Machine Learning for Big Data Analysis, Cybersecurity, and Privacy-Preserving Data Mining

Seiichi Ozawa

Center for Mathematical Data Science

Graduate School of Engineering

Kobe University

A Brief Introduction of Center for Mathematical and Data Science

Started from Dec. 1st, 2017

Missions

1. Development of standard and advanced programs to foster talent who has strong mathematics and data science background for students with any majors.
2. Facilitate research and education on Data Science, AI, Big Data Analytics, and ICT.
3. Collaboration with industries, regions, research & educational institutions on Data Science.

Structure of CMDS

Education

- **19 Faculties**
- **8 Coordinator Groups** (Mathematics, Statistics, Computer, Data Science, Advanced Program, Computer Simulation, ICT, etc.)

Research

- **62 Faculties**
- **11 Research Teams** (Mathematical Science, Machine Learning, Big Data/Security, Multi-Media Data Analysis, HPC, CPS, FinTech, Healthcare, etc.)

Collabo-ration

- **5 Faculties**
- **3 Sections** (Data Science Entrepreneurship, Open Innovation Consortium, International Relations)

Education System

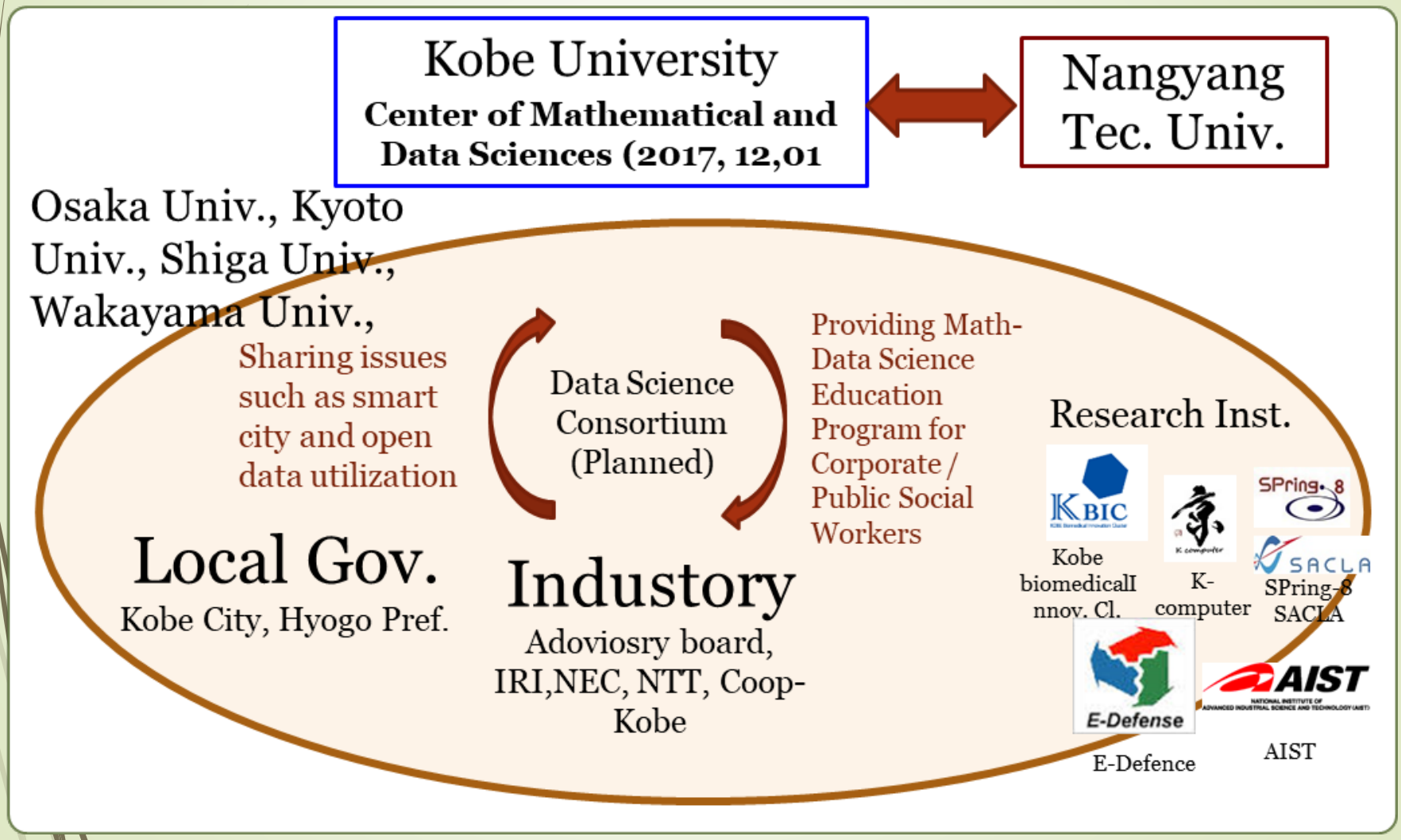
Standard Curriculum for Mathematics and Data Science

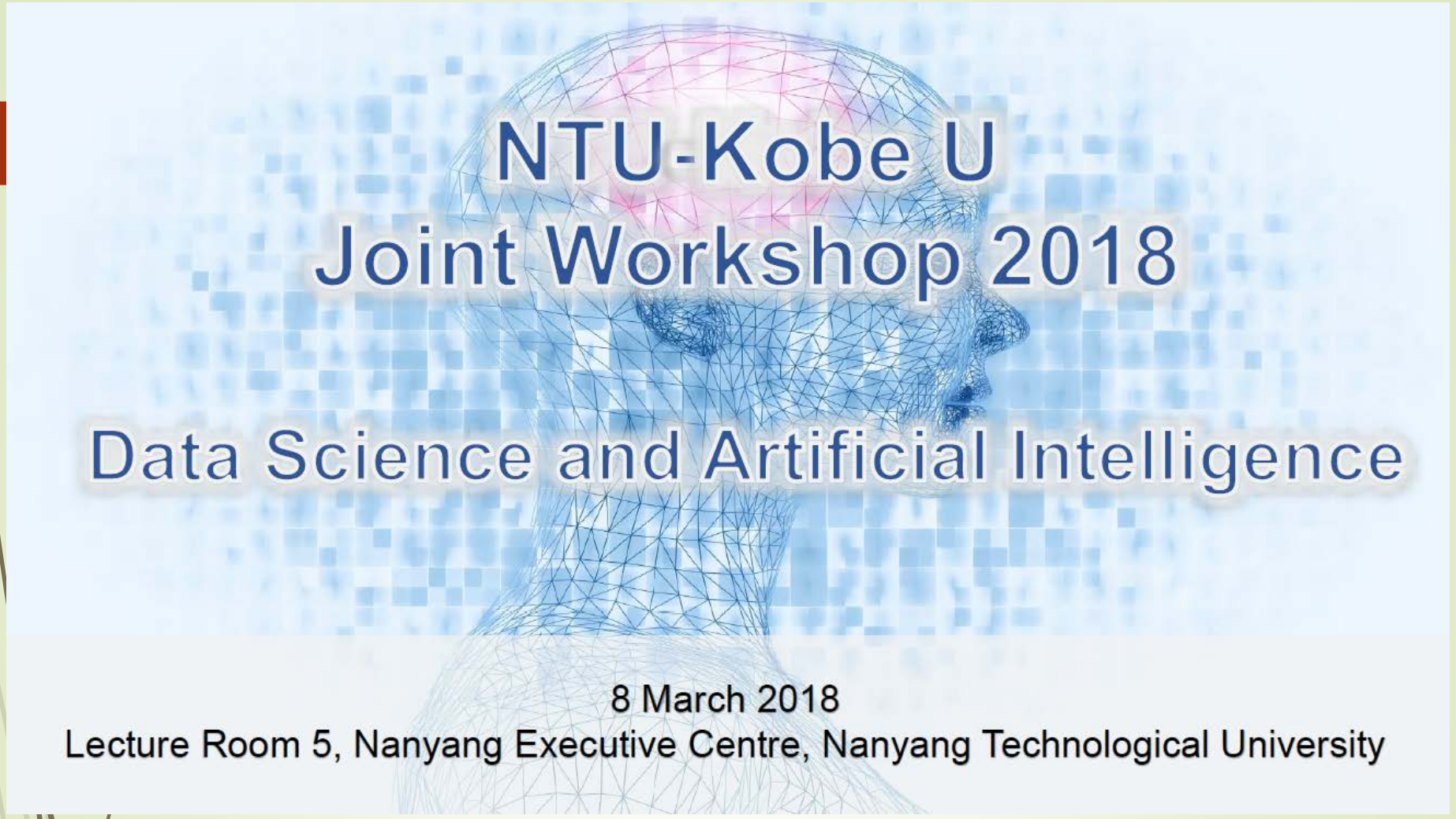
- **Level 1** (for 1st and 2nd grade undergraduate)
  - Mathematics, Statistics, Algorithm, Programming, Introduction of Data Science
- **Level 2** (for 3rd and 4th grade undergraduate)
  - Introduction of AI, Big Data Analytics, Computer Simulation, DS Practical Exercise
- **Level 3** (for master course students)
  - Statistical Machine Learning, Deep Learning, Cloud Computing, Information Security etc.

Data Science Advanced Course (PBL-based)

- **Problem Solving DS Advanced Course**
  - Date Science and Science / Engineering
  - Data Science and Finance / Marketing
  - Data Science and Medical / Healthcare, etc.
- **Value Innovation DS Advanced Course**
  - Data Science Open Innovation Workshop
  - Smart City project with IoT and Open Data

Collaborations





The workshop Objective is to share
- unique data sets of our own
- ways to generate unique data
- unique open problems
- up-to-date analytical methods (at top-conference level)

Machine Learning for Big Data Analysis, Cybersecurity, and Privacy-Preserving Data Mining

Seiichi Ozawa

Center for Mathematical Data Science

Graduate School of Engineering

Kobe University

Ongoing Projects

Cyber-Security

- Darknet Traffic Analysis for Detection and Observation of Cyberattacks
- Deep/Dark Web AI Crawler
- Detection of Malicious Java Scripts

SNS Analysis

- Sentiment Analysis for SNS Comments and Its Application to Flaming Detection
- Fast Bi-term Topic Model for Flaming Event Detection

Big Data Analysis

- Image Sensing Methods to Capture Growth Status of Agricultural Plants (*Smart Agriculture*)
- Machine Learning for Encrypted Data (*Privacy-Preserving Data Mining*)

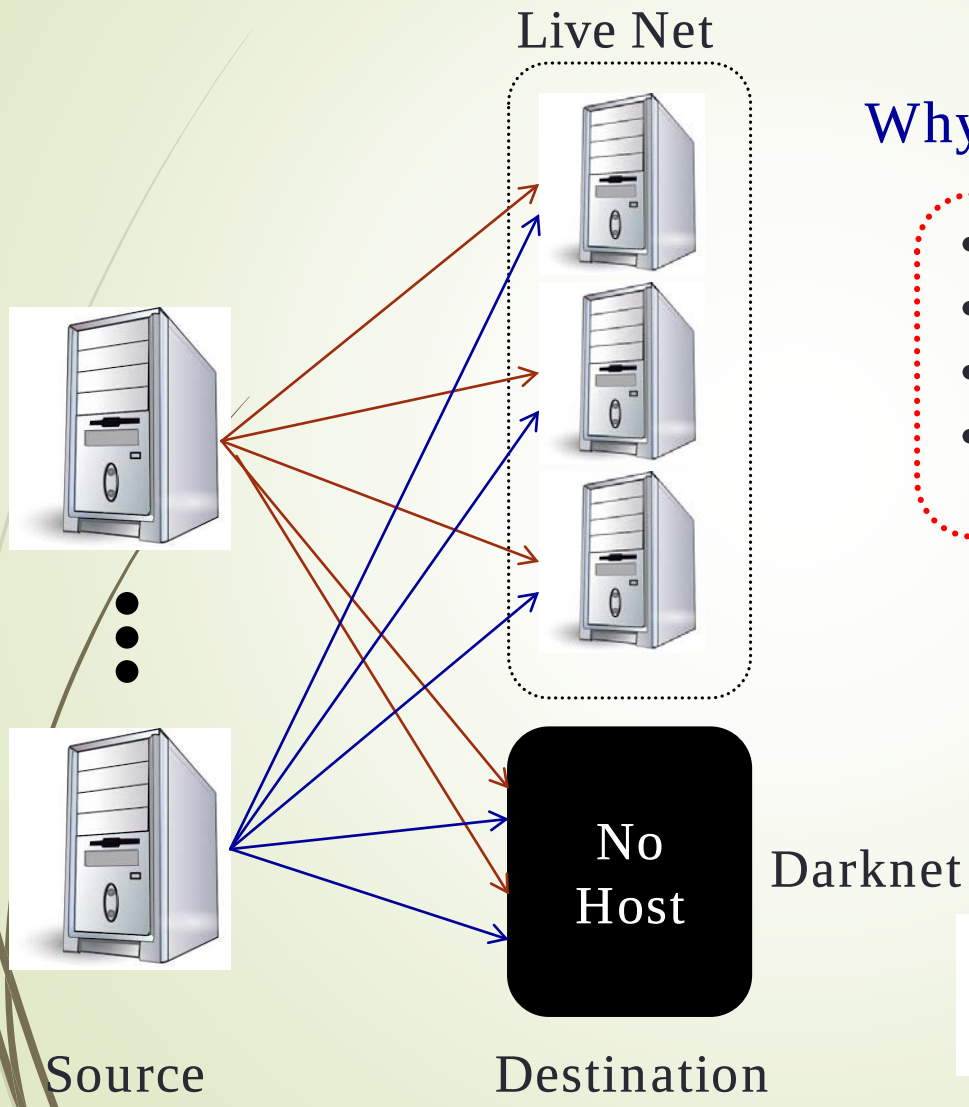
Machine Learning for Cybersecurity (1)

Darknet Monitoring for Detection of DDoS Attacks and Probing

Cybersecurity Projects Using ML

- **Malware Analysis (static and dynamic)**
  - ✓ Malware detection/classification
- **Network Intrusion Detection**
  - ✓ Anomaly detection
  - ✓ Monitoring malicious activities
  - ✓ Honeypot Data Analysis
- **Darknet Traffic Analysis**
  - ✓ Anomaly detection
  - ✓ Attack detection/classification
  - ✓ botnet activity detection and analysis
- **Web-Based Attack Detection**
  - ✓ Malicious websites and spam mail detection
  - ✓ Malicious JavaScript detection
- **Collection and Analysing Cyberattack Information**
  - ✓ Surface web analysis (SNS, security blogs/reports)
  - ✓ Deep/Dark web analysis (dark market, dark forum)

What is Darknet ?



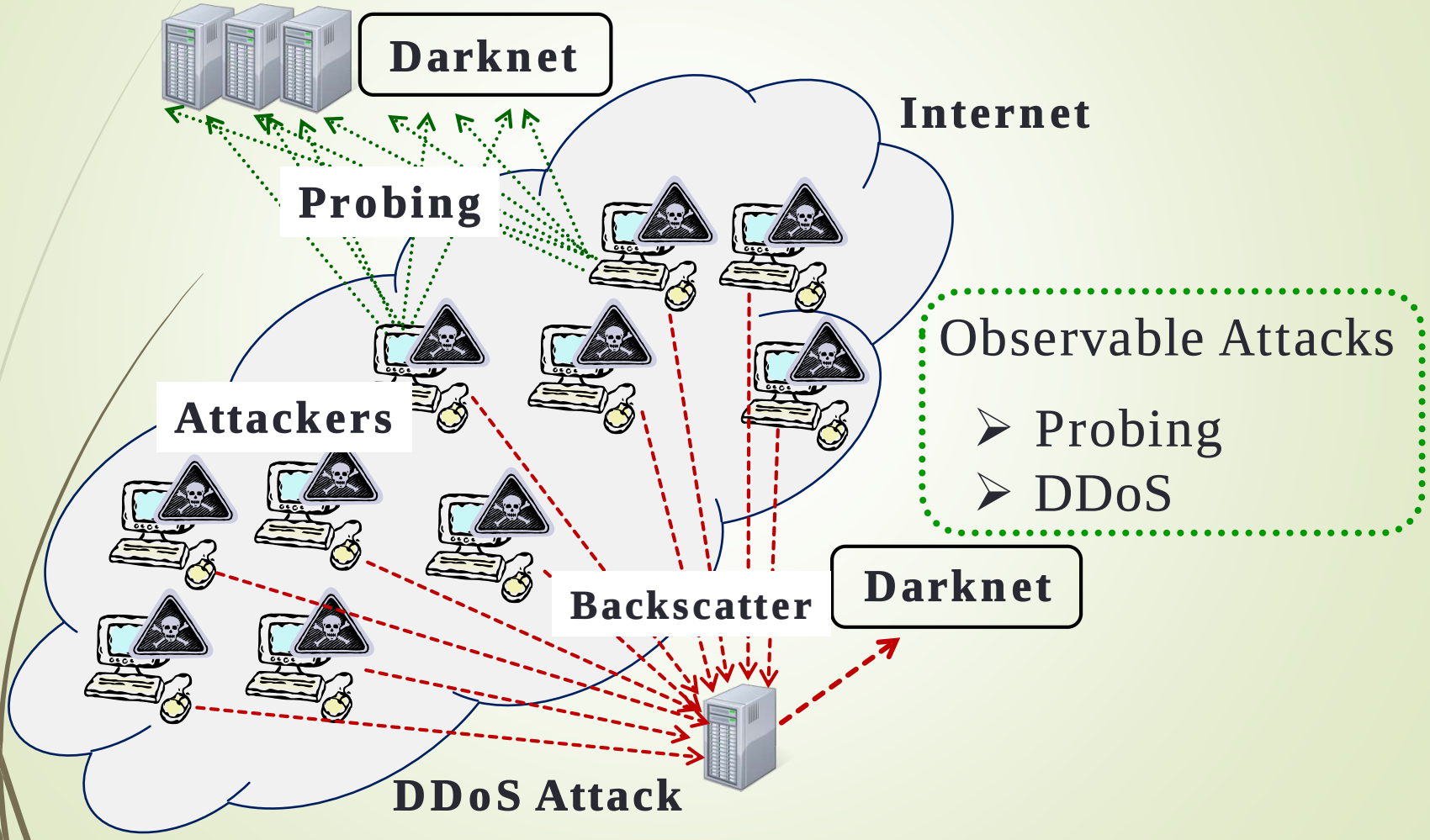
Why packets reach Darknet?

- Misconfiguration
- Scanning
- Exploit code
- Responses from targeted hosts by DDoS

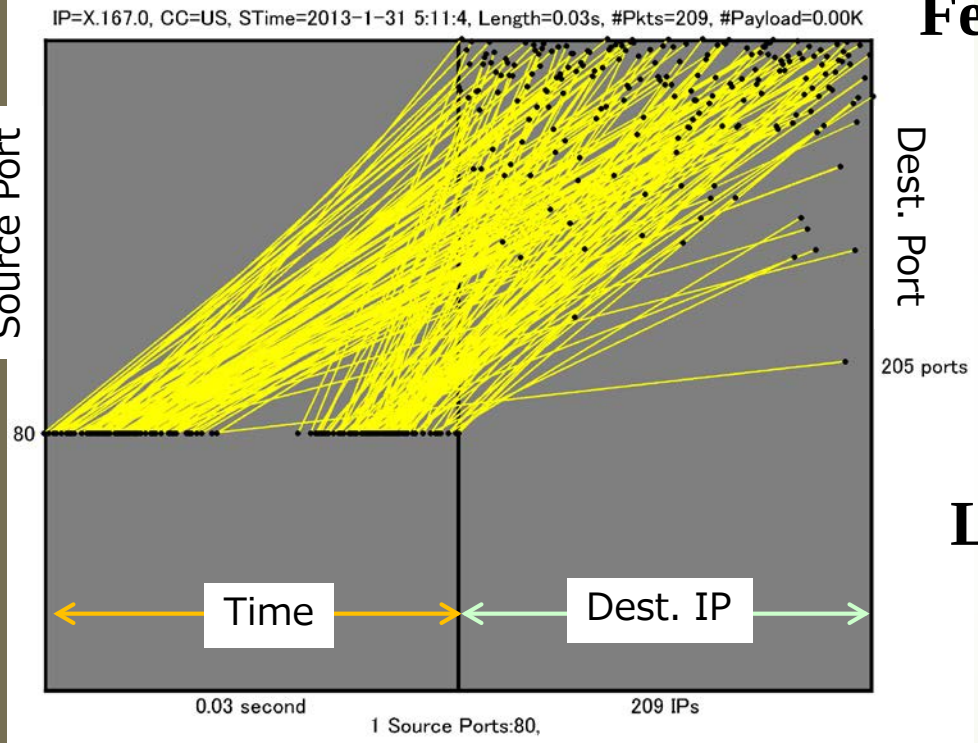
Malicious Activities

Large-scale monitoring for cyber attacks

What attacks are observed?



Traffic Features of Cyberattacks

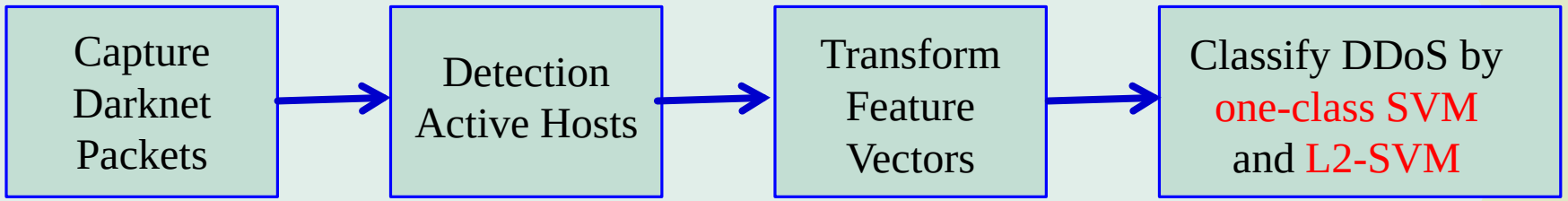


Feature Vectors

- #Packets
- Ave/STD Time Span of Packets
- Ave/STD #Packets from Src. Port
- Ave/STD #Packets to Dest. Port
- Ave/STD Payload Size, etc.

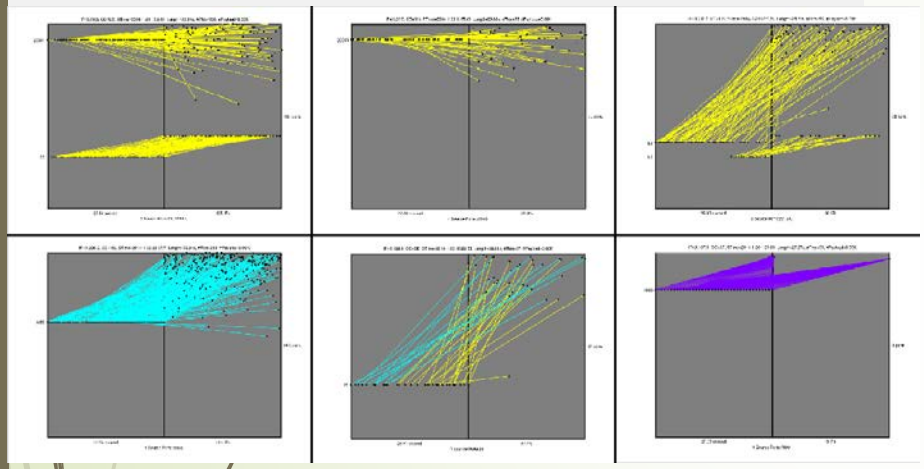
Labelling

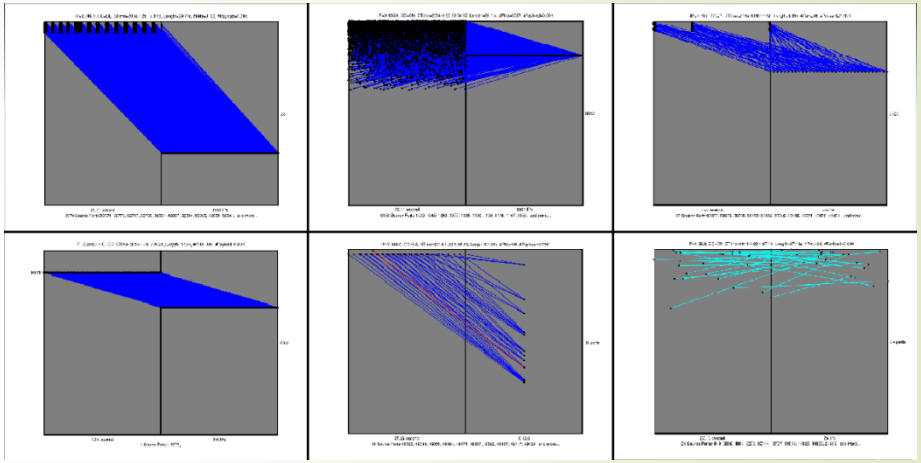
Based on Heuristic Rule for 80/TCP and 53/UDP packets

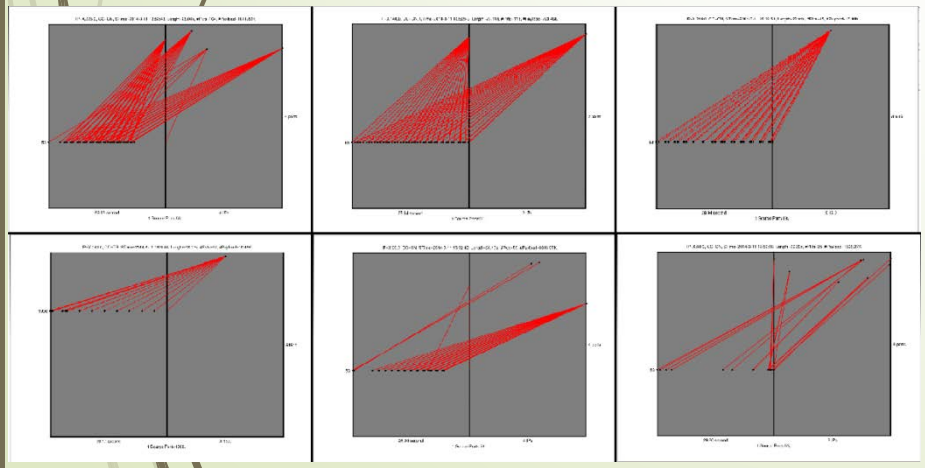


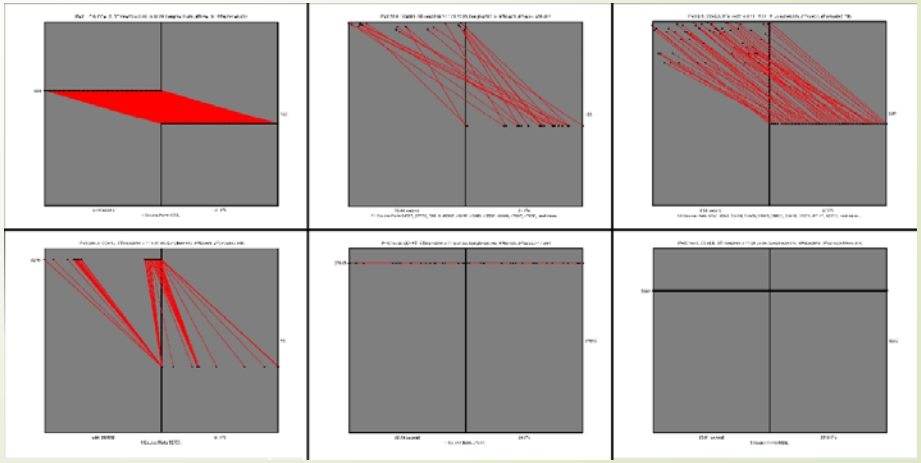
Variety of Darknet Traffic Features

Tiles (provided by NICT)

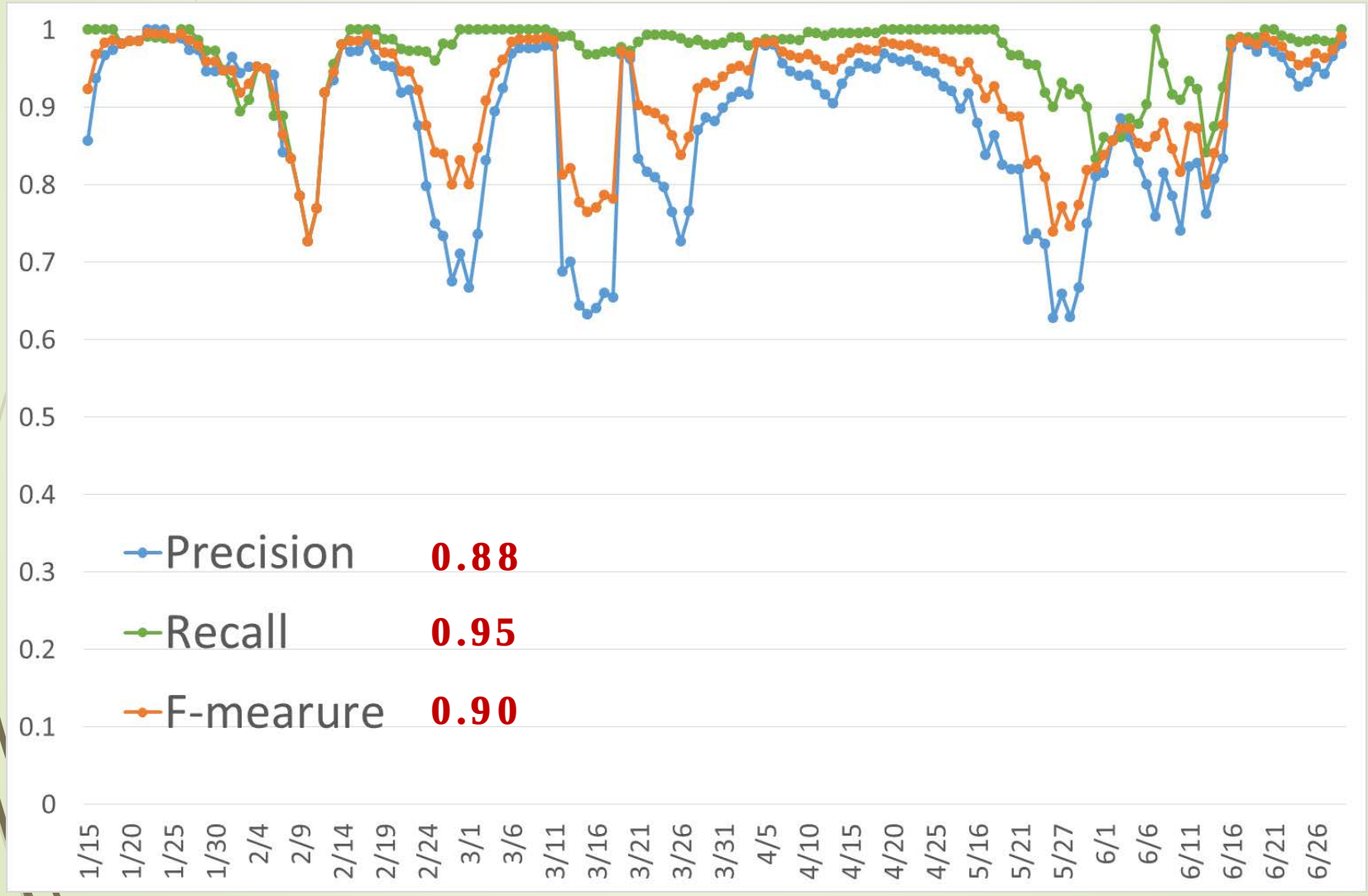








DDoS Detection Performance



Machine Learning for Cybersecurity (2)

Darknet Monitoring Using Association Rule Mining

Association Rule Mining

Association Rule

$$X \Rightarrow Y$$

Antecedent Consequent

When X is satisfied, Y is frequently satisfied.

Criteria of Rule Generation

Minimum Support :
The number of rules satisfying $X \Rightarrow Y$

Minimum Confidence :
Probability of satisfying $X \Rightarrow Y$

Case Example

Association Rule

$$X \Rightarrow Y$$

Antecedent Consequent

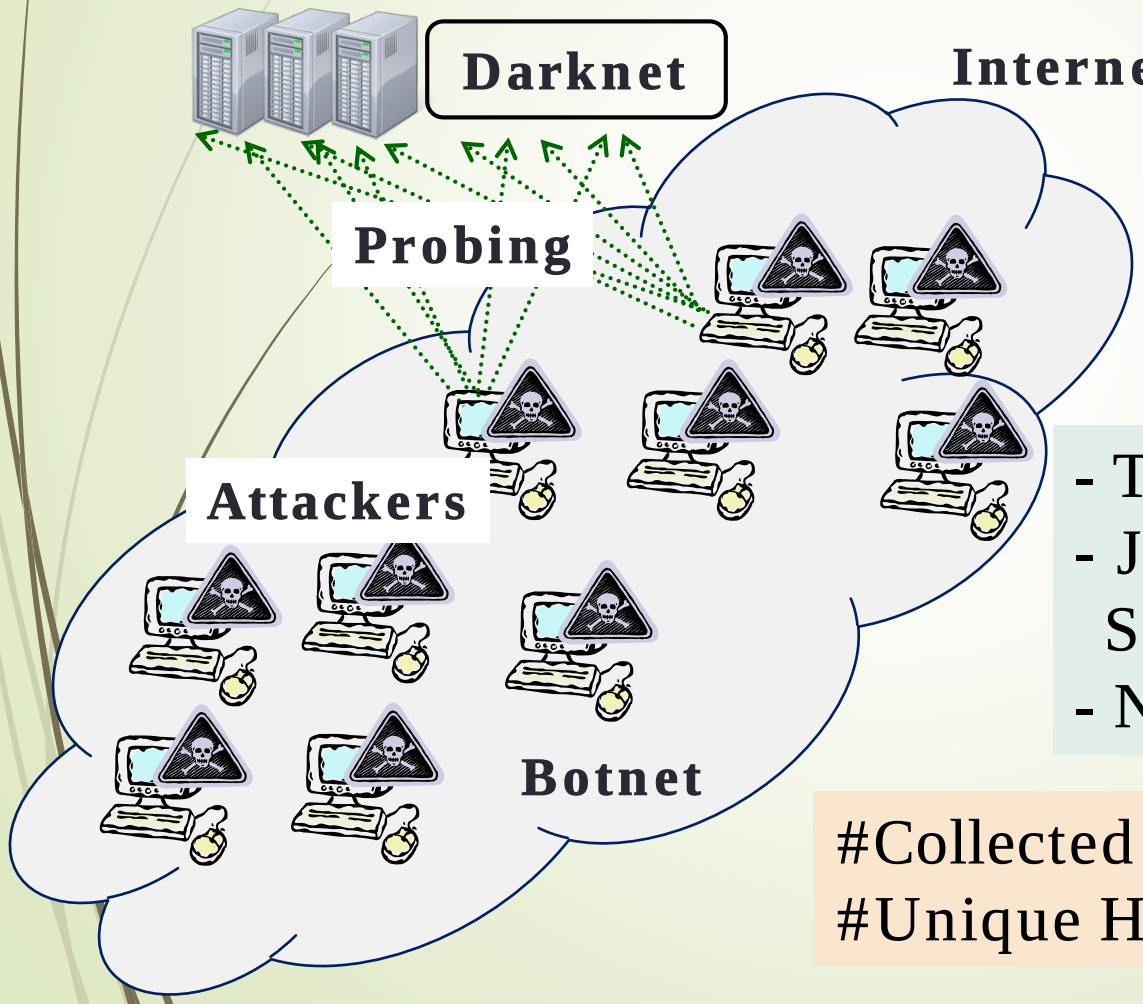
When X is satisfied, Y is frequently satisfied.

T1{Bread, Milk}
T2{Onigiri, Tea}
T3{Bread, Juce}
T4{Onigiri, Snack, Tea}
T5{Cup Nudle, Tea}
T6{Onigiri, Chicken, Tea}
T7{Onigiri, Chicken}

Ex.
{Onigiri} ⇒ {Tea}

The one who buys Onigiri frequently buy Tea as well.

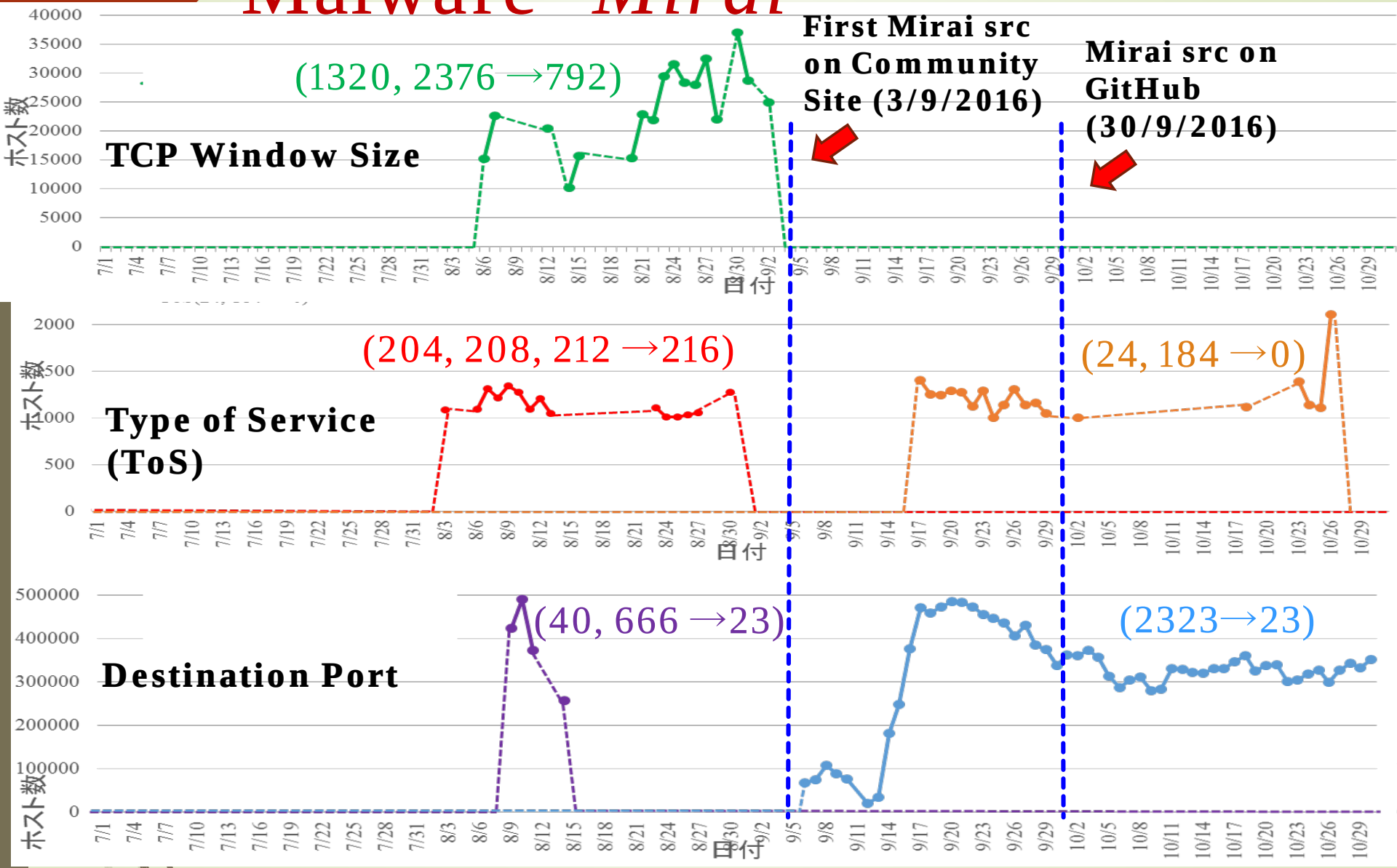
Darknet Packet Analysis Using Association Rule Mining



- TCP SYN packets
- July 1st, 2016 to September 15th, 2016
- NICT /16 darknet sensor

#Collected Packets: 1,840,973,403
#Unique Hosts: 17,928,006

Capturing Behavior of IoT Malware "*Mirai*"

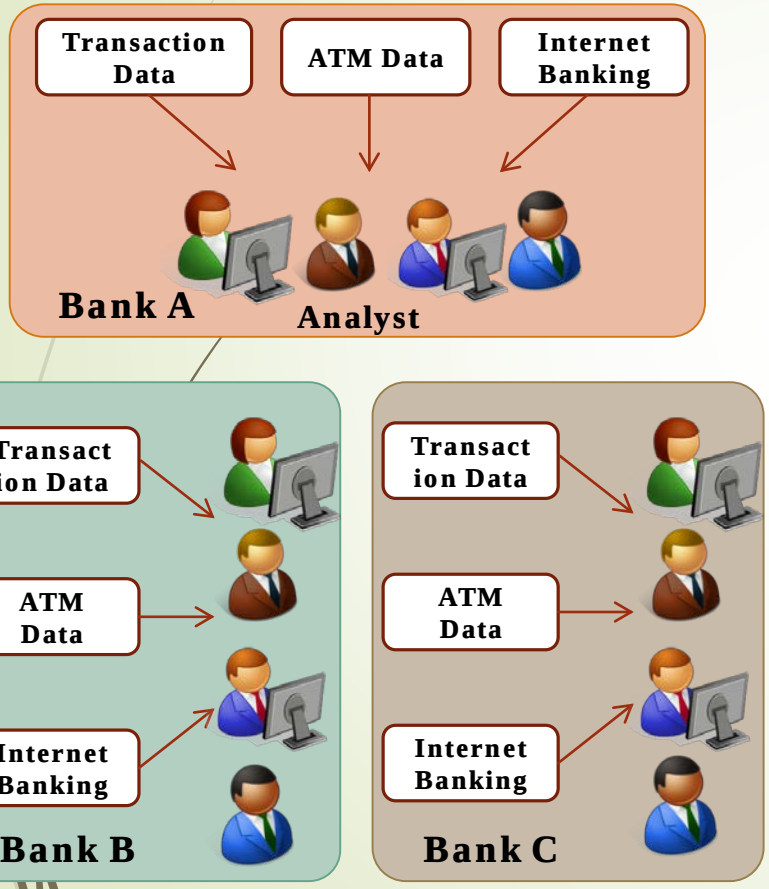


Machine Learning for Privacy-Preserving Data Mining (PPDM)

Darknet Monitoring for Detection of DDoS Attacks and Probing

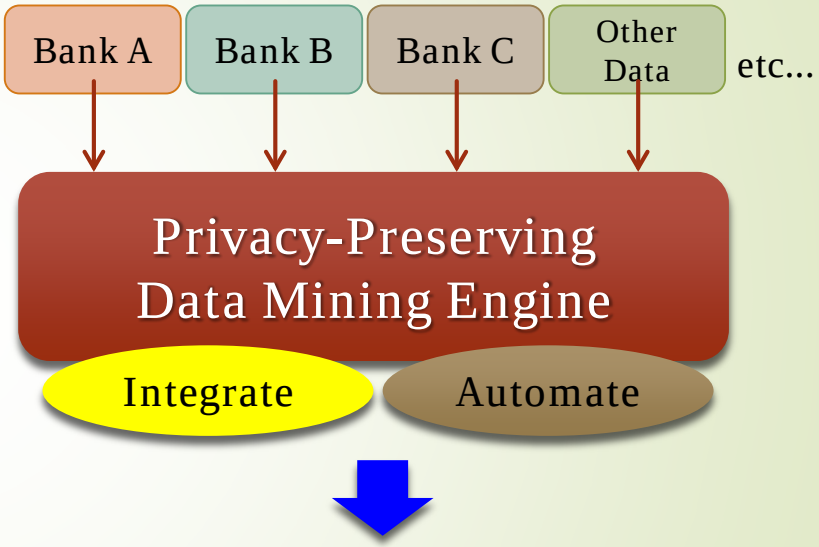
Mining from Sensitive Data

Current Approach



Individual Analysis

New Approach



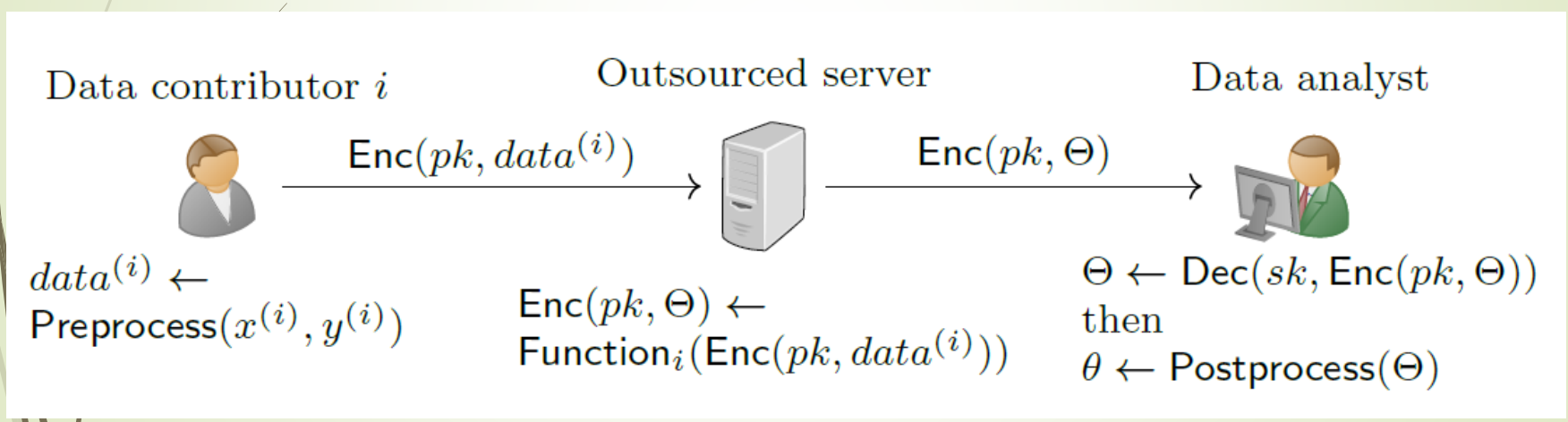
- Detection of illegal money transfer
- Calculate proper interest rate

Machine Learning over Encrypted Data

Privacy-Preserving Platform on Cloud Computing

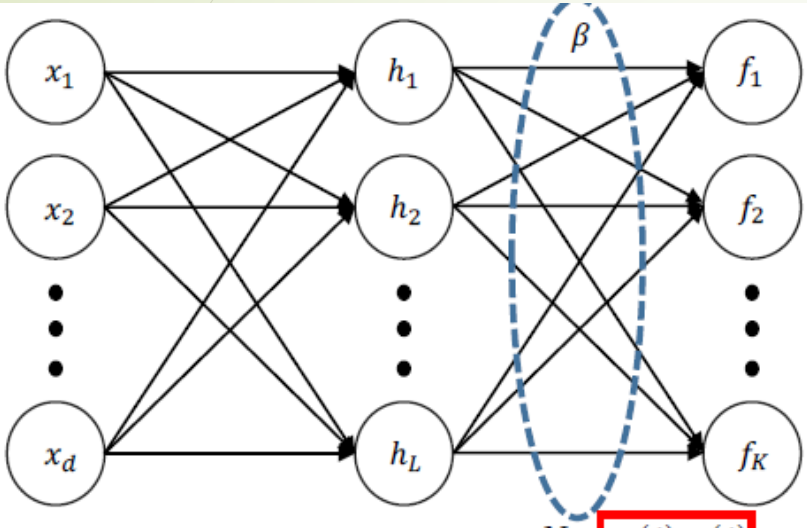
Additively Homomorphic Encryption

$$m_1 \cdot m_2 = \mathsf{Dec}(sk, \mathsf{Enc}(pk, m_1) \odot \mathsf{Enc}(pk, m_2)),$$



Privacy Preserving Extreme Learning Machine

Privacy Preserving Extreme Learning Machine



$$\beta = \left(\frac{\mathrm{I}}{\lambda} + H^T H\right)^{-1} H^T Y \qquad (N \gg L)$$

(N : *The number of data records,*
L : *The number of hidden nodes*)

$$H^T H = \begin{bmatrix} \sum_{i=1}^{N} h_1^{(i)} h_1^{(i)} & \cdots & \sum_{i=1}^{N} h_1^{(i)} h_L^{(i)} \\ \vdots & \cdots & \vdots \\ \sum_{i=1}^{N} h_L^{(i)} h_1^{(i)} & \cdots & \sum_{i=1}^{N} h_L^{(i)} h_L^{(i)} \end{bmatrix}$$

$$H^T Y = \begin{bmatrix} \sum_{i=1}^{N} h_1^{(i)} y_1^{(i)} & \cdots & \sum_{i=1}^{N} h_1^{(i)} y_K^{(i)} \\ \vdots & \cdots & \vdots \\ \sum_{i=1}^{N} h_L^{(i)} y_1^{(i)} & \cdots & \sum_{i=1}^{N} h_L^{(i)} y_K^{(i)} \end{bmatrix}$$

Performance Evaluation

Data Sets:
4 Bench Mark Datasets in Machine Leaning Repository

Encryption: LWE base Homomorphic Encryption

Datasets	PP-ELM L= 300	PP-Logistic ovr	Logistic ovr
Glass	0.684 +/- 0.089	0.596 +/- 0.099	0.604 +/- 0.070
Digits	0.965 +/- 0.021	0.889 +/- 0.037	0.925 +/- 0.027
Sattelite	0.875 +/- 0.007	0.758 +/- 0.019	0.827 +/- 0.018
Shuttle	0.997 +/- 0.001	0.873 +/- 0.002	0.933 +/- 0.002

(L：#Hidden Units)

+0.04～0.12