



マルウェア対策のための 研究用データセット ～MWS Datasets 2018～

高田 雄太, 寺田 真敏, 松木 隆宏
笠間 貴弘, 荒木 粧子, 畑田 充弘
(MWS 2018 実行/企画委員)



はじめに

- 本発表では、マルウェア対策研究コミュニティである MWS が提供する研究用データセット
～MWS Datasets 2018～を紹介させていただきます。

- 目次
 - 背景
 - MWS について
 - MWS データセット 2018 の内容/利用
 - MWS の活動
 - おわりに



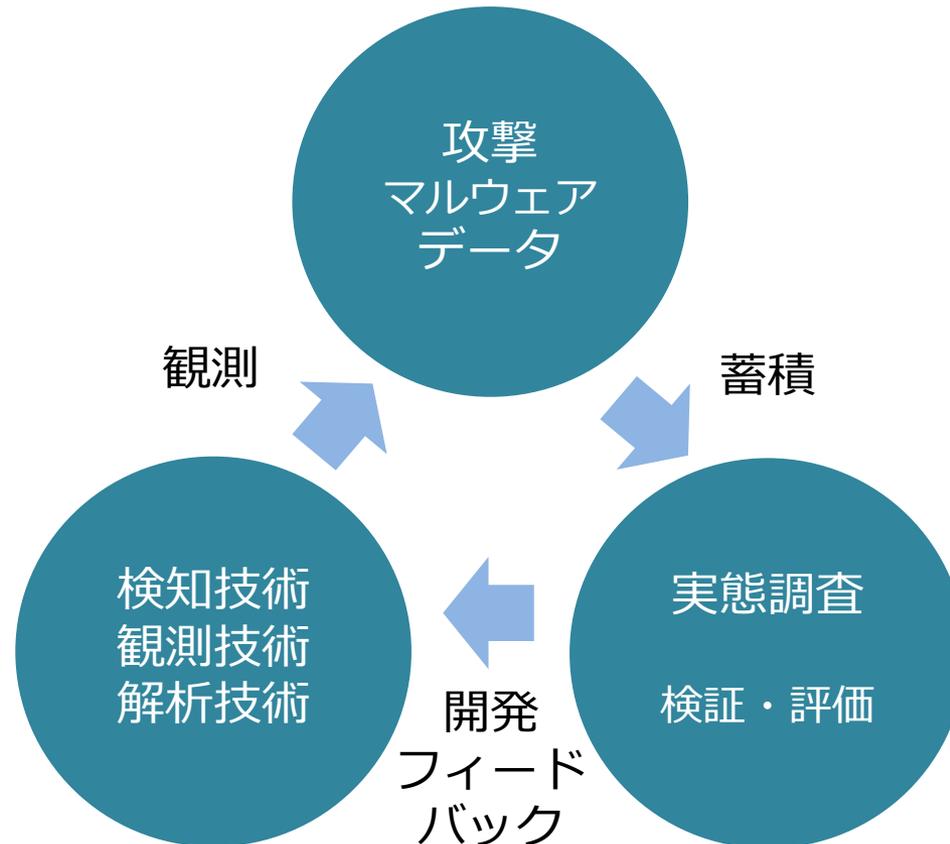
背景：複雑化するサイバー攻撃

- マルウェアを悪用したサイバー攻撃による脅威
 - Drive-by Download 攻撃
 - Advanced Persistent Threat (APT) 攻撃
 - ボットネットを利用した企業および国家間での DDoS 攻撃
 - IoT (Internet of Things) マルウェアからの攻撃 など
- マルウェア対策研究は盛んに行われているが、
攻撃の複雑化が進みサイバー攻撃の観測はより困難に



マルウェア対策研究

- 研究開発サイクルを加速させ、日々進化するサイバー攻撃に対抗
 - サイクルの循環を始めるには？ 加速させるには？

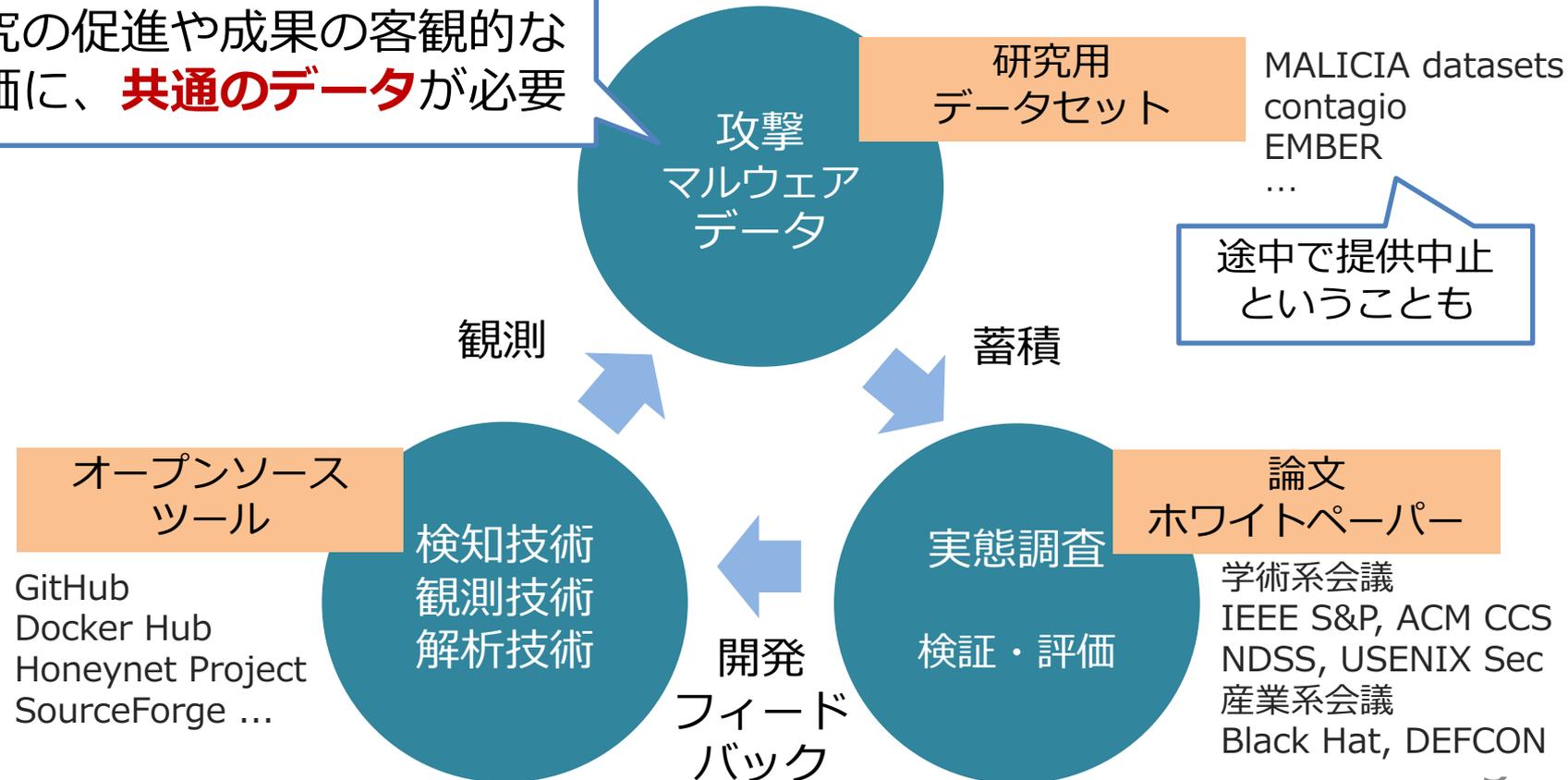




研究開発サイクルを加速させるために

- 各フェーズをサポートする情報やツールは充実化
 - 既存データセットは「継続性」や「網羅性」に欠けていたり、取得が困難であったり等の課題が存在

研究の促進や成果の客観的な評価に、**共通のデータ**が必要





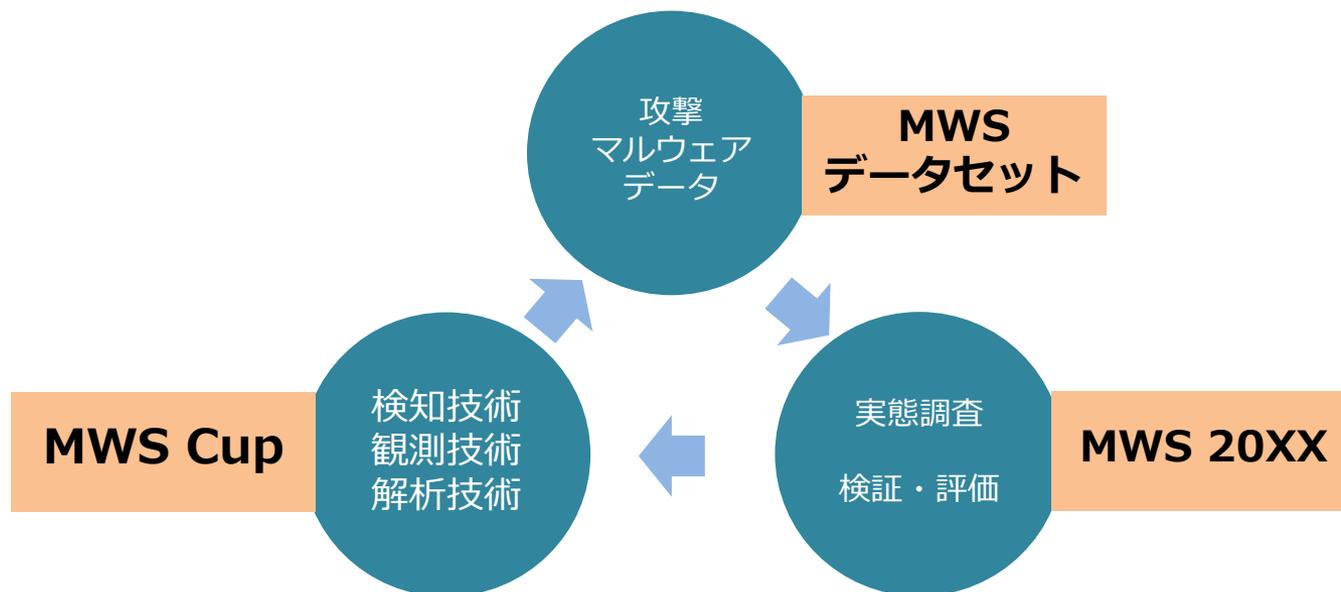
マルウェア対策研究人材育成 ワークショップ (MWS)

■ マルウェア対策研究コミュニティである MWS を組織

研究サイクルを継続的に回すことで研究活動を推進、研究開発した技術等の共有により人材育成を含む本研究分野の発展に寄与

- ✓ 研究用データセットの提供: **MWS データセット**
- ✓ 研究成果の共有: **MWS 20XX**
- ✓ 切磋琢磨する環境の提供: **MWS Cup**

本発表では**データセット**
を中心にご紹介





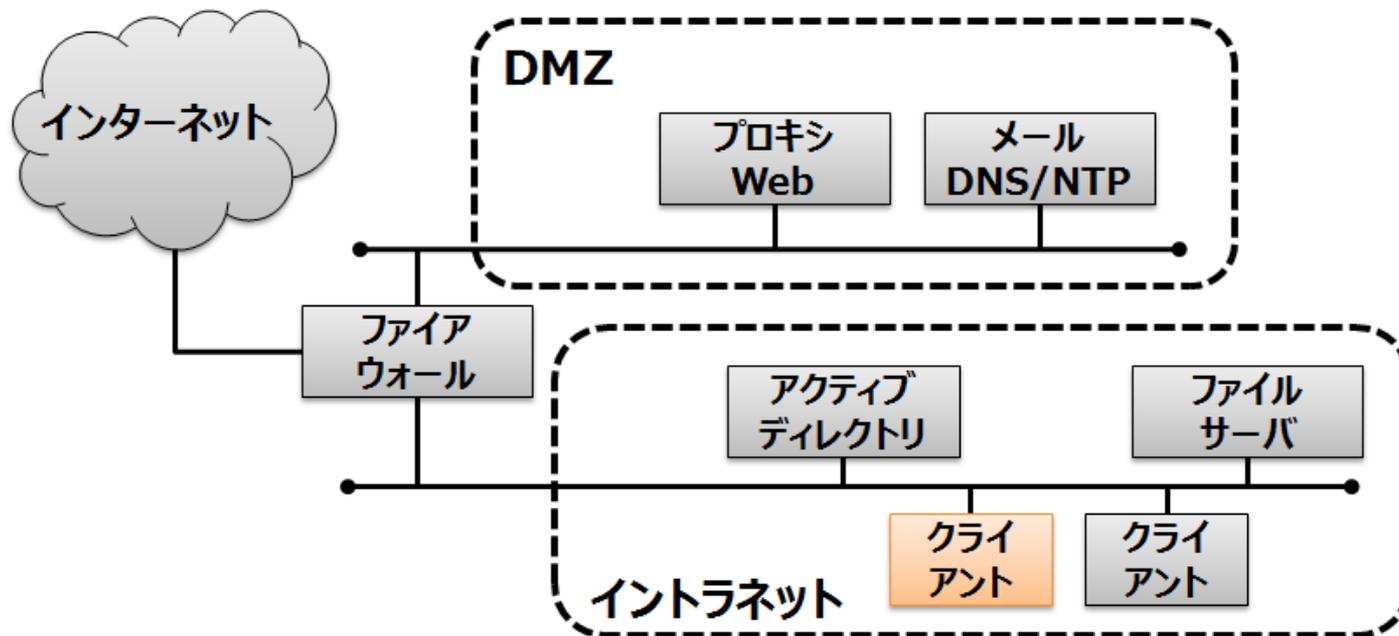
BOS Dataset 2018

■ 攻撃者行動視点で脅威を特徴付けるデータセット

- 攻撃者が標的組織内でどのような操作をしたのか、どのようなファイルにアクセスしたのかを監視可能

■ BOS の観測環境

- 組織内 NW を模擬した動的活動観測環境を構築





BOS Dataset の主な内容

■ マルウェア検体のハッシュ値

- 観測に使用したマルウェア検体のハッシュ値を STIX 形式 (Structured Threat Information eXpression; 脅威情報構造か記述形式) で記載したファイル

■ 通信観測データ

- マルウェア検体実行時の通信キャプチャデータ

■ プロセス観測データ

- マルウェア検体を実行したクライアントでのプロセスの稼働状況を記録したデータ

■ その他

- Windows のイベントログ、プロキシログ

※注：動的活動観測のケースごとに提供する観測データは異なる。



FFRI Dataset 2018

■ 2013 から 2017 まではマルウェアの動的解析ログ

- FFRI が収集したマルウェア検体の動的解析ログ

- ✓ 2013年:約2,600、2014年:約3,000、2015年:約3,000検体、
2016年:約8,000、2017年:約6,200



■ 2018 からはマルウェアの表層解析ログ

- ユーザアンケートを通じて、マルウェアの表層解析ログへ変更
- 悪性データ約29万検体、良性データ約21万検体
 - ✓ 良性データも提供することで、検知率 + 誤検知率の計算を可能に

FFRI Dataset の データ項目

- 1検体1行の csv ファイル
- 収集日
- 各種ハッシュ値
- 表層情報 (詳細は原稿参照)
 - 32bit/64bit
 - DLL or not
 - パッキング 有無
 - Anti-debug 有無
 - GUI/CUI
 - PEiD
 - ファイル種別
 - ヘッダダンプ

表 3 FFRI Dataset 2018 表層解析データ項目一覧

No.	項目名	概要
1	date	収集日 (マルウェアのみ)
2	MD5	ハッシュ値
3	SHA-1	ハッシュ値
4	SHA-256	ハッシュ値
5	ssdeep	ファジーハッシュ値
6	imphash	インポートテーブルから算出したハッシュ値
7	impfuzzy	インポートテーブルに ssdeep を適用したファジーハッシュ値
8	Totalhash	peHash の実装の一種
9	AnyMaster	peHash の実装の一種
10	AnyMaster_v1.0.1	peHash の実装の一種
11	EndGame	peHash の実装の一種
12	Crits	peHash の実装の一種
13	peHashNG	peHash の実装の一種
14	Platform	32bit または 64bit
15	GUI Program	GUI プログラムか否か
16	Console Program	Console プログラムか否か
17	DLL	DLL か否か
18	Packed	パッキングの有無
19	Anti-Debug	Anti-Debug の有無
20	mutex	mutex の有無
21	Contain base64	Base64 文字列の有無
22	Anti-DebugMethod	AntiDebug 手法 (デリミタは " ")
23	PEiD	マッチした PEiD シグネチャ名 (デリミタは " ")
24	TrID	ファイル種別推定結果



NICTER Dataset 2018

■ ダークネットトラフィックデータ

- /20 (約4,000アドレス) のダークネットトラフィック
- ダークネット = 未使用IPアドレス
 - ✓ 通常はダークネットにはトラフィックは届かない
- データ形式は pcap + DB
- 観測期間は2011年4月1日から現在までの **7年間 + α**

■ スпамメールデータ

- NICT のメールサーバに届いたダブルバウンスメール
- ダブルバウンスメール：
 - ✓ 送信元/宛先メールアドレスアカウントが存在しない場合に発生
 - ✓ エラーメールが二通やり取りされる
- データ形式はメールファイル
- 観測期間は2015年1月1日から現在までの **3年間 + α**



ダークネットの観測状況

10年間観測し続けていますが
基本的にずっと増加傾向です

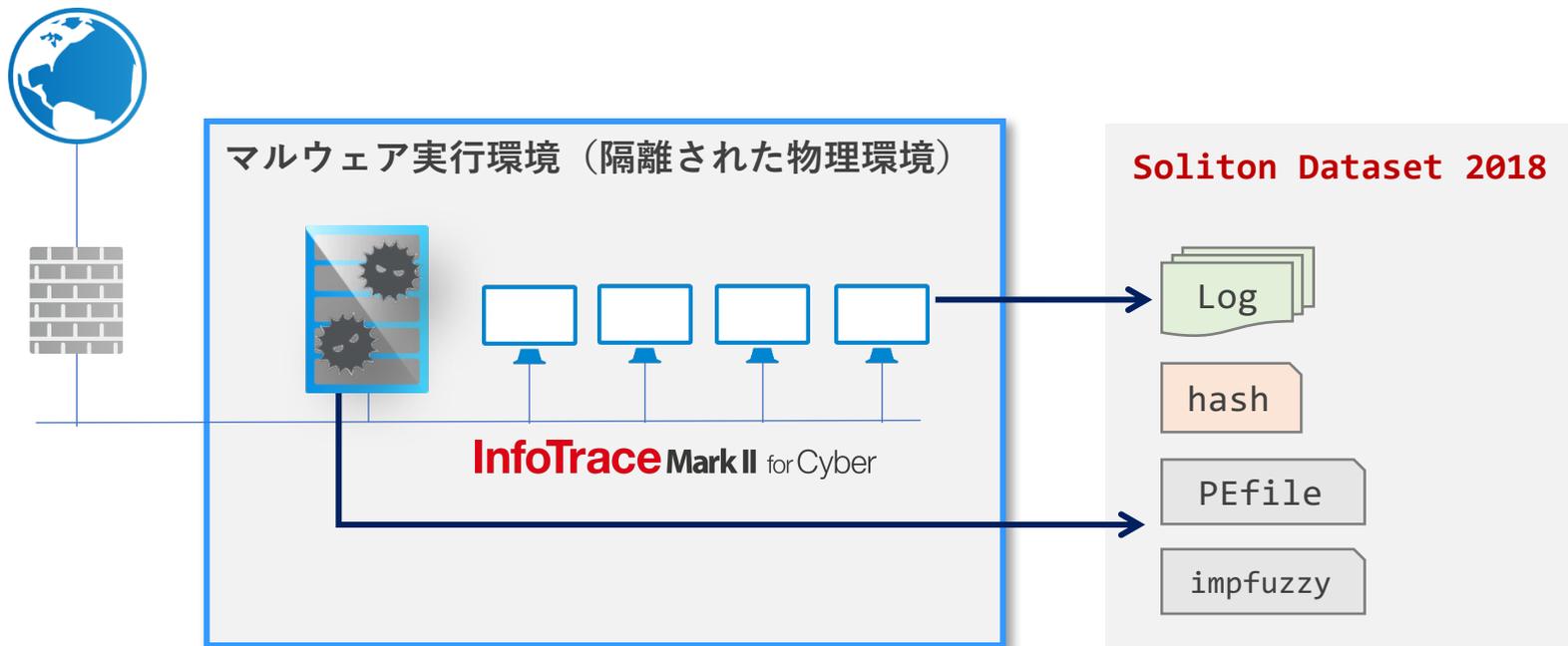




Soliton Dataset 2018

■ セキュリティログ取得製品導入環境におけるマルウェアの動的解析ログ

- Soliton が収集したマルウェア計117検体の動的解析ログ
- 2017年1月～2018年2月に話題になった OlympicDestroyer Petya, WannaCry 等のマルウェアが対象



Soliton Dataset の主な内容

■ 製品のログファイル

- Key=Value 形式の通常/詳細ログ
- 1 検体 30 分実行（延長含む）

■ sysinfo

- 各マルウェア実行時の環境情報

■ impfuzzy, PEfile

- 各マルウェア検体ごとの結果

■ その他

- ドキュメントやログ変換ツール



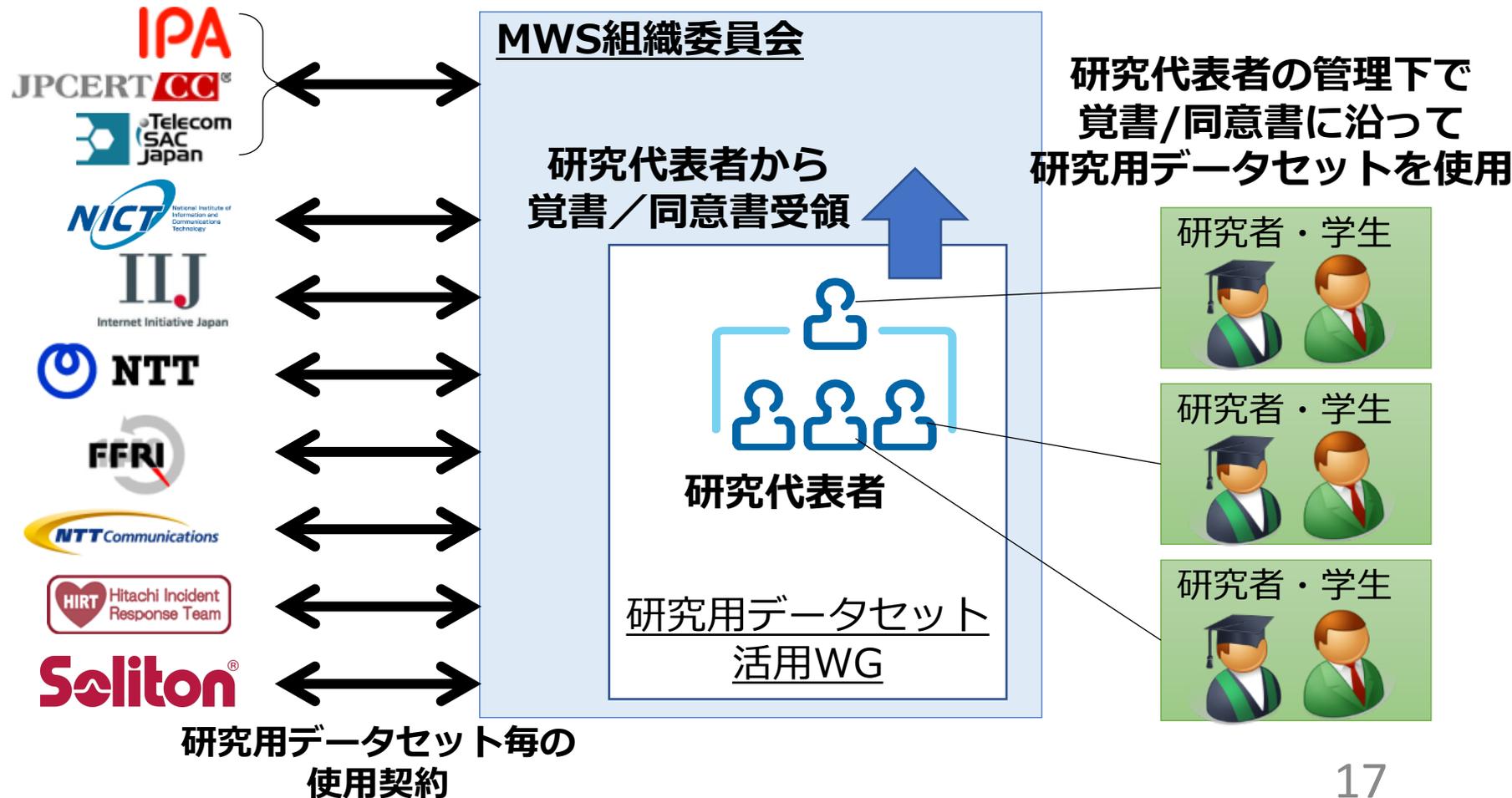
MWS Cup Dataset 2018

- 昨年の **MWS Cup 2017** に参加したチームが収集したデータセットも提供
- **Alkanet データセット**
 - システムコールトレサ “Alkanet” によるマルウェアの動的解析ログ
 - システムコールログと解析サマリー結果を含む
- **Jinkai Dataset 2017**
 - 独自ハニークライアントで収集した Drive-by Download 攻撃に関連するトラフィックデータ
 - 巡回 URL やスクリーンショット、pcap ファイルを含む



MWS データセットの利用

- 契約形態：MWS 組織委員会をハブとした利用手続き
 - 事務局 [csecreg@sdl.hitachi.co.jp] へコンタクト

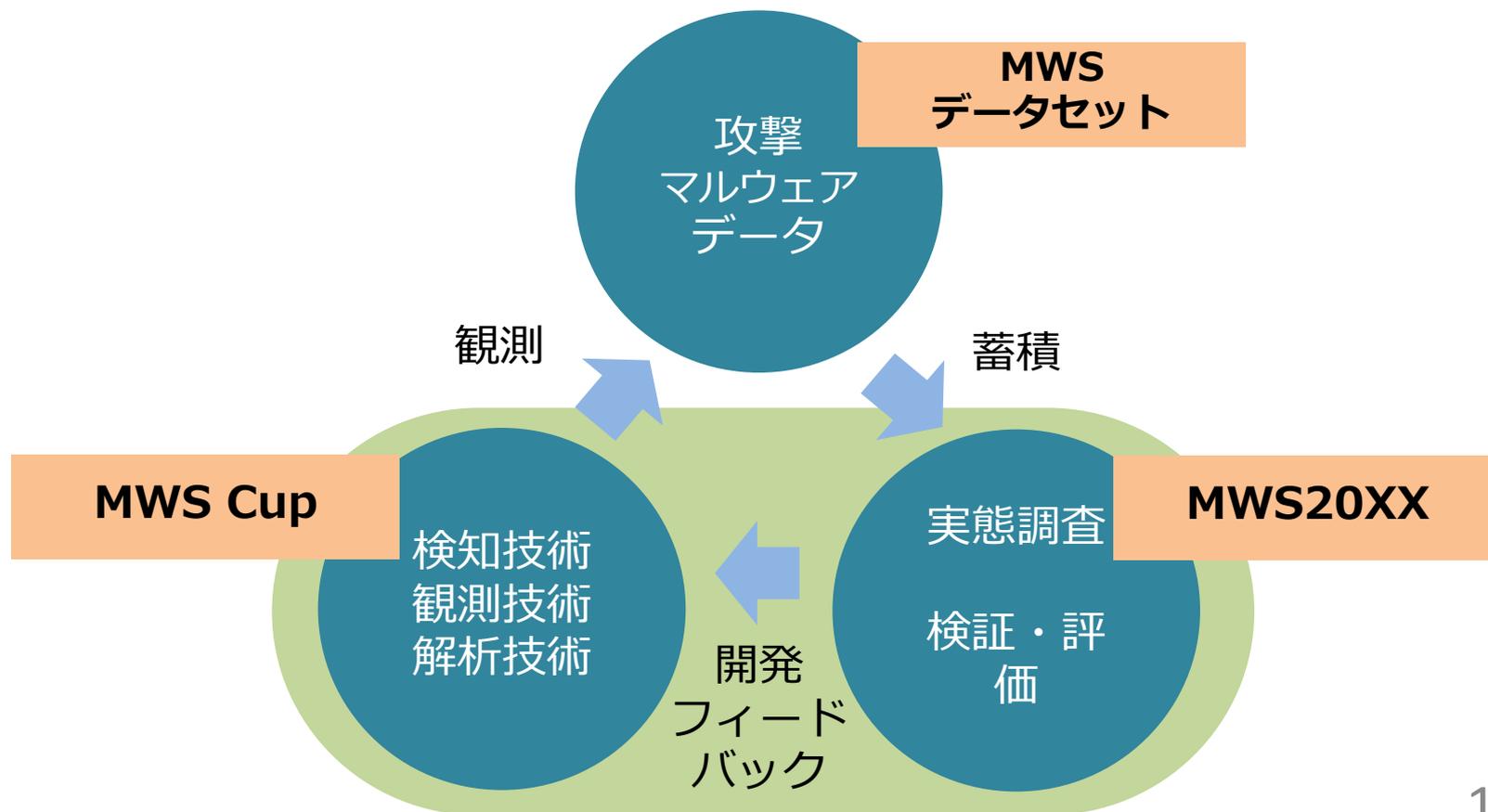




データセットを活用することで...MWS

■ 「技術」の「創出」および「検証・評価」を実施

- **MWS20XX**: 研究成果の共有 (論文の書き方、研究発表)
- **MWS Cup**: 切磋琢磨する環境 (実用的な技術やツールの発掘)





マルウェア対策研究人材育成 ワークショップ (MWS)

- **MWS20XX:** 研究者コミュニティが提供するデータセットを活用する産学官連携の学術系ワークショップ
 - **研究成果を共有する場**として2008年から開催
 - ✓ 攻撃解析、マルウェア解析、Android 解析、ダークネット解析とデータセットに関連する発表が多数
 - ✓ MWS2018 は、**2018年10月22日～10月25日長野県長野市**にて開催; <https://www.iwsec.org/mws/2018/>





MWS Cup

■ マルウェア対策に関するセキュリティコンテスト

- 日頃の研究で培ったノウハウやツール、データセットを基に創出した技術を活用しながら規定時間内で課題に取り組み、解析結果を競う「**切磋琢磨する場**」

<https://www.iwsec.org/mws/mwscup.html>

● 課題例

- ✓ Drive-by Download 攻撃解析
- ✓ マルウェア検体の静的解析
- ✓ マルウェア動的解析ログの分析 など

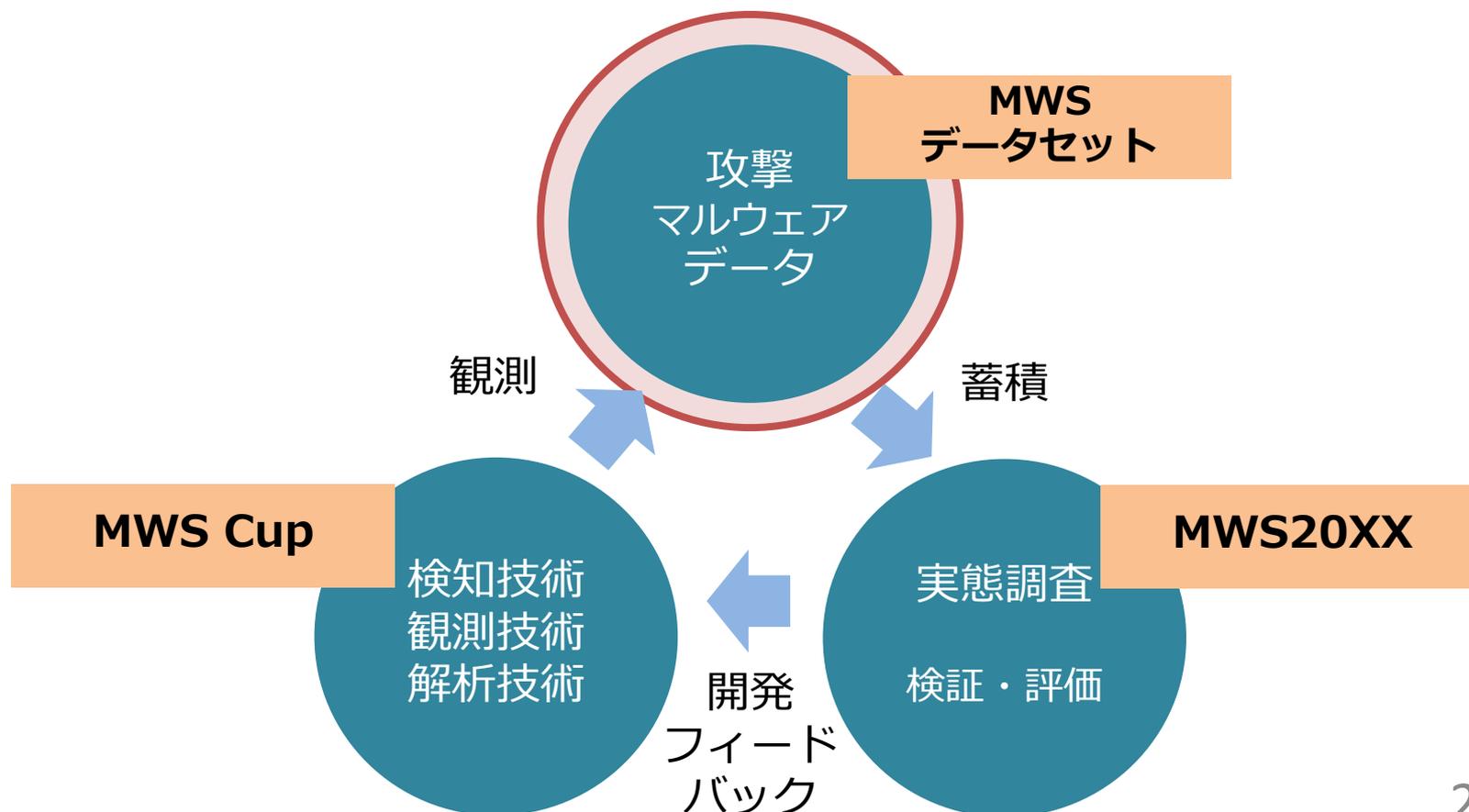
- 数時間の解析競技の後、**プレゼンも実施**





データセットの重要性

- 研究開発サイクルの加速に「データ」は重要
 - MWSでは「データセット」に加え「実用的な研究」にも価値があると考え、それらを適切に評価する仕組みを検討中





おわりに

- 複雑化するサイバー攻撃に対抗すべく、
マルウェア対策人材育成ワークショップ MWS では
MWS Datasets 2018 を提供中
 - 研究開発の推進／技術の共有により本研究分野の発展に寄与
 - MWS Datasets 2018 利用には、研究代表者の WG 参加とデータセット使用に関する契約が必要
 - ✓ MWS 組織委員会事務局
「csecreg@sdl.hitachi.co.jp」までご連絡を
- 宣伝
 - MWS 2018 は、**8/2 アブスト締切、8/20 原稿締切**
 - MWS では、**MWS Datasets へのデータ提供者**および
MWS Cup 参加者を随時募集中
 - ✓ <https://www.iwsec.org/mws/>
 - ✓ <https://www.iwsec.org/mws/mwscup.html>
 - ✓ 各データセットの説明資料も公開中
<https://www.iwsec.org/mws/2018/mws20180530.html>

參考資料



関連研究

■ IMPACT Dataset

- ネットワークデータ装置やセキュリティ装置，通信ログ等から得られるセキュリティ脅威に関するデータセット

■ MALICIA Dataset

- ドライブバイダウンロード攻撃を仕掛ける悪性ウェブサイトから収集したマルウェア検体のデータセット

■ Malware-Traffic-Analysis.net

- マルウェア感染およびエクスプロイトキットに関する通信データ

■ Contagio Malware Dump

- 各種ファイルフォーマットの正規ファイルおよび悪性ファイル

■ Android Malware Genome Project Dataset

- マルウェアファミリ毎に分類されたAndroid マルウェア検体

■ ACODE dataset

- Google Play とサードパーティマーケットから収集したAndroid アプリ20万個の説明文に関するデータセット



データセットを使用したい場合は？

- MWS データセットを使用するにあたって
 - 研究代表者の研究用データセット WG 参加とデータセットの使用に関する契約をお願いします。
 - 契約書に記載された**注意事項の遵守**（e.g., 各種情報の開示をしないこと）をお願いします。
 - ✓ その他問い合わせは「csecreg@sdl.hitachi.co.jp」まで
- MWS データセットを使用した研究論文を執筆する場合は、**本文献の引用**をお願いします。

高田雄太, 他: マルウェア対策のための研究用データセット
～ MWS 2018 Datasets ～,
情報処理学会, Vol.2018-CSEC-82, No.38, 2018年7月.