ANTI MALWARE ENGINEERING WORKSHOP

MWS

MWS2018意見交換会（プレミーティング）

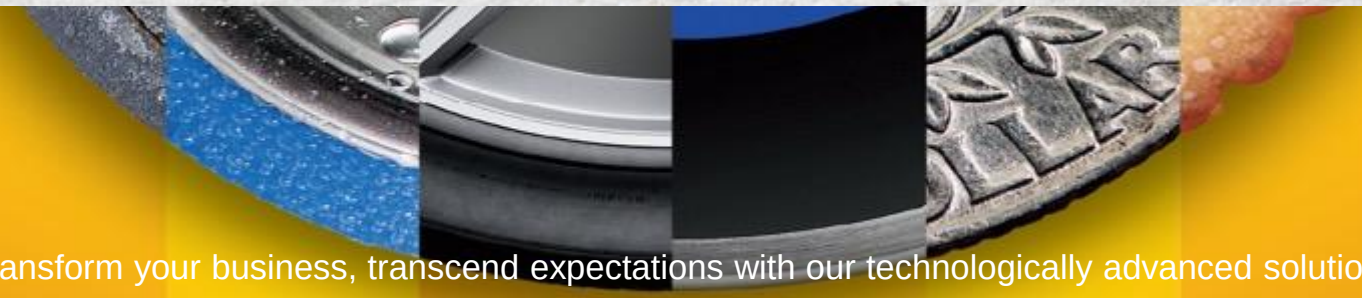
ntt.com

NTT Communications

Transform. Transcend.

マルウェア対策研究のデータセットで考慮すべきこと

2018年 5月 30日
NTTコミュニケーションズ株式会社
畑田充弘
m.hatada@ntt.com



Transform your business, transcend expectations with our technologically advanced solutions.

お題

- 評価のために使うデータセットに必要な要件って何だろう？
- どうやって準備すればいいんだろう？

といったことを自身の研究事例を交えながら紹介します

大規模データを活用した研究例

アクティブスキャンによるC2発見 – CCS '14

Zhaoyan Xu、他
"AUTOPROBE: Towards Automatic Active Malicious Server Probing Using Dynamic Binary Analysis"

- **各マルウェアファミリーのC2サーバを、アクティブスキャンするためのフィンガープリント自動生成**
  - **主要な56ファミリー389検体から105プローブ生成**
  - **悪性ドメインリストの9,500IPアドレスをもとに、28ファミリーについて260万の近隣アドレスを調査**

PUA流通調査 – USENIX ’16

Platon Kotzias、他

“Measuring PUP Prevalence and PUP Distribution through Pay-Per-Install Services”

- **世界中のSymantec製品による検知イベントをもとにした大規模調査**
  - **調査対象の390万ホストのうち54%にPUAがインストール、65%のPUAはさらに他のPUAをインストール**
  - **25%のPUAは23種類のPPIサービスによって流通**

大規模Mirai調査 – USENIX '17

Manos Antonakakis、他

"Understanding the Mirai Botnet"

- **多数の組織のデータをもとに、Miraiの発生、進化の過程、感染手法と対象機器などを明らかにし、発展するIoT社会に向けた警鐘**
  - **3700億パケット、2.9億DNSログ/日、27.8万攻撃元IP**

マルウェアドメインの長期分析 – S&P ’17

Chaz Lever、他

“A Lustrum of Malware Network Communication: Evolution and Insights”

- **マルウェア出現後、早ければ数ヶ月、少なくとも2週間前から悪用されるドメインはアクティブ**
  - **5年間で2,680万のマルウェアが名前解決したドメインと、VTやpDNS、公開BLの膨大なデータを分析**

漏洩した認証情報の追跡調査 – CCS '17

Kurt Thomas、他

"Data Breaches, Phishing, or Malware? Understanding the Risks of Stolen Credentials"

- **漏洩した認証情報と被害にあったGoogleアカウントを事例として1年間に渡る実態調査を行い、パスワード管理と2要素認証の教育を強化すべきと示唆**
  - **漏洩した認証情報19億件、認証情報を窃取するフィッシング被害380万件、etc.**

スピアフィッシング検知 – USENIX ’17

Grant Ho、他

“Detecting Credential Spearphishing Attack in Enterprise Settings”

- **誘うための特徴と攻撃に利用するURLの特徴を利用して、教師なし学習による独自のスコアリングで、低FPなスピアフィッシング検知を実現**
  - **5,000人規模のネットワークにおける4年間の3億7千万件のメール（SMTPヘッダ、本文中のURL）と、NIDSログ、LDAPログを利用**

（私の研究）

Mitsuhiro Hatada and Tatsuya Mori, “Finding New Varieties of Malware with the Classification of Network Behavior,” IEICE Transactions on Information and Systems, vol.E100-D, no.8, pp.1691-1702, August 2017.

マルウェアの通信パターンによる分類をもとに、新種を発見
（約28,000のマルウェア）

DNSクエリのみでAndroid PUAを識別し、亜種を分類
（約45万のアプリ）

Mitsuhiro Hatada and Tatsuya Mori, “Detecting and Classifying Android PUAs by similarity of DNS queries,” Proceedings of the 7th IEEE Inter- national COMPSAC Workshop on Network Technologies for Security, Administration and Protection (NETSAP 2017), pp.590-595, July 2017.

データセットがもたらす価値

信頼性（一般化）

- 検知、解析技術の評価結果の信頼性を高める
  - アクティブスキャンによるC2発見
  - スピアフィッシング検知
  - （私の研究）

- 調査で明らかにした事象の信頼性を高め、対策を促す
  - 大規模Mirai調査
  - マルウェアドメインの長期分析
  - PUA流通調査
  - 漏洩した認証情報の追跡

新規性

- データセットそのものが新規であり、有用なデータを提供することで他の研究にも貢献する
  - KDD Cup 1999 Data
    - http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
  - DARPA Intrusion Detection Data Sets
    - https://www.ll.mit.edu/ideval/data/
  - MWS Datasets
    - https://www.iwsec.org/mws/2017/about.html
- データの収集方法を提案、ツールを公開するものも
  - Zmap - USENIX ’13
    - https://github.com/zmap/zmap

データセットをどう準備するか？

パターン

- データセット First、研究ネタ Second
  - 所属組織（企業、研究室、etc.）または個人で保有している
  - せっかくあるので何かできないか？

- 研究ネタ First、データセット Second
  - こういう研究をしたい！
  - そのためにはこういうデータセットが必要
    - 探す
    - 作る

考えないといけないこと

- 何を評価するか？
  - 研究目的そのもの（例：マルウェアを検知したい）
- どう評価するか？
  - ノイズ除去、サンプリング、ラベリング、分割
- データを収集・蓄積・処理するシステムの開発・運用
  - インターネット接続環境
  - 監視
  - ストレージコスト（生データ vs メタデータ）
- 倫理

参考）P.10の研究における倫理

- 流出データを用いた研究そのものに倫理的な課題あり
- 論文において”Ethics”の節を設け下記を説明
  - 悪行に加担していないこと
  - 脆弱な状態のユーザを守ること
  - 他サイトで認証を試してないこと
  - 検知はあくまで機械的に行っていること

何を評価するか？

Mitsuhiro Hatada and Tatsuya Mori,

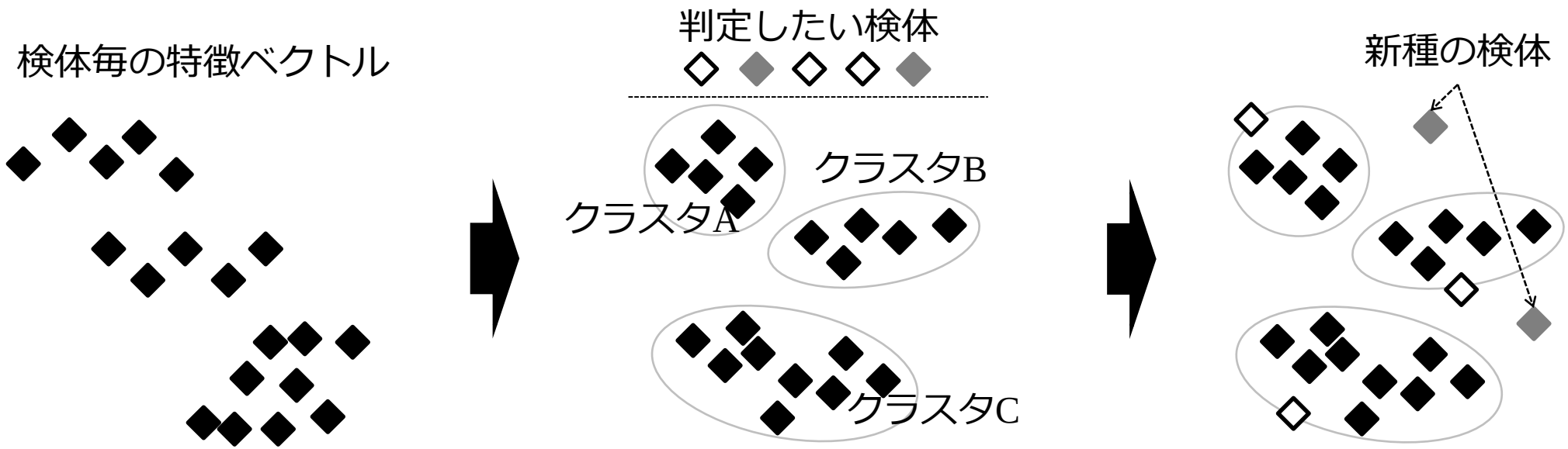
``Finding New Varieties of Malware with the Classification of Network Behavior,"

IEICE Transactions on Information and Systems, vol.E100-D, no.8, pp.1691-1702, August 2017.

- マルウェアのネットワーク挙動の分類をもとに新種のマルウェアを発見したい
  - 分類の精度
  - 新種発見の精度

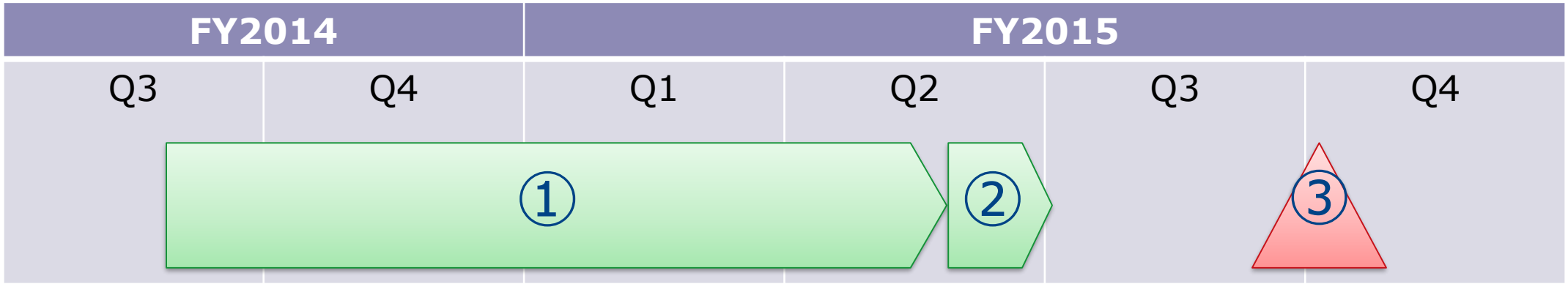
問題点とアプローチ

- アンチウイルス（AV）で検知できないマルウェアが多数
  - 分類や新種発見のための正解ラベル管理は現実的に困難
- 新種の詳細解析に優先度を上げたい
  - AVで検知できないもの、かつ挙動が異なるもの
    - パッキングなどで検知を逃れても実体が既知のものは既知と扱う

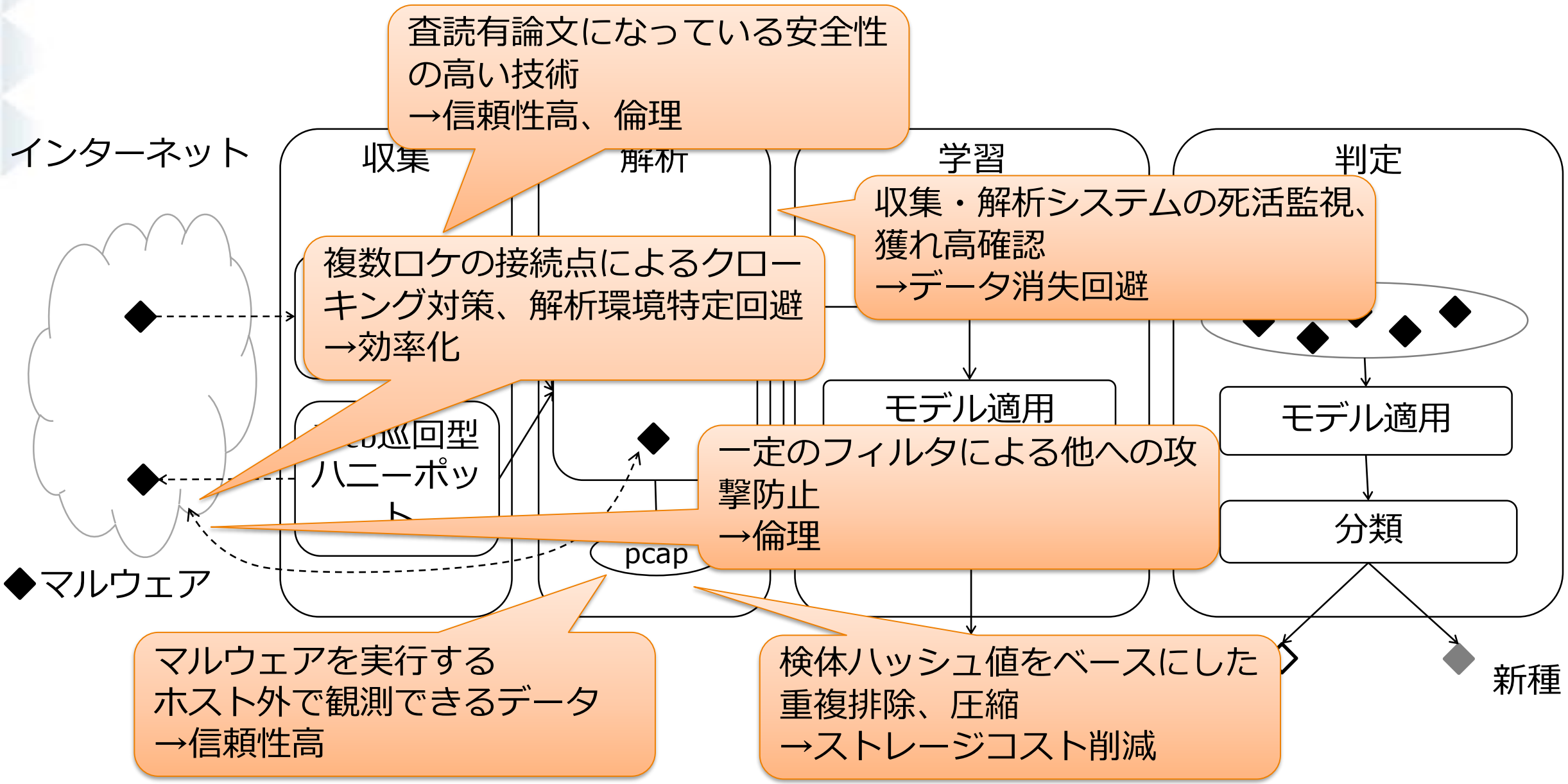


どう評価するか？

- ①学習用・②テスト用に異なる時期で収集
  - ①21,717検体（641種類）
  - ②6,078検体（237種類）
- 収集時点から3ヶ月以上後に③AV検知
  - シグネチャ更新によるAV検知率を高めるため
- ①には無くて②のみに存在する検知名の検体＝新種



システム



マルウェアの通信モデル（特徴ベクトル）

- 解析事例と従来研究から、マルウェアらしい特徴（25種類↓）と一般的な特徴（70種類）を定義

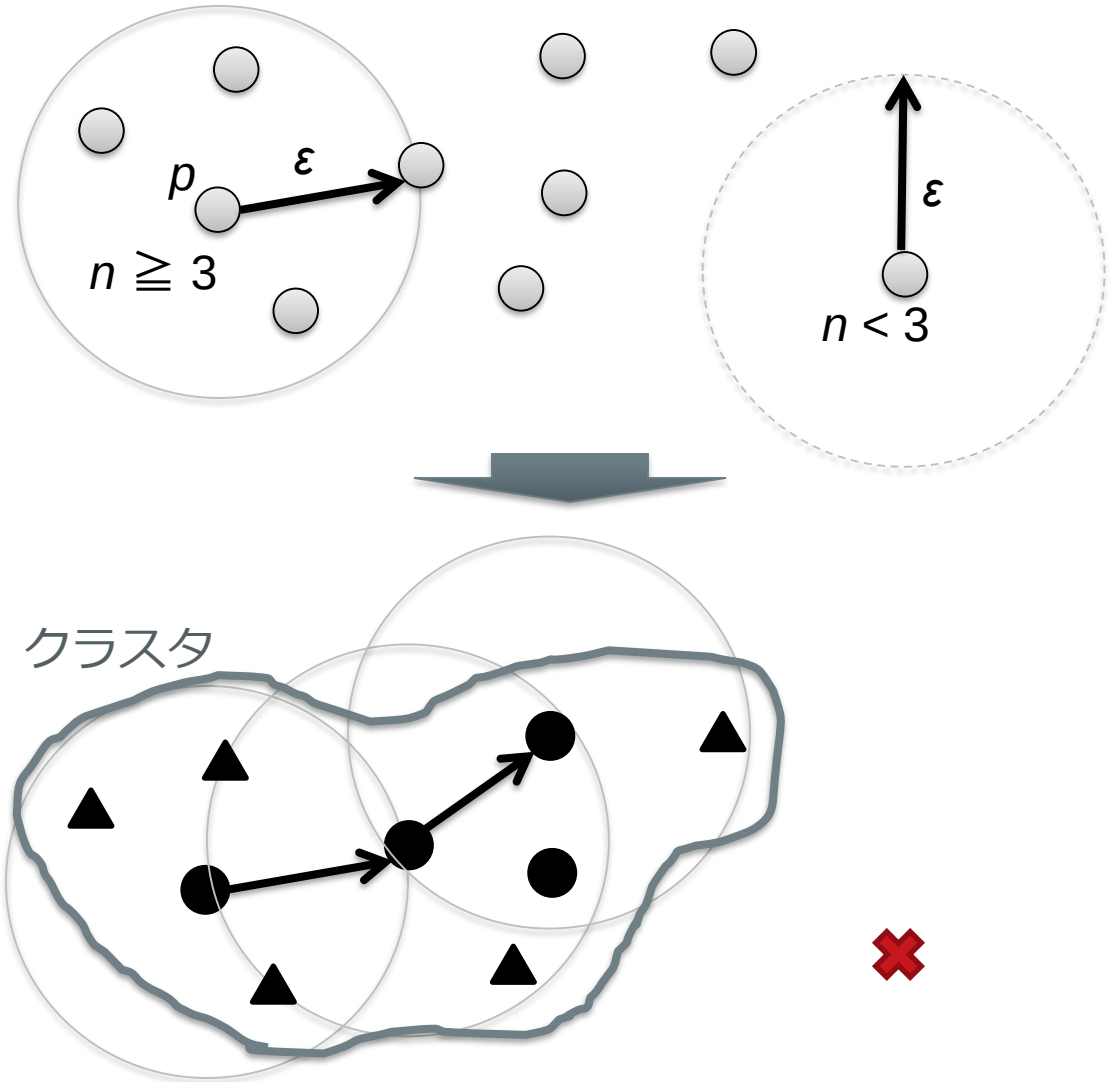
Class	Feature
Activation timing	Time to start communication after binary execution. [sec]
Checking	Number of DNS queries and Number of HTTP requests for major web sites. [27] Number of DNS queries and number of HTTP requests for global IP address check sites.
Information theft	Number of HTTP requests contained inherent host information [19]
C2 (Command and Control)	Number of sessions matched with [illegible] 5 public blacklists. [illegible]RC ses[illegible]ver TCP and UDP [28]. [illegible]essions over TCP and UDP [29]. [illegible]ssions with sinkhole hosts [30].
downloading	[illegible]wnloaded executable files.
Spam/scam email sending	Number of DNS queries for MX. Number of SMTP sessions [16, 25, 27].
Probing	Number of ICMP echo requests to the internal host and external hosts [31].
Monetizing	Number of HTTP requests contained advertising words.
Verification	Number of distinct HTTP User-Agent. Maximum and minimum length of HTTP User-Agent [25].

動的解析システムの固有情報のメンテナンス

収集期間に定期的に取得してないと特徴量にできない(>_<)

クラスタリング（教師なし学習） - DBSCAN

- **クラスタは、コアサンプルと非コアサンプルの極大集合**



● コアサンプル:
半径ε以内にnサンプルあるサンプル(p)

▲ 非コアサンプル:
コアサンプルではないが、コアサンプルから半径ε以内にあるサンプル

✖ エラー（ノイズ）

分類、新種発見

- 「最もユークリッド距離が近いサンプルが属するクラスタ」にテストデータ（判定したい検体）を分類
  - クラスタは複数の検知名を含むため、一意に代表検知名を決めずに複数の検知名を正解
    - クラスタ内の上位R%の検知名

- 「どのクラスタからも閾値Θ以上離れているサンプル」を新種検体として抽出

結果

- 分類
  - 77.4%（R=0.5、θ=3.0）

- 新種発見
  - 649/(649+106)=85%を新種として抽出
    - 検知名だと90/120
    - 15%はほとんど特徴のない既知クラスタ
    - （AV検知できた検体じゃないと確認しようがないので）検知できるものに限定

データセットは
研究そのものといっても過言ではない！

おわり