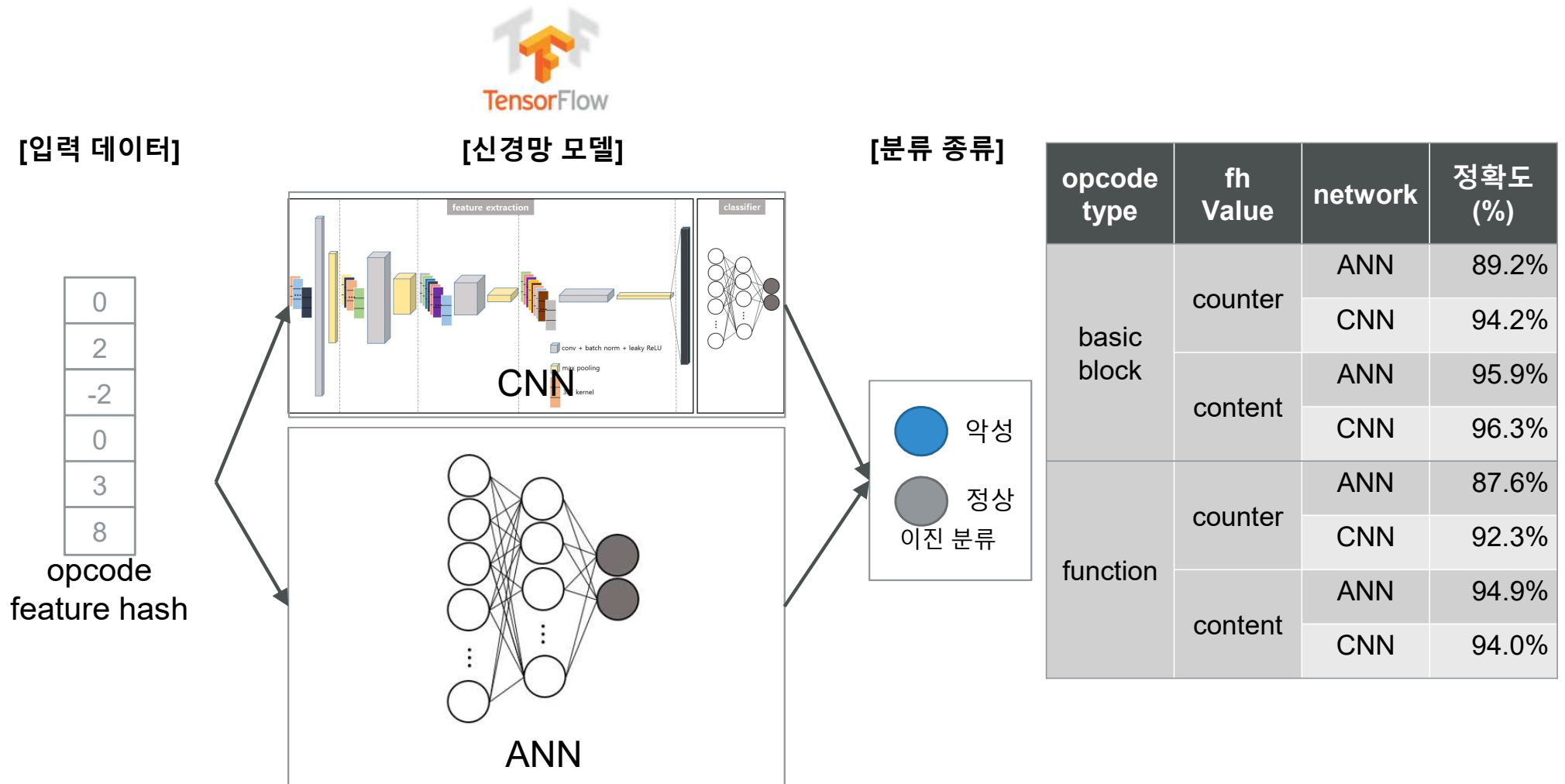
- PE HEADER → 88.9%
- OPTAIONAL HEADER → 94.0%



pFile	Data	Description
00000078	010B	Magic
0000007A	00	Major Linker Version
0000007B	26	Minor Linker Version
0000007C	00008200	Size of Code
00000080	00008200	Size of Initialized Data
00000084	00000000	Size of Uninitialized Data
00000088	00001000	Address of Entry Point
0000008C	00001000	Base of Code
00000090	0000A000	Base of Data
00000094	00400000	Image Base
00000098	00001000	Section Alignment
0000009C	00000200	File Alignment
000000A0	0004	Major O/S Version
000000A2	0000	Minor O/S Version
000000A4	0000	Major Image Version
000000A6	0000	Minor Image Version
000000A8	0004	Major Subsystem Version
000000AA	0000	Minor Subsystem Version
000000AC	00000000	Win32 Version Value

악성코드 분석 사례

•딥러닝 모델

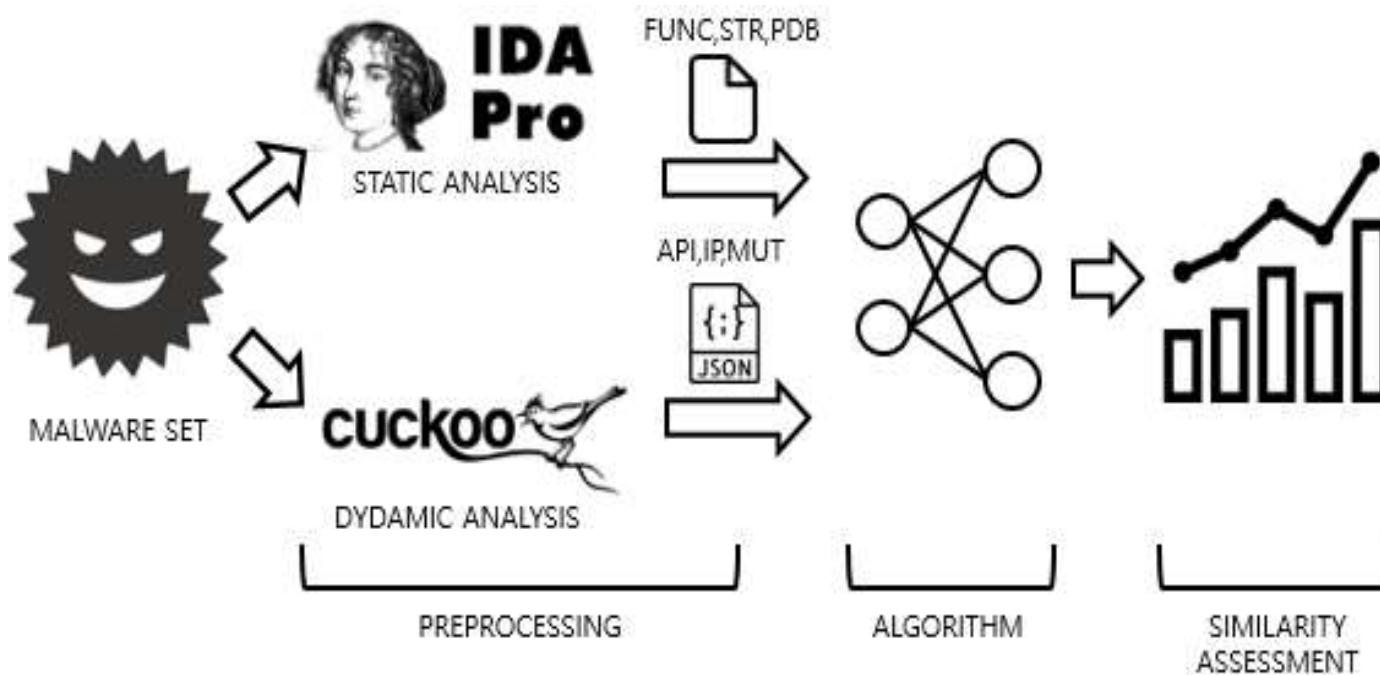


악성코드 분석 사례

• 악성코드 유사도 측정 시스템

■ 현재 6개 특징(feature) 기준 비교 가능

- ✓ 정적 특징 정보: 함수 대표값, PDB 정보, 문자열 정보
- ✓ 동적 특징 정보: API 집합, 뮤텍스(Mutex) 정보, IP주소

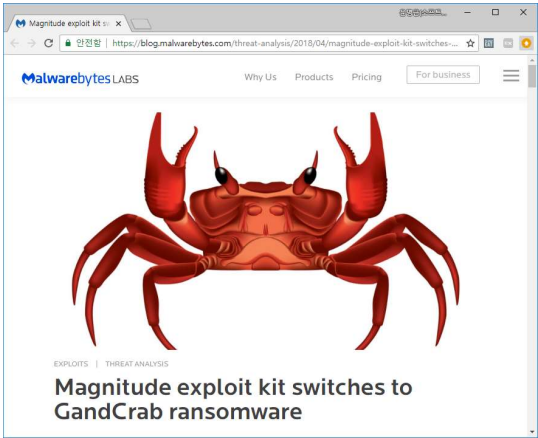


악성코드 분석 사례

•악성코드 유사도 측정 시스템

SMUF	백신 라벨		전문가 분류 그룹	
	파일1	파일2	파일1	파일2
함수 대표값, API, Pdb	none	Trojan	A	A
함수 대표값, API, Mutex	Trojan-Downloader	none	S	S
API, Mutex, IP	Trojan	none	G1	G1
함수 대표값, API, IP	none	none	M	M
API, Mutex, IP	Trojan	none	S	A
API, Mutex, IP	Backdoor	none	B2	S
API, Mutex, IP	Backdoor	none	B2	A

SMUF (finding Similar Malware Using file Factorization)

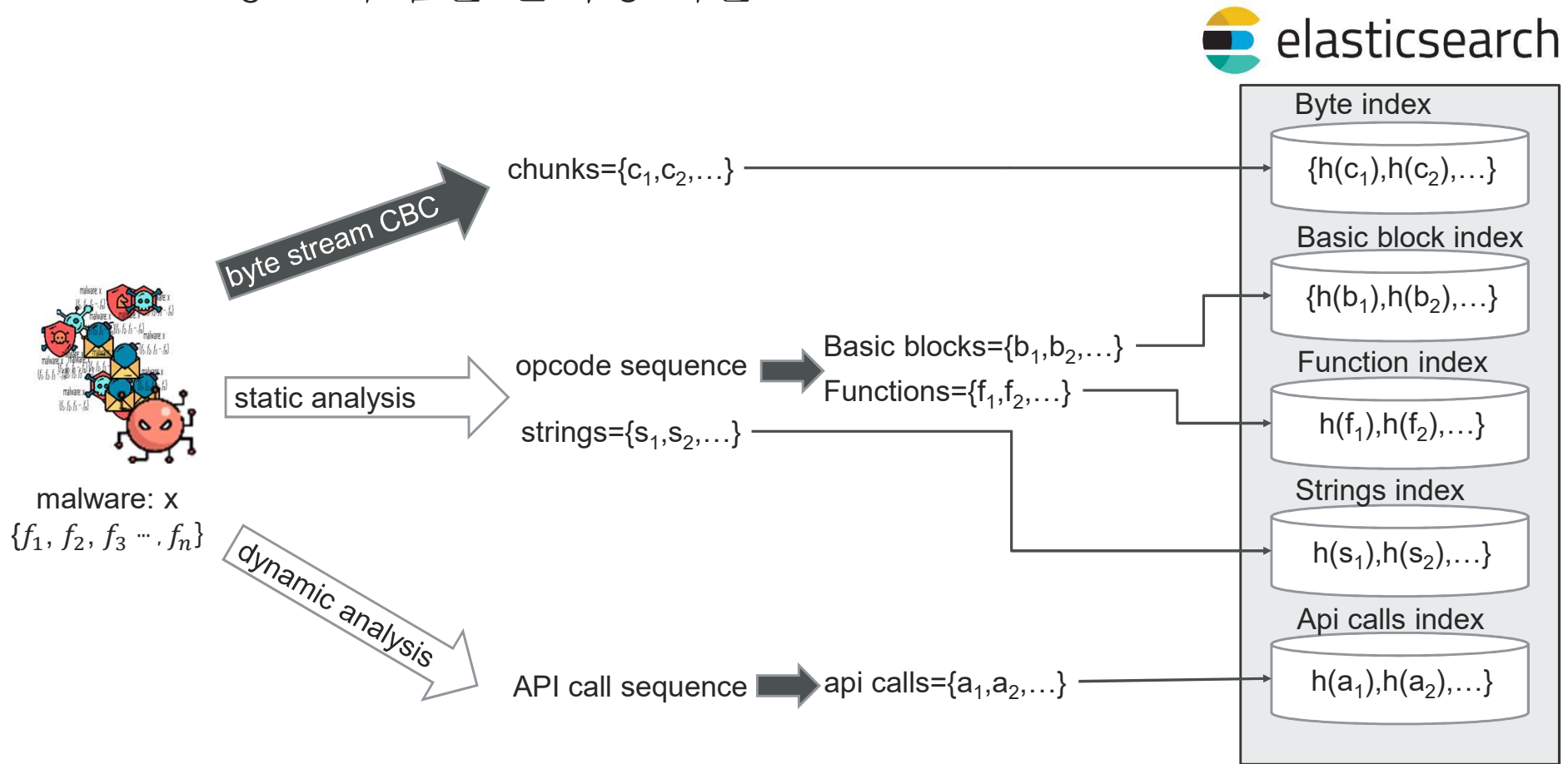


<https://blog.malwarebytes.com/threat-analysis/2018/04/magnitude-exploit-kit-switches-gandcrab-ransomware/>

악성코드 분석 사례

•Elasticsearch

▪고정 크기 샘플 인덱싱 기술



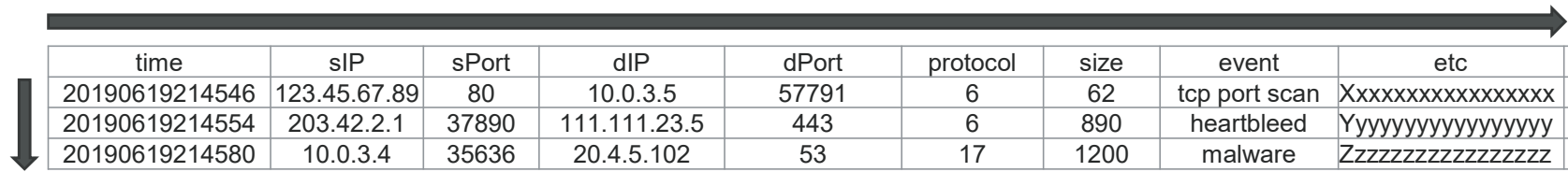
[illegible]

KMU KOOKMIN UNIVERSITY

보안관제 데이터 분석 사례

•보안관제

■intrusion detection/prevention system logs



time	sIP	sPort	dIP	dPort	protocol	size	event	etc	label
20190619214546	123.45.67.89	80	10.0.3.5	57791	6	62	tcp port scan	Xxxxxxxxxxxxxxxxxx	False
20190619214554	203.42.2.1	37890	111.111.23.5	443	6	890	heartbleed	Yyyyyyyyyyyyyyyy	Flase
20190619214580	10.0.3.4	35636	20.4.5.102	53	17	1200	malware	Zzzzzzzzzzzzzzzz	True

■ $e_i := ['time', 'sIP', 'sPort', 'dIP', 'dPort', 'protocol', 'size', 'event', 'etc']$

■ 연속된 이벤트 $:= e_0 e_1 e_2 e_3 \dots e_{i-1} e_i e_{i+1} \dots e_{n-1}$

✓ false positives vs true positives

✓ $label(e_i) = \text{true}$, or false

■ Why e_i is true?


✓ $['time', 'sIP', 'dPort', 'size', 'event'] \rightarrow$ intra-log analysis

✓ $e_{i-3} e_{i-2} e_{i-1} \rightarrow$ inter-log analysis

보안관제 데이터 분석 사례

•보안관제

- intrusion detection/prevention system logs
- Intra-log analysis



time	sIP	sPort	dIP	dPort	protocol	size	event	etc	label
20190619214546	123.45.67.89	80	10.0.3.5	57791	6	62	tcp port scan	Xxxxxxxxxxxxxxxxxx	False
20190619214554	203.42.2.1	37890	111.111.23.5	443	6	890	heartbleed	Yyyyyyyyyyyyyyyy	Flase
20190619214580	10.0.3.4	35636	20.4.5.102	53	17	1200	malware	Zzzzzzzzzzzzzzzz	True

- Decision tree, RF, XGBoosting
- 'etc' → 피쳐 추출 및 가공 (악성코드 문제와 유사해짐)
 - ✓ TF (Term Frequency), feature hashing, embedding...
- 높은 정확도 가능

보안관제 데이터 분석 사례

•보안관제

- intrusion detection/prevention system logs
- Inter-log analysis

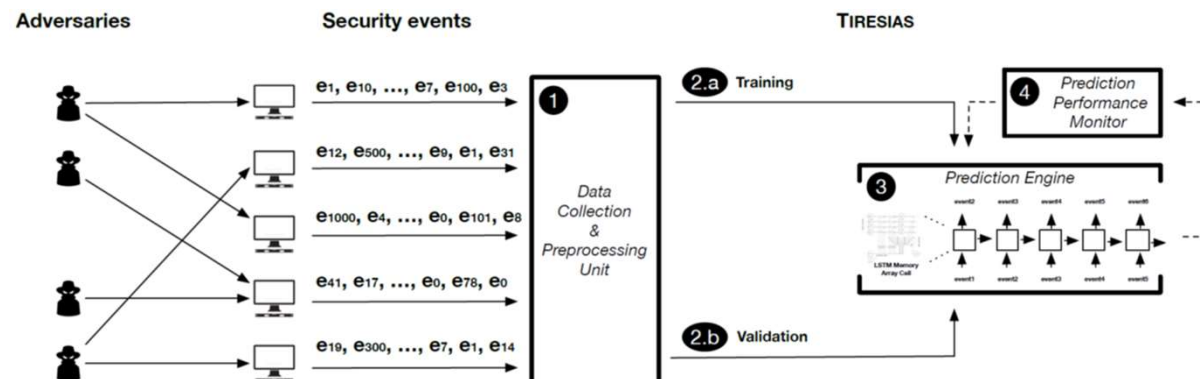
time	sIP	sPort	dIP	dPort	protocol	size	event	etc	label
20190619214546	123.45.67.89	80	10.0.3.5	57791	6	62	tcp port scan	Xxxxxxxxxxxxxxxxxx	False
20190619214554	203.42.2.1	37890	111.111.23.5	443	6	890	heartbleed	Yyyyyyyyyyyyyyyyyy	Flase
20190619214580	10.0.3.4	35636	20.4.5.102	53	17	1200	malware	Zzzzzzzzzzzzzzzzzz	True

- Y. Shen, etc., "Tiresias: Predicting Security Events Through Deep Learning," ACM CCS'18

✓ 시만텍 침입방지 제품 로그 예측

- 3.4 billion events (4,495 unique events)

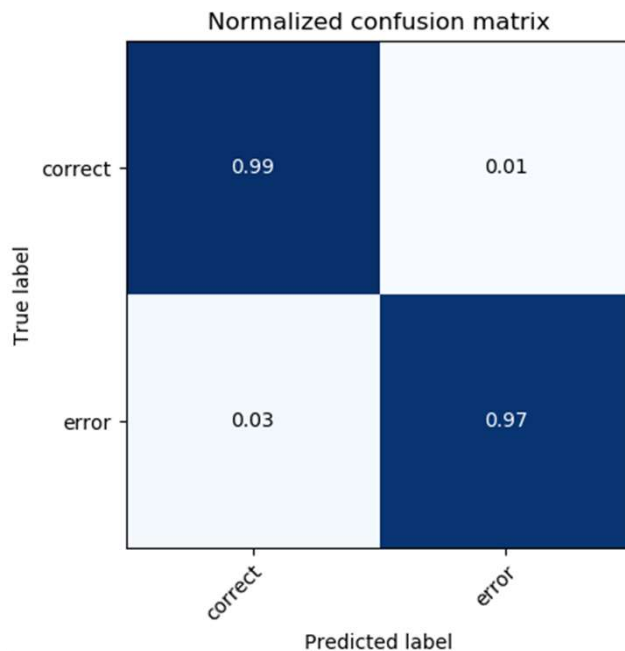
✓ Recurrent Neural Network



보안관제 데이터 분석 사례

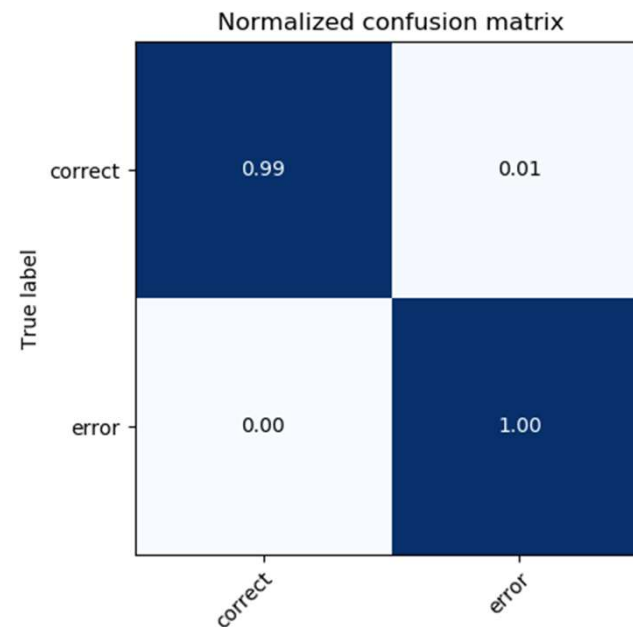
- 보안관제 데이터 분석 사례
 - 침입탐지시스템 로그 (2018년, 4백만 건)
 - ✓ 모델1: "etc" 제외 모든 피쳐 + RF
 - ✓ 모델2: "etc" tabular 작성 + 모든 피쳐 + RF

time	sIP	sPort	dIP	dPort	protocol	size	event	etc	label
20190619214546	123.45.67.89	80	10.0.3.5	57791	6	62	tcp port scan	Xxxxxxxxxxxxxxxxxx	False
20190619214554	203.42.2.1	37890	111.111.23.5	443	6	890	heartbleed	Yyyyyyyyyyyyyyyy	Flase
20190619214580	10.0.3.4	35636	20.4.5.102	53	17	1200	malware	Zzzzzzzzzzzzzzzz	True



Predicted label

[모델1]



Predicted label

[모델2]

사이버 보안 빅데이터 활용 공유 세미나

2019-06-21

Summary

- 보안을 위한 인공지능 → 머신러닝 기반 보안 빅데이터 분석
- 유사성 기반의 빅데이터 분류 문제는 기술 완성도 높음
- 보안빅데이터 자동 피쳐 추출과 딥러닝 적용은 시작 단계
- 보안 도메인 지식과 데이터 가공 및 해석 능력 중요

Q & A

- 감사합니다!
- mkyoon@kookmin.ac.kr

