マルウェア対策のための研究用データセット
～MWS Datasets 2019 ～

荒木 粧子，**笠間 貴弘**, 押場 博光,
千葉 大紀，畑田 充弘，寺田 真敏
(MWS 2019 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 2019～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 2019 の内容/利用
  - MWS の活動
  - おわりに

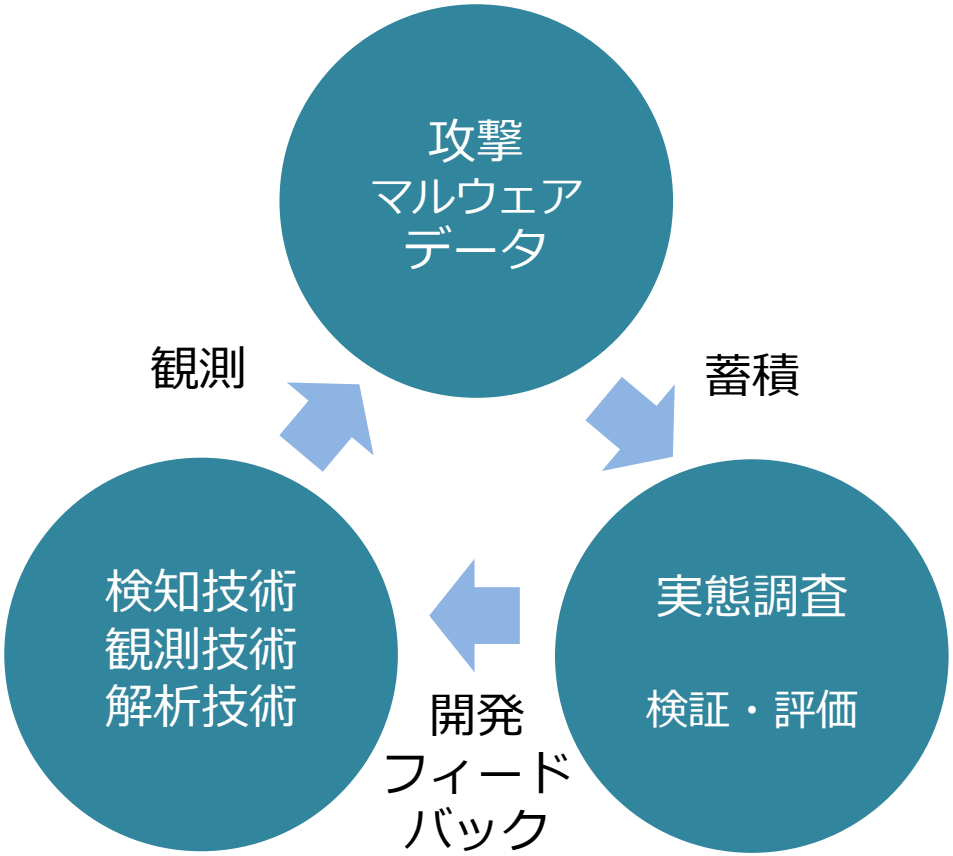
背景：複雑化するサイバー攻撃

- 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

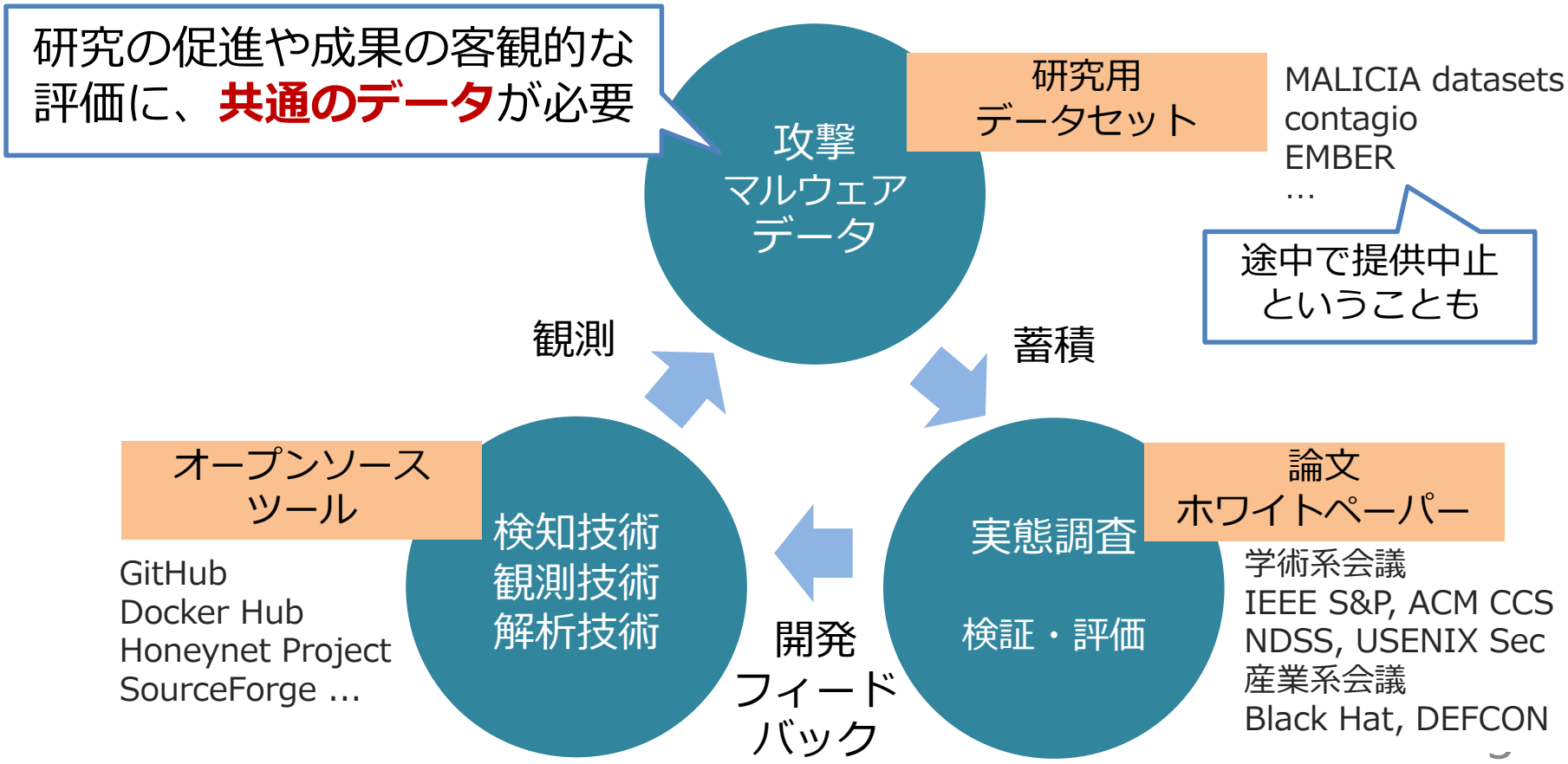
- 研究開発サイクルを加速させ、日々進化するサイバー攻撃に対抗
  - **サイクルの循環を始めるには？加速させるには？**



研究開発サイクルを加速させるために

ANTI MALWARE ENGINEERING WORKSHOP MWS

- 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**



マルウェア対策研究人材育成ワークショップ (MWS)

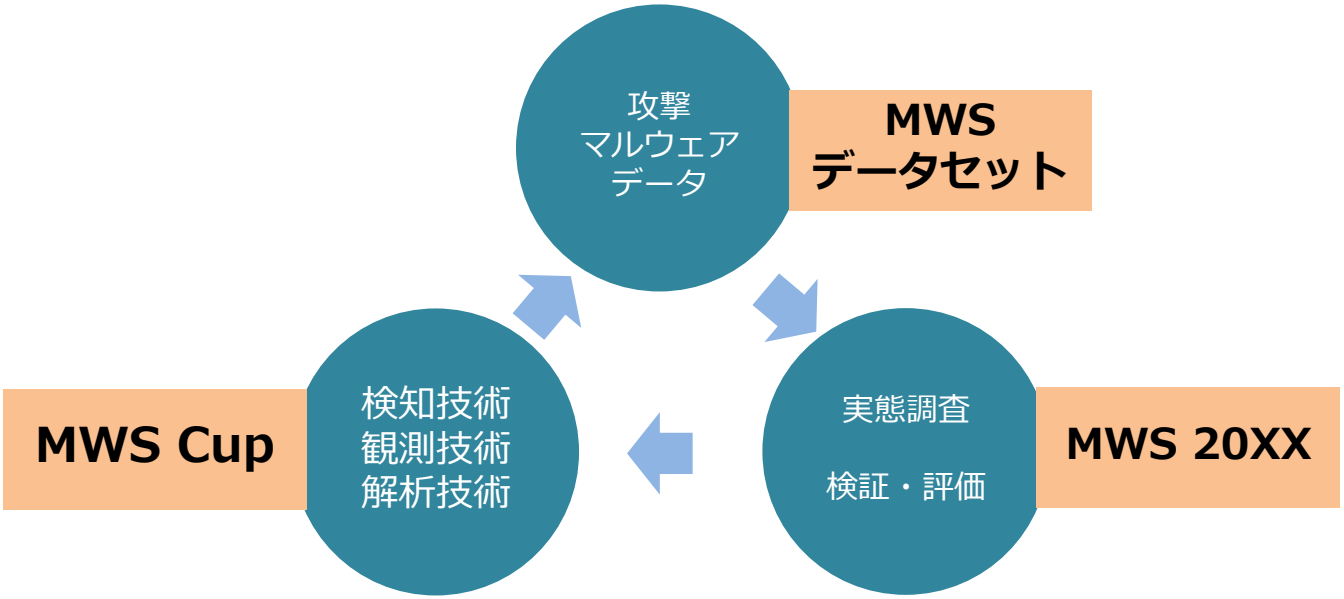
ANTI MALWARE ENGINEERING WORKSHOP
MWS

- 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**

本発表では**データセット**を中心にご紹介



MWS データセット 2019 の内容

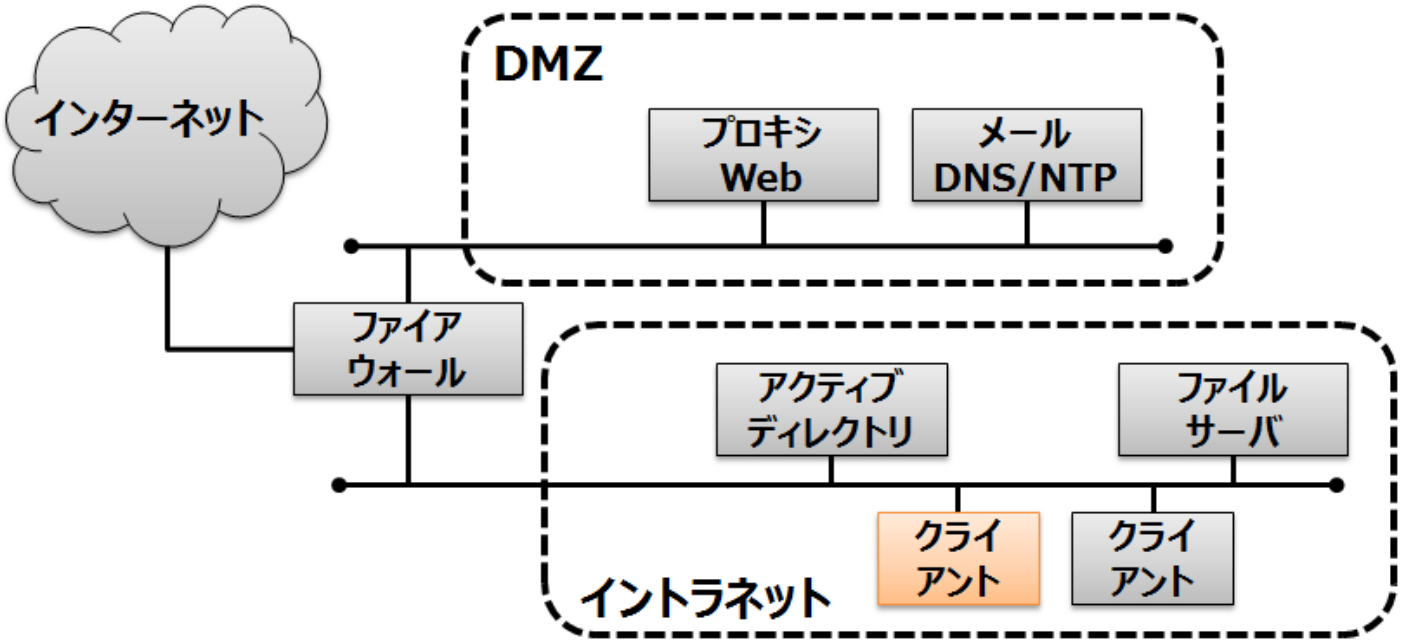
- 提供されるデータセットは１０種類
- 2019年に更新のあるデータセットは５種類

データセット名	09	10	11	12	13	14	15	16	17	18	19
ボット観測用攻撃通信/攻撃元/マルウェア検体 CCC DATAset (サイバークリーンセンター)	✓	✓	✓	✓	✓						
ウェブ感染型マルウェアデータ D3M Dataset (NTT)		✓	✓	✓	✓	✓	✓				
マルウェア感染後の通信データ PRACTICE Dataset					✓						
DRDoS 攻撃の観測データ PRACTICE (AmpPot) Dataset								✓			
マルウェア動的解析ログデータ FFRI Dataset (FFRI)					✓	✓	✓	✓	✓	✓	✓
ダークネットトラフィックデータ NICTER Dataset (NICT)					✓	✓	✓	✓	✓	✓	✓
攻撃者活動観測データ BOS Dataset (日立)						✓	✓	✓	✓	✓	✓
NCD in MWS Cup 2014(MWS) 一般的な通信を想定したデータ							✓				
マルウェア動的解析ログデータ Soliton Dataset (ソリトン)										✓	✓
MWS Cup Dataset										✓	✓

BOS Dataset 2019

- **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 2019

- **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
  - FFRI Dataset 2019では、悪性データ約25万検体、良性データ約25万検体
    - ✓ 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)
  - 検体のsha256ハッシュ値
  - ファイルサイズ
  - マルウェアか良性ファイルか(マルウェア: 1, 良性ファイル: 0)
  - 収集日
  - ハッシュ値（md5,sha1, sha256, ssdeep, impfuzzy, tlsh, anymaster, endgame, crits, pehashngのハッシュ値
  - PEiDによる表層解析結果
  - LIEFによる表層解析結果
  - TrIDによるファイル種別推定結果
  - Strings（検体中に含まれる文字列情報)

表 3　FFRI Dataset 2019 表層解析データ項目一覧

No.	項目名	概要
1	id	id （検体の SHA-256）
2	file_size	ファイルサイズ
3	label	マルウェアは 1，クリーンウェアは 0
4	date	収集日 (マルウェアのみ)
5	hashes	ハッシュ値．5.1 から 5.13 を含む
5.1	md5	ハッシュ値
5.2	sha1	ハッシュ値
5.3	sha256	ハッシュ値
5.4	ssdeep	ファジーハッシュ値
5.5	imphash	インポートテーブルから算出したハッシュ値
5.6	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
5.7	tlsh	ファジーハッシュ値
5.8	totalhash	peHash の実装の一種
5.9	anymaster	peHash の実装の一種
5.10	anymaster_v1.0.1	peHash の実装の一種
5.11	endgame	peHash の実装の一種
5.12	crits	peHash の実装の一種
5.13	pehashng	peHash の実装の一種
6	peid	シグネチャのスキャン結果．6.1 から 6.10 を含む
6.1	PE	32bit または 64bit
6.2	GUI Program	GUI プログラムか否か
6.3	Console Program	Console プログラムか否か
6.4	DLL	DLL か否か
6.5	Packed	パッキングの有無
6.6	Anti-Debug	Anti-Debug の有無
6.7	mutex	mutex の有無
6.8	contains base64	Base64 文字列の有無
6.9	AntiDebug	AntiDebug 手法
6.10	PEiD	マッチした PEiD シグネチャ名
7	lief	ファイル表層情報
8	TrID	ファイル種別推定結果
9	strings	文字列出力結果

マルウェア対策のための研究用データセット ～MWS Datasets 2019～正誤表

ANTI MALWARE ENGINEERING WORKSHOP MWS

表 3　FFRI Dataset 2019 表層解析データ項目一覧

No.	項目名	概要
1	id	id （検体の SHA-256）
2	file_size	ファイルサイズ
3	label	マルウェアは 1，クリーンウェアは 0
4	date	収集日 (マルウェアのみ)
5	hashes	ハッシュ値．5.1 から 5.13 を含む
5.1	md5	ハッシュ値
5.2	sha1	ハッシュ値
5.3	sha256	ハッシュ値
5.4	ssdeep	ファジーハッシュ値
5.5	imphash	インポートテーブルから算出した ハッシュ値
5.6	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
5.7	tlsh	ファジーハッシュ値
5.8	totalhash	peHash の実装の一種
5.9	anymaster	peHash の実装の一種
5.10	anymaster_v1.0.1	peHash の実装の一種
5.11	endgame	peHash の実装の一種
5.12	crits	peHash の実装の一種
5.13	pehashng	peHash の実装の一種
6	peid	シグネチャのスキャン結果．6.1 から 6.10 を含む
6.1	Platform	32bit または 64bit
6.2	GUI Program	GUI プログラムか否か
6.3	Console Program	Console プログラムか否か
6.4	DLL	DLL か否か
6.5	Packed	パッキングの有無
6.6	Anti-Debug	Anti-Debug の有無
6.7	mutex	mutex の有無
6.8	Contain base64	Base64 文字列の有無
6.9	AntiDebug	AntiDebug 手法
6.10	PEiD	マッチした PEiD シグネチャ名
7	lief	ファイル表層情報
8	TrID	ファイル種別推定結果
9	strings	文字列出力結果

訂正後

表 3　FFRI Dataset 2019 表層解析データ項目一覧

No.	項目名	概要
1	id	id （検体の SHA-256）
2	file_size	ファイルサイズ
3	label	マルウェアは 1，クリーンウェアは 0
4	date	収集日 (マルウェアのみ)
5	hashes	ハッシュ値．5.1 から 5.13 を含む
5.1	md5	ハッシュ値
5.2	sha1	ハッシュ値
5.3	sha256	ハッシュ値
5.4	ssdeep	ファジーハッシュ値
5.5	imphash	インポートテーブルから算出した ハッシュ値
5.6	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
5.7	tlsh	ファジーハッシュ値
5.8	totalhash	peHash の実装の一種
5.9	anymaster	peHash の実装の一種
5.10	anymaster_v1.0.1	peHash の実装の一種
5.11	endgame	peHash の実装の一種
5.12	crits	peHash の実装の一種
5.13	pehashng	peHash の実装の一種
6	peid	シグネチャのスキャン結果．6.1 から 6.10 を含む
6.1	PE	32bit または 64bit
6.2	GUI Program	GUI プログラムか否か
6.3	Console Program	Console プログラムか否か
6.4	DLL	DLL か否か
6.5	Packed	パッキングの有無
6.6	Anti-Debug	Anti-Debug の有無
6.7	mutex	mutex の有無
6.8	contains base64	Base64 文字列の有無
6.9	AntiDebug	AntiDebug 手法
6.10	PEiD	マッチした PEiD シグネチャ名
7	lief	ファイル表層情報
8	TrID	ファイル種別推定結果
9	strings	文字列出力結果

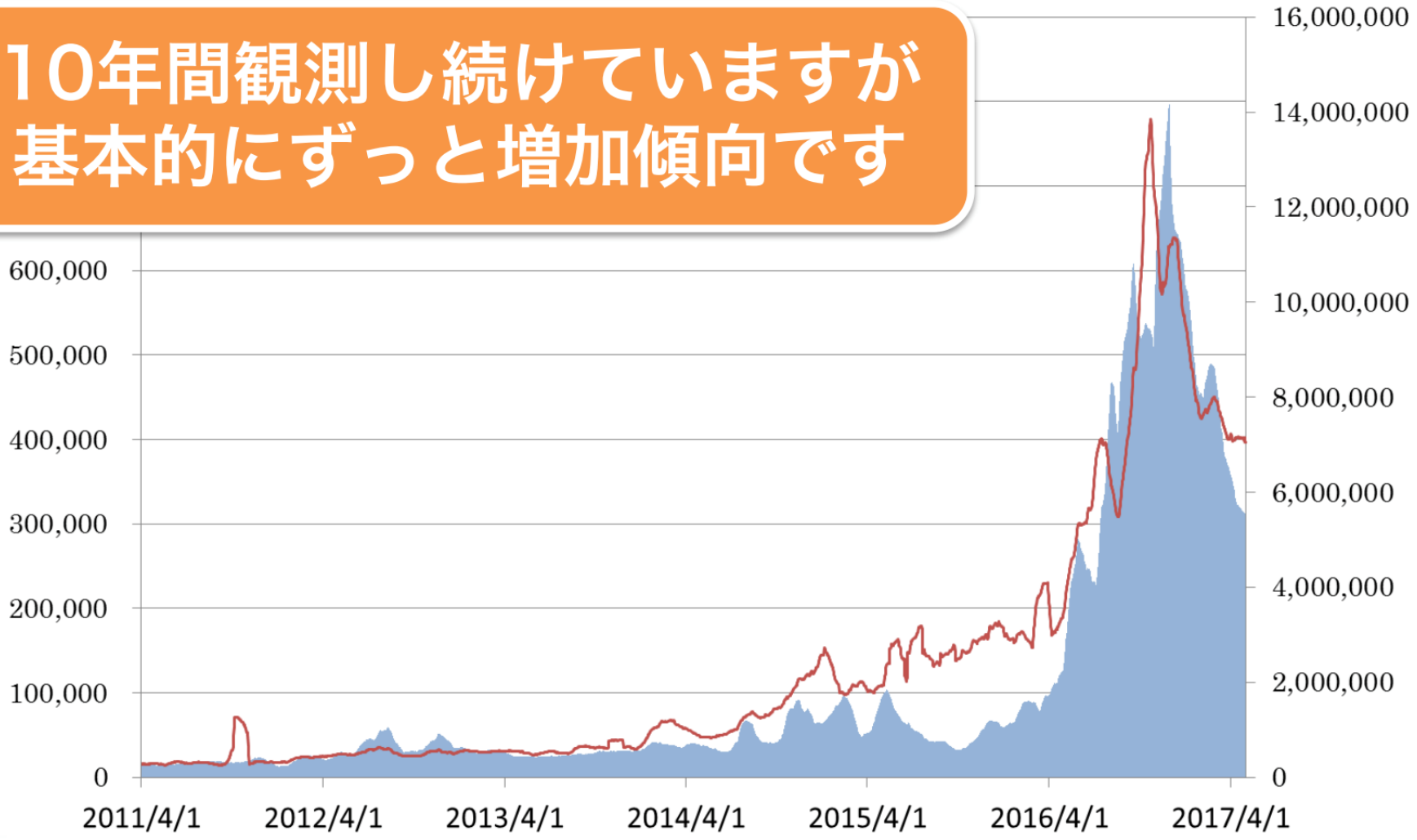
NICTER Dataset 2019

- **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ヒック
  - 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**8年間＋α**

- **スパムメールデータ（要望あれば）**
  - 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**4年間＋α**

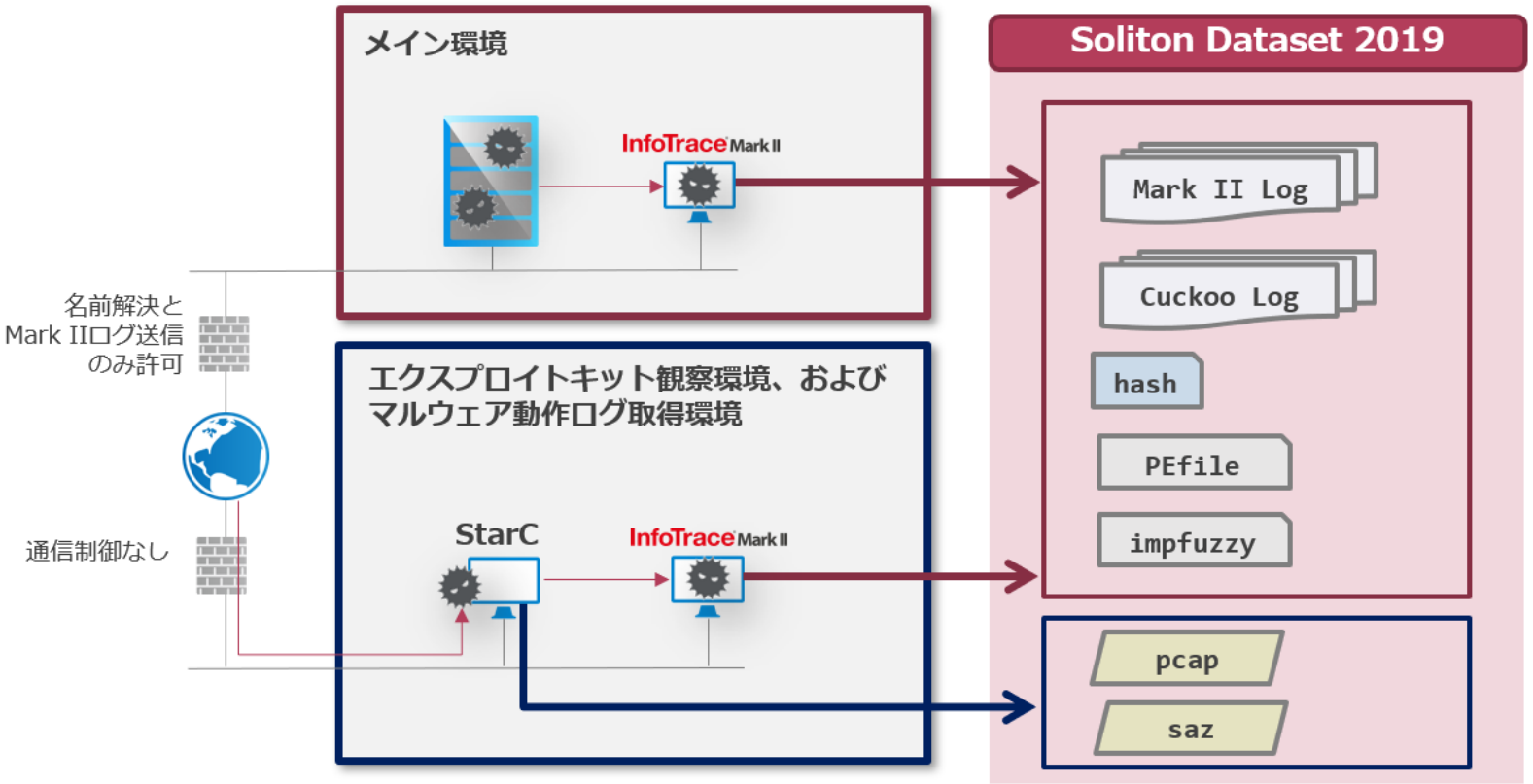
ダークネットの観測状況

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 2019

- **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**
  - 2018年に話題になったマルウェアの検体実行（485検体）
  - エクスプロイトキット観測＆入手検体を実行（3検体）



※セキュリティログ取得製品は、InfoTrace Mark II for Cyber のことです。

Soliton Dataset の主な内容

- **メイン環境**
  - **製品のログファイル**（Key=Value 形式）
  - **Cuckooログ**
- **Exploit Kit観測環境＆実行環境**
  - **製品のログファイル**（Key=Value 形式）
  - **Cuckooログ**
  - **saz/pcap(Exploit Kit観測環境で取得したデータ)**
- **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 2019

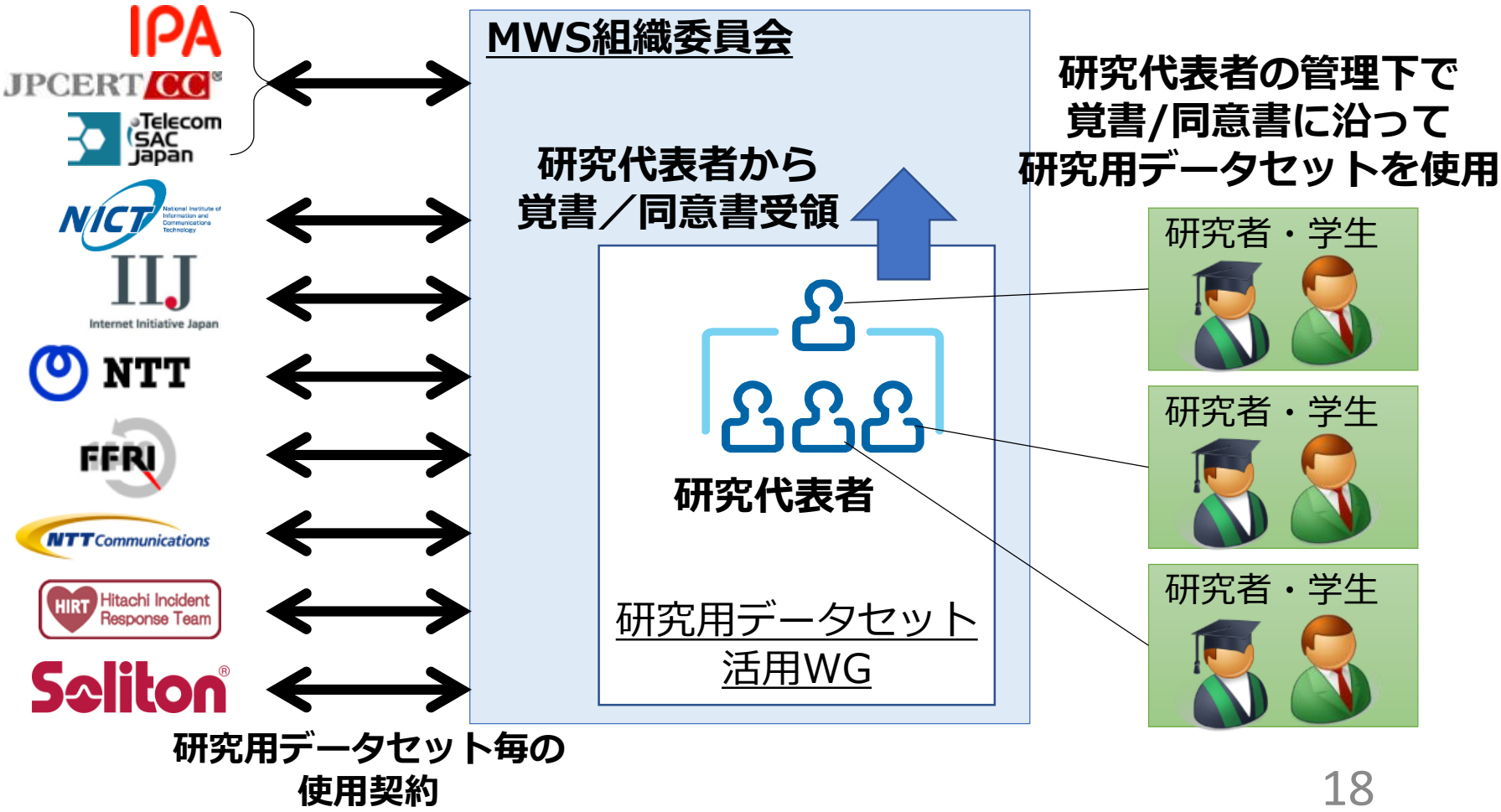
- 昨年の **MWS Cup 2018 に参加したチーム**が収集・作成したデータセットも提供

- **UN頼みデータセット**
  - Web ブラウザ拡張機能のデータセット
  - 作成に利用されたスクリプト

- **たこ焼きLabデータセット**
  - マルウェアの典型的な挙動を模擬するソースコード
  - 当該実行可能ファイルの動的解析結果

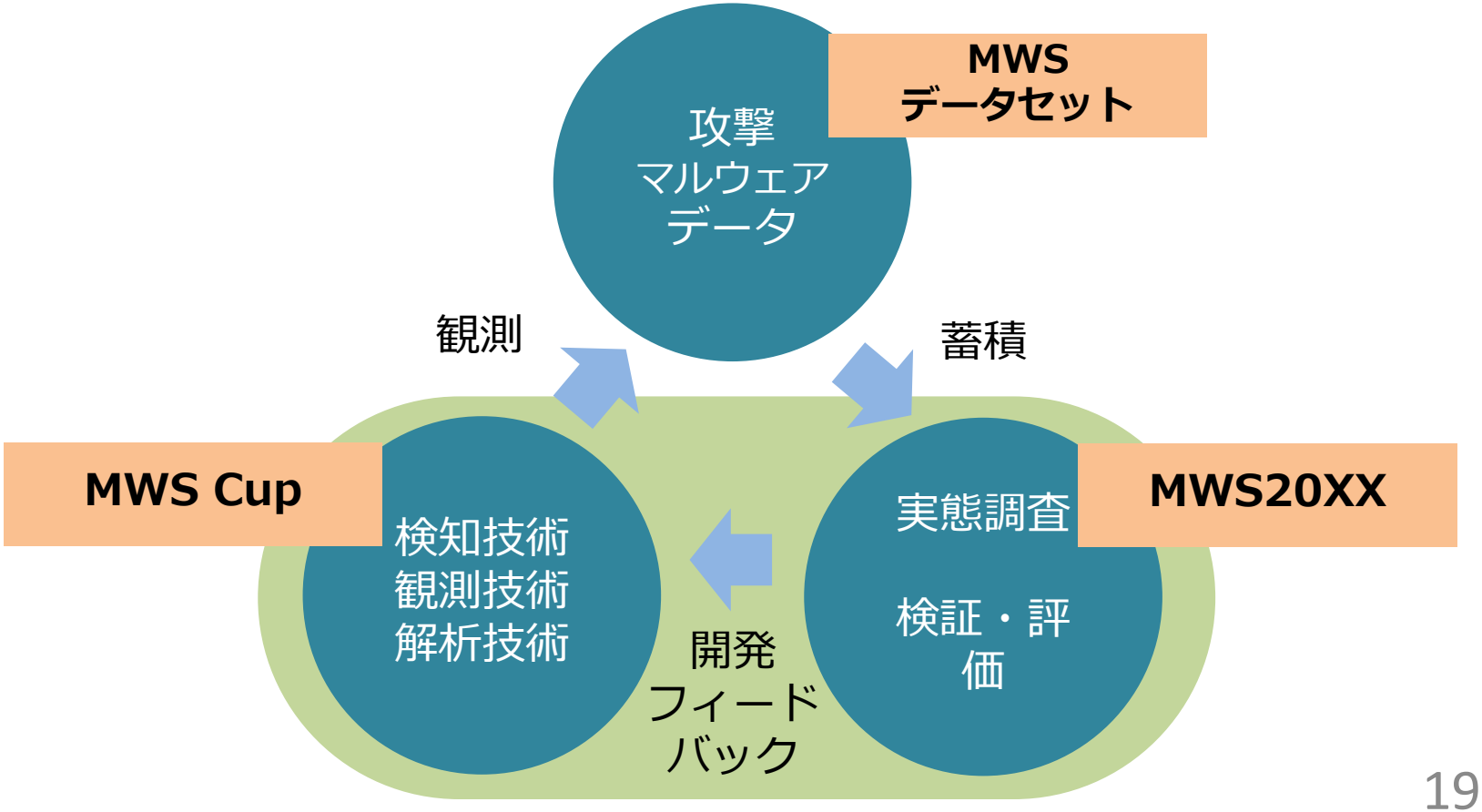
MWS データセットの利用

- 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...

- **「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
    - ✓ MWS2019 は、**2019年10月21日～10月24日長崎県ハウステンボス**にて開催; https://www.iwsec.org/mws/2019/

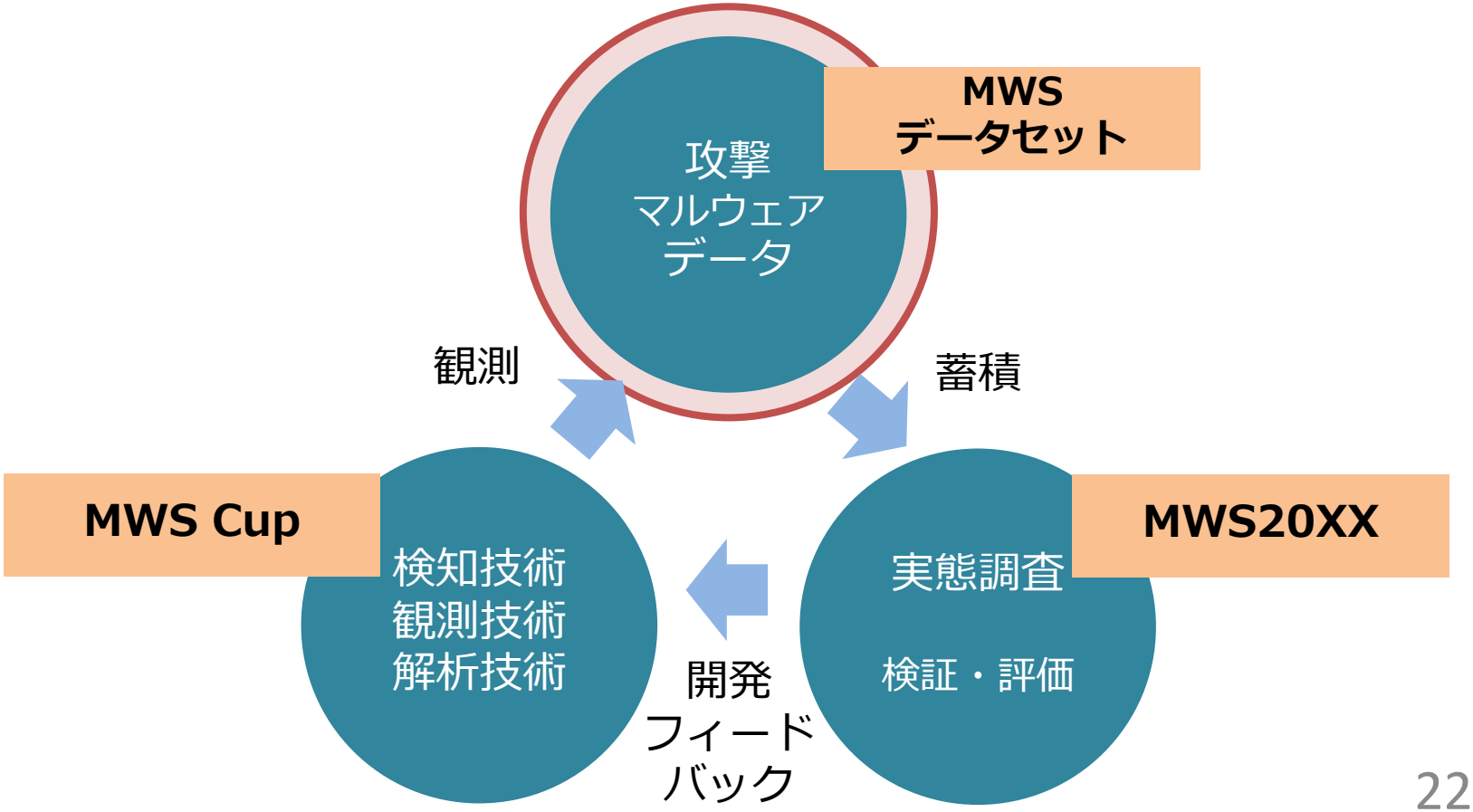
MWS Cup

- 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
    - ✓ マルウェアの動的解析・静的解析・表層解析
    - ✓ 解析競技の後、自由課題の成果物について**プレゼンも実施**

http://www.iwsec.org/mws/2018/photo.html

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 2019 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 2019 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 2019 は、**8/1 アブスト締切、8/22 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/2019/mws20190604.html

参考資料

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。

荒木粧子, 他: マルウェア対策のための研究用データセット ～ MWS Datasets 2019 ～, 情報処理学会, Vol.2019-CSEC-86, No.8, 2019年7月.