Application of Machine Learning on APT Defense

Tian Tian,
Zhenyu Qiu,
ZTE Corporation

Advanced Attack Vs. Defense

Spear-Phishing

0-Day Exploits

Custom Malware

Social Engineering

Malware Variants metamorphism &packer

Convert/Encrypted Tunnel

Firewall

IDS/IPS

Email-GW

Anti-Virus

Unknown Threats

Anti- Unknown

- **C&C Communication**
- Internal abnormal access
- Abnormal data transmission

Network Traffic Analysis

File Analysis

- Email attachments
- Web download files
- Other files

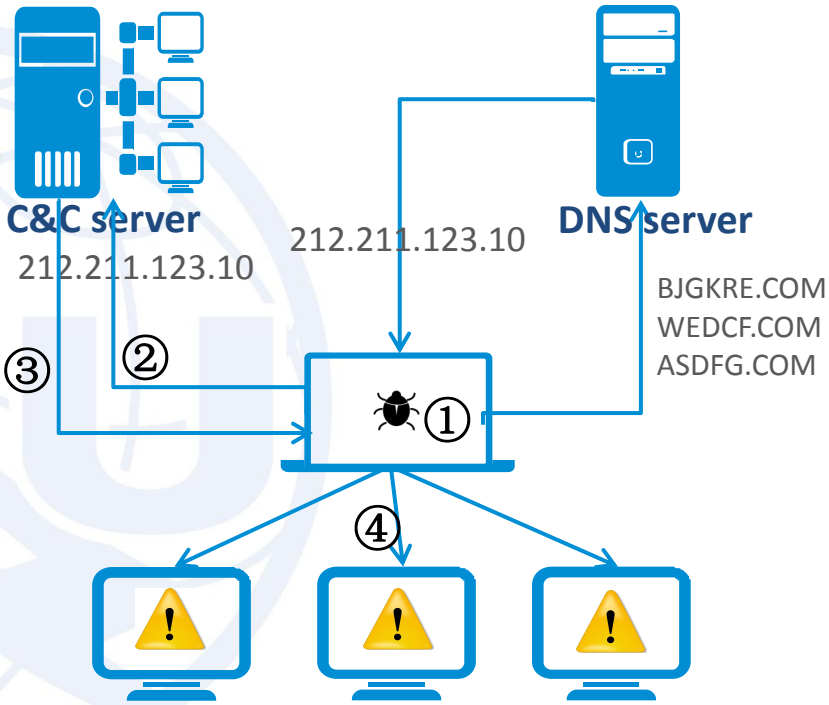
Machine Learning --- Simplify and assist the experts' work.

Application Case 1:
Improvement of DGA Detection

C&C Communication

C&C server (Command & Control Server)

- Communicate with hosts which are infected with malwares in the botnet and direct their attack behavior



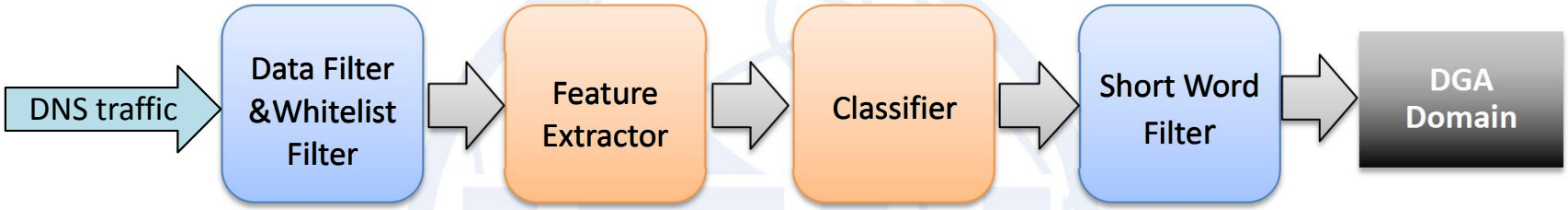
DGA (Domain Generating Algorithms)

- Common technique used in **C&C Location**
- Many sophisticated malware families use DGA instead of IP addresses thus avoid the blacklist detection and defense

① Infected with malware
② Locate C&C Server
③ Receive instructions
④ Excute instructions

DGA Detection Method 1

- **Original workflow**



- **Feature extraction**

Feature Set	Feature Examples
Distribution Template Feature	• Domain valuation • Appearance probability
Structure Template Feature	• Transition probability • TLDs probability
Pronounce Feature	• Number of word-like unit
Common Feature	• Domain length • Number of Repeat letters

- **Sample set**

Category	DGA family	Size
ALGORITHM	Conficker/Kraken/ Necurs/Tinba/...	100,000 +
HASH	Murofet/Pushdo/Zeus/...	100,000 +
WORD	Gozi/Matsnu/Rovix/...	100,000 +

Problems of Method 1

Defects of Method 1

- High false positive rate on domains especially for chinese corporations
- Current features and classifiers have limit improvement on detection accuracy
- Need expert knowledge on related domain
- Complicated procedure of training and inference
- Bottlenecks in training and testing efficiency

Examples of legitimate domains detected as dga:

Gouqi001.com
18183.com
500px.me
17zwd.com
.....

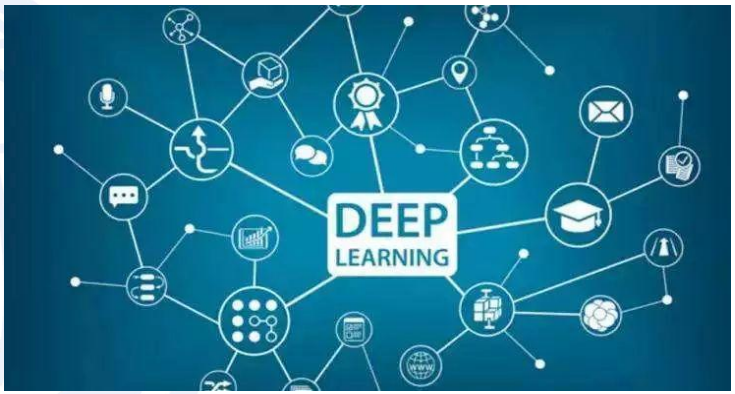
Method	Precision	Recall	FPR
RF	0.936	0.938	0.08

- Huge numbers of raw features > 100,000 +
- Expert knowledge needed for data analysis and linguistics ...
- Big effort for feature selection
- Long training time

Method 2
Deep Learning on DGA Detection

URL domains belong to natural language text

Transform DGA detection to character-level text classification



Pros:

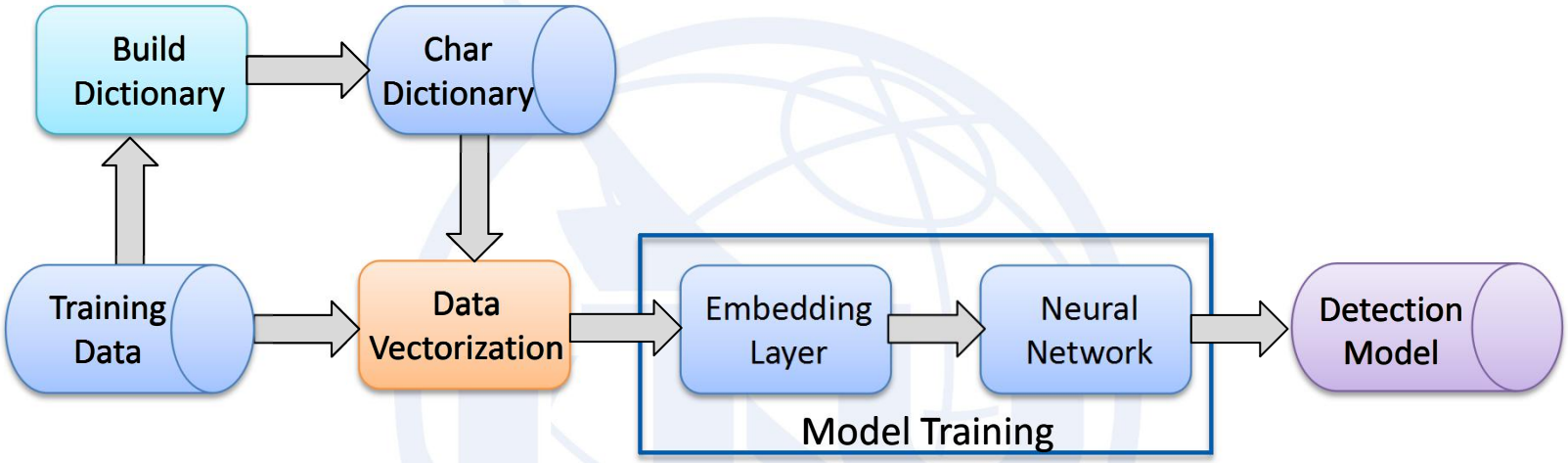
- High efficiency: Char-level dictionary contains much less element
- End-to-end procedure: No need to extract features manually
- Improve Performance: Make full use of big data and extract implicit complicate features, leading to accuracy improvement

Three neural network structures used:

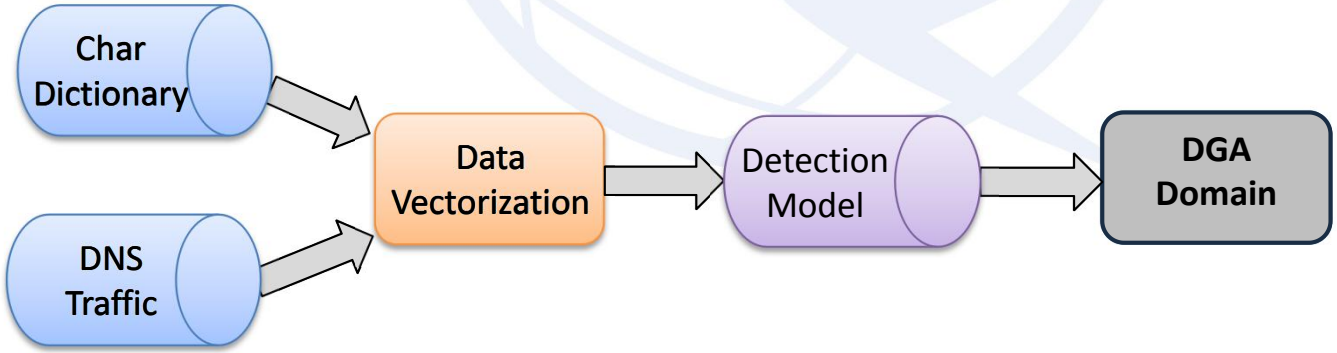
- *Convolutional neural network*
- *Recurrent nerual network with gate mechanism(LSTM)*
- *Attention mechanism(based on RNN network structure)*

Workflow of Method 2

- **Training workflow**



- **Detection workflow**



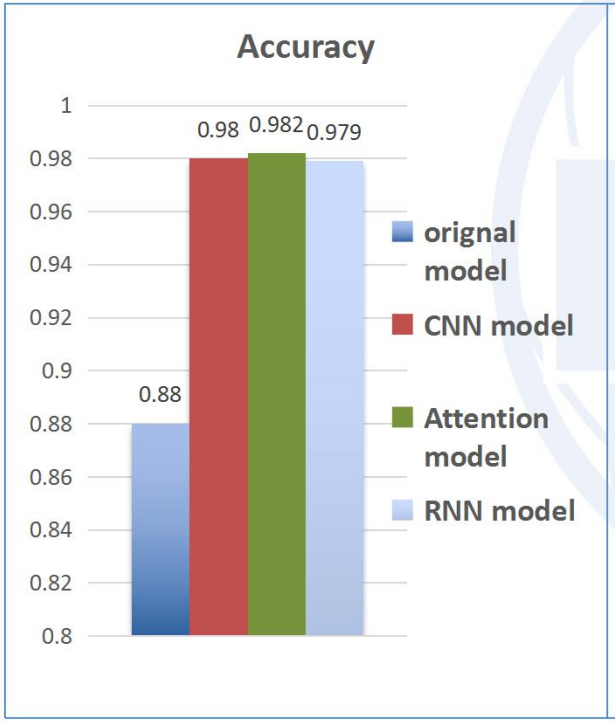
Results of Deep Learning on DGA

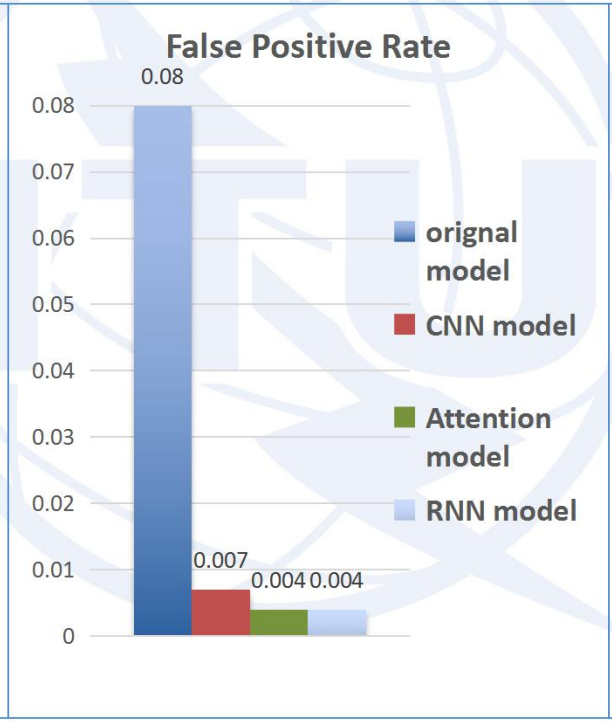
Positive Samples (DGA Domains)

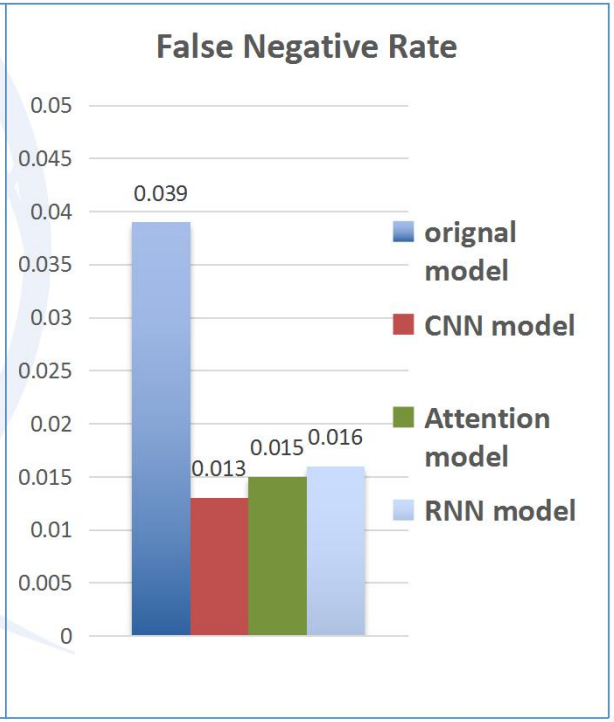
- Remove duplicate samples
- 600 thousands samples from cooperative partners
- 200 thousands samples from Bambenek[1]

Negative Samples (Normal Domains)

- 1 million domains from Alexa
- 50 thousand domains from China not in Alexa's top ranking list
- Remove duplicate samples







- Average accuracy improvement rate: 11.4%
- Average FPR improvement rate: 93.7%
- Average FNR improvement rate: 62.3%

✓ **Final pick: CNN model (high efficiency and acceptable performance)**

Reference: [1] osint.bambenekconsulting.com

Application Case 2: Malware Detection

Method of Malware Detection

Unknown malware

An effective weapon for avoiding detection in the initial compromise, delivery and exploitation phases of advanced cyber attacks

General Method

- Omit parameters of all behaviors, only consider behavior sequence itself
- Based on expert knowledge, conduct detection rules to analyze file dynamic behaviors

Problem

- Overall accuracy of system is not so good
- False positive Rate (set malicious as Positive) is high, too many normal files being detected as malicious
- Expert knowledge base is still lack of rules for all kinds of files and security scenes

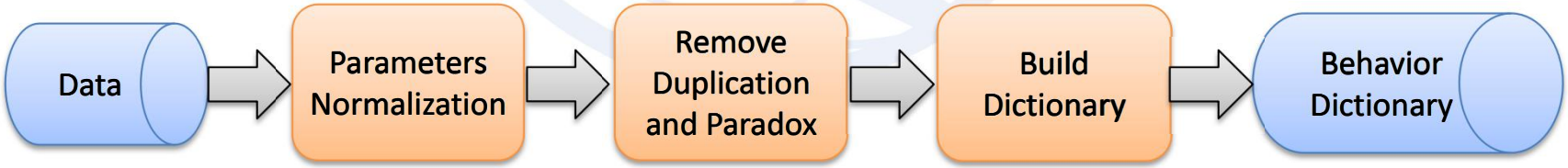
File Dynamic Behaviors → **New method: Machine learning + NLP techniques**

Parameter Abstraction & Data Preprocessing

- **Rich parameters**

Behavior	Parameters
readFile	[filePath:”c:/system32”,fileType:exe]
copyFile	[srcPath:”c:/windows”,dstPath:”d:/”,fileType:dll]
modifyRegistry	[regPath:”HKEY/.....”,key:”...”,value:”....”]
.....	.....

- **Data preprocessing procedure**



Feature Abstraction & Classification

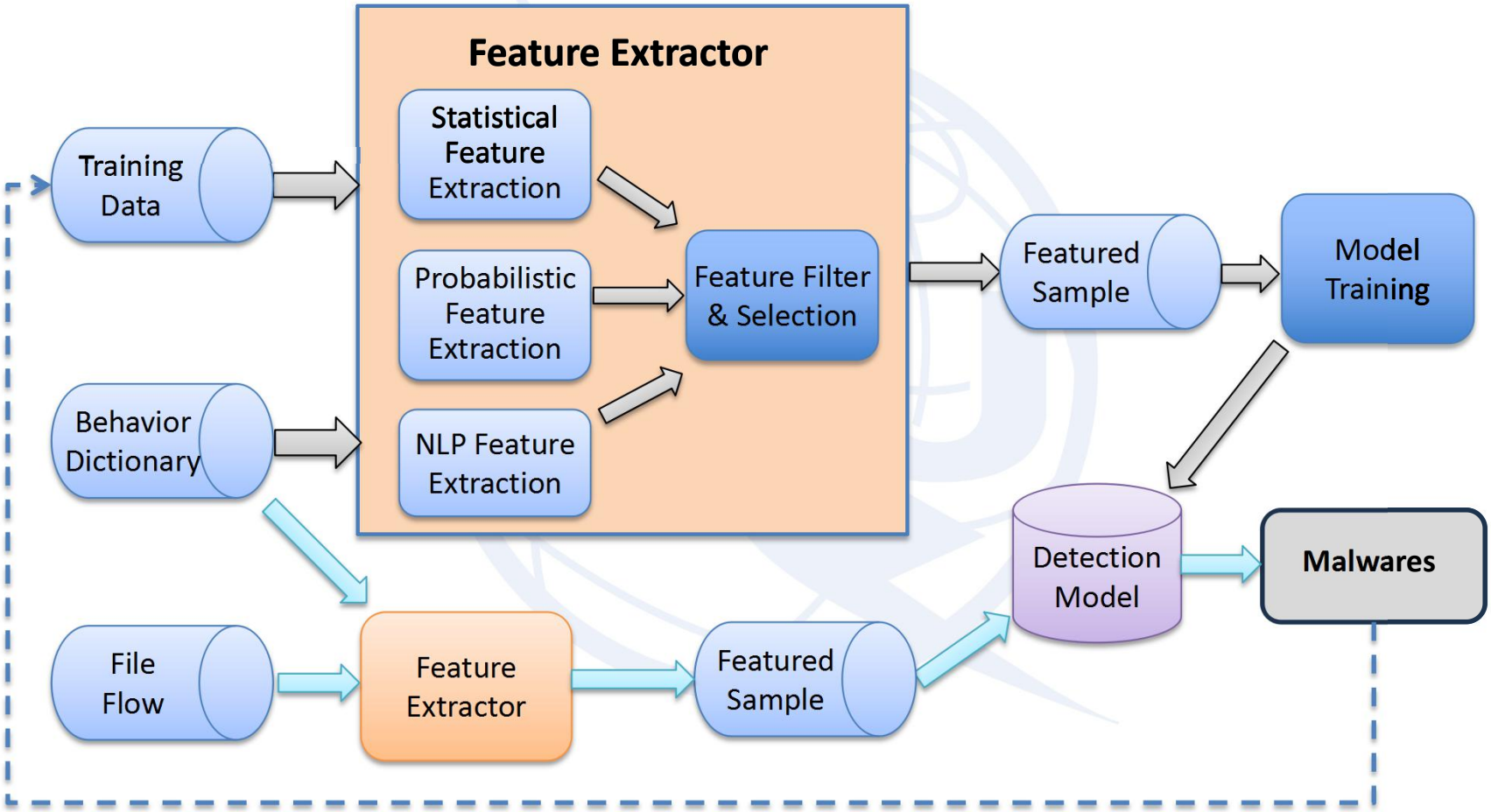
- **Feature abstraction**

Feature Type	Features
Statistical Features	behavior sequence length, consecutive repetitive behavior sequence ratio, write-file behaviors' count and ratio on sensitive file....
Probabilistic Features	behavior+parameter entropy, information gain...
NLP Features	tf-idf features, isi features, ida features, nmf features....

- **Classification**

Classification Algorithm Type	Classification Algorithm
Linear Model	logistic regression, SVM(linear kernel), ...
Tree Model	random forest, ...
Boosting Model	xgboost, ...

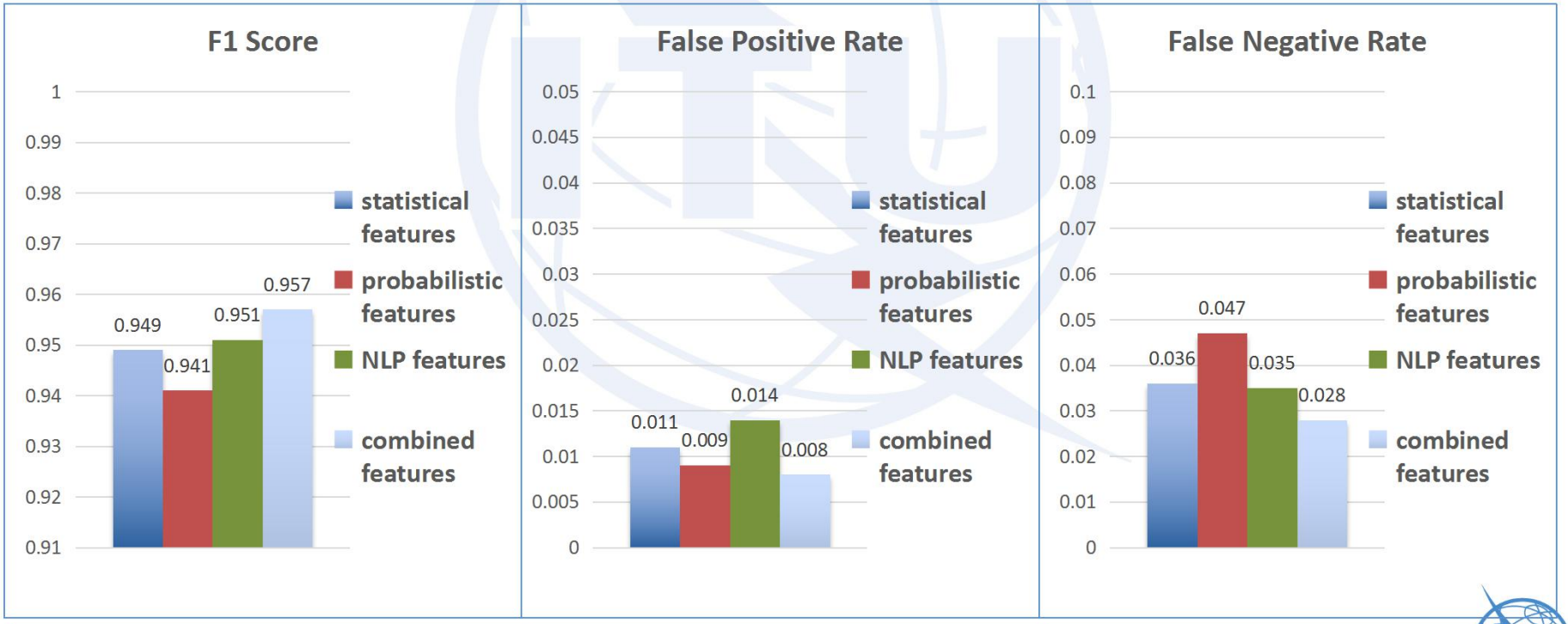
Training & Detection Workflow



Classification Results

Positive Samples (Malicious Files)
- 1,000,000+ malicious from real scene
- Remove duplicate and paradox samples

Negative Samples (Normal Files)
- 500,000+ normal files from real scene
- Remove duplicate and similar samples



** Results are based on xgboost algorithom.*

Conclusion

- In cyber security, machine learning ,especially deep learning can improve efficiency of some tasks, e.g. DGA detection and malware detection.
- However, machine learning techniques are still not ready for practice application for its lack of interpretability and reliable accuracy. Also, it's being vulnerable to some GAN attack,which is another research area in cyber security.
- Expert knowledge should be combined with maching learning techniques, which could improve accuracy of the models. In some tasks, we still need to use mannual rules like detecting specific malicious behaviors of a malware, which could help us extract more effective features for training models.

Thanks !