ANTI MALWARE ENGINEERING WORKSHOP

MWS

マルウェア対策のための研究用データセット ～MWS Datasets 2016～

高田 雄太, 寺田 真敏, 村上 純一,
笠間 貴弘, 吉岡 克成, 畑田 充弘

2016年7月14日

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 2016 を紹介させていただきます。

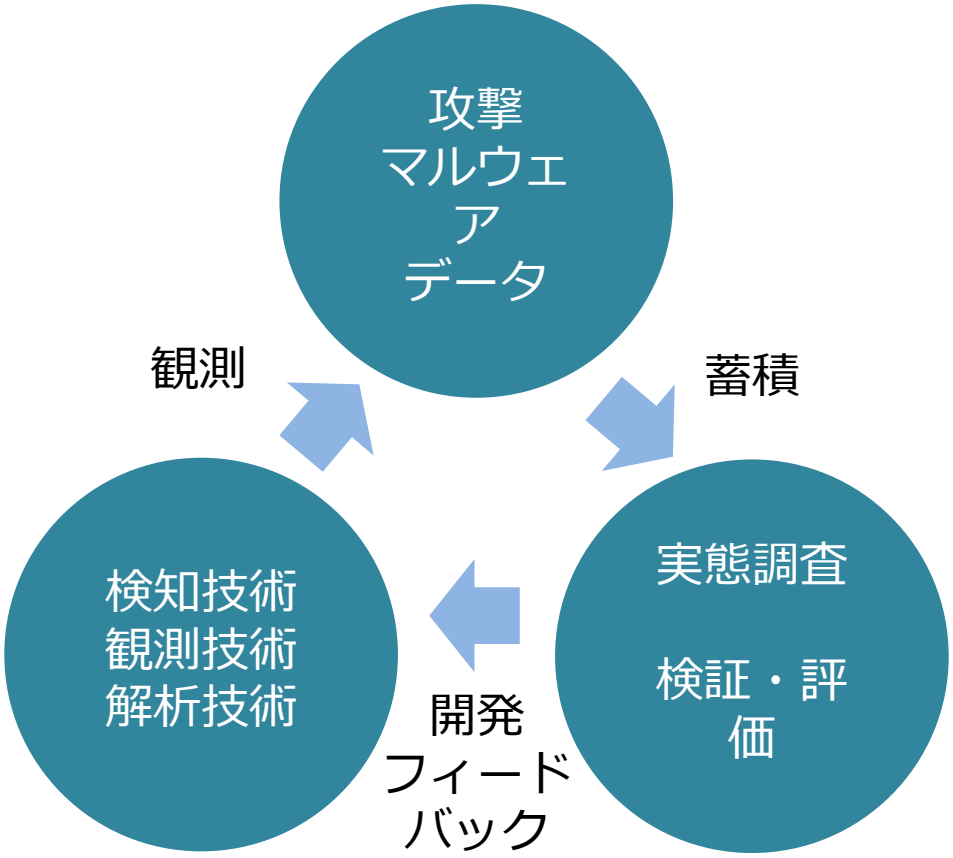
背景：複雑化するサイバー攻撃

- 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

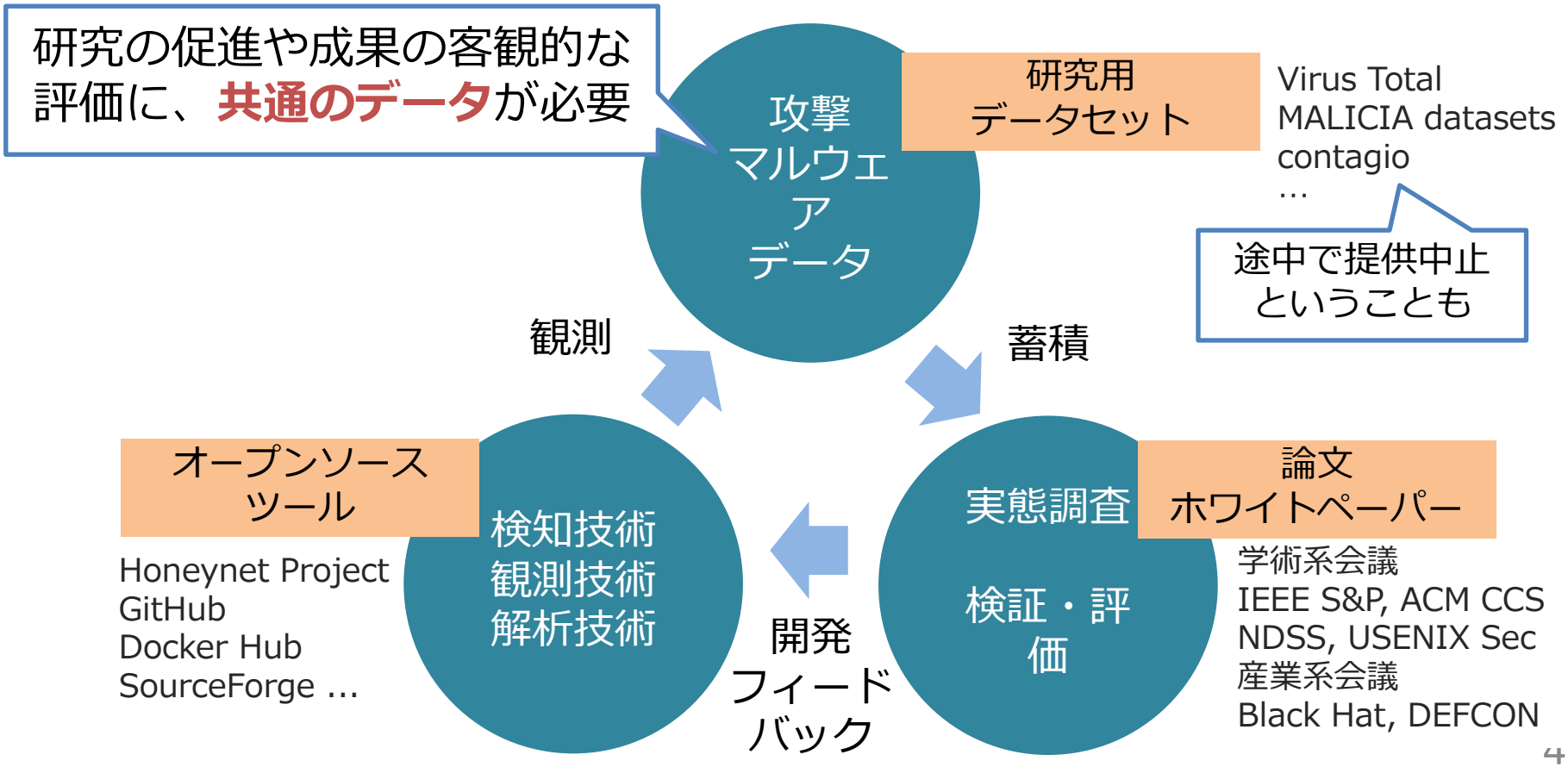
マルウェア対策研究

- 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
  - **しかしながら、既存のデータセットは「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といった課題が存在**



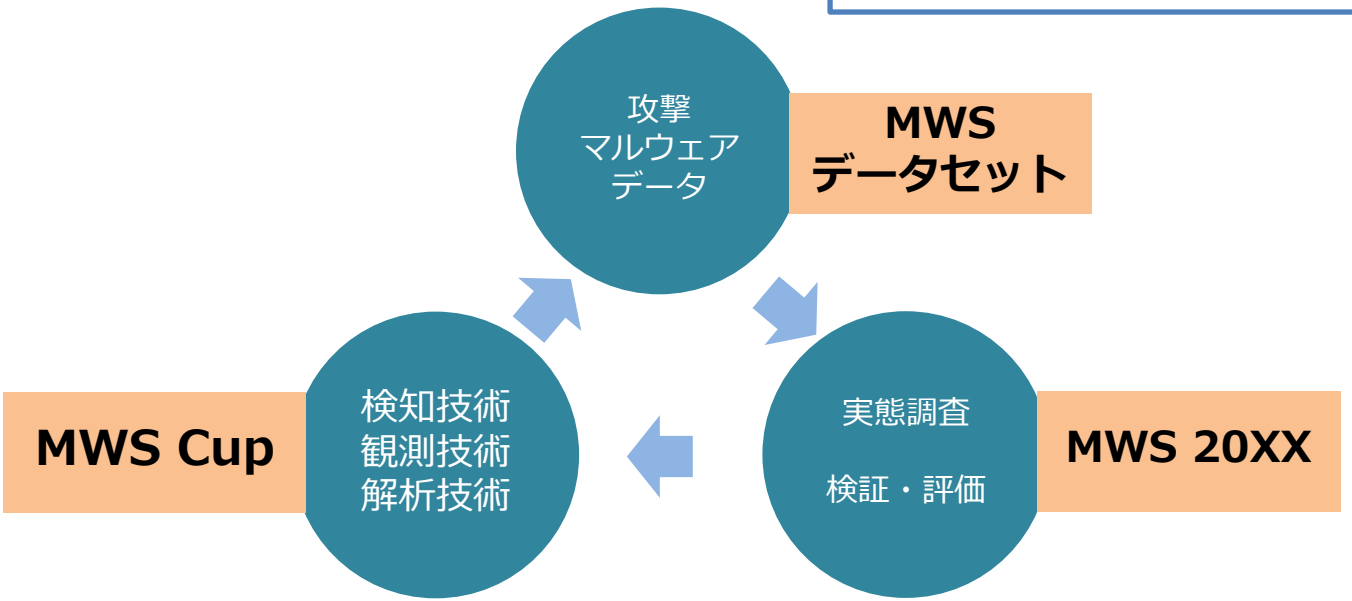
マルウェア対策研究人材育成ワークショップ (MWS)

- 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



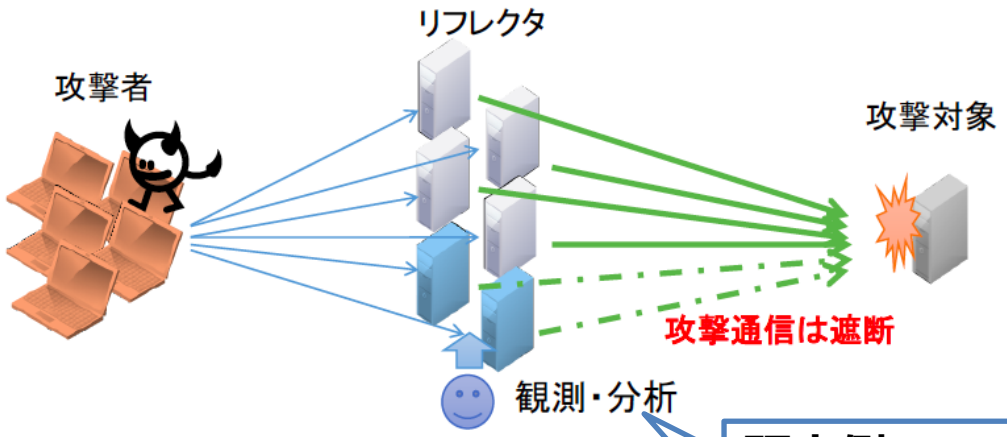
MWS データセット 2016

- 提供されるデータセットは 8種類
  - e.g., マルウェア検体 / 実行ログ, 悪性通信データ
- 今年更新のあるデータセットは 4種類

データセット名	08	09	10	11	12	13	14	15	16
ボット観測用攻撃通信、攻撃元データ、マルウェア検体 (サイバークリーンセンター) CCC DATAset	✓	✓	✓	✓	✓	✓			
ウェブ感染型マルウェア観測データ(NTT) D3M			✓	✓	✓	✓	✓	✓	
マルウェア感染後の通信挙動データ 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) FFRI Dataset						✓	✓	✓	✓
ダークネットトラフィックデータ(NICT) NICTER Darknet Dataset						✓	✓	✓	✓
攻撃者活動観測データ(日立) Behavior Observable System (BOS)							✓	✓	✓
NCD in MWS Cup 2014(MWS) 一般的な通信を想定したデータ								✓	

PRACTICE (AmpPot) Dataset の概要

- 横浜国立大学で運用している **AmpPot**[RAID2015] **(DRDoSハニーポット)** で観測したトラフィックデータ
  - おとりのリフレクタをインターネット上に設置

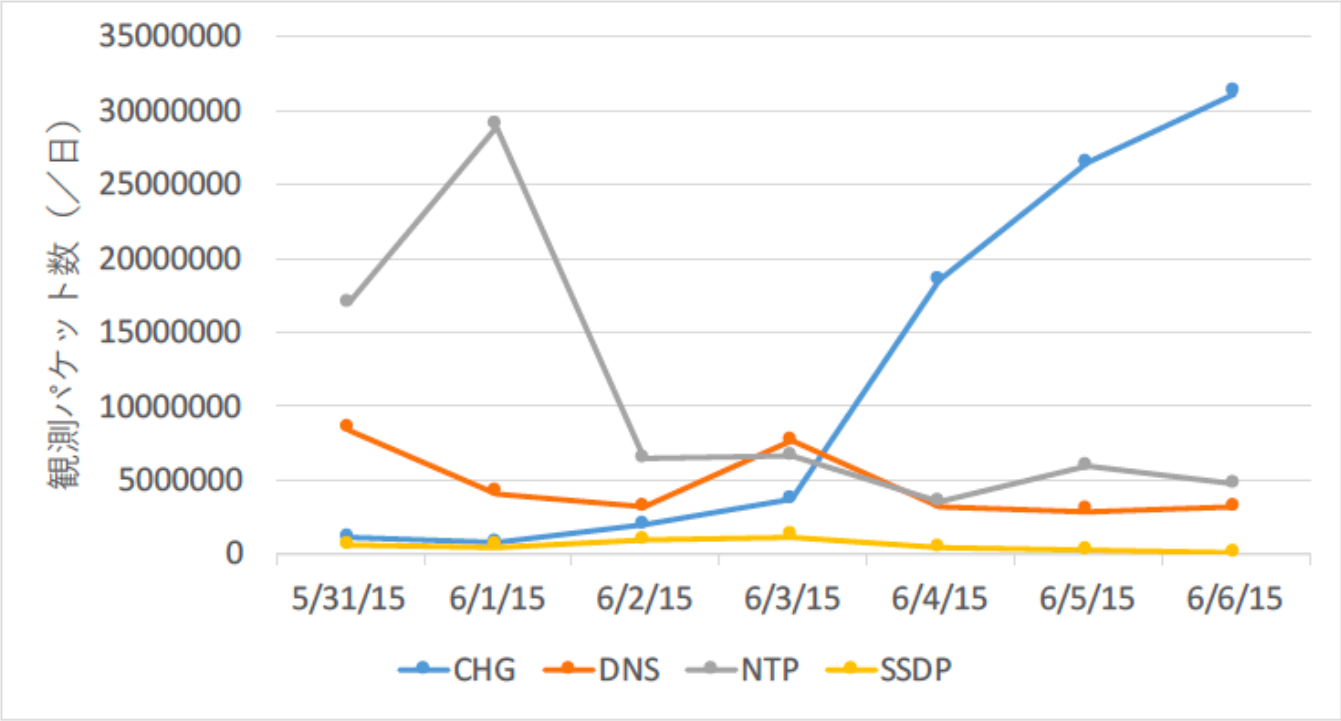


研究例：
リアルタイム DoS 検知システム
ダークネット観測都の相関分析　等

[RAID2015] Lukas Kramer, Johannes Krupp, Daisuke Makita, Tomomi Nishizoe, Takashi Koide, Katsunari Yoshioka, Christian Rossow, "AmpPot: Monitoring and Defending Amplification DDoS Attacks."

PRACTICE (AmpPot) Dataset の統計情報

- 現在は悪用される頻度の高い ４つのサービスに対応



（1週間の合計）	CHG	DNS	NTP	SSDP
パケット数	83,021,625	32,057,243	72,449,159	3,223,331
通信先アドレス数	4,304	405	104,446	6,415

17

FFRI Dataset の概要

- 株式会社 FFRI で収集した**マルウェアの動的解析ログ**
  - 2016/1～2016/3に収集された検体計8,243件を対象に、Cuckoo sandbox 上で実行した際の**ログを収録**
    - マルウェア検体自体は含まない

- **データ活用例**
  - マルウェア検知・分類、悪性通信の検出
  - 自身の自動解析システムとの比較、有効性検証
  - 動作プラットフォームにおける振る舞いの差異　等

FFRI Dataset の具体的なデータ項目

- 1検体1ログファイル (.json), 詳細は原稿参照

項目（大見出し）	内容
info	解析の開始、終了時刻、id等（idは1から順に採番）
signatures	ユーザー定義シグニチャとの照合結果（今回は使用無）
virustotal	VirusTotalの検査履歴との照合結果（検体のMD5値に基づく）
static	検体のファイル情報（インポートAPI、セクション構造等）
dropped	検体が実行時に生成したファイル
behavior	検体実行時のAPIログ（PID、TID、API名、引数、返り値等）
processtree	検体実行時のプロセスツリー（親子関係）
summary	検体が実行時にアクセスしたファイル、レジストリ等の概要情報
target	解析対象検体のファイル情報（ハッシュ値等）
debug	検体解析時のCuckoo Sandboxのデバッグログ
strings	検体中に含まれる文字列情報
network	検体が実行時に行った通信の概要情報

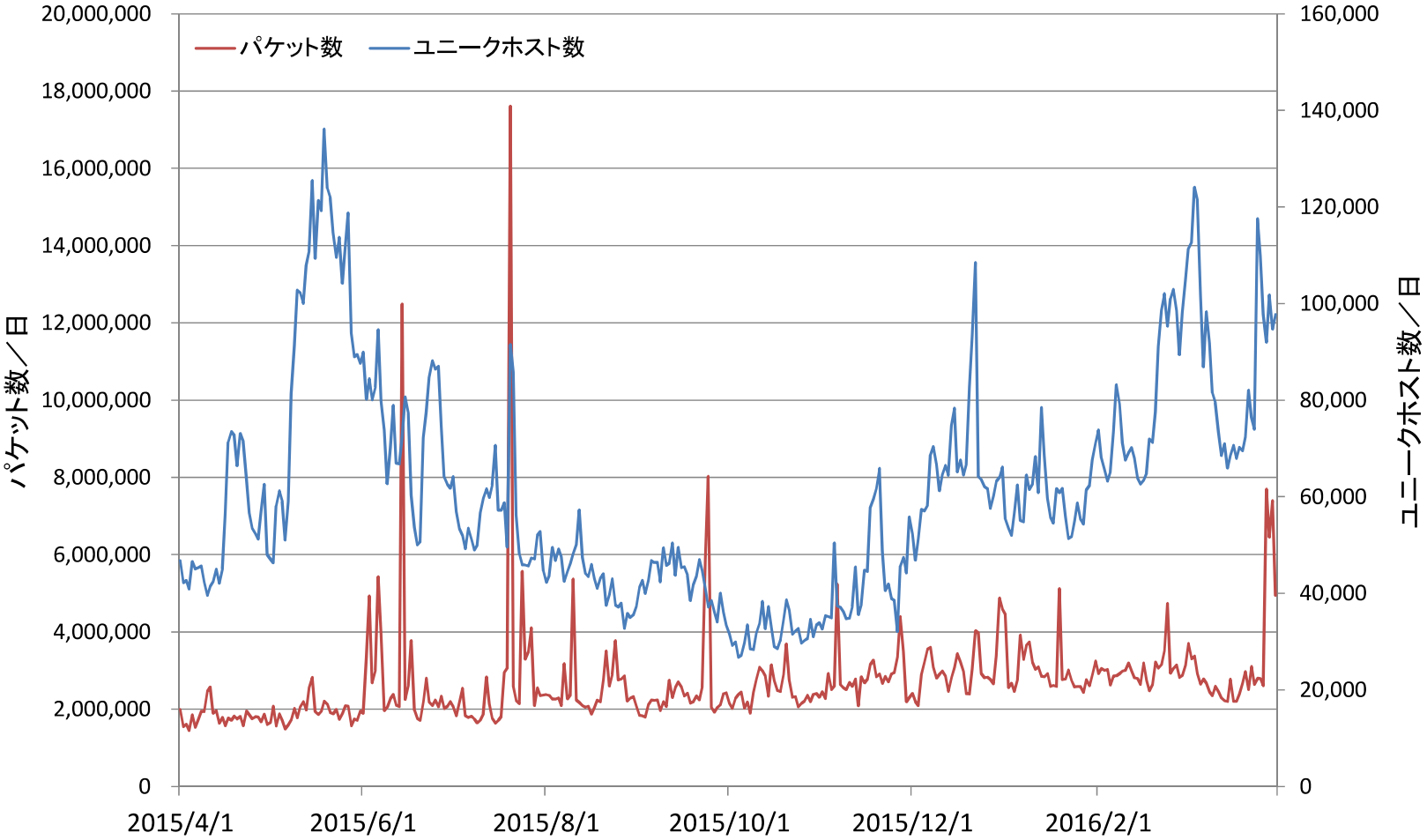
NICTER Darknet Dataset の概要

- /20 の**ダークネット宛て**のトラフィックデータ
  - 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
  - 観測期間は2011年4月1日～の**５年間＋α**

- 実際は**多くのトラフィックが届いている**
  - マルウェア（ワーム、ボット）によるスキャン活動
  - DDoS 攻撃の跳ね返り
  - リフレクション攻撃の準備活動
  - 設定ミス
  - Zmap, DNS Water Torture, IoT デバイス, etc.

ダークネットの観測状況

- 2015/04/01～2016/03/31 における観測
  - DRDoS のための端末探索や IoT を狙ったスキャンを多く観測



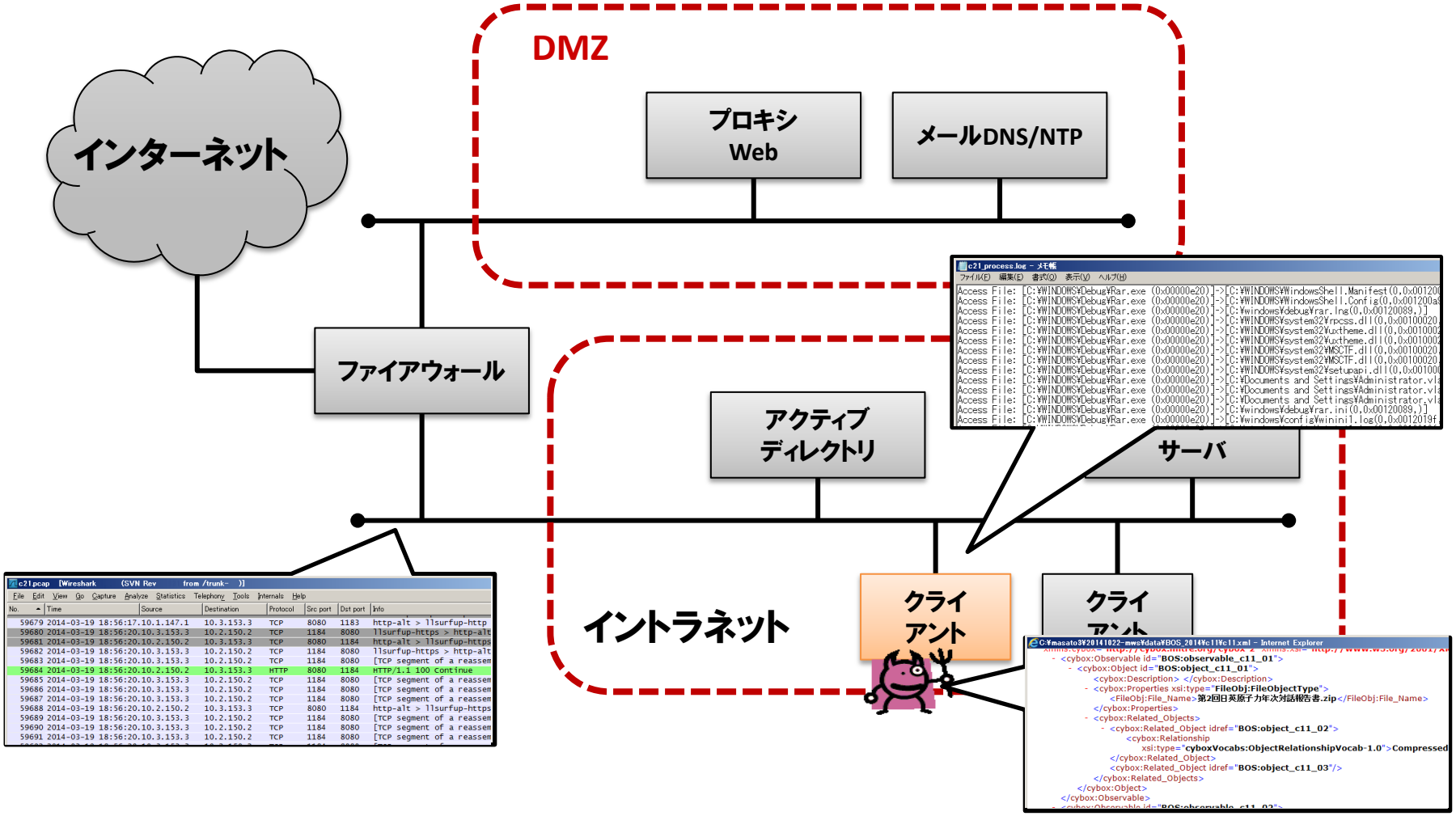
BOS Dataset の概要

- **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

- データセット構成
  - マルウェア検体
    - 観測に使用したマルウェア検体のハッシュ値を CybOX 形式 (Cyber Observable eXpression；サイバー攻撃観測記述形式) 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

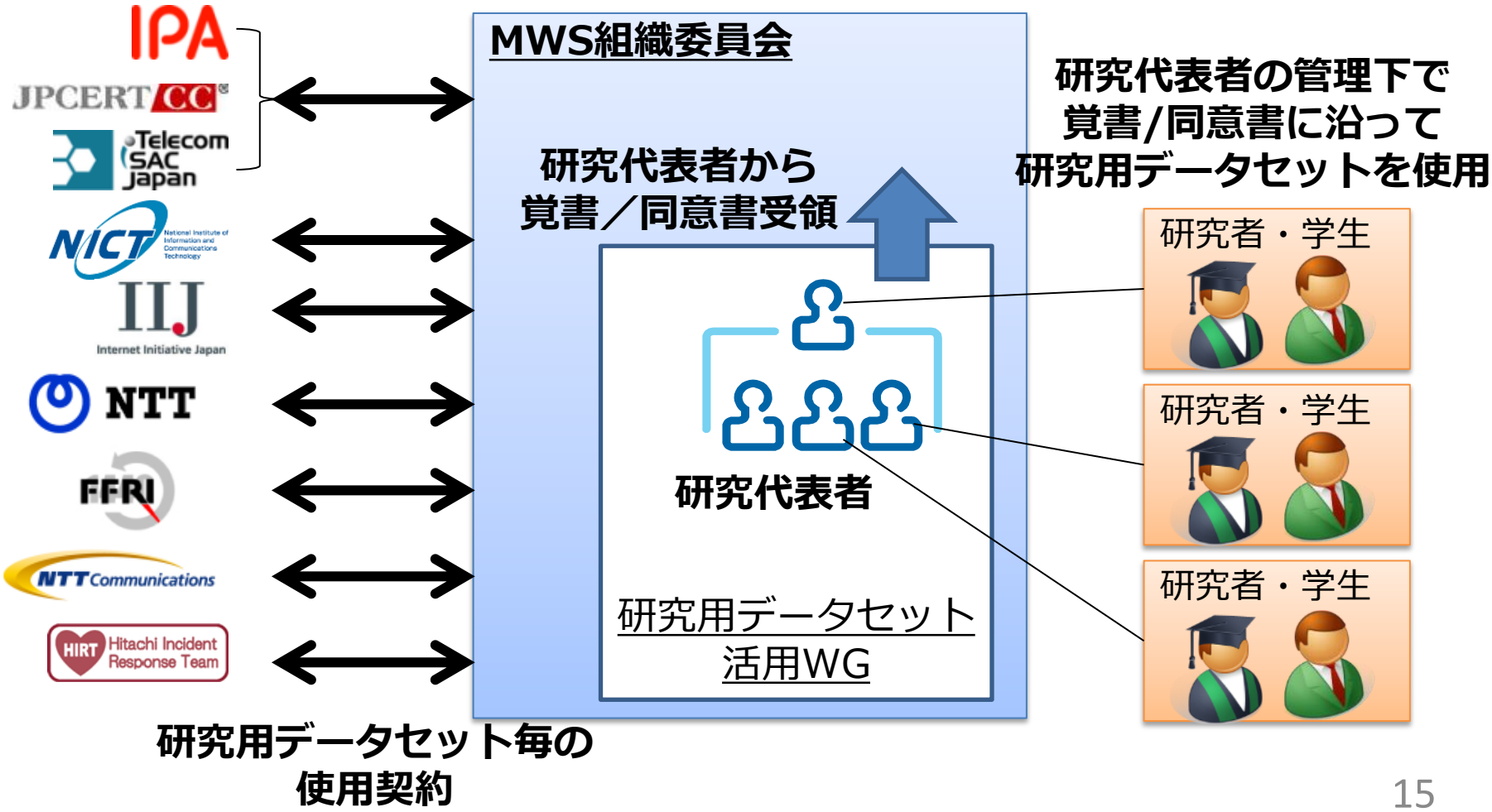
BOS の観測環境

- 組織内 NW を模擬した動的活動観測環境を構築



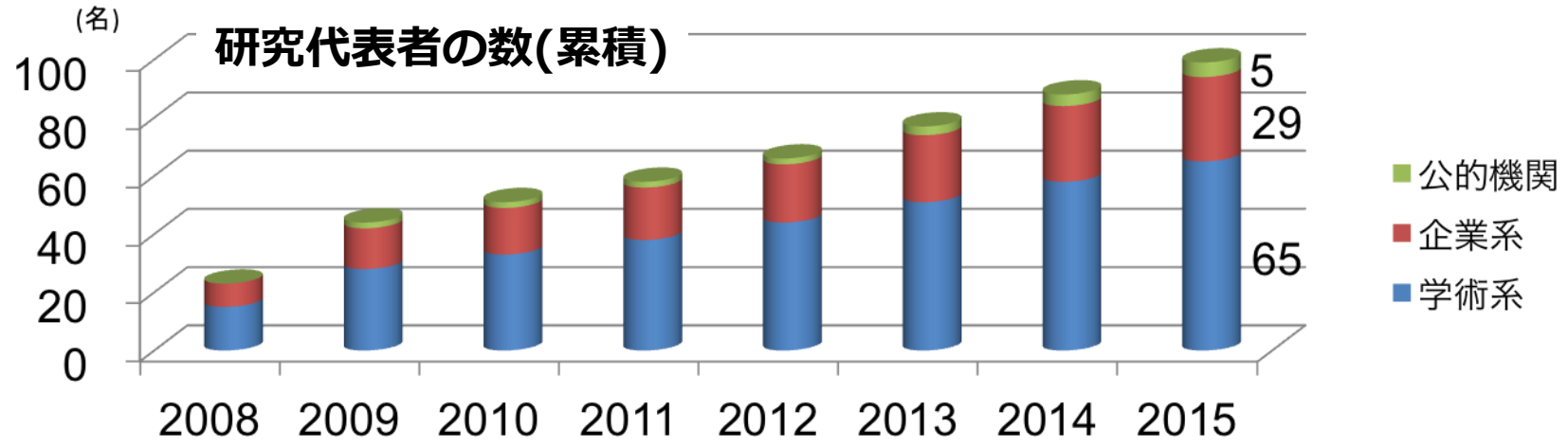
MWS データセット契約形態

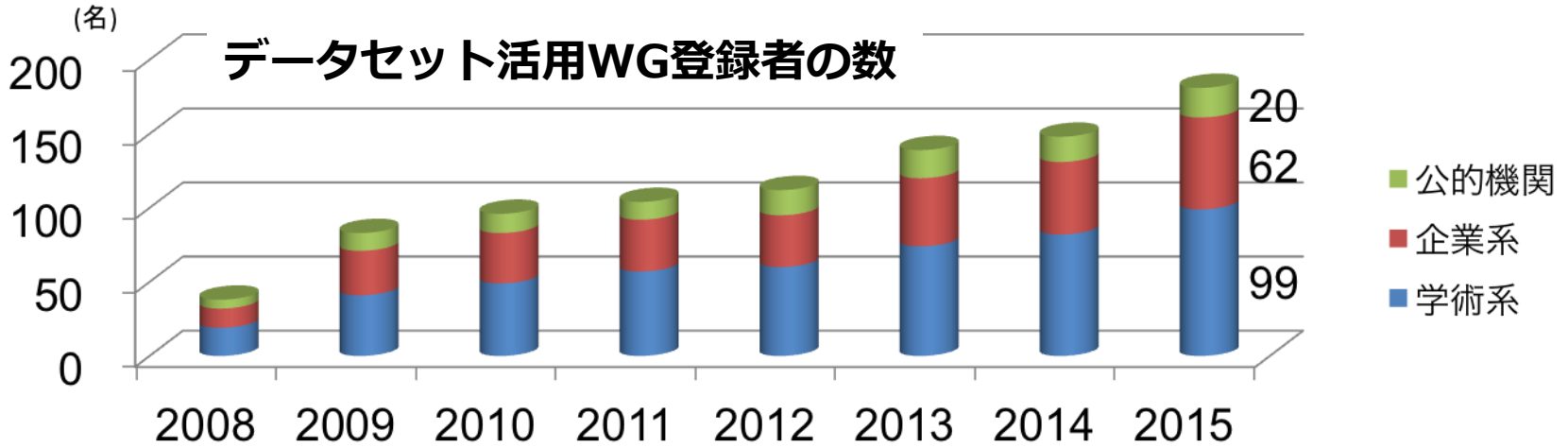
- MWS 組織委員会をハブとした利用手続き
  - 事務局 [csecreg@sdl.hitachi.co.jp] へコンタクト



MWS データセットの利用者数

- 約100の組織にデータセットは活用されている。





主なデータセットの利用組織

- セキュリティ分野以外からの研究用データセット利用（＝研究のきっかけとして活用）が微増

(独)情報処理推進機構
情報通信研究機構
産業技術総合研究所
JPCERT コーディネーションセンター
Telecom-ISAC Japan
警察庁

(株)インターネットイニシアティブ、
NRIセキュアテクノロジーズ(株)、
NTTコミュニケーションズ(株)、(株)NTTデータ、
NTT西日本電信電話(株)、(株)FFRI、
(株)MCセキュリティ、KDDI(株)、
(株)セキュアブレイン、セコム(株)、
SECCON実行委員会、デジタルアーツ(株)、
トレンドマイクロ(株)、日本電気(株)、
日本電信電話(株) セキュアプラットフォーム研究、
(株)PFU、(株)日立製作所、(株)富士通研究所、
(株)富士通システム統合研究所、
富士通ソーシアルサイエンスラボラトリ、
日本マイクロソフト(株)、三菱電機(株)、楽天(株)、
(株)ラック、(株)リクルートテクノロジーズ

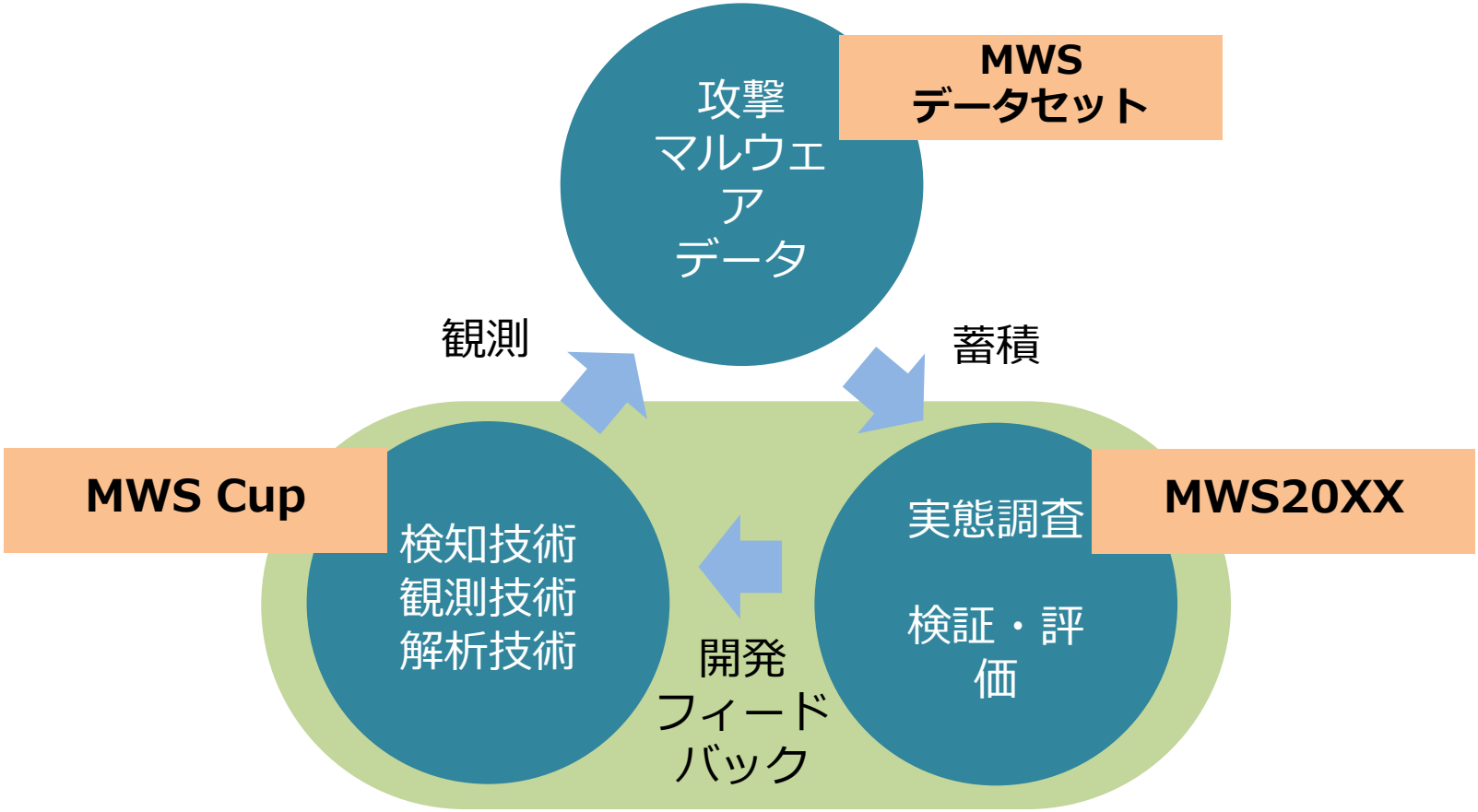
岩手県立大学
岡山大学
金沢大学
関西大学
九州大学
九州工業大学
京都産業大学
熊本高等専門学校
慶應義塾大学
神戸大学
佐賀大学
静岡大学
芝浦工業大学
情報セキュリティ大学院大学
千葉大学
中央大学
筑波大学
電気通信大学
東海大学
東京大学
東京工科大学
東京工業大学

東京情報大学
東京電機大学
東京理科大学
東邦大学
東北大学
東北工業大学
東北文化学園大学
富山大学
長崎大学
名古屋大学
名古屋工大学
奈良先端科学技術大学院大学
南山大学
新潟県立大学
はこだて未来大学
弘前大学
防衛大学校
法政大学
北陸先端科学技術大学院大学
明治大学
横浜国立大学
立命館大学
早稲田大学

http://www.iwsec.org/mws/2015/committee.html

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
    - ドライブバイ解析、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
    - MWS2016 は、2016年10月11日～10月13日に秋田にて開催
      http://www.iwsec.org/mws/2016/

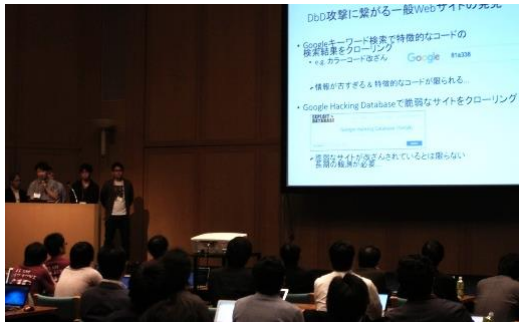
MWS データセットを用いた発表件数

MWS Datasets			08	09	10	11	12	13	14	15
MWS CSS での発表 [2015/10 時点]	CCC Dataset	マルウェア検体	5	7	6	5	7	3	3	0
		攻撃通信データ	9	14	5	6	2	-	-	2
		攻撃元データ	8	6	5	4	-	-	-	1
	MARS Dataset		-	-	1	1	-	-	-	-
	D3M		-	-	4	3	3	9	14	9
	IIJ MITF Dataset		-	-	-	-	1	-	-	-
	FFRI Dataset		-	-	-	-	-	5	2	4
	PRACTICE Dataset		-	-	-	-	-	3	1	0
	NICTER Darknet Dataset		-	-	-	-	-	6	2	3
	BOS		-	-	-	-	-	-	1	4
	NCD in MWS Cup 2014		-	-	-	-	-	-	-	0
	データセット概説		0	1	1	1	0	1	0	0
	合計 ()内は学生発表		22 (8)	28(15)	22(10)	20(9)	13(9)	27(10)	23(10)	23(14)
MWS/CSS 以外での発表/論文投稿 [2015年1月時点] ()内は海外発表			0	1	10(5)	10(3)	12(6)	16(3)	12(4)	1

MWS 以外の国際会議等における
データセットを用いた発表や論文も多数！

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」**
  - 課題例
    - Drive-by Download 攻撃解析
    - マルウェア静的解析／動的解析
    - ダークネットのパケット分析　など
  - 数時間の解析競技の後、手法や戦略をまとめた**プレゼンも実施**
    - 他チームからも学べる！



http://www.iwsec.org/mws/2015/photo.html

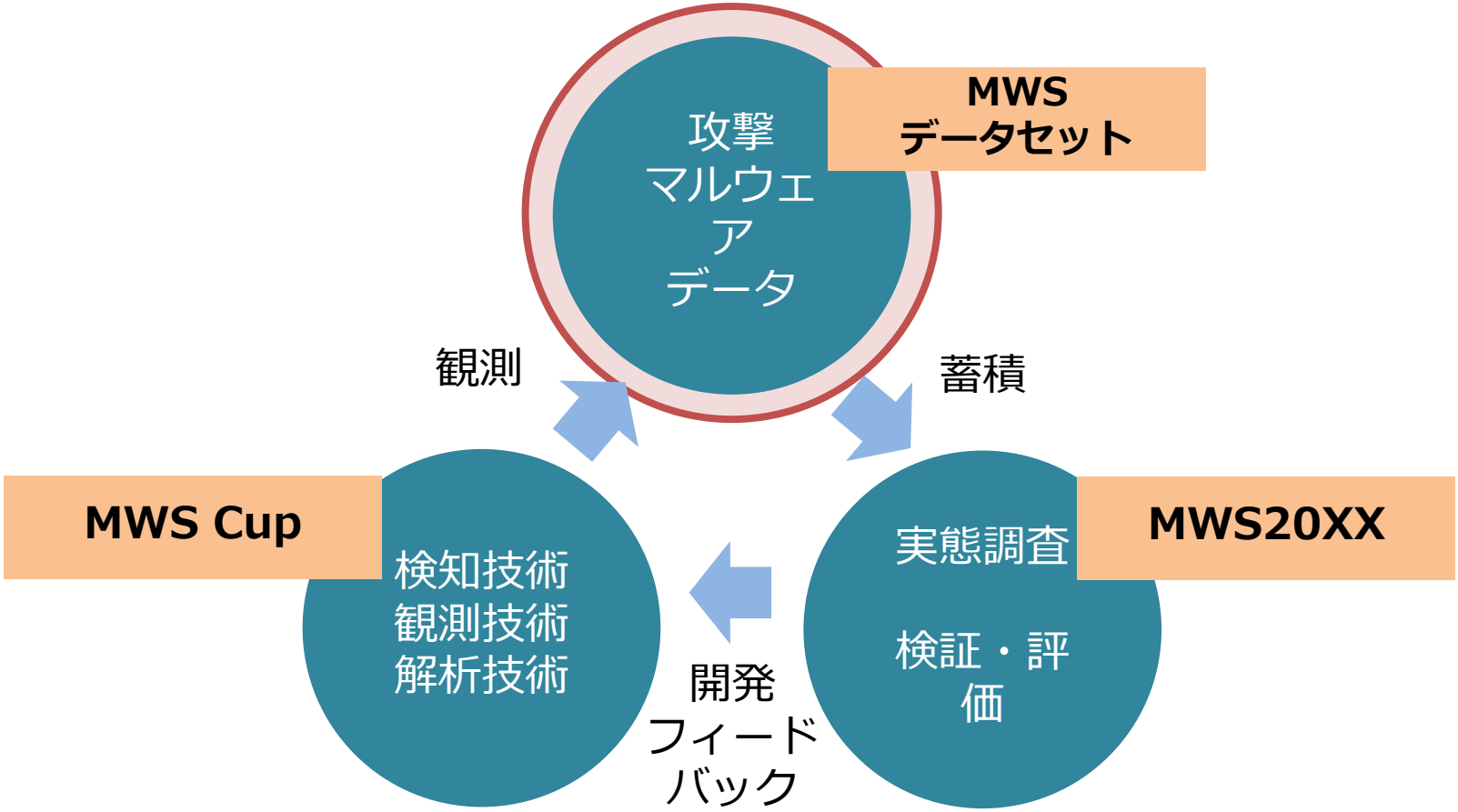
MWS Cup の様子

- 昨年は15チーム81名が参加

http://www.iwsec.org/mws/2015/about.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 2016 を提供中
  - 研究開発の推進／開発技術の共有により本研究分野の発展に寄与
  - MWS Datasets 2016 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 2016 実行委員会事務局
      「csecreg@sdl.hitachi.co.jp」までご連絡を

- 宣伝
  - MWS 2016 は、**7/19 アブスト締切、8/12 原稿締切**
  - MWS Cup は、**問題出題協力者**を募集中
  - http://www.iwsec.org/mws/2016/

参考資料

関連研究

- CAIDA Data
  - ネットワーク運用に関わる通信ログのデータセット
- MAWILab
  - サンプリングで保存された通信リポジトリにラベル付けしたデータセット
- 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。

高田他, ”マルウェア対策のための研究用データセット ～MWS Datasets 2016～,” 情報処理学会 研究報告コンピュータセキュリティ (CSEC), 2016年7月.

MWS の研究動向

- 巡回 URL リスト生成
  - 既知の悪性URL群と類似した特徴を持つURLの検索 *[MWS2014]* (早大 孫ら)
- JavaScript 解析
  - 抽象構文解析木による不正なJavaScriptの特徴点抽出手法の提案 *[MWS 2011]* (セキュアブレイン 神薗ら, *MWS2011優秀論文賞*)
  - 難読化されたスクリプトにおける特徴的な構文構造のサブツリーマッチングによる同定 *[MWS 2011]* (奈良先端大 *Gregory*ら)
- PDF 解析
  - 動的解析を利用した難読化JavaScriptコード解析システムの実装と評価 *[MWS 2010]* (セキュアブレイン 神薗ら, *MWS2010優秀論文賞*)
  - PDF の構造検査による悪性 PDF の検知 *[MWS2013]* (*NISC* 大坪ら)
- Exploit kit 解析
  - Drive-by-Download攻撃における通信の定性的特徴とその遷移を捉えた検知方式 *[MWS 2013]* (NTT データ 北野ら)
  - Exploit kit の特徴を用いた悪性 Web サイトの検知手法 *[MWS 2013]* (*NICT* 笠間ら)
- リダイレクト解析
  - 検知を目指した不正リダイレクトの分析 *[MWS 2010]* (富士通研 寺田ら)
  - パスシーケンスに基づく *Drive-by-Download* 攻撃の分類 *[MWS 2010]* (東海大 桑原ら)

D3M の概要

- Drive-by Download Data by Marionette[*]
  - 高対話型ハニークライアントで観測したドライブバイダウンロード攻撃に関連する**「攻撃通信データ」「マルウェア情報」「マルウェア通信データ」**を収録
  - pcap ファイルで提供
    - 攻撃を行う URL, ドメイン名, IP アドレス
    - ウェブコンテンツ（HTML, JS, PDF, Jar, …）
    - C&C サーバとの通信