



# 이상 SDS

정보보호 R&D 데이터 챌린지

악성코드 탐지 트랙

중앙대학교 산업보안학과 송원호

중앙대학교 산업보안학과 박진경

중앙대학교 산업보안학과 강현수

중앙대학교 산업보안학과 이지은

중앙대학교 산업보안학과 손정우

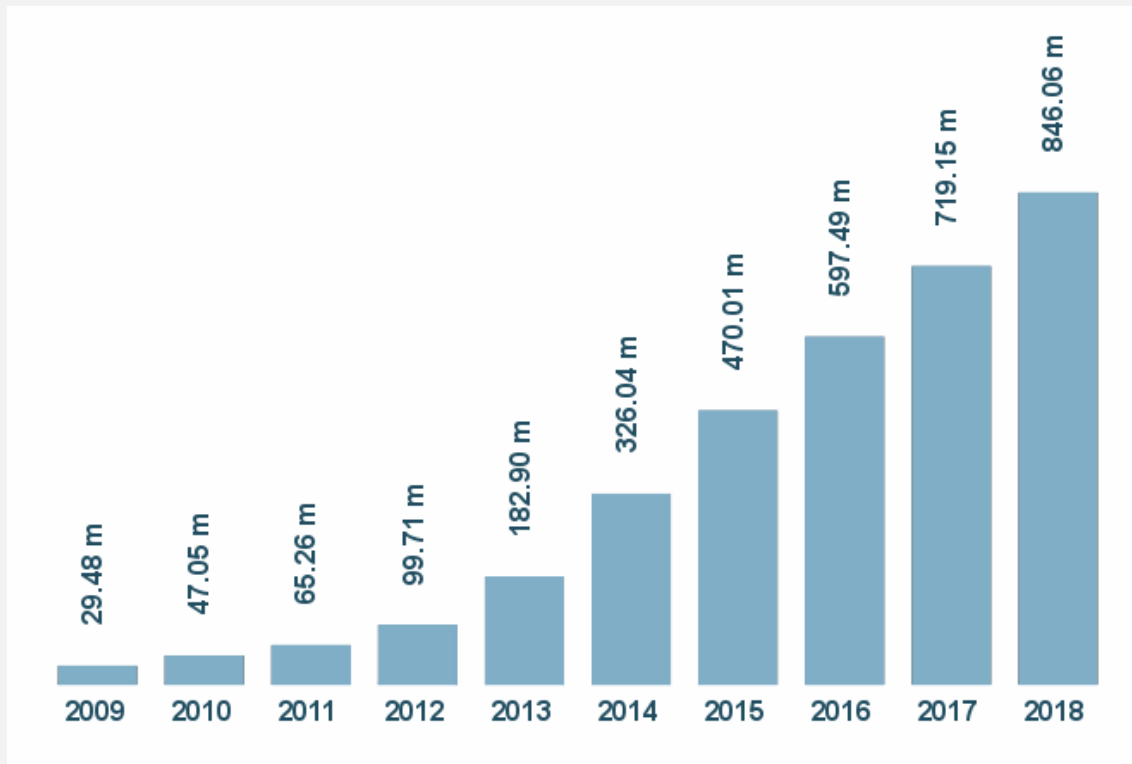


# 이상 SDS

- 개요
- 방법론
- 모델
- 평가
- 향후 활용방안

## ❖ 악성코드의 출현

- 악성코드의 수는 매년 기하급수적으로 증가: 10년 전과 비교하여 약 30배 증가
- AV-TEST, 매일 39만 개 이상의 새로운 악성 프로그램 발견



AV-TEST, Total Malware

## ❖ 도전과제

- 막대한 양의 바이너리를 분석하는 것은 물리적으로 불가능
- 난독화, 패킹, 안티 디버깅, 안티 가상화, 스케줄링 등을 활용한 악성코드 급증
- 기존에 정의된 악성코드 탐지 규칙만으로는 신종/변종 악성코드의 효과적 탐지가 어려움

👉 머신러닝을 활용한 새로운 악성코드의 탐지율을 높이는 것이 중요

👉 머신러닝을 활용한 악성코드 탐지의 자동화 필요

## ❖ 데이터 셋 분석

- 구성 : PE / Win32 / dll, exe, driver
- 비율 : 악성파일 7000개 - 정상파일 3000개

```
In [31]: pd.crosstab(index=label.is_malware, columns="cnt")
```

```
Out [31]:
```

	col_0	cnt
is_malware		
0	0	3000
1	1	7000

## ❖ Feature Extraction

- 지난 3년간 국내 PE 분석 관련 논문 전체
- 구글 드라이브를 통해 자료 공유

논문 이름	날짜	중요도	사람	정리여부	채택여부
1차					
4차 산업혁명을 대비한 딥러닝 기술의 금융보안 적용 연구	2017-12-22	상	이지은	⊖	✖
Section, DLL feature 기반 악성코드 분석 기술 연구	2017-10-01	상	송원호	○	○
딥러닝 기반의 R-CNN을 이용한 악성코드 탐지 기법	2018-06-01	상	송원호	○	○
머신러닝을 이용한 지능형 악성코드 분석기술 동향	2018-04-01	상	강현수	○	○
바이너리 시각화와 기계학습을 이용한 악성코드 분류	2018-04-01	상	박진경	○	○
신경망 기반 자연어 처리를 이용한 악성코드 분류	2018-02-01	상	송원호	○	○
심볼 기반 악성코드 정적분석 및 분류 시스템	2018-02-01	상	이지은	○	○
악성코드 탐지를 위한 바이너리 실행 파일 분석 기법	2017-02-04	상	박진경	⊖	✖
악성코드에 자주 사용되는 API 정보를 이용한 유사도 비교 방법	2016-12-09	상	박진경	○	○
연관규칙 마이닝과 나이브베이지 분류를 이용한 악성코드 탐지	2017-11-01	상	손정우	○	○
심층학습과 악성코드 분석 연구	2018-02-04	중	손정우	⊖	✖
윈도우 이벤트 로그 기반 기업 보안 감사 및 악성코드 행위 탐지 연구	2018-06-01	중	손정우	○	○
기계학습 기반의 악성코드 탐지 기법 분석	2018-08-04	하	강현수	⊖	✖
기계학습을 활용한 변종 악성코드 식별 연구 동향 분석	2017-06-04	하	강현수	⊖	✖
머신러닝을 이용한 악성코드 분류	2017-11-04	하	강현수	⊖	✖
2차					
API 콜시퀀스와 Locality Sensitive Hashing을 이용한 악성코드 클러스터링 기법에 관한 연구	2017-02-01	상	박진경	○	○
Multisize N-gram과 개선된 카이제곱을 이용한 API 호출 시퀀스 기반 악성코드 탐지에 관한 연구	2015-12-01	상	박진경	○	○
Registry 분석을 통한 악성코드 감염여부 탐지 방법 연구	2017-08-05	상	송원호	○	○
기계학습을 이용한 Operation Code Frequency 분석을 통한 악성코드 탐지에 관한 연구	2017-04-01	상	이지은	○	○
실행파일 이미지화와 Word2Vec 을 이용한 딥러닝 기반 악성코드 탐지 방법에 대한 연구	2017-05-01	상	송원호	○	○
연관규칙 탐사 기법과 SVM을 이용한 악성코드 탐지 방법	2018-02-01	상	손정우	○	○
침입탐지 시스템과 샌드박스를 이용한 악성파일 탐지 시스템 구현	2017-06-04	상	송원호	⊖	✖
필수 모듈코드 데이터를 활용한 힙 메모리 상주형 악성코드 탐지 기법 연구	2015-12-04	중	송원호	⊖	✖
지능형 악성코드 분석을 한 리얼머신 기반의 바이너리 자동실행 환경	2016-03-04	중	강현수	⊖	✖
프로세스 마이닝을 이용한 변종 악성코드 탐지 기법	2018-02-04	중	손정우	⊖	✖

50개의 논문 중 정리한 자료 리스트

## 2. Feature Extraction

### ❖ IDA Pro Python Script








#### ■ IDA Pro 사용

 analysis.py	2018-10-19 오후 1:10	JetBrains PyCharm	9KB
 main.py	2018-10-21 오후 1:10	JetBrains PyCharm	1KB
 utils_c.py	2018-10-16 오후 1:10	JetBrains PyCharm	1KB

디렉터리 구조도

- main.py 악성코드를 아이다 핸들러에 넘겨줌
- utils.\_c.py 디렉터리 생성, 파일 이름 얻기 등 전체적인 유틸리티 기능을 커스텀한 모듈
- analysis.py 아이다 프로 파이썬 스크립트

#### ■ 추출 결과

 dll.json	2018-10-22 오후 1:10	JSON 파일	1KB
 function.json	2018-10-22 오후 1:10	JSON 파일	1KB
 hash.txt	2018-10-22 오후 1:10	텍스트 문서	1KB
 opcode.json	2018-10-22 오후 1:10	JSON 파일	1KB
 pe.json	2018-10-22 오후 1:10	JSON 파일	2KB
 sample.txt	2018-10-22 오후 1:10	텍스트 문서	1KB
 section.csv	2018-10-22 오후 1:10	Microsoft Excel...	1KB

IDA Pro를 통해 추출한 Raw Data

## 2. Feature Extraction

### ❖ IDA Pro Python Script






- IDA Pro 사용

 analysis.py	2018-10-19 오후 1:10	JetBrains PyCharm	9KB
 main.py	2018-10-21 오후 1:10	JetBrains PyCharm	1KB
 utils_c.py	2018-10-16 오후 1:10	JetBrains PyCharm	1KB

디렉터리 구조도

## Feature 분석 자동화

- 추출 결과

 dll.json	2018-10-22 오후 1:10	JSON 파일	1KB
 function.json	2018-10-22 오후 1:10	JSON 파일	1KB
 hash.txt	2018-10-22 오후 1:10	텍스트 문서	1KB
 opcode.json	2018-10-22 오후 1:10	JSON 파일	1KB
 pe.json	2018-10-22 오후 1:10	JSON 파일	2KB
 sample.txt	2018-10-22 오후 1:10	텍스트 문서	1KB
 section.csv	2018-10-22 오후 1:10	Microsoft Excel...	1KB

IDA Pro를 통해 추출한 Raw Data





## 2. Feature Extraction

### ❖ Bytes

Symbol

String  
Length

Keyword

Byte  
N-gram

Byte to  
Image

### ❖ ASM

API

파일  
메타데이터

ASM to  
Image

PE 파일

Data  
Definition

섹션

Instruction

Register

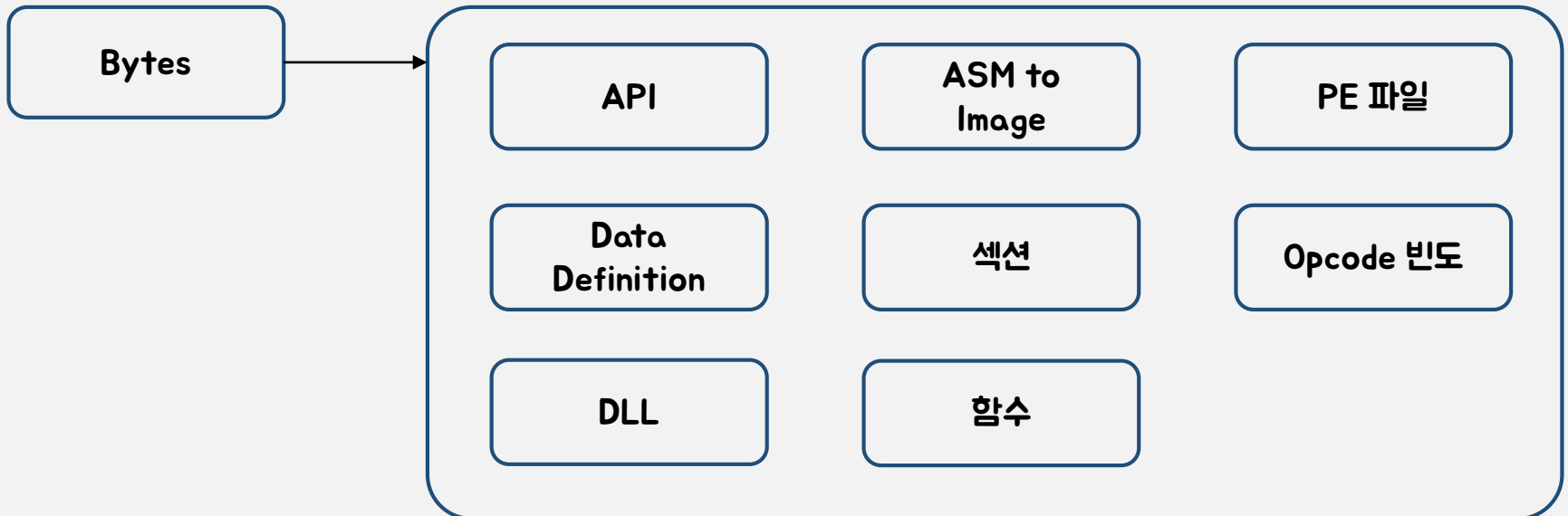
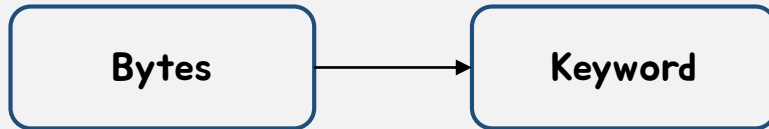
DLL

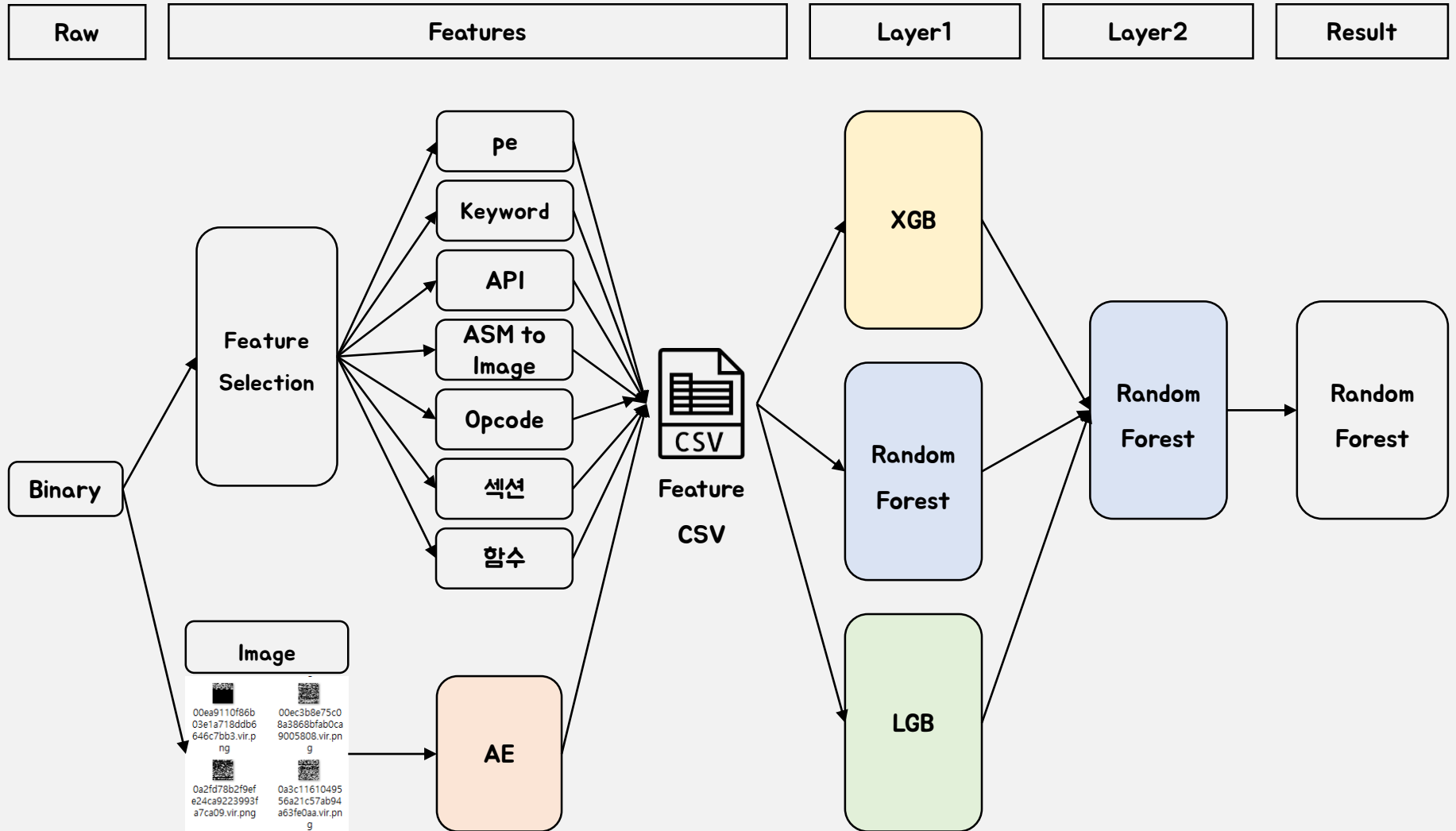
함수

Opcode  
빈도, N-gram

엔트로피

## ❖ 최종 Feature





## ❖ 평가 기준

- 5 Fold Accuracy

Column	Column+ New Feature	Accuracy
c1	PE	0.934
c2	C1 + section	0.9398
c3	C2 + register	0.9485
c4	C3 + function	0.9519
C5	C4 + opcode	0.9527
C6	C5 + AE	0.9538



## ❖ 탐지 결과

1차

95.04

2차

95.88



## ❖ 한계점

- H/W의 한계: CPU, GPU 성능 한계
- 통계적 접근의 부족

## ❖ 향후 활용

- 12월 3일까지 백신 프로토타입 제작
- 동적 분석 Feature 활용



**Thank You**

**Q & A**