블렌딩 기법을 활용한 악성코드 탐지

KAISA 문다민 김영재 손현기 오예린

2019.11.22

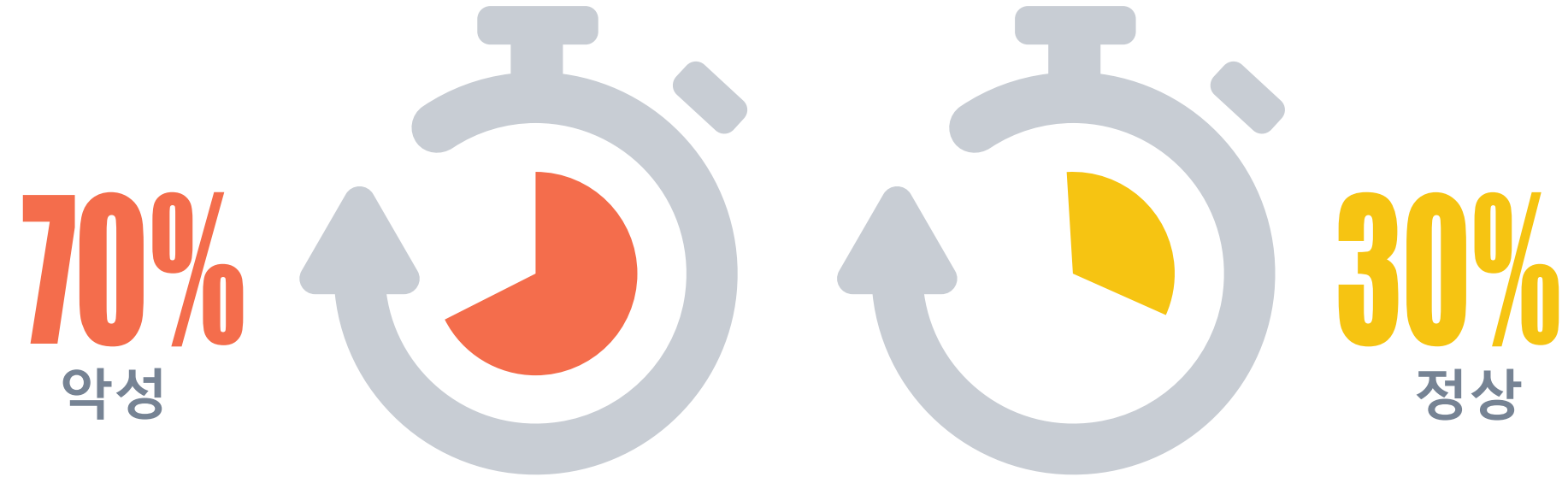
CONTENTS

/01 데이터 분석

Team KAISA

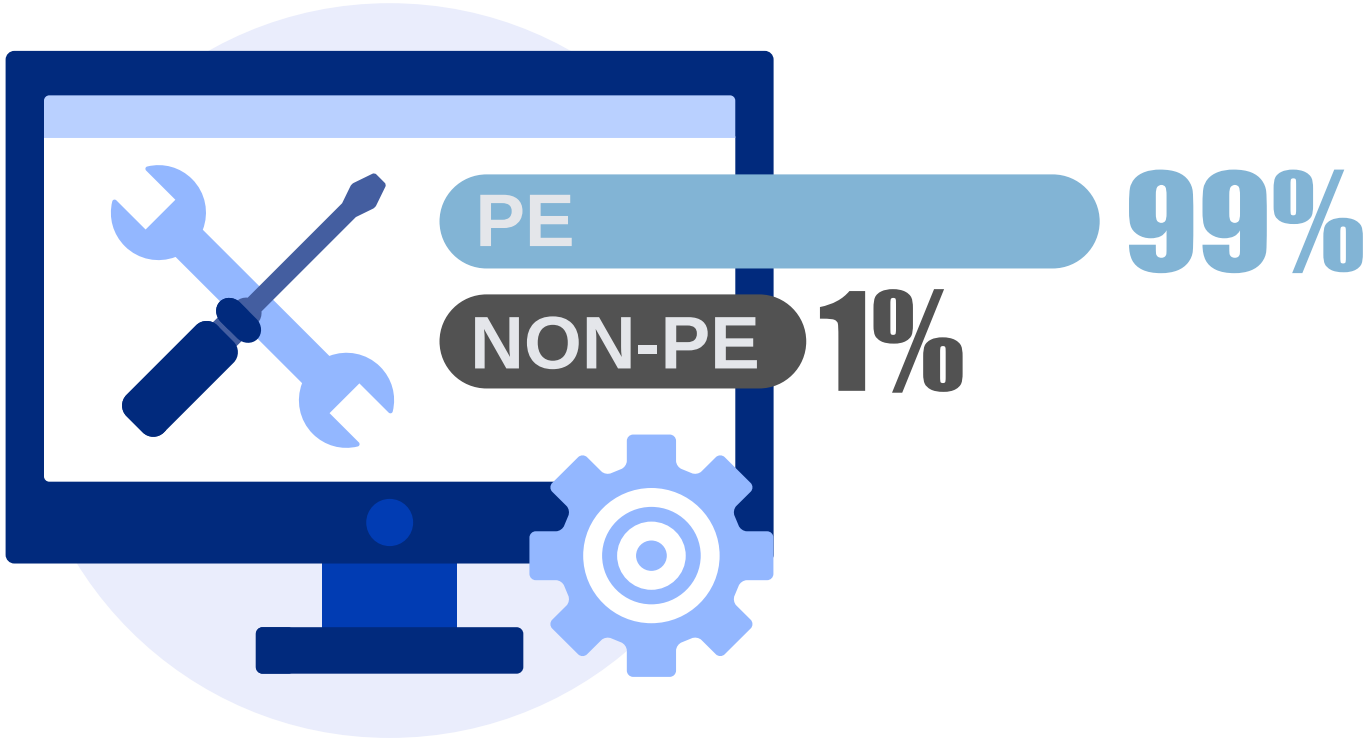
데이터 분석

학습 데이터셋



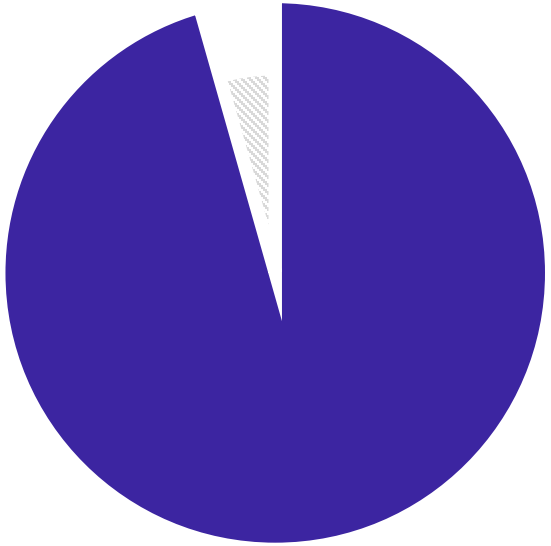
데이터 분석

학습 데이터셋 → PE

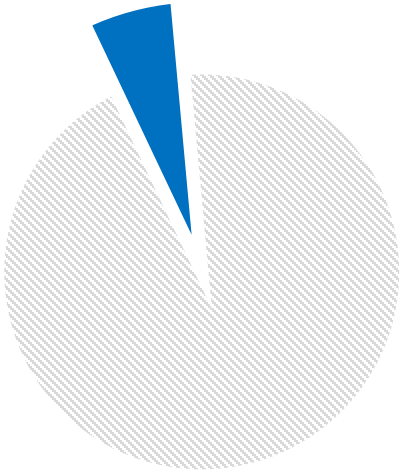


데이터 분석

학습 데이터셋 → PE



90% PE32

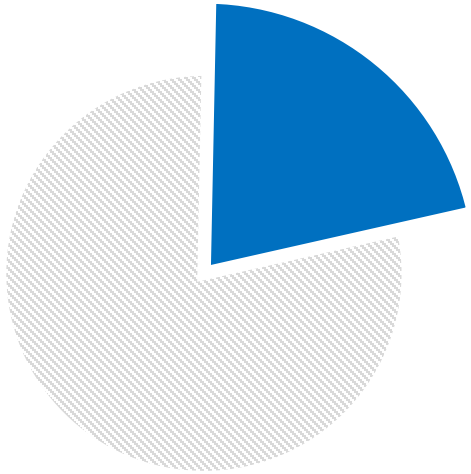


10% PE64

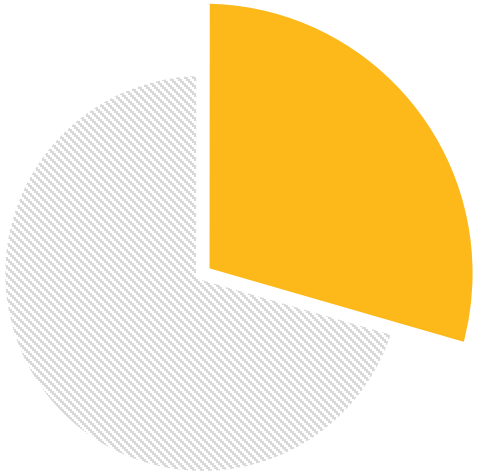
데이터 분석

학습 데이터셋→ PE → Non-PE

44% HTML



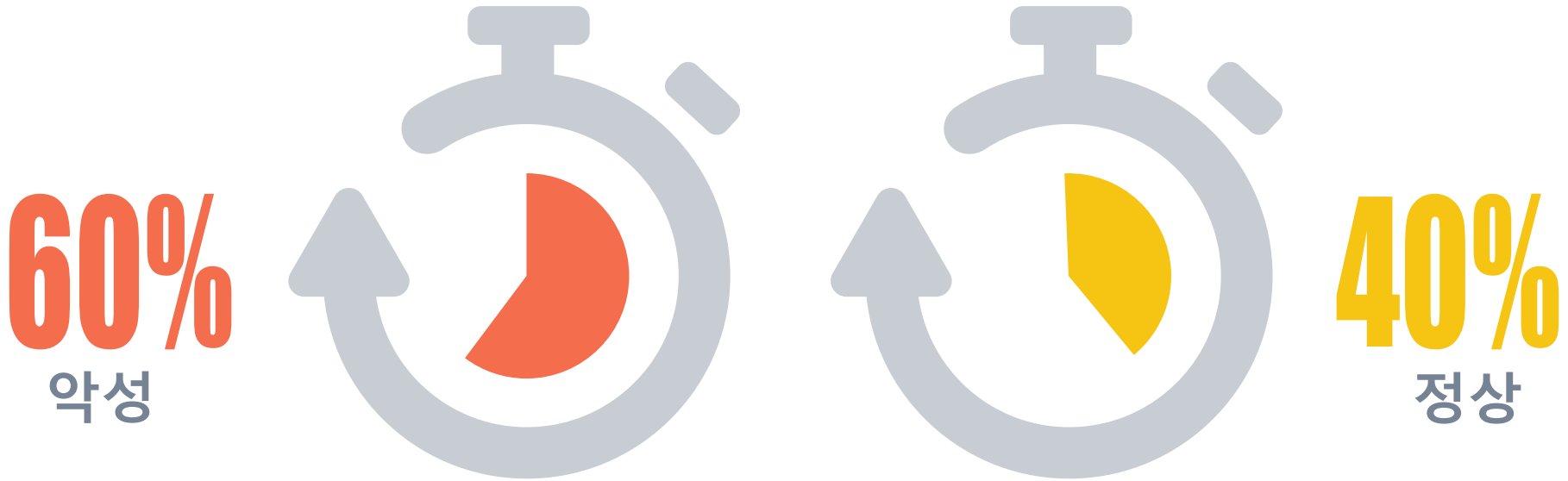
22% HWP



34% ETC

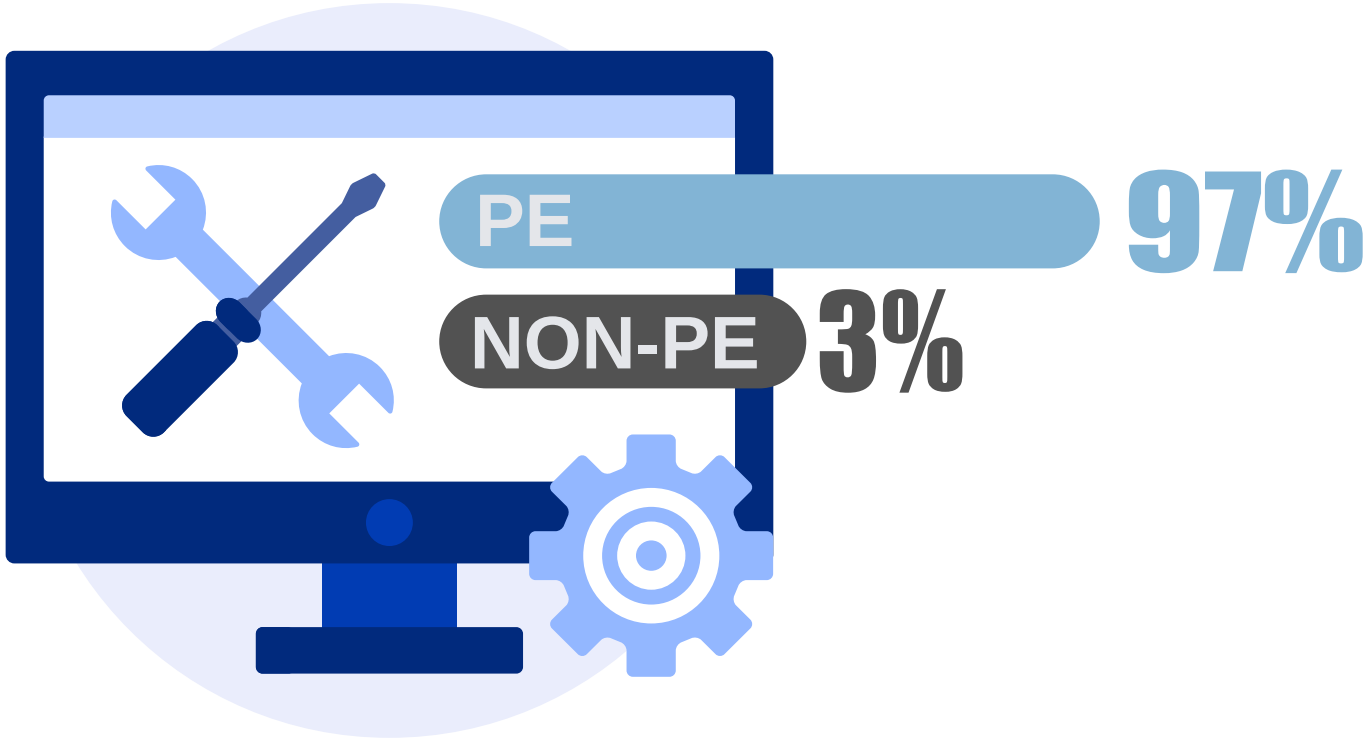
데이터 분석

예선 데이터셋



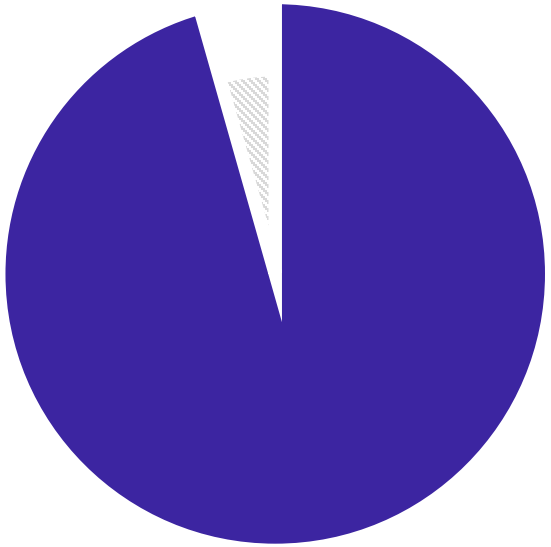
데이터 분석

예선 데이터셋 → PE



데이터 분석

예선 데이터셋 → PE

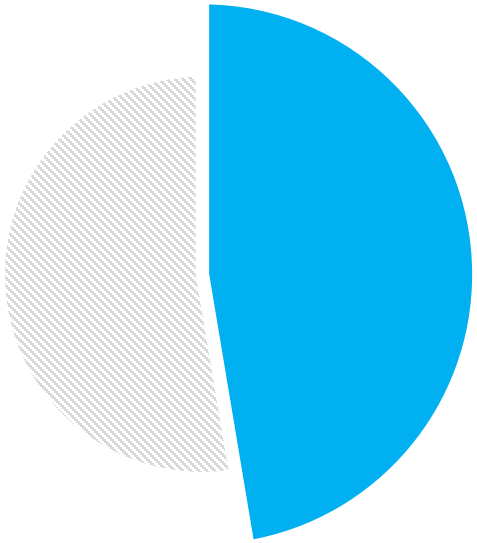


90% PE32

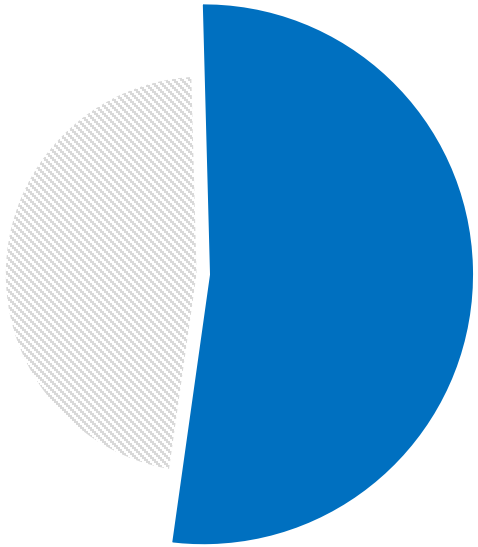
10% PE64

데이터 분석

예선 데이터셋 → PE → Non-PE



48% HTML



52% ETC

/02 데이터 수집 및 라벨링

Team KAISA

데이터 수집 및 라벨링

01

데이터 수집

- VirusTotal, VirusShare, VirusSign

02

라벨링

- VirusTotal 을 활용하여 라벨링

03

학습에 사용한 데이터 수

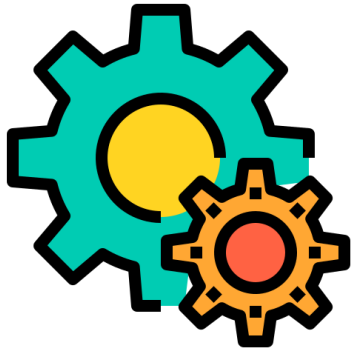
- PE32: 약 50만 개
- PE64: 약 30만 개

/03 탐지 알고리즘

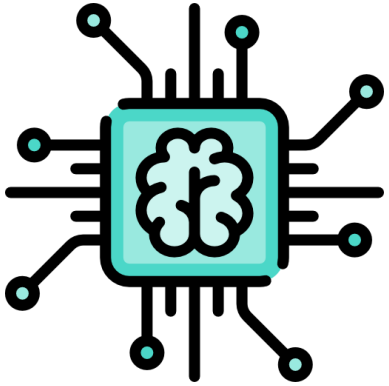
Team KAISA

탐지 알고리즘

파일 분류



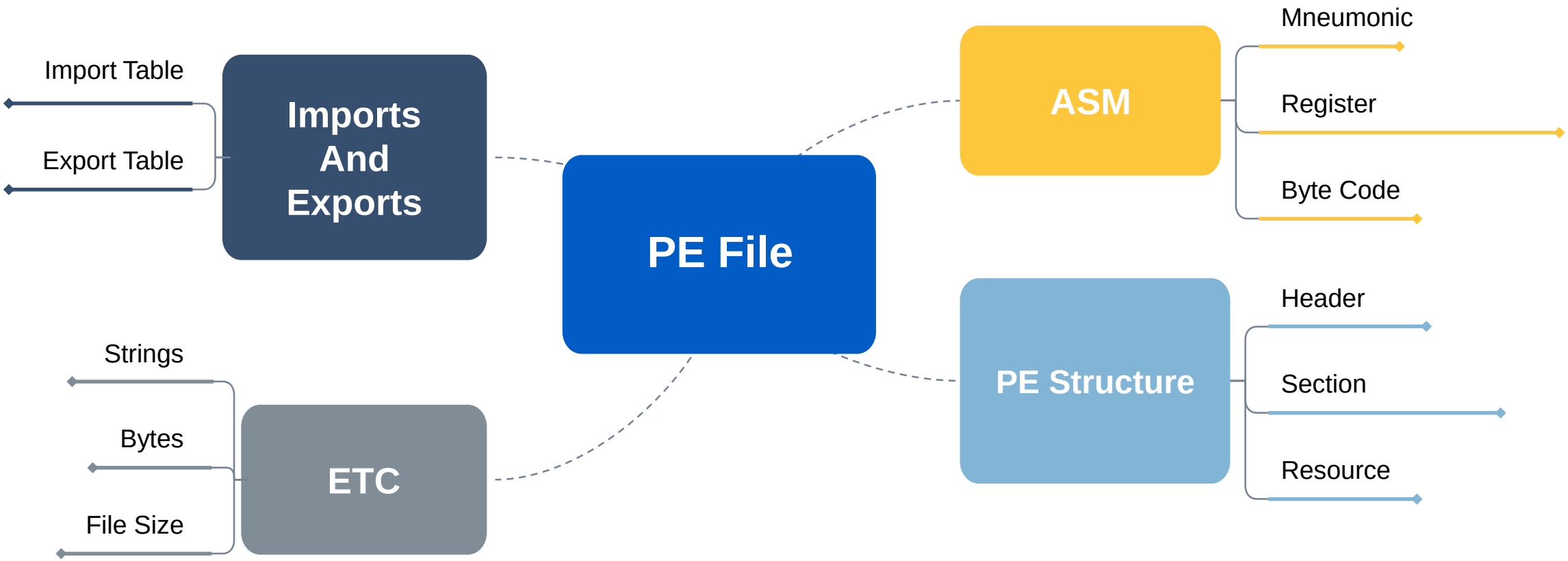
파일 별 특징 추출



탐지

탐지 알고리즘

PE


Import Table
Export Table
Imports And Exports
PE File
ASM
Mneumonic
Register
Byte Code
Strings
Bytes
File Size
ETC
PE Structure
Header
Section
Resource

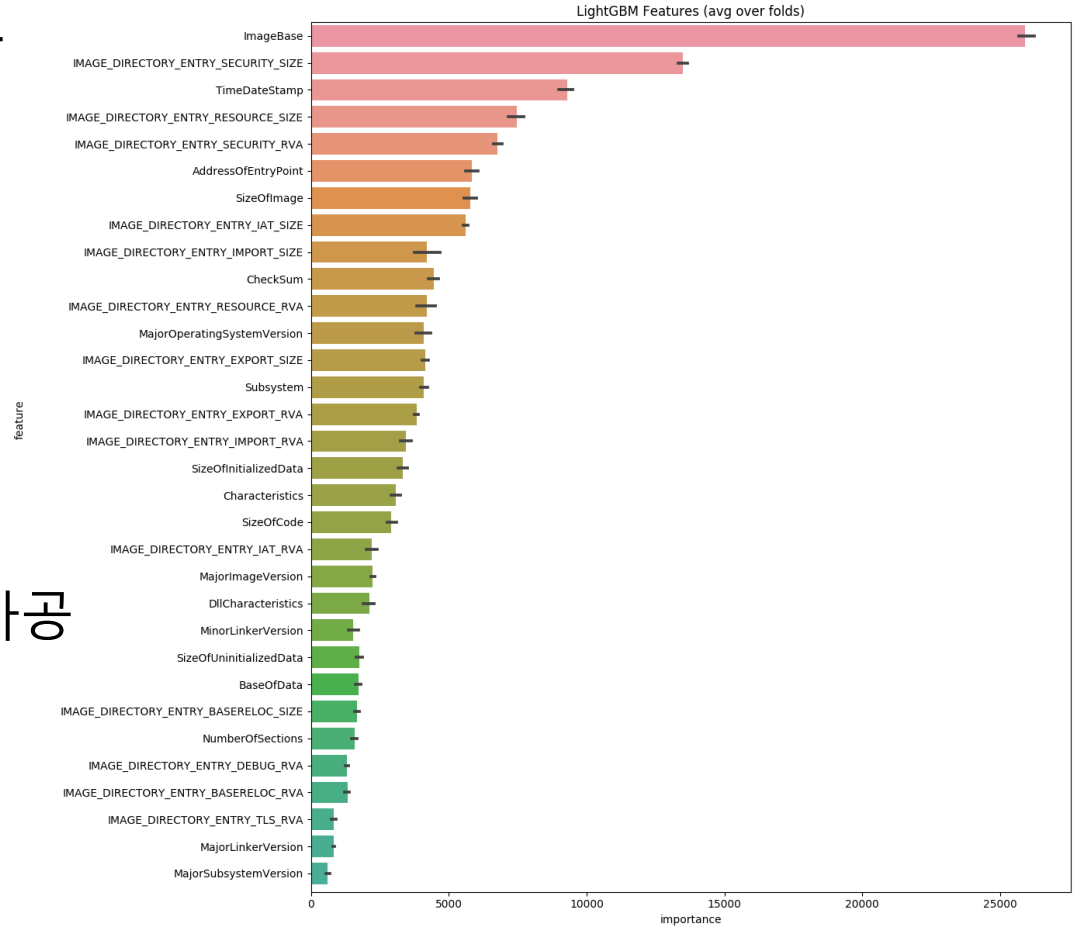
탐지 알고리즘

PE

정적 분석으로 추출할 수 있는 모든 특징 추출

판단은 모델에게

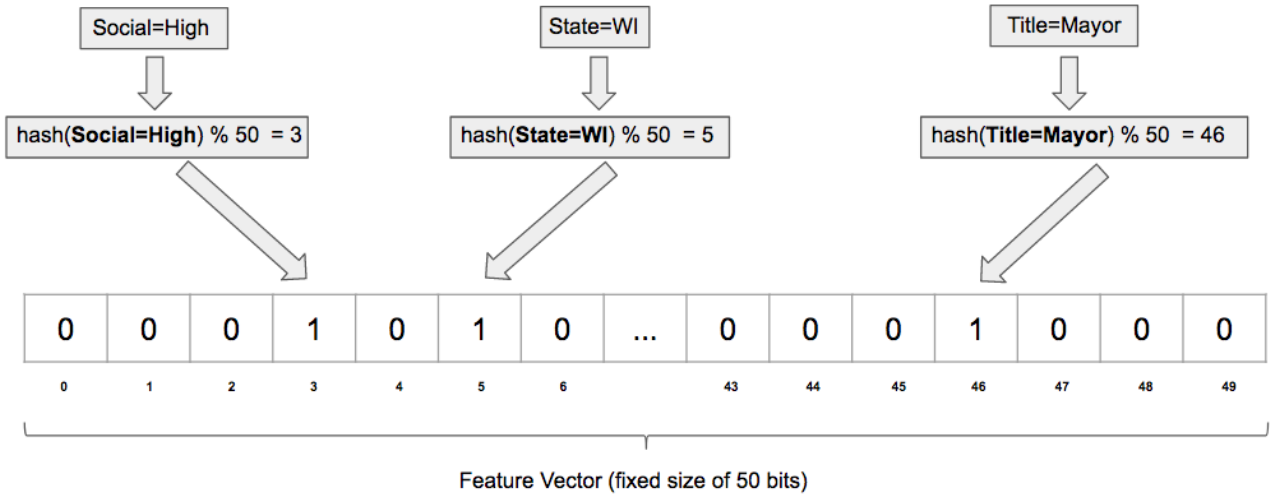
- 정형화 데이터는 정형화 데이터의 값 자체 사용
- 가변적 데이터는 해싱 트릭(Hashing Trick)으로 가공
- 특징 벡터 크기: 10,191차원



탐지 알고리즘

해싱 트릭

- Kilian Weinberger, Anirban Dasgupta, Josh Attenberg et al. “Feature Hashing for Large Scale Multitask Learning,” International Conference on Machine Learning (ICML) 2009
- 아래의 특징들을 해싱 트릭으로 벡터화
  - ✓ Import Table, Export Table, Rich Header, Strings, Opcode 3-gram



출처: https://bit.ly/2NUL4U1

탐지 알고리즘

PE

사용한 머신 러닝 모델

- LightGBM
- XGBoost
- Random Forest
- DNN

학습 데이터

- KISA 데이터셋
- KISA 데이터셋 + 자체 수집 **악성/정상** 데이터
- KISA 데이터셋 + 자체 수집 **정상** 데이터

탐지 알고리즘

PE

5 폴드 교차 검증 (5 Fold Cross Validation)

```
seclab@seclab:~/kisa_2019$ python3 tmp.py cv xgb KISA_TRAIN_FEATURE.csv
Generate Dataframe
CV Start...
Fold 1 >> acc score 0.93505039193729
Fold 2 >> acc score 0.9394957983193277
Fold 3 >> acc score 0.9316526610644258
Fold 4 >> acc score 0.9417366946778711
Fold 5 >> acc score 0.945627802690583
5 Fold Average Score:  0.9387126697378996
CV End, Elapsed time: 597.6390080451965
```

```
seclab@seclab:~/kisa_2019$ python3 tmp.py cv xgb KISA_TRAIN_FEATURE.csv
Generate Dataframe
CV Start...
Fold 1 >> acc score 0.9624860022396416
Fold 2 >> acc score 0.969187675070028
Fold 3 >> acc score 0.9647058823529412
Fold 4 >> acc score 0.9697478991596639
Fold 5 >> acc score 0.9714125560538116
5 Fold Average Score:  0.9675080029752173
CV End, Elapsed time: 3228.9058578014374
```

탐지 알고리즘

PE

- 탐지
- 모델 별 탐지 결과 블렌딩

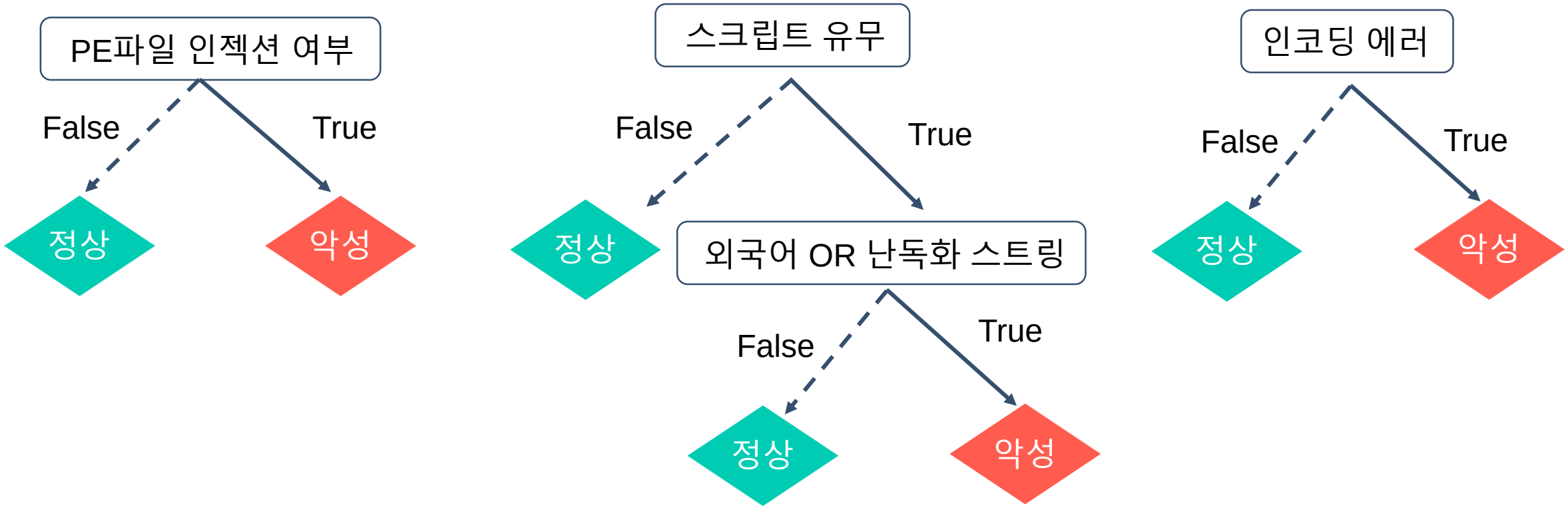
✓보팅(Voting) 방법 사용
✓모델 조합과 악성 기준 설정

$$y_{pred} = y_{1_{pred}} \circledcirc y_{2_{pred}} \circledcirc \cdots \circledcirc y_{n_{pred}}$$

탐지 알고리즘

HTML JS

아래 알고리즘의 결과와 추출한 값을 피처화 ➔ 의사 결정 트리 학습



탐지 알고리즘

HTML JS

사용한 피처

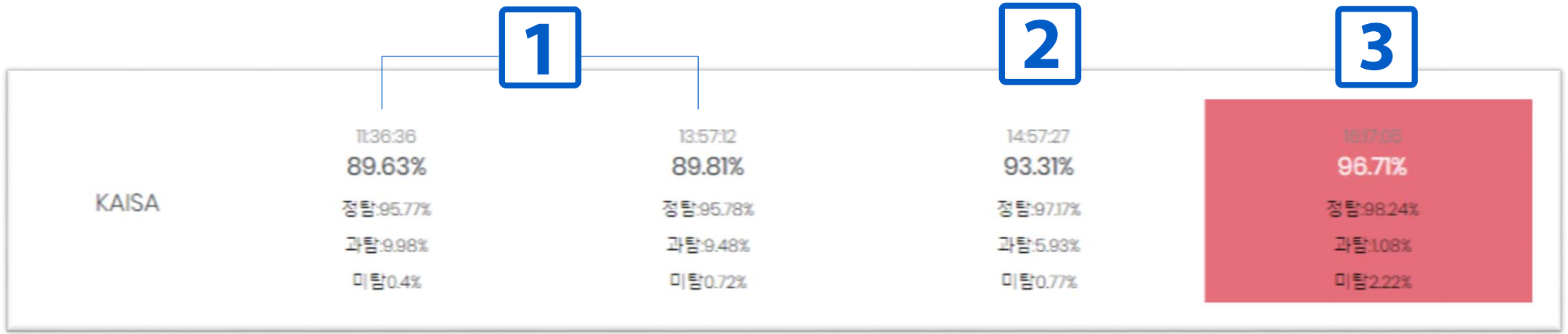
PE	PE 파일 인젝션 유무
Script	스크립트 유무
Strange_lan	HTML 에 포함된 국가 언어 개수 (중국어,아랍어,라틴어 등)
Array_no10	'2138329423 ' 와 같은 난독화 된 문자 유무
Error1	파일 읽기 에러 유무
Error2	인코딩 에러 유무
Entropy	HTML 엔트로피
Tag_count	HTML 태그 개수

탐지 알고리즘

예선

- LightGBM 단일 모델 사용

1. KISA 학습 데이터 + 자체 수집 악성/정상 데이터
2. 첫 번째 모델, KISA 학습 데이터 모델 블렌딩(AND)
3. 두 번째 모델, KISA 학습 데이터 + 자체 수집 정상 데이터 모델 블렌딩(AND)



Q&A

Team KAISA
문다민 김영재 손현기 오예린

데이터 분석 – Train Set

파일 타입	개수	비율	악성 개수	정상 개수	악성 비율
PE32	8925	89.25	6929	2656	70.24
PE64	978	9.78	681	297	69.63
HTML	42	0.42	36	6	85.71
HWP	20	0.2	10	10	50.00
XML	**13**	**0.13**	**0**	**13**	**0.0**
UNKNOWN	13	0.13	1	12	7.69
MS-DOS	**3**	**0.03**	**3**	**0**	**100.00**
MACH-O 64	**1**	**0.01**	**0**	**1**	**0.00**
WOFF v2	**1**	**0.01**	**0**	**1**	**0.00**
PHP	**1**	**0.01**	**0**	**1**	**0.00**
PY	**1**	**0.01**	**0**	**1**	**0.00**
PYC	**1**	**0.01**	**0**	**1**	**0.00**
TXT	**1**	**0.01**	**0**	**1**	**0.00**