セキュリティ分野における AI活用の現状と期待

小澤 誠一
神戸大学 数理・データサイエンスセンター
工学研究科 電気電子工学専攻
ozawasei@kobe-u.ac.jp
http://www2.kobe-u.ac.jp/~ozawasei/

自己紹介

小澤 誠一 (ozawasei@kobe-u.ac.jp)
所　属： 神戸大学 数理・データサイエンスセンター・研究部門長
　　工学研究科電気電子工学専攻・教授
ホームページ: http://www2.kobe-u.ac.jp/~ozawasei/

◎ **研究内容**（第2次ブームのときからずっと人工知能を研究）
ニューラルネット，機械学習，パターン認識，画像認識，
ビッグデータ解析，セキュリティ，スマート農業，文書解析

◎ **主な研究テーマ**

1) 暗号データに対する機械学習アルゴリズムの開発
プライバシー保護データマイングと異常検知

2) 機械学習のサイバーセキュリティへの応用
サイバー攻撃検知・観測・可視化，攻撃情報の収集など

3) 機械学習の文書解析への応用
SNS炎上検知，金融文書解析、悪性JaveScript判定

4) スマート農業に向けた農作物の画像センシング手法の開発
大豆の花・子実検出

AI×セキュリティ

- AIの強みと限界を知り、セキュリティについて考える。

AI×セキュリティとは？

- AIによるサイバーセキュリティ
  - 攻撃の検知・分類
  - 攻撃の観測・可視化
  - 悪性コンテンツ検知 (悪性サイトや悪性コード)
  - サイバー攻撃関連の情報収集 (SNS, ダークウェッブ) など
- AIのためのセキュリティ
  - Machine Learning as a Service (MLaaS)への攻撃
  - スマートIoTへの攻撃
- AIとセキュリティの新しい展開
  - プライバシー保護データマイニング (PPDM)

AI×セキュリティ（その１）

- AIによるサイバーセキュリティ

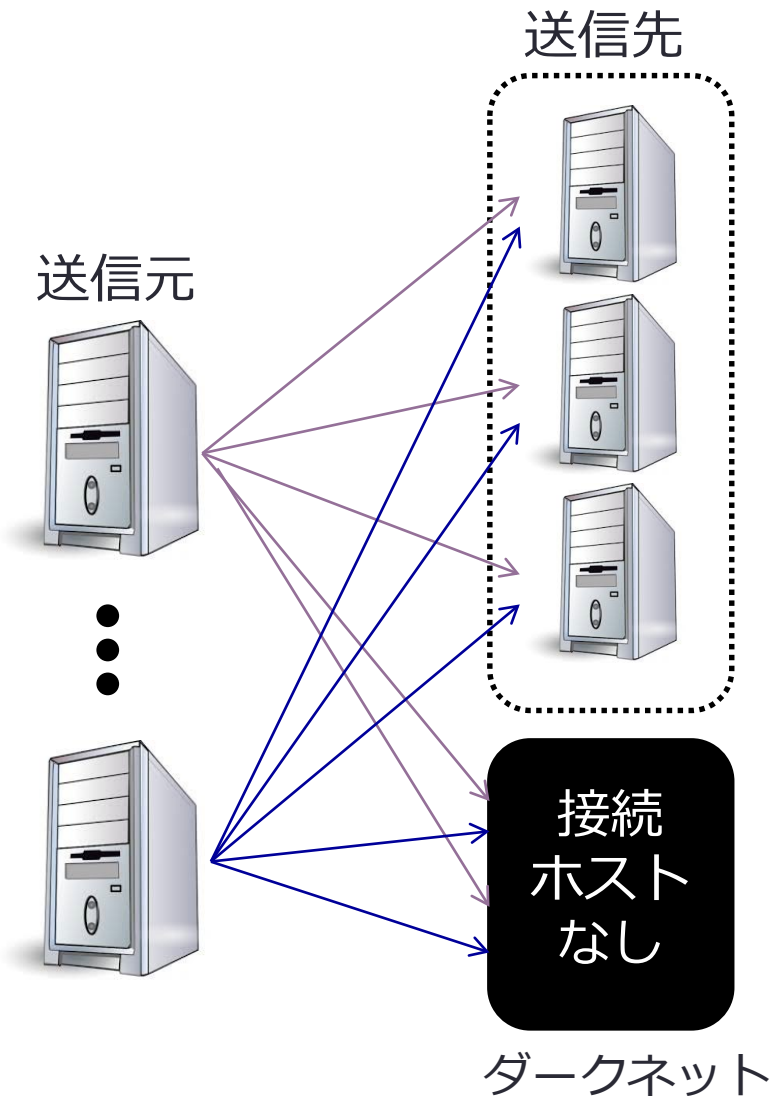
サイバーセキュリティの現状

- サイバー攻撃の多様化と巧妙化
  - ✓ ランサムウェア（WanaCryなど）
  - ✓ IoTマルウェア（Mirai、Hajimeなど）
  - ✓ Web媒介型攻撃（Drive-by-Download攻撃、
  - ✓ ソーシャルエンジニアリング攻撃（人間の持つ脆弱性に対するエクスプロイト）
- AIへのサイバー攻撃
- ➢ Adversarial Examples
  - - Evasion attack：難読化（暗号化，画像ベース）
  - - Poisoning attack：訓練データの操作，ラベルの反転など

サイバーセキュリティにおける機械学習の応用

- **マルウエア解析(静的/動的)**
  - ✓ マルウエア検知・分類
- **不正侵入検知(IDS/IPS)**
  - ✓ 異常検知
  - ✓ 攻撃分類・検知
  - ✓ ハニーポット観測・分析
- **広域攻撃観測（ダークネット分析）**
  - ✓ 異常検知
  - ✓ 攻撃分類・検知
  - ✓ ボットネットの活動検知・分析
- **Webベース攻撃検知**
  - ✓ 悪性サイト・悪性スパムメール検知
  - ✓ 悪性JavaScript検知
- **サイバー攻撃情報の収集・分析**
  - ✓ 表層Web解析（SNS、セキュリティブログ・レポートなど）
  - ✓ 深層Web解析（ダークマーケット、ダークフォーラムなど）

ダークネット ～ サイバー攻撃の広域観測



ダークネット＝未使用IP群

＝ 訓練データを自動収集する仕組み

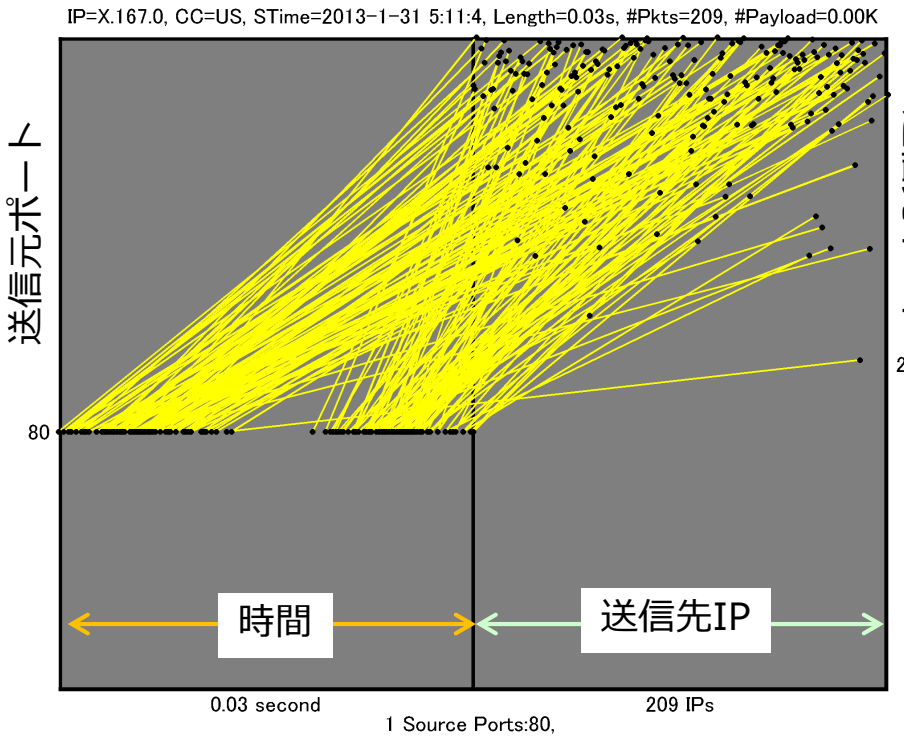
Network Telescopeとも言われる。

ダークネットにパケットが届く理由

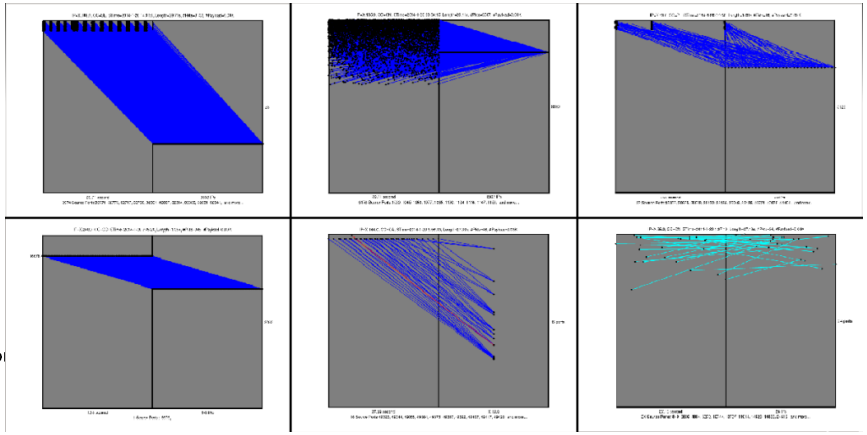
- 設定ミス
- スキャン
- DDoS攻撃ターゲット
  ホストからの返信

サイバー攻撃に関連した通信

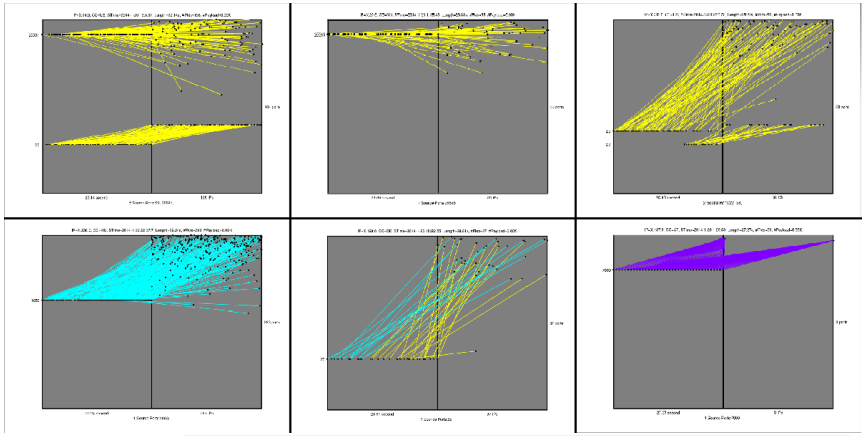
ダークネット・トラフィックの特徴



スキャン



DDoSバックスキャッタ



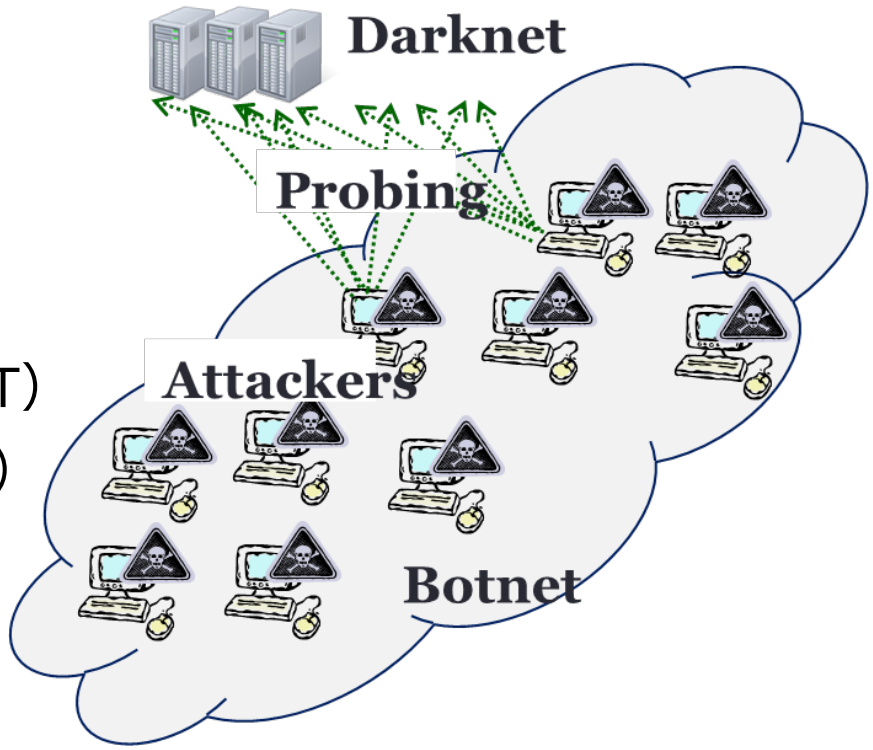
攻撃の種類によって、パケット送信に違いがある。

NICT サイバーセキュリティ研究所 提供

頻出パターンマイニングによるスキャン攻撃観測

期間：2016年7月1日～9月15日
観測パケット数:　1,840,973,403
ユニークホスト数: 17,928,006

- /16 ダークネットセンサー（NICT）
- TCP SYN パケット （ヘッダ情報）



ホストごとに送信パケットのヘッダ情報に相関パターンが見つかるか？

頻出パターンマイニング

相関ルール

$$X \Rightarrow Y$$

前提部　　結論部

Xを満たすときYの条件も頻繁に満たす.

評価指標

最低支持数:X ⇒ Yの条件を満たすルールの数

最低確信度:X ⇒ Yの条件が成立する確率

相関ルール学習の例

$X \Rightarrow Y$

前提部　　結論部

Xを満たすときYの条件も頻繁に満たす．

T1{パン, 牛乳}
T2{おにぎり, お茶}
T3{パン, ジュース}
T4{おにぎり, お菓子, お茶}
T5{カップ麺, お茶}
T6{おにぎり, チキン, お茶}
T7{おにぎり, チキン}

例：
{おにぎり} ⇒ {お茶}

おにぎりを買う人は、
高い頻度でお茶を買う

観測方法・実験条件

2016年の7月～9月の3ヶ月間 (Miraiソースコード公開周辺)

相関ルール学習の条件： 最低ホスト数: 1000
最低確信度: 90%

▼

TCP ウィンドウサイズ
パケット優先度 (ToS)
送信先ポート番号

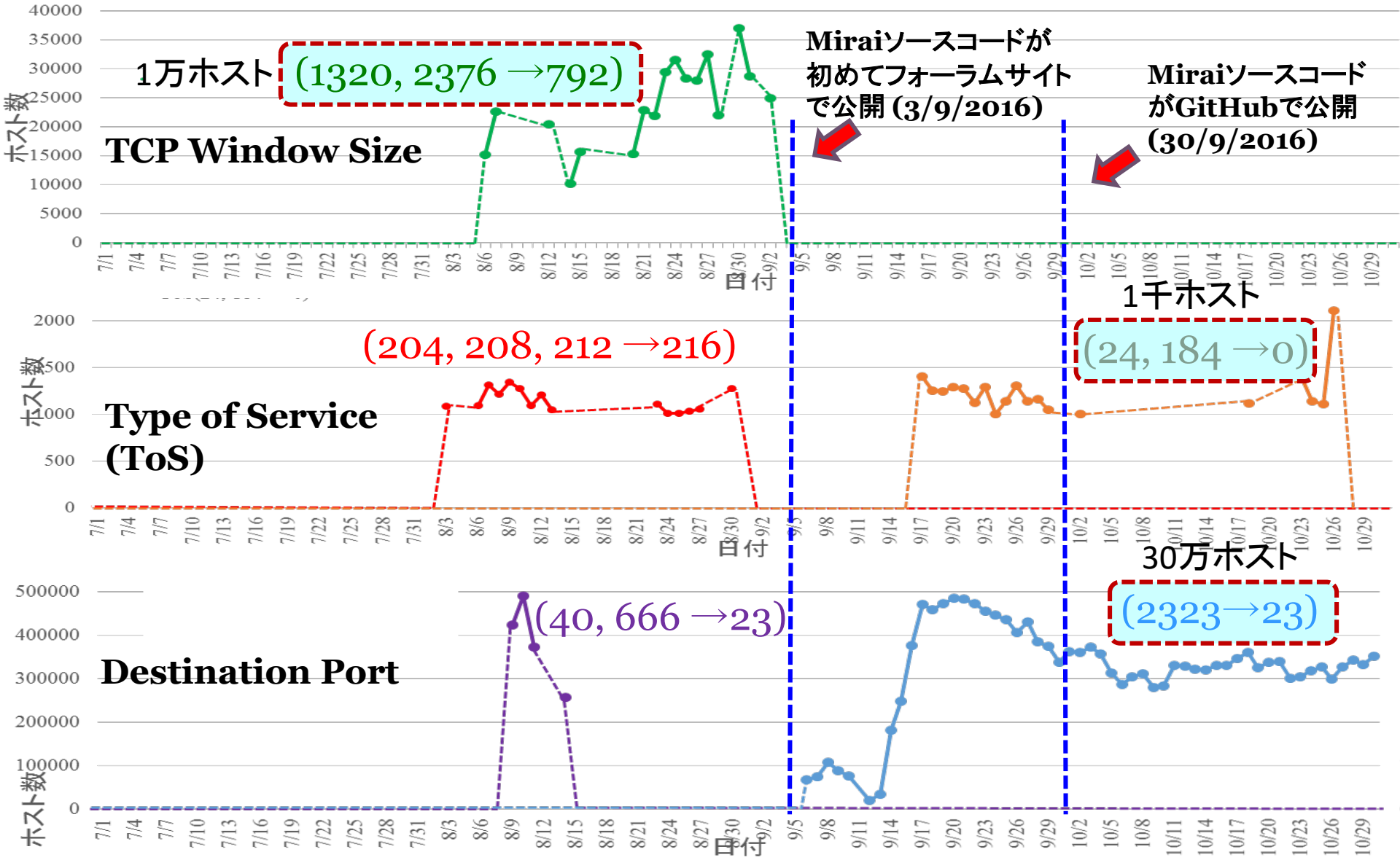
で特徴的な相関ルールが出現.

Mirai 判定基準

シーケンス番号＝宛先IPアドレス

を全て満たすパケットが90%を超えたホストを「Mirai感染ホスト」と定義

IoT マルウエア*Mirai* の特徴変化



国立研究開発法人 情報通信研究機構
National Institute of Information and Communications Technology

WarpDrive

Web-based **A**ttack **R**esponse with **P**ractical and **D**eployable **R**esearch **I**nitiati**VE**

NICT委託研究『Web媒介型攻撃対策技術の実用化に向けた研究開発』

電脳空間における「タチコマ・リアライズ」

実証実験にご参加下さい！

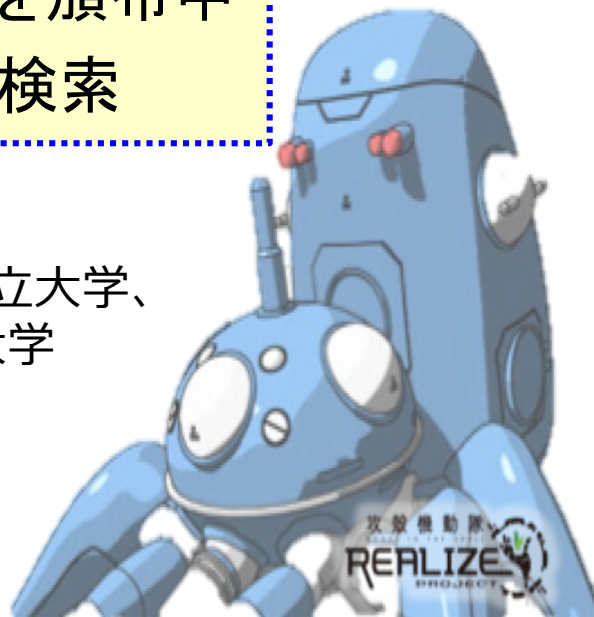
◎無料セキュリティアプリ『タチコマSA』を頒布中

『WarpDrive』または『タチコマSA 』で検索

参画機関：

KDDI総合研究所、セキュアブレイン、横浜国立大学、
神戸大学、構造計画研究所、金沢大学、岡山大学

WarpDrive Webサイトから一部引用



Web媒介型攻撃の脅威増大

悪性Webサイトによるサイバー攻撃の巧妙化・多様化

- **ドライブ・バイ・ダウンロード攻撃の高度化**
  - 戦略的な悪性サイトの設置(水飲み場攻撃)
  - Webトラヒック売買やSEO(検索エンジン最適化)悪用による誘導
- **ソーシャルエンジニアリング(ユーザクリック型)ソフトウェアダウンロード攻撃の増加**
- **フィッシング等従来からの脅威の継続**

Webに関する新たなサイバー攻撃

- **IoT機器のWebインターフェイス(機器管理・設定用画面)への攻撃**
- **リフレクション型DoS、L7DoSなどWebサイトへのサービス妨害攻撃の多様化、高度化**

WarpDrive中間評価ヒアリング資料から引用

AI×セキュリティ（その２）

- AIのためのセキュリティ ～ AIが攻撃者に狙われる？

機械学習がサイバー攻撃の標的に！

Machine Learning as a Service (MLaaS)

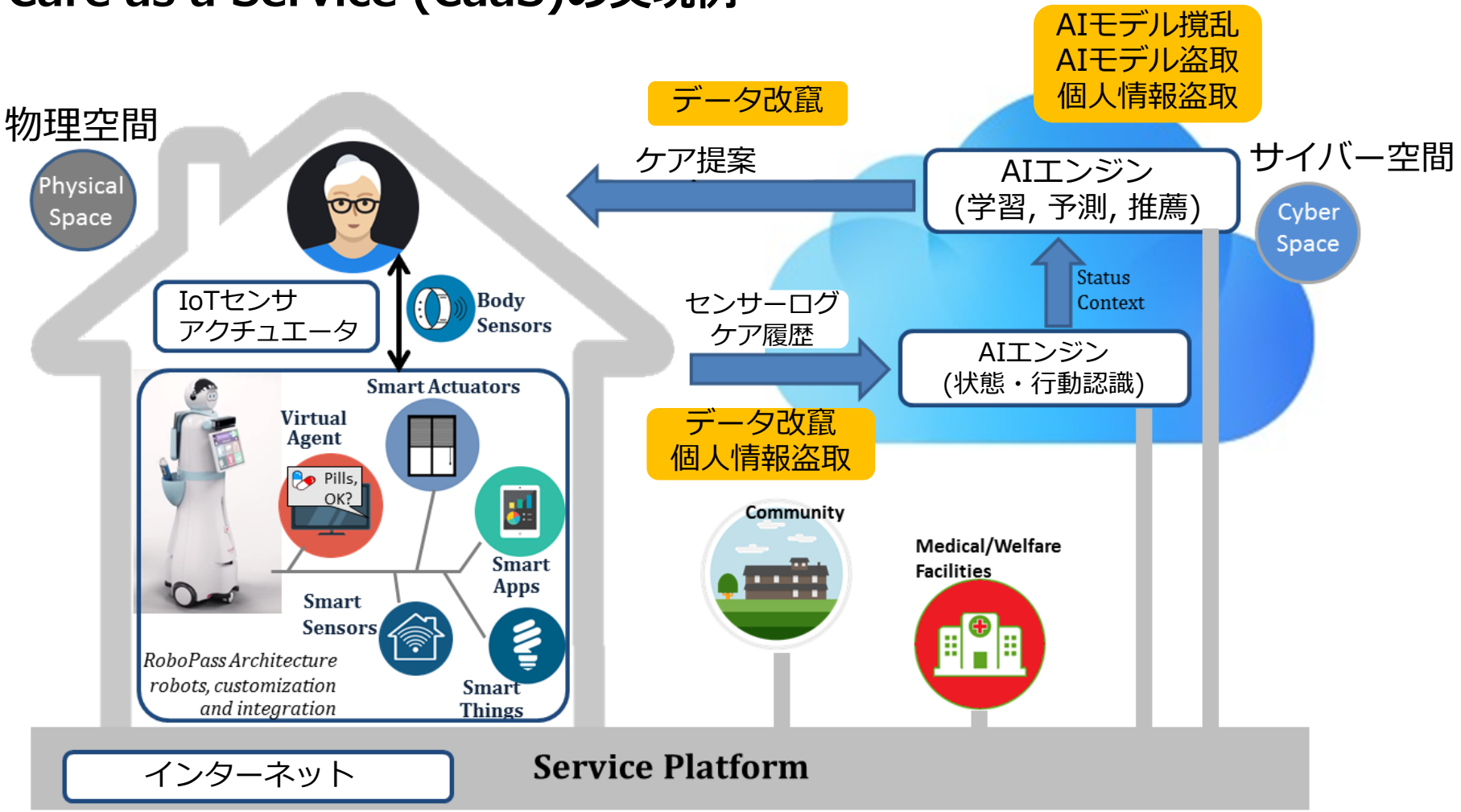
Google Cloud Platform

AIをビジネス展開するときの落とし穴

- AIモデルの撹乱（営業妨害）
- AIモデルの盗取（知財の漏洩）
- 訓練データの盗取（個人情報の漏洩）

サイバーフィジカルシステム (CPS)への脅威

Care as a Service (CaaS)の実現例



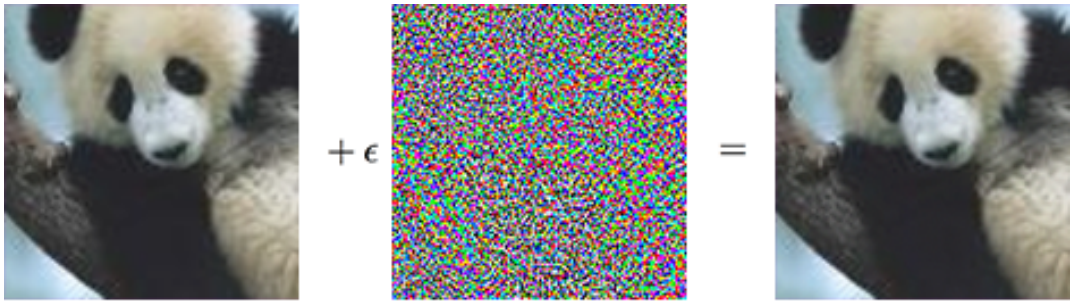
AIモデルへの撹乱攻撃

(1)Evasion attack：難読化（難読化，画像化）

スパムメールやマルウェアの難読化や画像ベーステキストへの置き換え

(2)Poisoning attack：訓練データの操作，ラベルの反転など

arXiv:1801.09573



パンダ

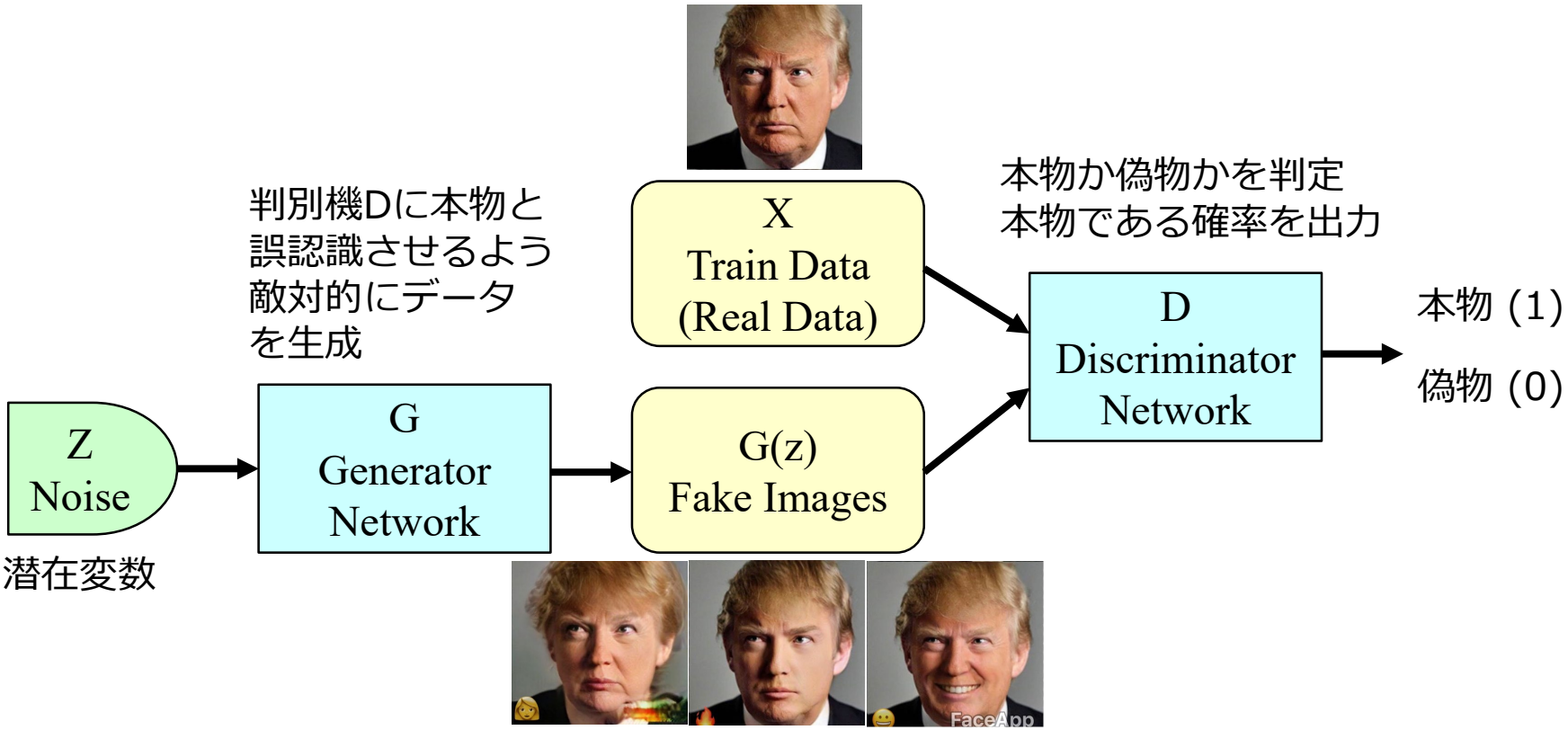
テナガザル

arXiv:1412.6572

敵対的生成ネットワーク

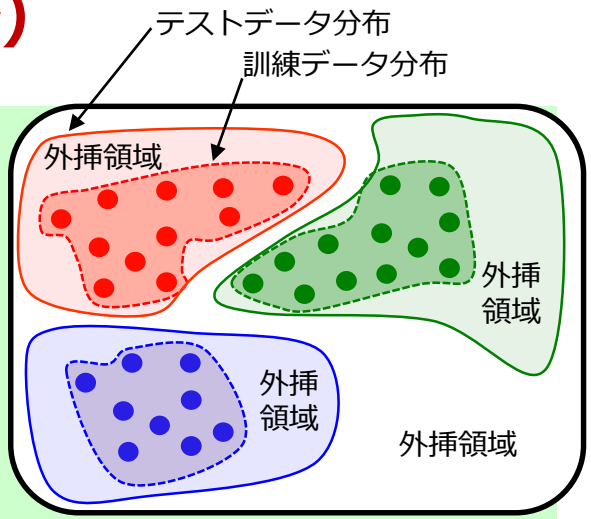
(Generative Adversarial Networks: GANs)

少ないラベル付きデータを補う． → AIを高性能化する方法



Goodfellow, Ian J.; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; Bengio, Yoshua (2014). "Generative Adversarial Networks".

Adversarial Examples
（AIモデルへの撹乱攻撃）



機械学習の仮定

訓練データ分布 ≈ テストデータ分布

現実には、

訓練データ分布 ≠ テストデータ分布

勾配が最大化される方向に、意図的に誤った訓練データを作成すれば機械学習を混乱させることができる。

→ 限られた訓練データで複雑のモデルを学習する

CNNなどの特性を脆弱性として利用した攻撃

特に**外挿能力の低さ**を利用した攻撃

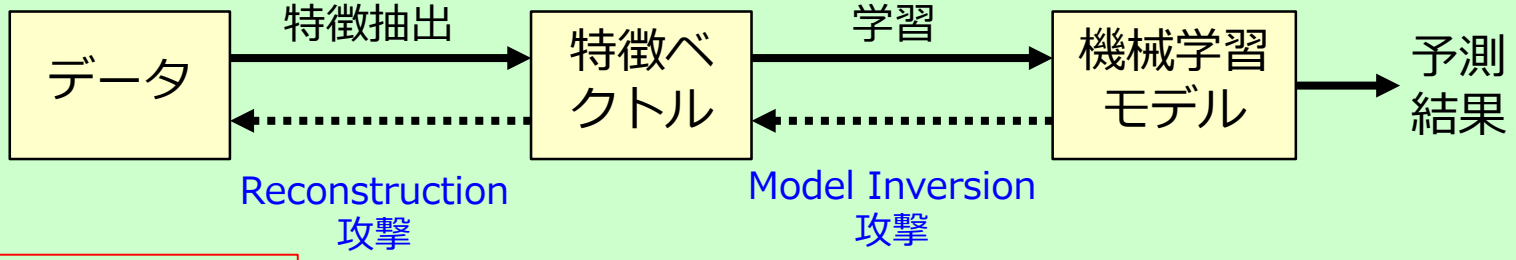
White-box攻撃でもBlack-box攻撃でも可

Adversarial Examplesに対する防御

1. 勾配情報を出力しない（ラベルだけ返す）
2. adversarial exampleも訓練データに混ぜて学習させる。
   → 適度な割合にすると、汎化能力が上がる。
3. 訓練データのauditを実施する。
   訓練データにadversarial exampleが知らないうちに混ざっていることを想定し、入出力関係が大きく変わる訓練データは学習しない。
   → Sandboxの導入
4. Defensive Distillationの導入
   教師ラベルを {0, 1} とせず、(0,1)のソフトターゲットにする。

AIへの脅威モデル～情報盗取

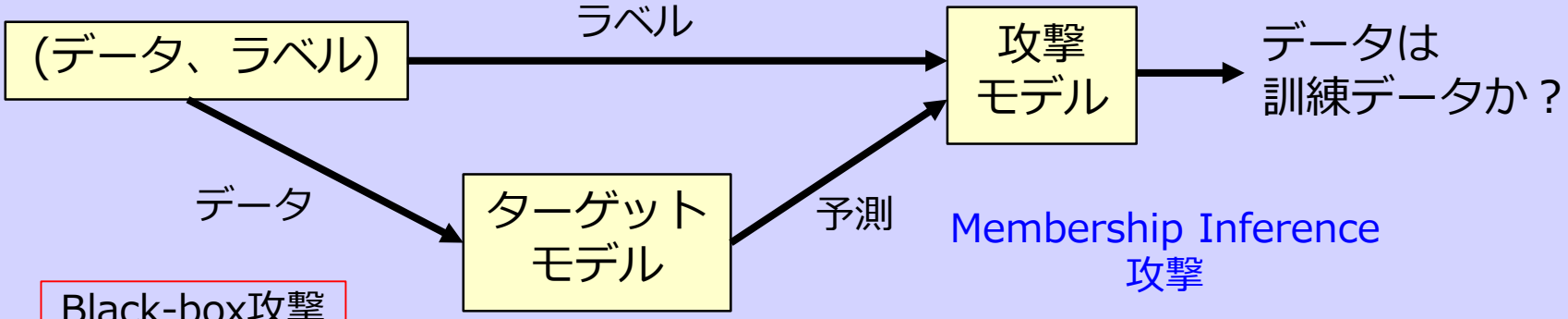
(a) 機械学習モデルと訓練データの推定



White-box攻撃が前提

知的財産・個人情報の盗取が目的

(b) 訓練データかどうかの推定

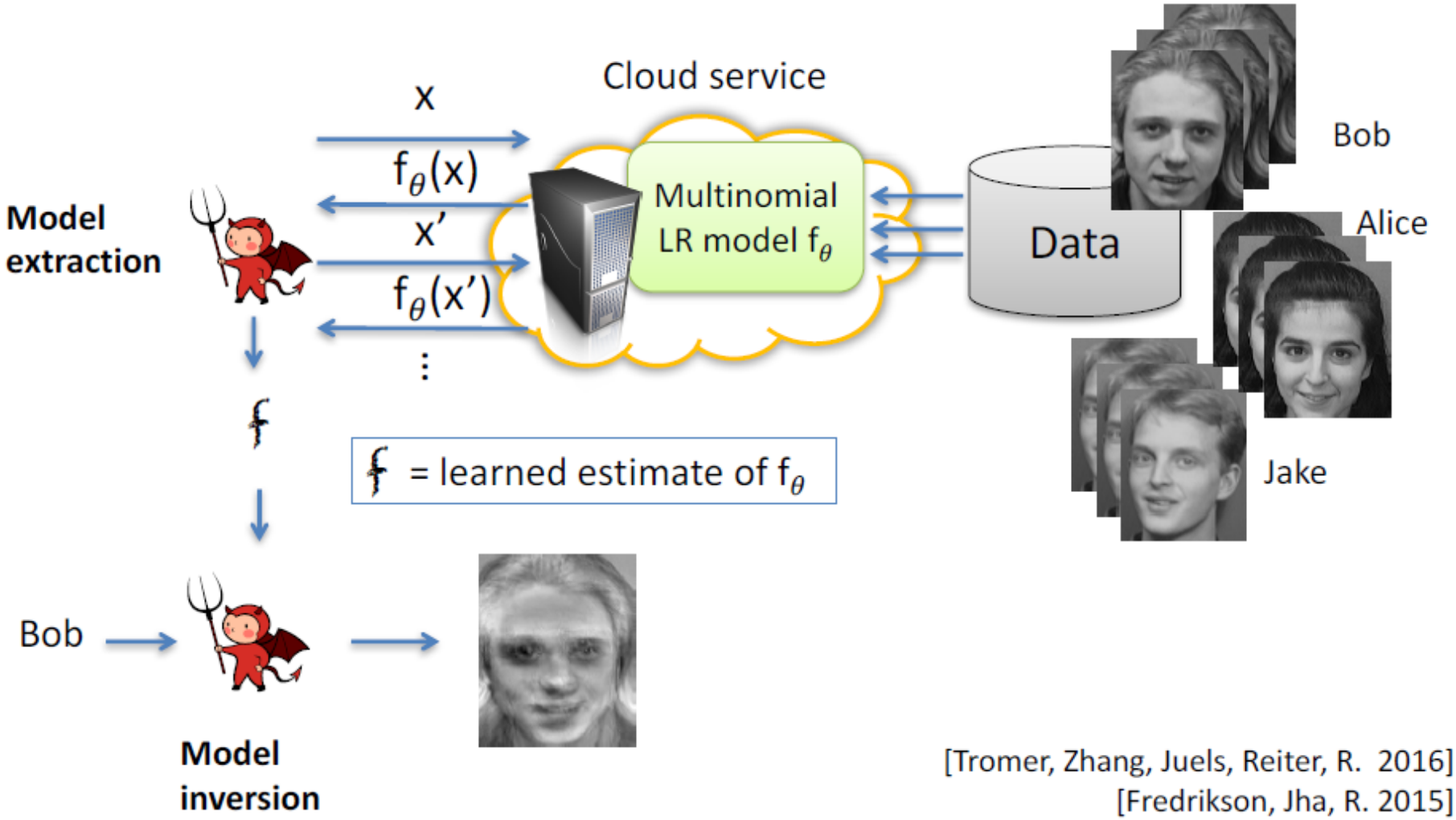


Black-box攻撃でも適用可

Model Inversion 攻撃で入手可

個人データの盗取が目的

訓練データの推定攻撃



[Tromer, Zhang, Juels, Reiter, R. 2016]
[Fredrikson, Jha, R. 2015]

AIのためのセキュリティ ～ 参考情報

日銀金融研究所　ディスカッションペーパー

https://www.imes.boj.or.jp/research/dps-j.html

機械学習システムのセキュリティに関する研究動向と課題 (2018.8)

宇根正志

キーワード：機械学習、人工知能、脆弱性、セキュリティ

概要：機械学習の脆弱性とMLaaSに対する攻撃をタイプ別に整理

金融分野で活用される機械学習システムのセキュリティ分析 (2019.1)

井上紫織、宇根正志

キーワード：機械学習、人工知能、脆弱性、セキュリティ

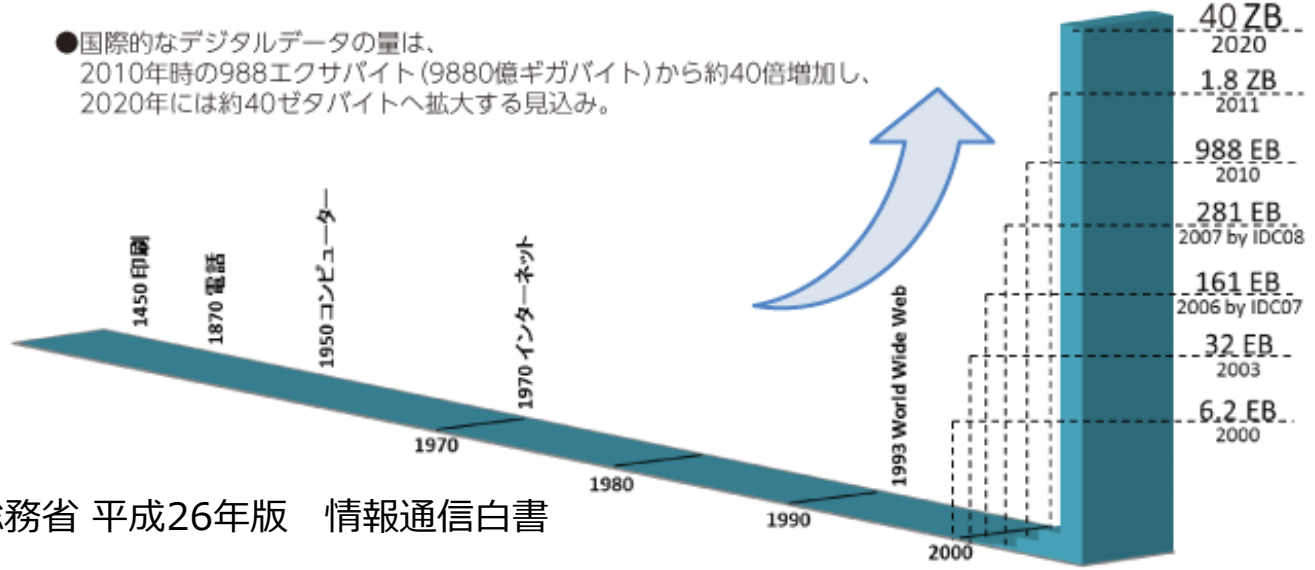
概要：金融システムに実装される機械学習システムの攻撃を機密性、完全性、可溶性の観点から整理し、その対策を議論

AI×セキュリティ（その３）

- AIとセキュリティの新しい展開

ビッグデータ分析の現状

- デジタルデータ量　飛躍的増大
  ICT、パーソナルデバイスやIoTの普及
- 個人情報を含むパーソナルデータの増大
- パーソナルデータの利活用促進
  → 改正個人情報保護法 (H29.5.30)
  匿名加工情報の流通促進 **しかし、…**



総務省 平成26年版　情報通信白書

ビッグデータ分析の現状

- 個人情報の漏洩リスクにより、データ利活用が進まない。
- 名前など個人を直接特定できる情報をマスキングしても、その他の情報から個人が特定できる場合がある。
- 法律や規制のもとで適切に処理されなければいけない。
- 個人のデータそのものでなくても、集団の統計情報や傾向などがわかればよい。

→ プライバシー保護データマイニング (PPDM)

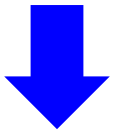
- 匿名化 (*k*-匿名化など)
- 差分プライバシー
- 準同型暗号（一定の演算が可能な暗号）
- 秘密分散
- マルチパーティ計算

匿名加工（マスキング）

適切な基準のもとで匿名加工処理が必要
(たとえば　k-匿名性，差分プライバシー)

- 効用と匿名性のトレードオフ
- 完全な個人情報保護が保証されるわけではない

どのようにして，有効性を失わず有効なルールをビッグデータから抽出できるか？



プライバシー保護データマイニング (PPDM)

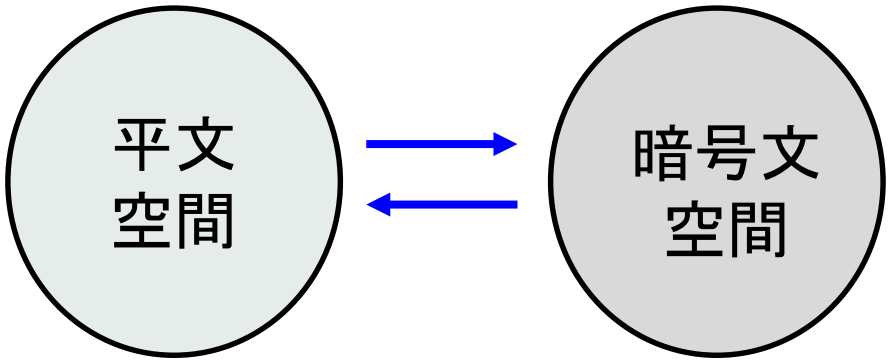
プライバシー保護機械学習 (PPML)

PPDMの手法 (1)

1.準同型暗号 (Homomorphic Encryption)

暗号文での計算を可能にする暗号方式

- 加法準同型 (*Additive HE*)： Pailier
- 乗法準同型 (*Multiplicative HE*)： Unpadded RSA, ElGamal
- 完全準同型 (*Fully HE*)： SHE+bootstrapping, etc.



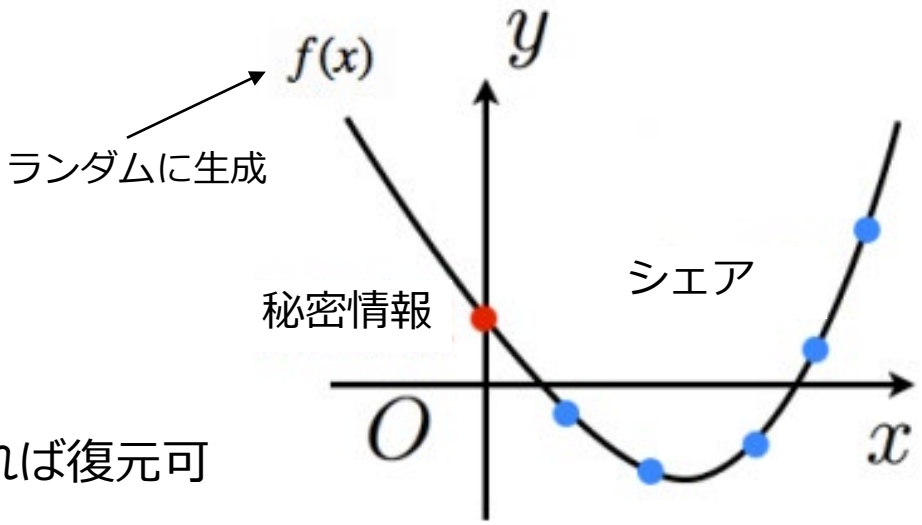
$2 + 3 = 5$

$E(2) + E(3) = E(5)$

PPDMの手法 (2)

2. 秘密分散法

秘密情報を*N*個の「分散情報（シェア）」に分けて*N*台の計算機に与え、そのうち任意の*k* (≦*N*)個を使えば秘密情報を復元できることを保証する。K-1台以下の計算機からシェアが盗まれても秘密情報は漏れず、N-k台以下の計算機であれば故障しても復元できるロバスト性をもつ。このシェアを複数台の計算機で特定のプロトコルに従って演算させることで、秘密情報を知らなくても加算や乗算、排他的論理和などの結果を得ることができる。



シャミアの秘密分散法

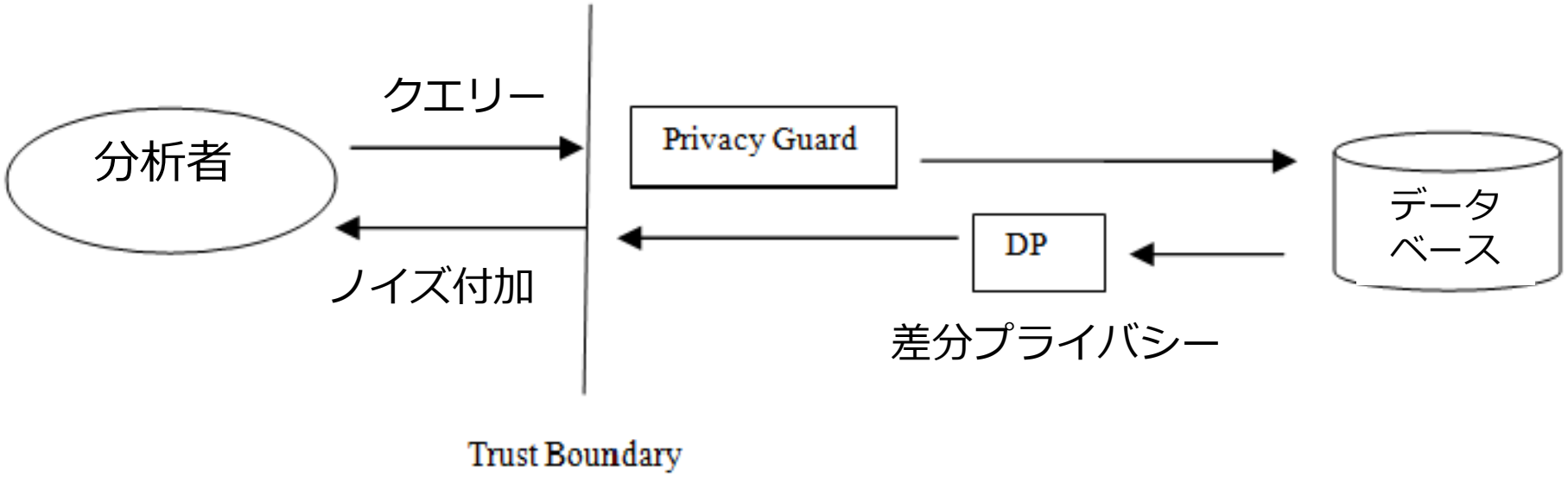
5つのシェアのうち3つ分かれば復元可

PPDMの手法(4)

3. 摂動法(Perturbation Approaches)

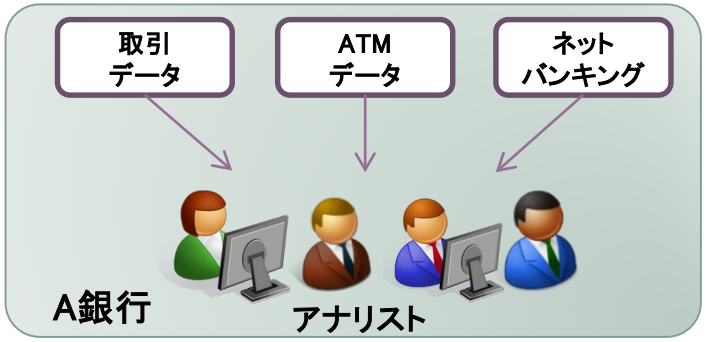
差分プライバシーを保証するメカニズムによって、情報漏洩しないようランダムノイズを付加するアプローチ

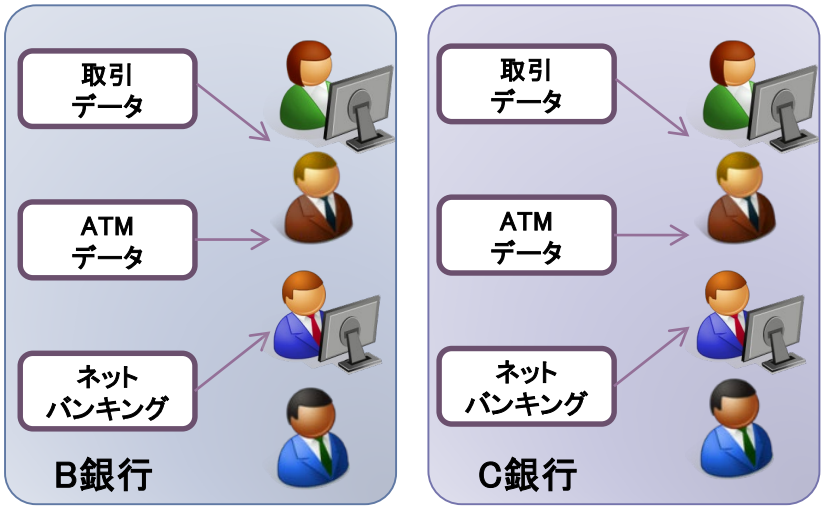
- 入力摂動法
- アルゴリズム摂動法
- 出力摂動法



プライバシー保護データマイニング
～ セキュリティ×機械学習の新たな方向性

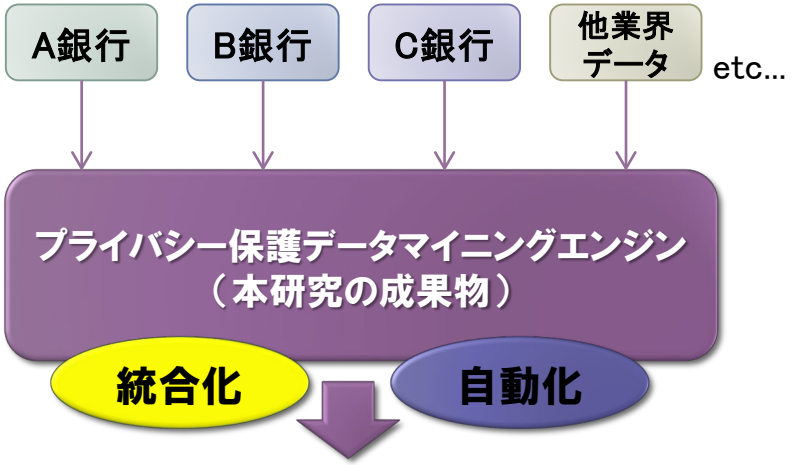
現状





個々の銀行内で分析

めざす構想



不正取引の検知，
与信管理，マーケティング

- 調査コスト削減
- 調査属人化の回避
- 調査精度の向上
  - ➢ 今まで見つからなかった検知が可能に！

研究領域「イノベーション創発に資する人工知能基盤技術の創出と統合化」
研究課題名「複数組織データ利活用を促進するプライバシー保護データマイニング」（盛合志帆）

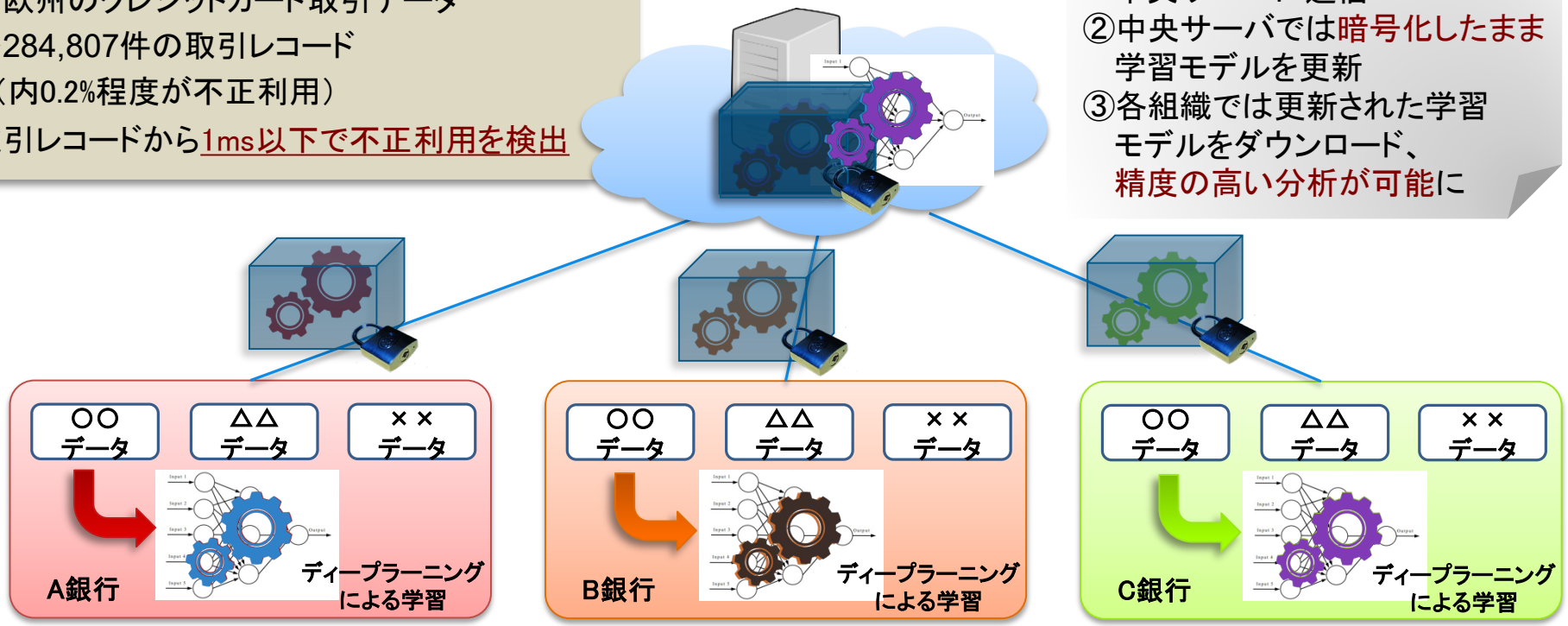
プライバシー保護ディープラーニング

複数の組織が持つデータを外部に開示することなく深層学習を行うプライバシー保護深層学習システム

オープンデータセットを用いた実用性検証
- ◆欧州のクレジットカード取引データ
- ◆284,807件の取引レコード
  （内0.2%程度が不正利用）

取引レコードから1ms以下で不正利用を検出

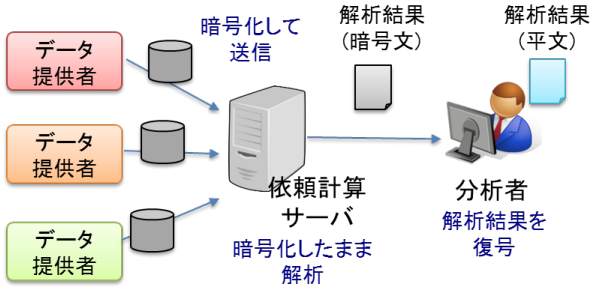
①各組織から学習済モデルのパラメータを暗号化して中央サーバに送信
②中央サーバでは暗号化したまま学習モデルを更新
③各組織では更新された学習モデルをダウンロード、精度の高い分析が可能に

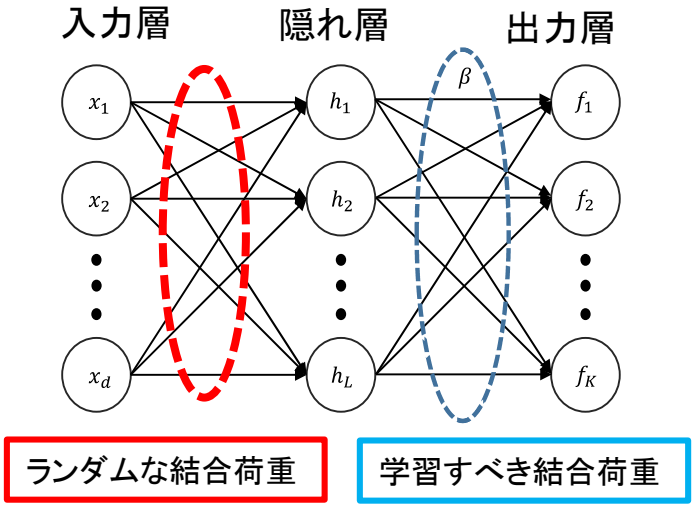


複数組織で連携した分散協調型の深層学習

多入力依頼計算型プライバシー保護機械学習

複数の組織がもつデータを加法準同型暗号で秘匿し、依頼計算サーバで学習・予測が可能なプライバシー保護ELM（Extreme Learning Machine)の提案





分類精度(既存研究との比較)

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:隠れ層のノード数）

+0.04～0.12

1. Single-hidden-layer neural networksの一種
2. 隠れ層の結合荷重はランダムに決め，学習しない
3. 出力層の結合荷重は解析的に求められる

提案した PP-ELM には近似が導入されておらず，ニューラルネット本来の高い精度を示す

JST CREST プロジェクト

[人工知能] イノベーション創発に資する人工知能基盤技術の創出と統合化

領域総括： 栄藤 稔(大阪大学 先導的学際研究機構 教授)

「複数組織データ利活用を促進するプライバシー保護データマイニング」

代表者：盛合 志帆（NICT サイバーセキュリティ研究所 室長）

主たる共同研究者：小澤 誠一（神戸大学）

菅原 貴弘（株式会社エルテス）

複数の異なる業種・組織が有する実社会の膨大なデータを統合して利活用する際に、プライバシー保護やデータ機密性の確保が課題となっています。本研究課題では、暗号技術や人工知能技術を活用し、プライバシーを保護した状態で高速にデータ分析や異常検知を行う技術の研究開発を行います。この技術を金融分野における不正送金検知や顧客に合わせた金利決定の支援に応用し、フィンテックにおけるイノベーション創出を目指します。

金融機関２行と特殊詐欺口座の検知に関する実証実験を開始！

さいごに

1. ニューラルネットや機械学習の手法は様々。（教師あり・なし学習，強化学習，追加学習など）**強みも限界もある**。
2. 問題をよく理解し、正しくアプローチするために、セキュリティと機械学習の**専門家とのコラボ**が必須
3. 攻撃データを**継続的に確保し、ラベルを付ける工夫**が必要（異常検知や攻撃分類などの教師あり学習のタスクの場合）
4. 得られた結果の精査やラベル付けなどで**専門家の介在は不可避**　→ 現状では，完全自動化は困難
5. Adversarial Examplesに**騙されない仕組み**が必要
6. AI・機械学習エンジンが**攻撃の対象**になる.
7. セキュリティ×AIによる新しい**ビッグデータ解析**