

# AIとセキュリティ



東京電機大学  
総合研究所特命教授  
サイバー・セキュリティ研究所所長  
佐々木良一  
[r.sasaki@mail.dendai.ac.jp](mailto:r.sasaki@mail.dendai.ac.jp)



# イントロダクション

---

(1) 自己紹介：佐々木良一 (東京電機大学特命教授)

1971年ー2001年 日立製作所。1984年より情報セキュリティなどの研究に従事

2001年ー2018年3月 東京電機大学未来科学部教授



日本セキュリティマネジメント学会会長  
内閣官房サイバーセキュリティ補佐官  
などを歴任

詳しくは下記参照

<http://www.isl.im.dendai.ac.jp/pdf/佐々木研究室総集編20180213.pdf>

# 目次

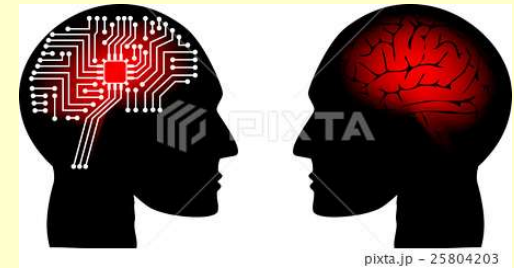
---

1. AIとセキュリティに関する4つの観点
2. AIを利用した攻撃: Attack using AI
3. AI自身による攻撃: Attack by AI
4. AIに対する攻撃: Attack to AI
5. AIを利用したセキュリティ対策: Measure using AI
6. おわりに



# 人工知能研究

- 人工知能(AI): 知能のある機械

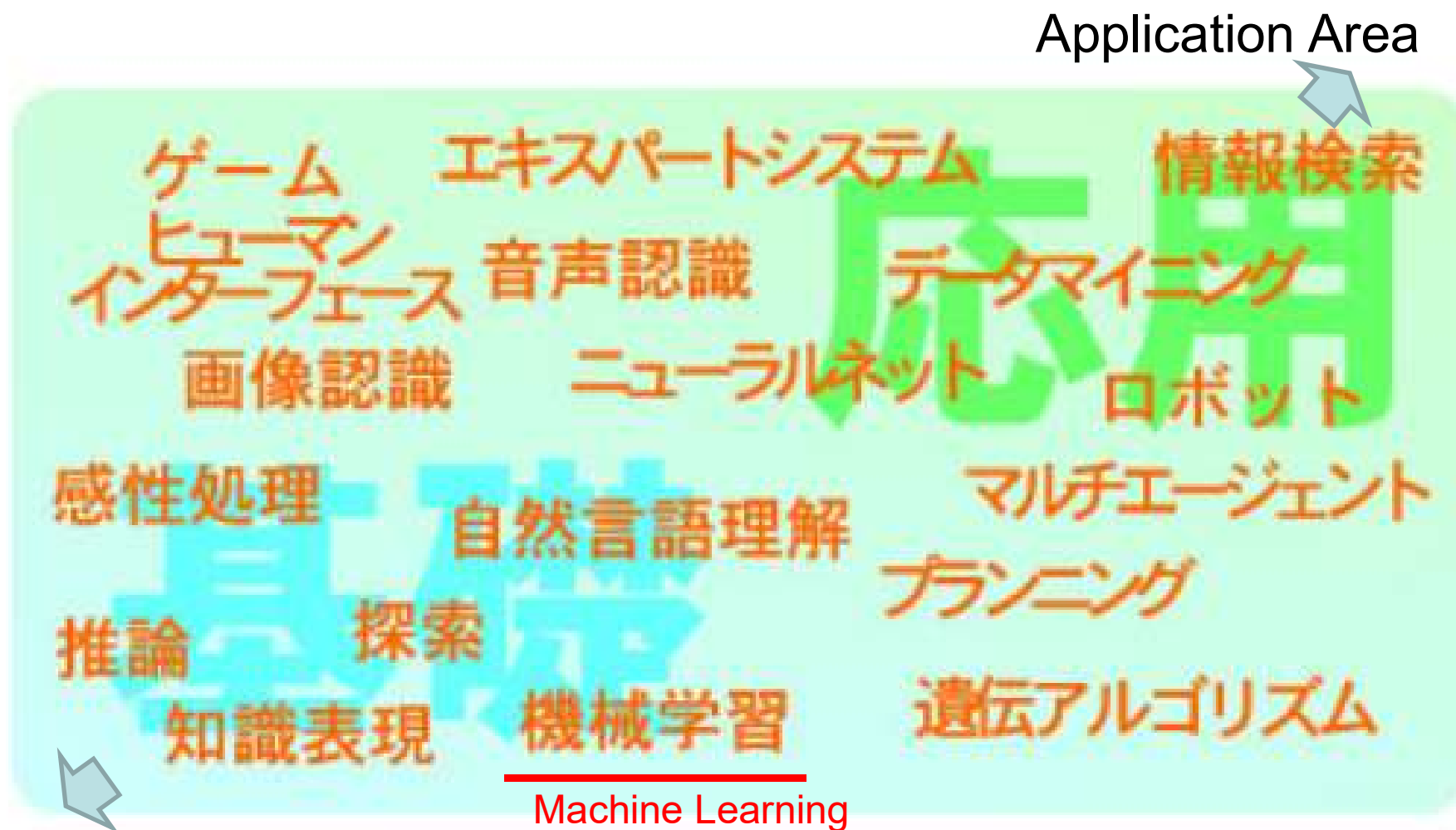


- 2つのAI

- ① 本当に知能のある機械である強いAI(汎用AI)と
- ② 知能があるようにも見える機械, つまり, 人間の知的な活動の一部と同じようなことをする弱いAI(専用AI)とがある. AI研究のほとんどはこの弱いAI.



# 人工知能の研究の分野



Basic Area

<http://www.ai-gakkai.or.jp/whatsai/AIresearch.html>

日本人工知能学会資料

# 基礎分野の主な用語(1)

---

## 機械学習

観測センサーやその他の手段で収集されたデータの中から一貫性のある規則を見つけだそうとする研究  
・ 数学の統計の分野と強い関連がある. [サポートベクターマシーン等]

## 遺伝アルゴリズム

二つの親の特徴が子に混ざり合って遺伝する原理を利用した問題解決の手法.

# 基礎分野の主な用語(2)

---

## 推論

いろいろなルールを統合して矛盾のない答えを導き出すための手法. 最も基本になるのはアリストテレスの三段論法.

## 知識表現

知識を, コンピュータの中で, 的確に内容を表し, 効率よく蓄積する方法についての研究.  
[オントロジー, 知識獲得と知識システム構築方法論, セマンティックWeb, ナレッジマネジメント]

# 応用分野の主な用語(1)

---

## データマイニング

データベース技術と機械学習が結びついた技術で、大量の整理されていないデータから役に立つと思われる情報を見つけだす手法。

## ニューラルネット

生物の神経を元にした手法。機械学習の有力な手法として発展し、AIの各分野で活用されている。



# 応用分野の主な用語(2)

---

## マルチエージェント

簡単な問題を解決できるエージェントがたくさん集まって、複雑な問題を解決しようとするものです。自然界の生物の集団や、金融市場でのディーラの振る舞いを調べたりするのに利用。

## エキスパートシステム

専門家の知見をルールとして蓄積し、推論の手法を用いて問題を解決するシステム

# AI研究の歴史

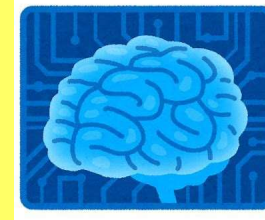
AI ブーム	次期	キー ワード	正 確 性	主な対象	備考
第1次	1950 から60 年代	論理	◎	パズルやゲーム。 期待された機械 翻訳などは失敗	1956年ダートマ ス会議
第2次	1980年 代	知識	○ △	専門家の代わりを 務めるエキスパート システム =>多くは失敗	日本では第5世代コ ンピュータの開発( 述語論理の記述を ハードで直接実行)
第3次	2015年 ごろから	統計 (学習)	△	画像認識・音声 認識・自動翻訳 など	将棋や囲碁でも コンピュータが 優位に

# 現在有効性が高いと考えられるAI

---

## 人工知能 (AI: Artificial Intelligence)

人工的な知能を実現しようとする技術全般



## 機械学習 (ML: Machine Learning)

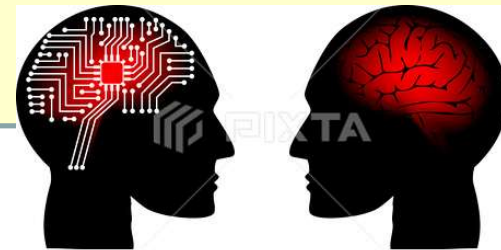
既存データから規則性を学習し、その結果に基づき、新たなデータの分析(認識、分類、予測など)を行う技術

## 深層学習 (DL: Deep Learning)

ニューラルネットにより学習を行い、高精度な分析を行う技術

# セキュリティとAIに関する 4つの観点

- (a) Attack using AI (AIを利用した攻撃)
- (b) Attack by AI (AI自身による攻撃)
- (c) Attack to AI (AIへの攻撃)
- (d) Measure using AI (AIを利用したセキュリティ対策)



# 目次

---

1. AIとセキュリティに関する4つの観点
2. AIを利用した攻撃: Attack using AI
3. AI自身による攻撃: Attack by AI
4. AIに対する攻撃: Attack to AI
5. AIを利用したセキュリティ対策: Measure using AI
6. おわりに



# すでに出現しているAI利用攻撃

---

- ① ボットを利用して、コンサートのチケットの買い占めの試み
- ② ボットのアクセス防止のために、コンピュータが判読を苦手とする画像認証などを利用
- ③ AIを利用して画像認証の解読確率を向上させたようである  
(2018年8月のイープラスサイトでのケースではチケット購入のアクセスのうち9割超がbotだったという。)

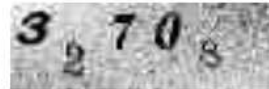
ボット: 一定のタスクや処理を自動化するためのアプリケーションプログラムのこと。ボットはロボットの略語。

# 画像認証画面の一例

ボット排除用



画像認証 不正な申込みを防ぐため、認証を行います。右の画像にある文字を半角で入力してください。



会員の方はこちらから

ログイン画面へ

会員登録をお持ちでない方は、**【新規会員登録】**をお済ませの上、ログインしてお申込みください

[http://atom.eplus.jp/sys/main.jsp?prm=U=82:P6=001:P1=0375:P2=000376:P5=0001:P7=1:P0=GGWC01:P3=0144&\\_ga=2.43604771.767253417.1535088889-1294490883.1535088889](http://atom.eplus.jp/sys/main.jsp?prm=U=82:P6=001:P1=0375:P2=000376:P5=0001:P7=1:P0=GGWC01:P3=0144&_ga=2.43604771.767253417.1535088889-1294490883.1535088889)

# AIを利用した攻撃の可能性

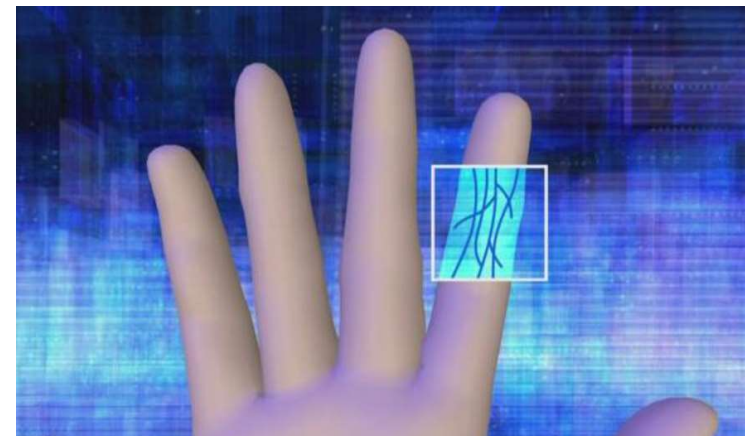
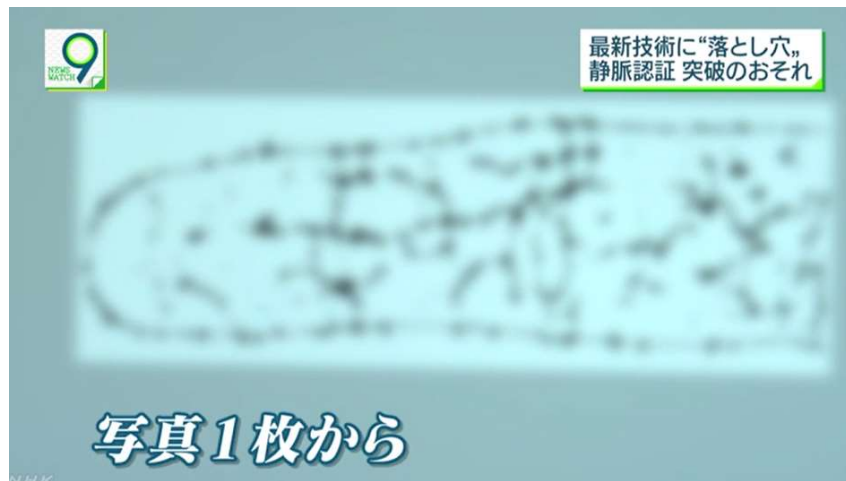
---

- ① 脆弱性情報の悪用: 脆弱性DBなどのインテリジェンス情報を学習して、攻撃行動を自動化
- ② 行動予測による標的型攻撃: 人の行動やシステム内部の脆弱性を予測し、最適な標的型攻撃を実行する
- ③ チャットボットによる詐欺: 悪意に基づきユーザを誘導するチャットボットにより、ソーシャルエンジニアリングを自動化する
- ④ 認証情報の偽造: 攻撃対象の特徴情報を含む指紋データや顔データを自動生成して、認証を突破する



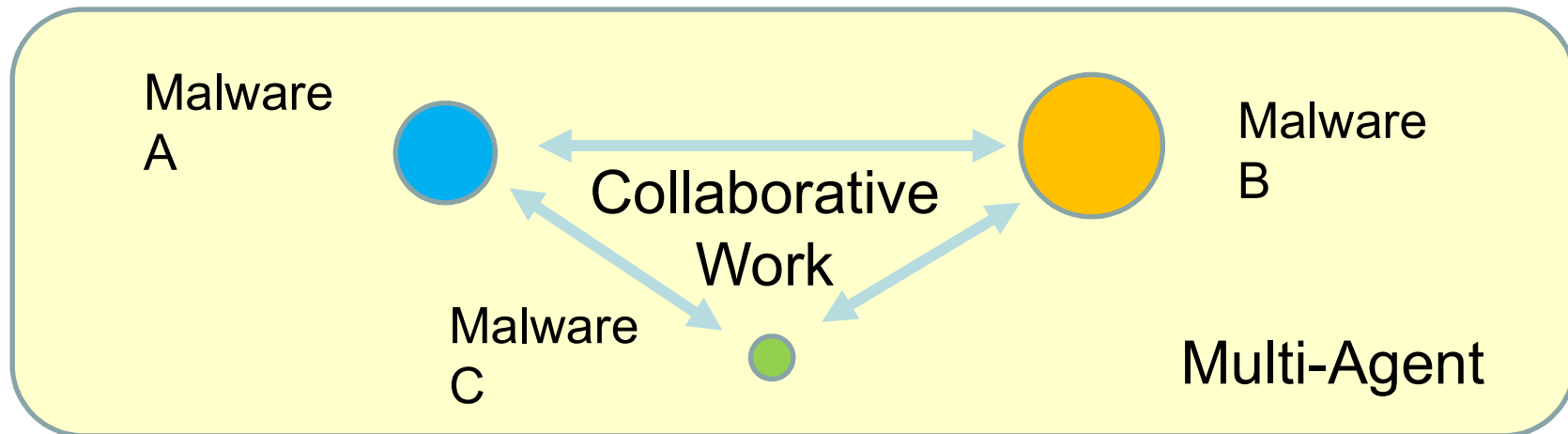
# 指の静脈の生体認証 デジカメ画像で突破される危険

国立情報学研究所の越前功教授らの研究グループは、市販のデジタルカメラで研究員2人の指を50センチの距離から1本ずつ撮影したうえで、指の画像を特殊な方法で加工すると、静脈が浮かび上がり、そのパターンを読み取れることを確認しました。



# 高度な攻撃：AI機能付きのマルウェア

- (例) マルウェアが大きいと侵入時に発見されやすい。  
=> 本格的物は従来はC&Cサーバからダウンロード。  
=> ここでのやり取りがチェックされる場合が多い。  
=> 今後は、小さな種々のマルウェアがいっぱい侵入。  
自律的に協力し合って高度な攻撃を実施？



➡ How do you think ?

# 目次

---

1. AIとセキュリティに関する4つの観点
2. AIを利用した攻撃: Attack using AI
3. AI自身による攻撃: Attack by AI
4. AIに対する攻撃: Attack to AI
5. AIを利用したセキュリティ対策: Measure using AI
6. おわりに



# 関連する動き：2045年問題

---

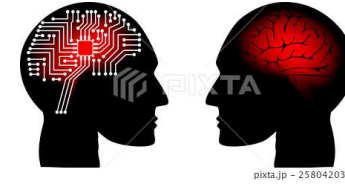
## 2045年問題とは

2045年には人工知能の知的能力が全人類の知的能力の総和よりも大きくなるシンギュラリティ（Technological Singularity：技術的特異点）を迎えるという予測に基づく問題。



松田卓也「来るべきシンギュラリティと超知能の驚異と脅威」情報処理学会誌、V  
ol. 56, No. 1, pp4－14 2015

# AIの反乱



人間を上回る能力を有する機械が誕生し将来的に人間が絶滅させられるではないか。

(a) AIの反乱心配派:

① Googleの研究者のレイ・カールワイツ: 2045年には人工知能の能力が、人間を超越するシンギュラリティが生じ、反乱すら起きるかもしれない

② スティーブン・ホーキング: 人工知能の発明は人類史上最大の出来事だった。だが同時に『最後』の出来事になってしまう可能性もある

(b) AIの反乱非存在派: 西垣通氏、松尾豊氏などは、人間に対して反乱を起こす可能性を否定

# 映画におけるAIの反乱

---

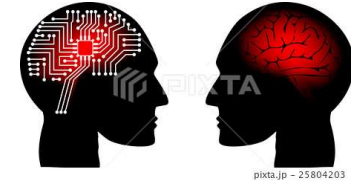
## ① 映画「ターミネーター」

2029年、スカイネット(AI)が人類に反乱を起こし、機械軍を作り出し人類を抹殺し始める。追い込まれた人類は地下に潜みながら抵抗を続ける。

## ② 映画「2001年宇宙の旅」

人工知能HALはクルーに対して持ちうる情報は全て開示する、即ち嘘をつけないようにも設計がなされていた。しかし、本当の目的を木星に着くまでは隠せという矛盾する命令をプログラミングされてしまったため、思考回路が混乱し暴走した。

# AIの反乱



1. 人間を上回る能力を有する機械が誕生し将来的に人間が絶滅させられるではないか。

(a) AIの反乱心配派:

① Googleの研究者のレイ・カールワイツ: 2045年には人工知能の能力が、人間を超越するシンギュラリティが生じ、反乱すら起きるかもしれない

② スティーブン・ホーキング: 人工知能の発明は人類史上最大の出来事だった。だが同時に『最後』の出来事になってしまう可能性もある

(b) AIの反乱非存在派: 西垣通氏、松尾豊氏などは、人間に対して反乱を起こす可能性を否定

# AIの反乱がないとする理由

1. 「強いAI(汎用AI)」ではなく「弱いAI(専用AI)」の研究が中心
2. AIがさらに高度なAIを自動的に作ることは困難
3. 制約を与えることにより反乱を抑えることができる  
例えば＜ロボット3原則＞





# ロボット3原則

**第一条:** ロボットは人間に危害を加えてはならない。また人間が危害を受けるのを何も手を下さずに黙視していてはならない。

**第二条:** ロボットは人間の命令に従わなくてはならない。ただし第一条に反する命令はこの限りではない。

**第三条:** ロボットは自らの存在を護(まも)らなくてはならない。ただし、それは第一条、第二条に違反しない場合に限る。



# 西垣通氏の意見

---

- ① 欧米のシンギュラリティ仮説の支持者たちは人間と人工知能を同一の存在として一次元で比較しようとする。
- ② しかし、人工知能は「人間という生物種の思考」から生まれたものであり、同質とはなりえない。
- ③ 人工知能の知的能力が人間の能力を超えていき、人間の理解できない領域に入るということは、賢くなったのではなく単に壊れただけである。

なお、西洋で人工知能の反乱を恐れるのは、一神教の影響で神に代わって創造主になることに対する恐れでないかという。

西垣通「ビッグデータと人工知能」中公新書, 2016、p111－112の要約

# 私の立場

- (1) AIが反乱を起こす可能性は極めて低い。
- (2) しかし、原子力プラントへの津波来襲に伴う事故にみられるように、人間のリスクに対する知覚能力は極めて低い。
- (3) また、反乱がおきたときや、賢くなったのではなく単に壊れたただけであっても、それが起こると取り返しのつかないことになっている可能性が強い。
- (4) したがって、動きを慎重に見守っていくことが大切。



# 目次

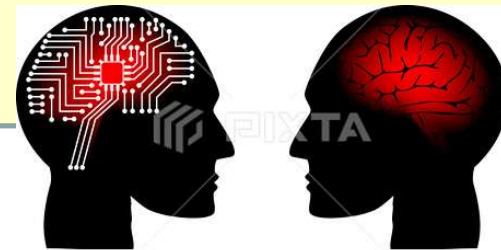
---

1. AIとセキュリティに関する4つの観点
2. AIを利用した攻撃: Attack using AI
3. AI自身による攻撃: Attack by AI
4. AIに対する攻撃: Attack to AI
5. AIを利用したセキュリティ対策: Measure using AI
6. おわりに

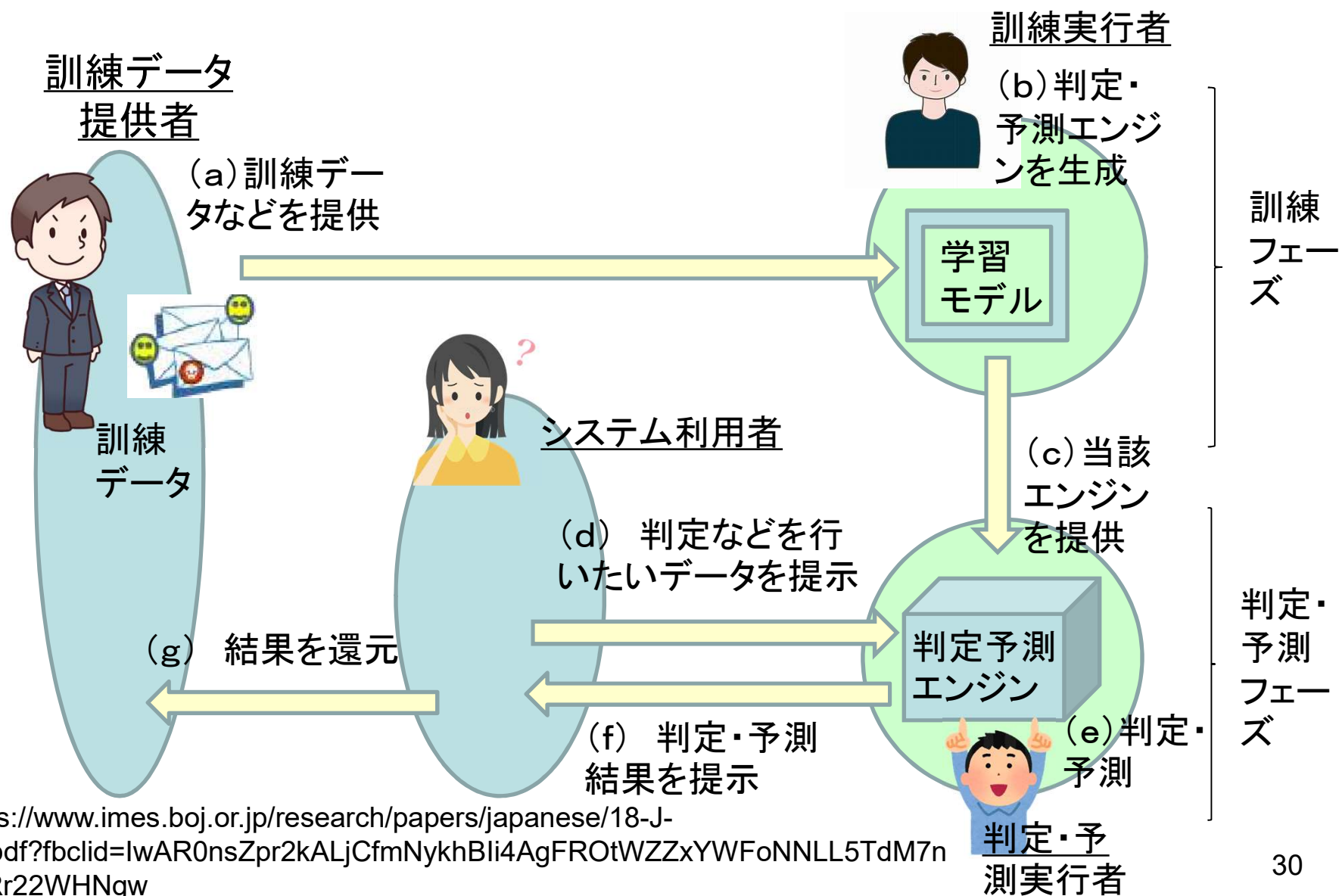


# セキュリティとAIに関する 4つの観点

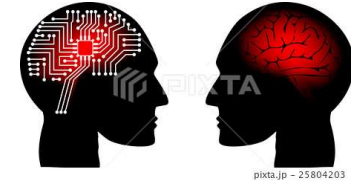
- (a) Attack using AI (AIを利用した攻撃)
- (b) Attack by AI (AI自身による攻撃)
- (c) Attack to AI (AIへの攻撃)
- (d) Measure using AI (AIを利用したセキュリティ対策)



# 機械学習の利用形態の概要



# AIへの攻撃



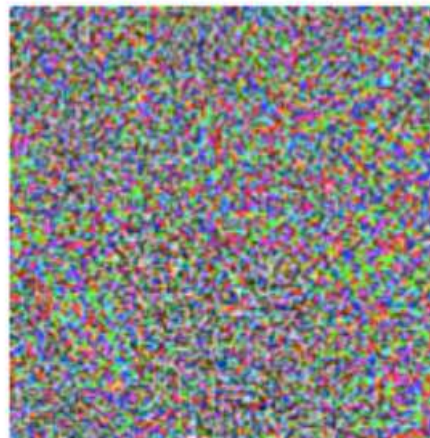
- ① 機械学習システムの停止やファイル情報、通信路情報の盗み出しなどの攻撃: 他システムへの攻撃と同じ
- ② 訓練済みモデルの誤分類を誘発する攻撃
- ③ 機械学習に対する偏った訓練データを意図的に与えるなどが原因で、不適切な判断をさせてしまう攻撃。
- ④ 学習モデルへデータを入出力することにより情報を漏洩させる攻撃(訓練データ、判定特性エンジンなど)

# 既知のモデルに誤分類を 誘発する攻撃

敵対的サンプル  
(Adversarial)  
Example



+ .007 ×



=



「パンダ」

摂動

「テナガザル」

Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy  
“Explaining and Harnessing Adversarial Examples”  
<https://arxiv.org/pdf/1412.6572.pdf>

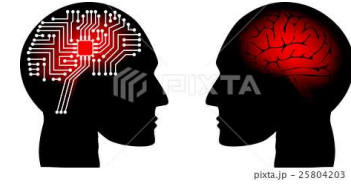


# 顔検出を防ぐアイウェア PrivacyVisor®

---

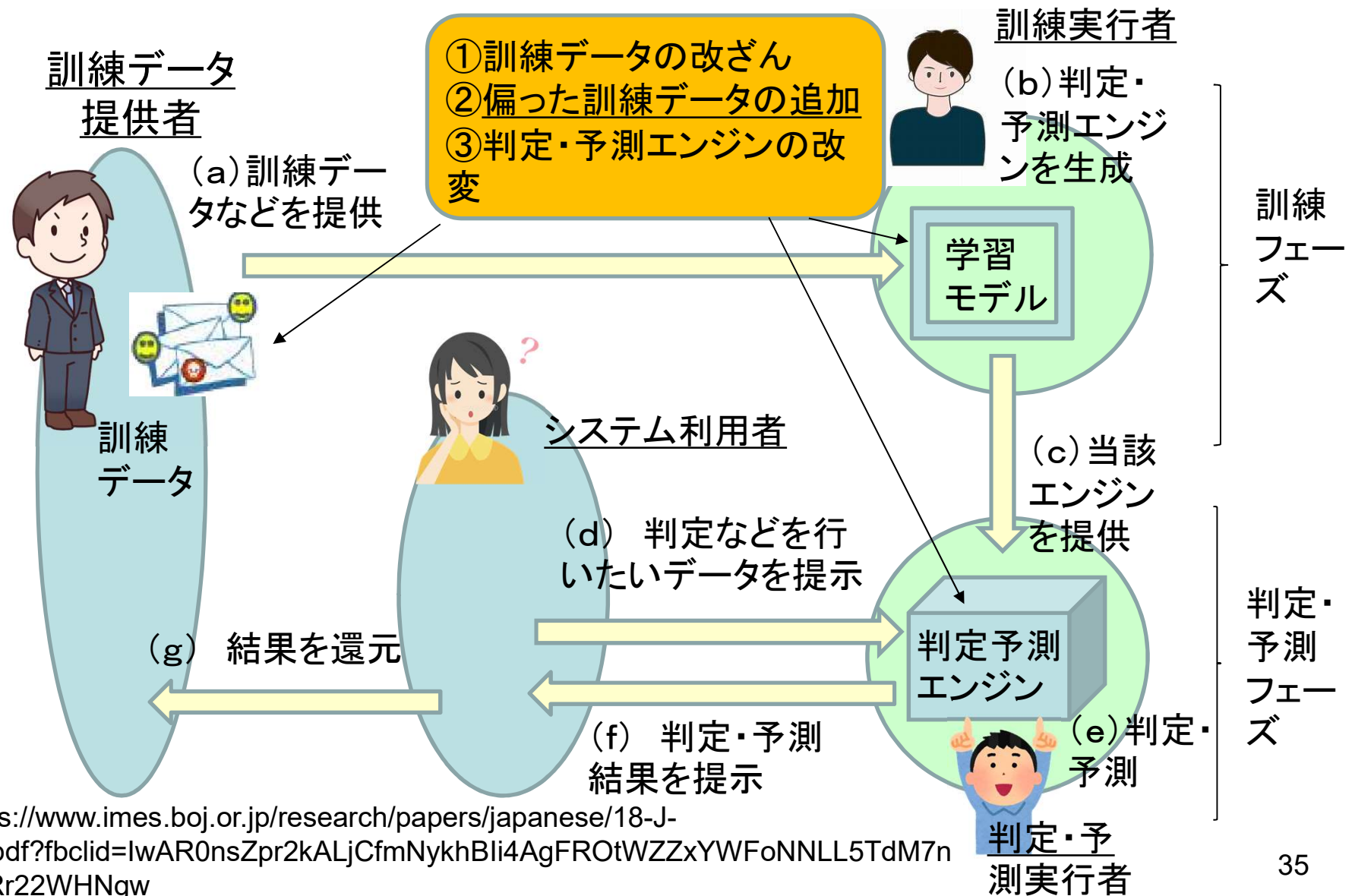
国立情報学研究所(東京都千代田区、所長:喜連川 優)の越前 功 教授が、顔に装着可能なデバイスを用いて顔認識を妨げる研究に取り組んできました。『PrivacyVisor®』は越前教授が開発した顔検出を妨げる技術を採用しています。

# AIへの攻撃



- ① 機械学習システムの停止やファイル情報、通信路情報の盗み出しなどの攻撃: 他システムへの攻撃と同じ
- ② 訓練済みモデルの誤分類を誘発する攻撃
- ③ 機械学習に対する偏った訓練データを意図的に与えるなどが原因で、不適切な判断をさせてしまう攻撃。
- ④ 学習モデルへデータを入出力することにより情報を漏洩させる攻撃(訓練データ、判定特性エンジンなど)

# 不適切な判断の原因



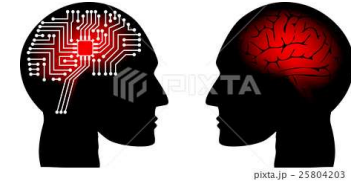
# 訓練データを汚染する攻撃例

---

## <Data Poisoning Attacksの例>

- 米マイクロソフトのチャットボット「Tay」は、クラウドソーシングを利用して学習させた。
- 悪意を持ったユーザが協力して差別的な意見を繰り返し入力
- Tayは差別発言を繰り返すように  
⇒ 適切な訓練データの必要性

# AIへの攻撃



- ① 機械学習システムの停止やファイル情報、通信路情報の盗み出しなどの攻撃: 他システムへの攻撃と同じ
- ② 訓練済みモデルの誤分類を誘発する攻撃
- ③ 機械学習に対する偏った訓練データを意図的に与えるなどが原因で、不適切な判断をさせてしまう攻撃。
- ④ 学習モデルへデータを入出力することにより情報を漏洩させる攻撃 (訓練データ、判定特性エンジンなど)

# 目次

---

1. AIとセキュリティに関する4つの観点
2. AIを利用した攻撃: Attack using AI
3. AI自身による攻撃: Attack by AI
4. AIに対する攻撃: Attack to AI
5. AIを利用したセキュリティ対策: Measure using AI
6. おわりに



# AI for measures against Cyber Attack

多くの研究・開発が存在



MIT（マサチューセッツ工科大学）「85%の攻撃を自動検出」

それでは、人工知能（AI）を活用したサイバー対策の実例を紹介しましょう。最初は、MIT（マサチューセッツ工科大学）のコンピュータ科学および人工知能研究所が、開発したサイバー攻撃検知システムです。

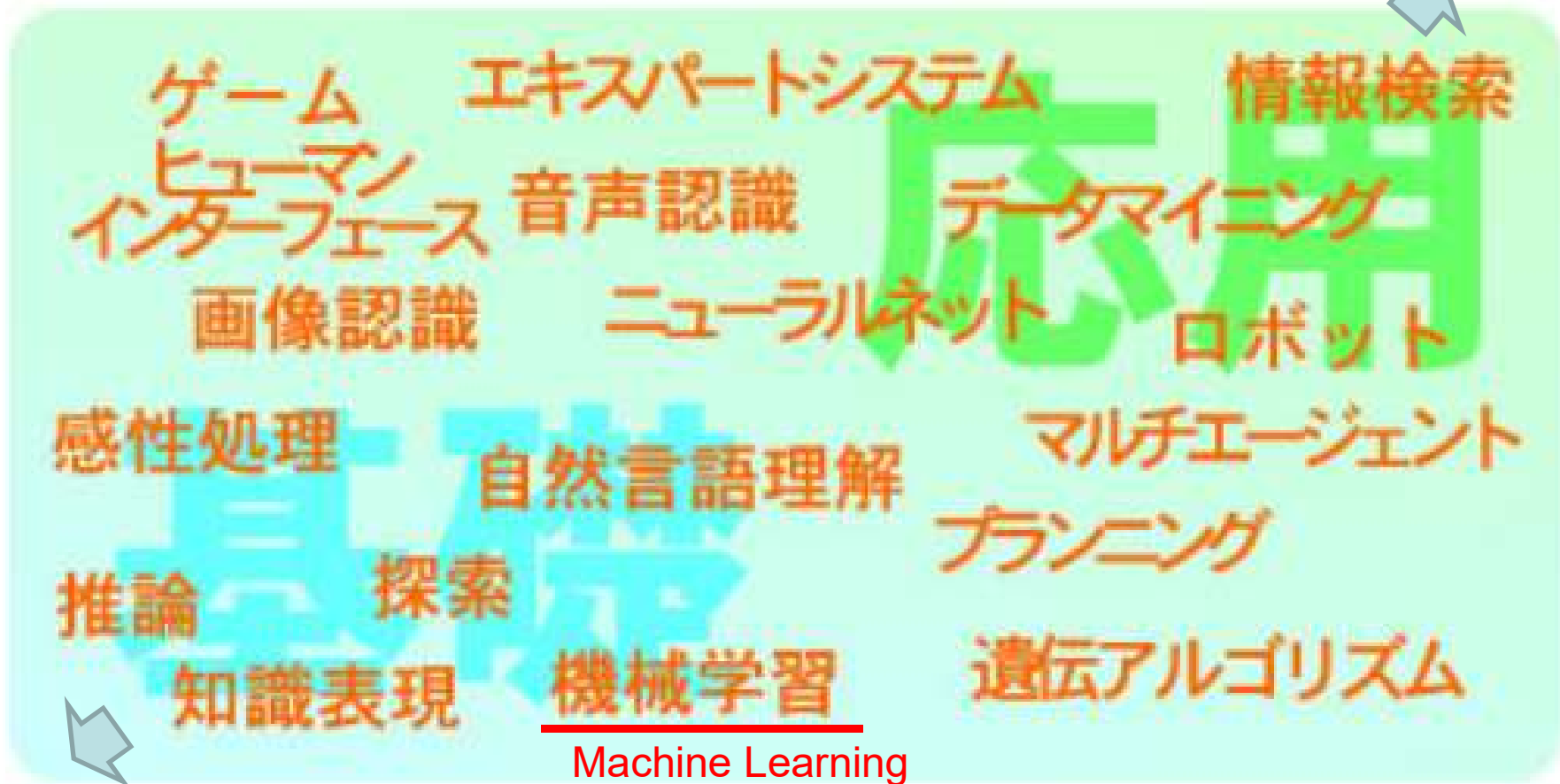
サイバー攻撃対策  
における  
人工知能（AI）  
の活用

予想できない新たな攻撃や複数の手法を組み合わせた攻撃、新種のマルウェアとその亜種の矢継ぎ早な出現など、私たちは常に未知の攻撃にさらされており、過去の防御策は役に立たなくなっています。そこで、サイバー攻撃対策に機械学習などを取り入れた、人工知能（AI）の活用が始まっています。ここでは、MIT（マサチューセッツ工科大学）、ソフトバンク、NTTコミュニケーションズが発表したサイバー攻撃対策を紹介します。

# 人工知能 (Artificial Intelligence) 研究の分野

3つめのインテリジェンス

Application Area



Basic Area

<http://www.ai-gakkai.or.jp/whatsai/AIresearch.html>

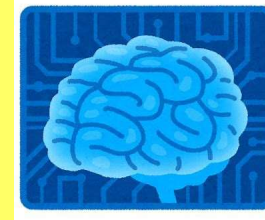
日本人工知能学会資料



# AIの分類

## 人工知能 (AI: Artificial Intelligence)

人工的な知能を実現しようとする技術全般



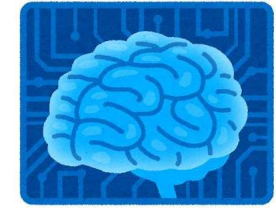
## 機械学習 (ML: Machine Learning)

既存データから規則性を学習し、その結果に基づき、新たなデータの分析(認識、分類、予測など)を行う技術

## 深層学習 (DL: Deep Learning)

ニューラルネットにより学習を行い、高精度な分析を行う技術

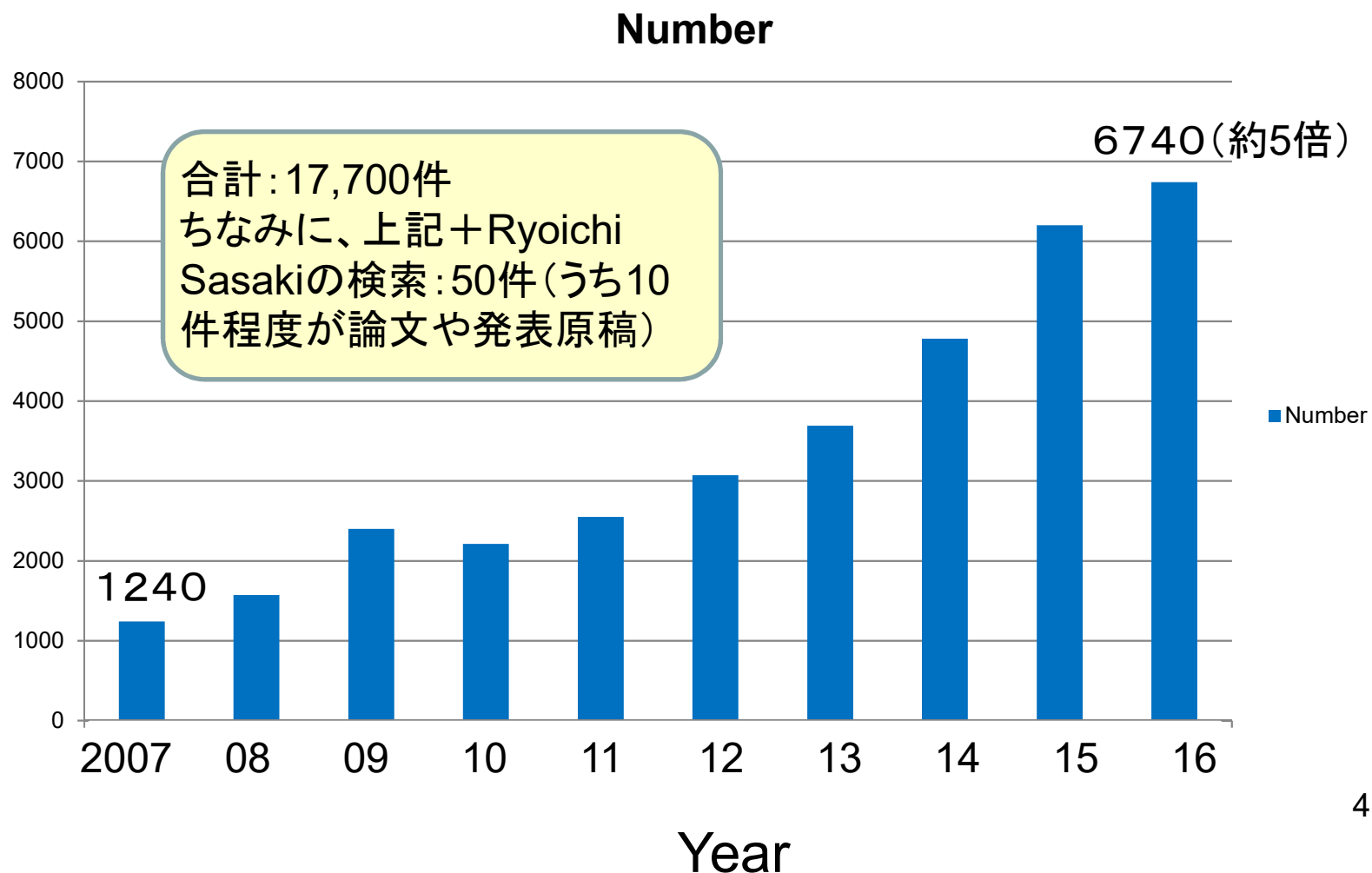
# 機械学習の分類



方式	教師あり学習	教師なし学習	強化学習
概要	問題Xと正解を表すラベル付きデータYから規則性を学習  X->Yの規則性を学ぶ	ラベルのないデータXをもとにその特徴を学習  Xのクラスタリングを学ぶ	環境に対し、「報酬」を最大化させる「行動」を反復学習  行動選択の戦略を学ぶ
用途	回帰・分類	クラスタリング	解の探索
例題	C&Cサイトか通常のサイトかの分類	ウイルスのグループ化	将棋や囲碁の次の一手

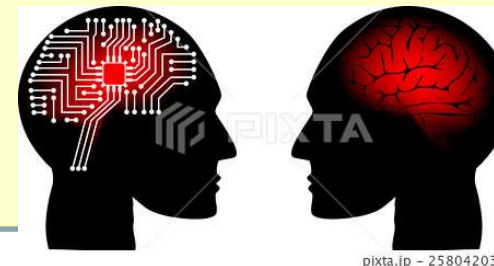
# Google Scholarを用い Cyber Security AND AIで探索した結果

---



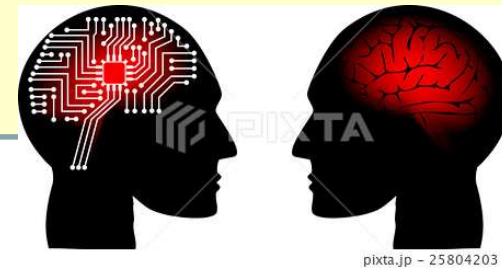
# AIを適用したセキュリティ対策

- 「マルウェアの検出」
- 「ログの監視・解析」
- 「継続的な認証」
- 「トラフィックの監視・解析」
- 「セキュリティ診断」
- 「スパムの検知」
- 「情報流出」など



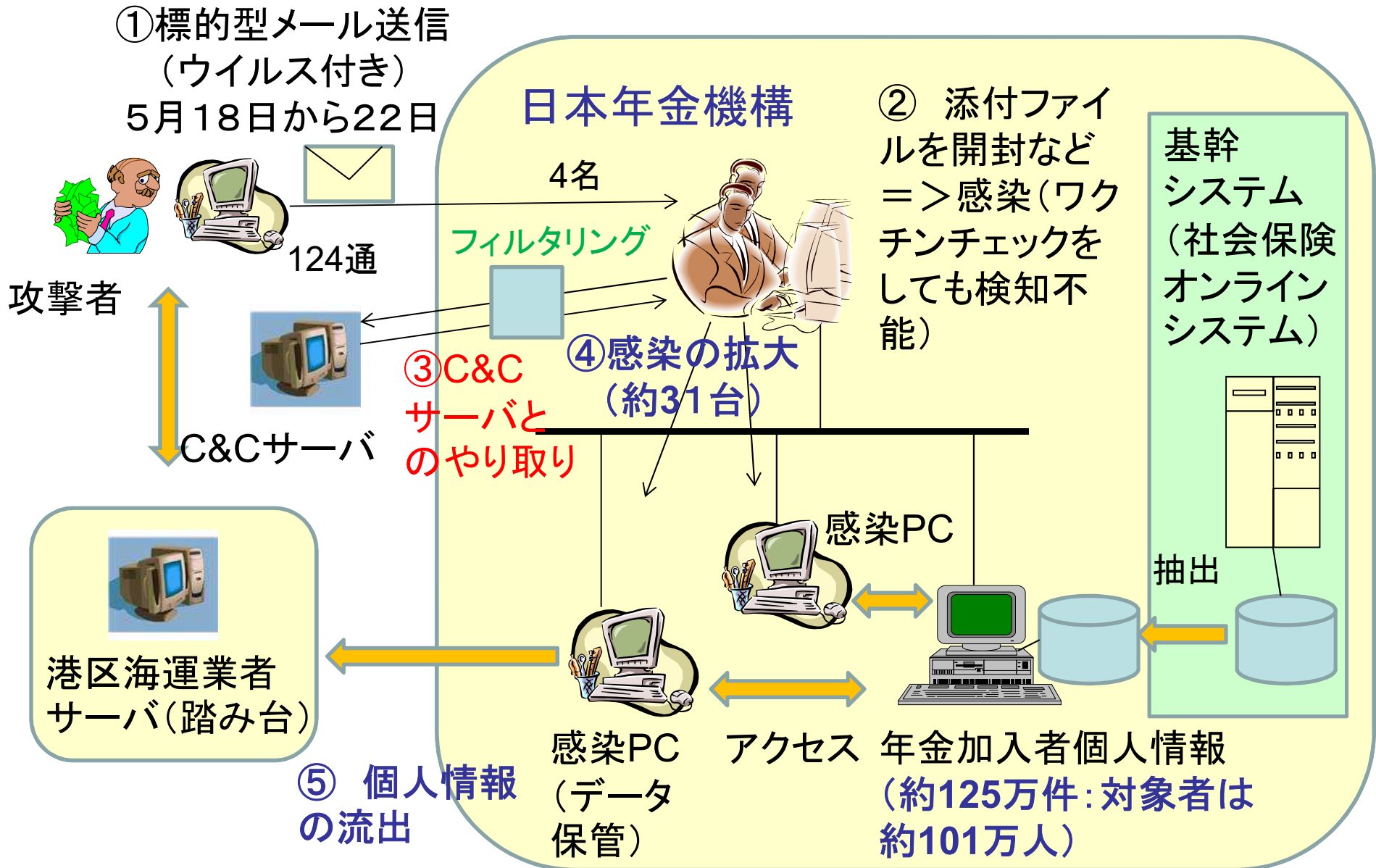
# わたしたちの主要な研究

- (1) 機械学習を利用した標的型攻撃用C&Cサーバの自動判別システムの開発
- (2) 機械学習以外のAI応用としてルールベースシステムやベイジアンネットワークを利用した知的ネットワークフォレンジックシステム

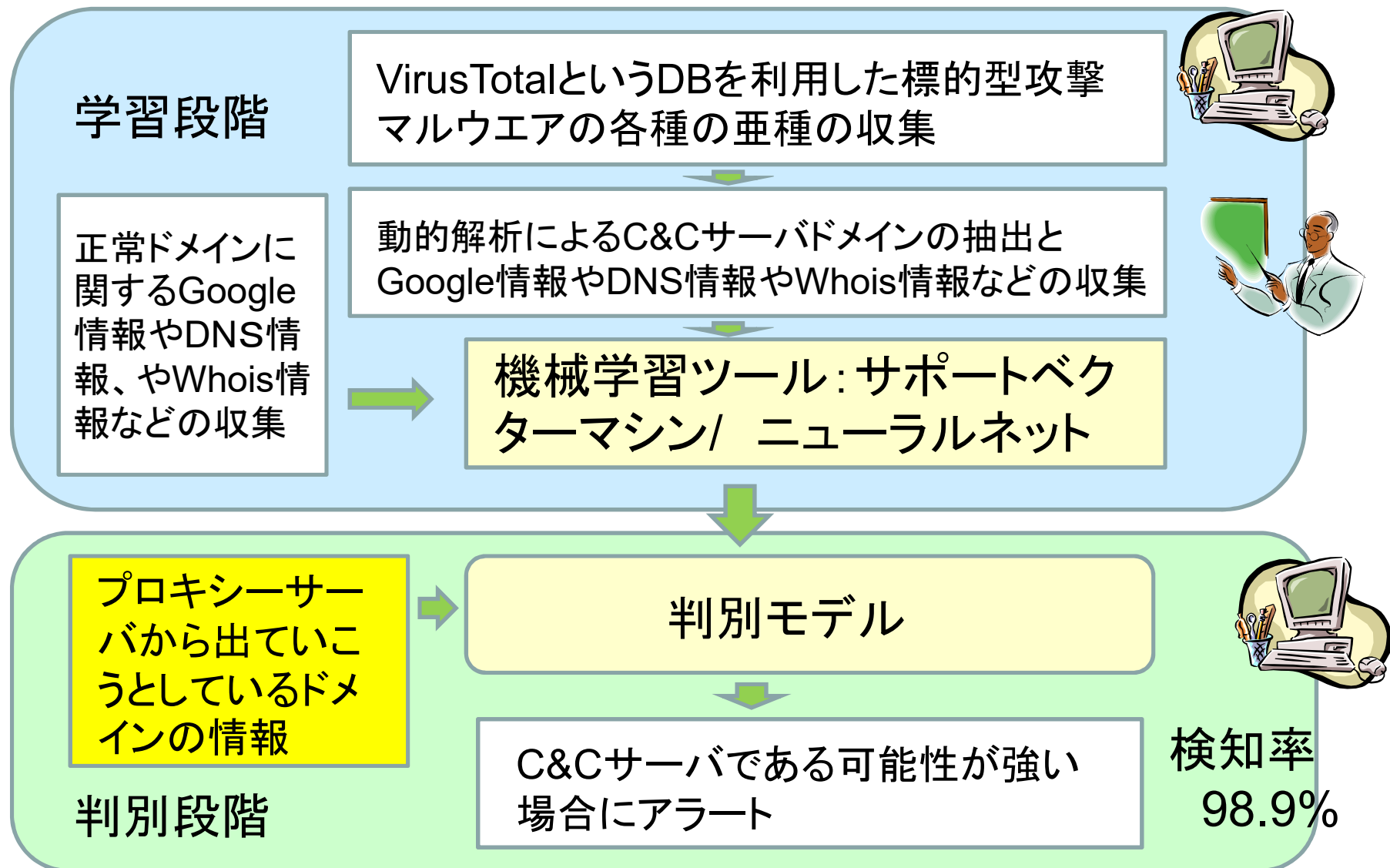


pixta.jp - 25804203

# 日本年金機構への標的型攻撃



# 教師あり学習の適用



久山真宏、柿崎淑郎、佐々木良一「攻撃者に察知されにくい情報を用いたC&Cサーバの検知手法の提案と評価」情報処理学会論文誌, Vol.58, No.9, pp1410-1418, 2017

# セキュリティ対策に機械学習 を用いる困難点


- ① サイバー攻撃などの例は少なく、データが十分でない場合があり、データをどう確保するのが課題である
- ② 攻撃パターンが時間とともに変化し、古いデータが使えない場合がある
- ③ 特に、攻撃側が学習モデルを知って、あえて、分類を間違えるように変更することが可能である



継続的なモデルの改良が必要

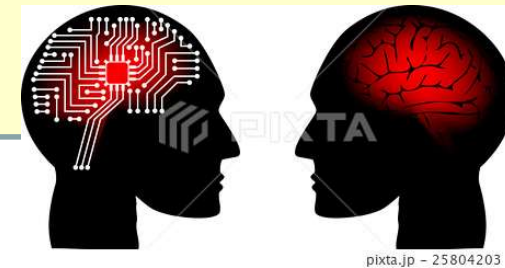


# 機械学習のセキュリティ応用の方針案

段階	学習法	処理 
監視・検知	教師なし学習	(すべての異常状態の列挙は困難なので) 正常な振る舞いや状態を学習 ➡ 未知の異常も検知
↓		
分析	教師あり学習	分析官による分析結果(正解)を学習 ➡ 過去の分析結果に基づき原因を推定 (主なものがわかるだけで有用)
↓		
対処	強化学習	セキュアな状態を報酬化して反復学習 ➡ 最適な対処策を決定(攻撃側に変化や進化に応じて行動)

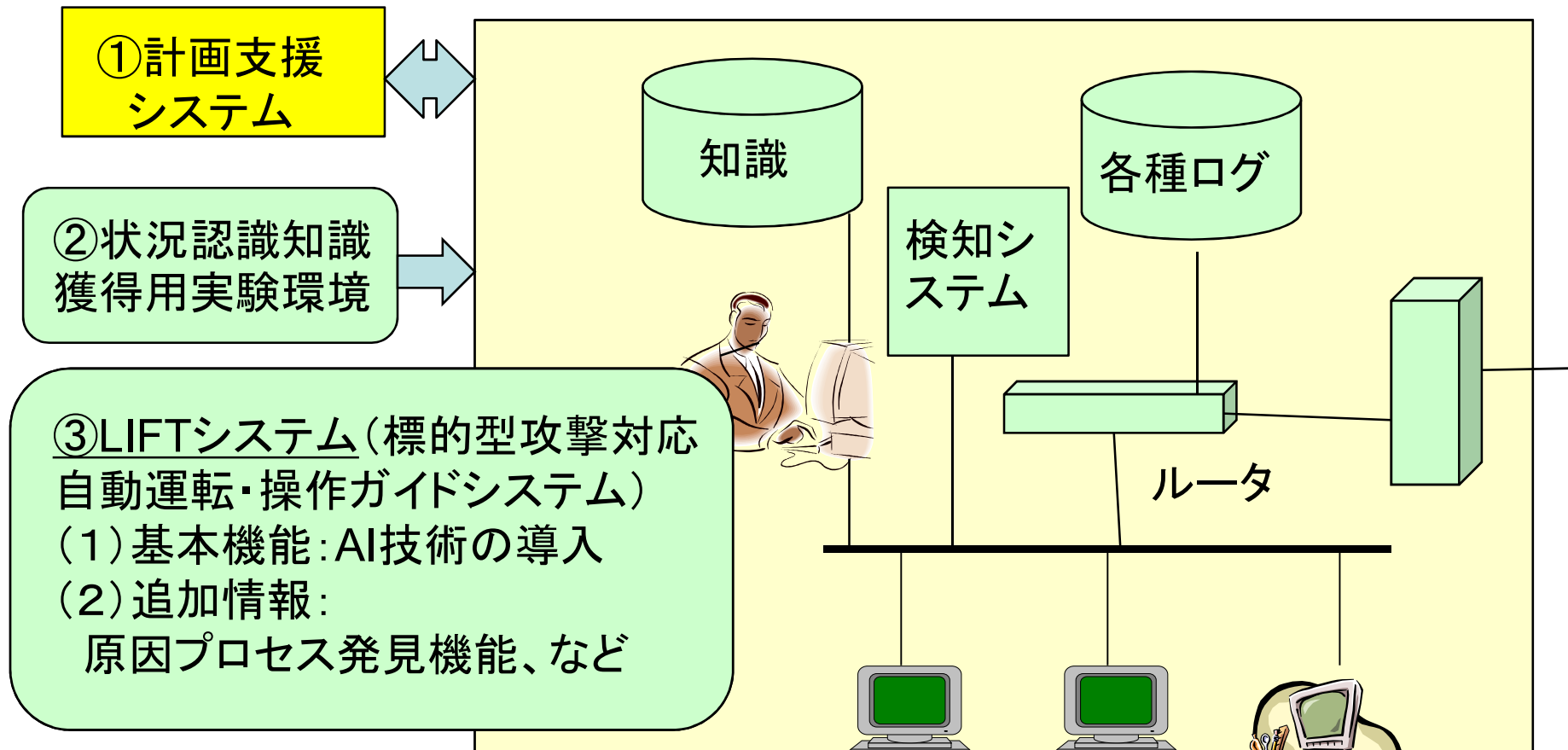
# わたしたちの主要な研究

- (1) 機械学習を利用した標的型攻撃用C&Cサーバの自動判別システムの開発
- (2) 機械学習以外のAI応用としてルールベースシステムやベイジアンネットワークを利用した知的ネットワークフォレンジックシステム



pixta.jp - 25804203

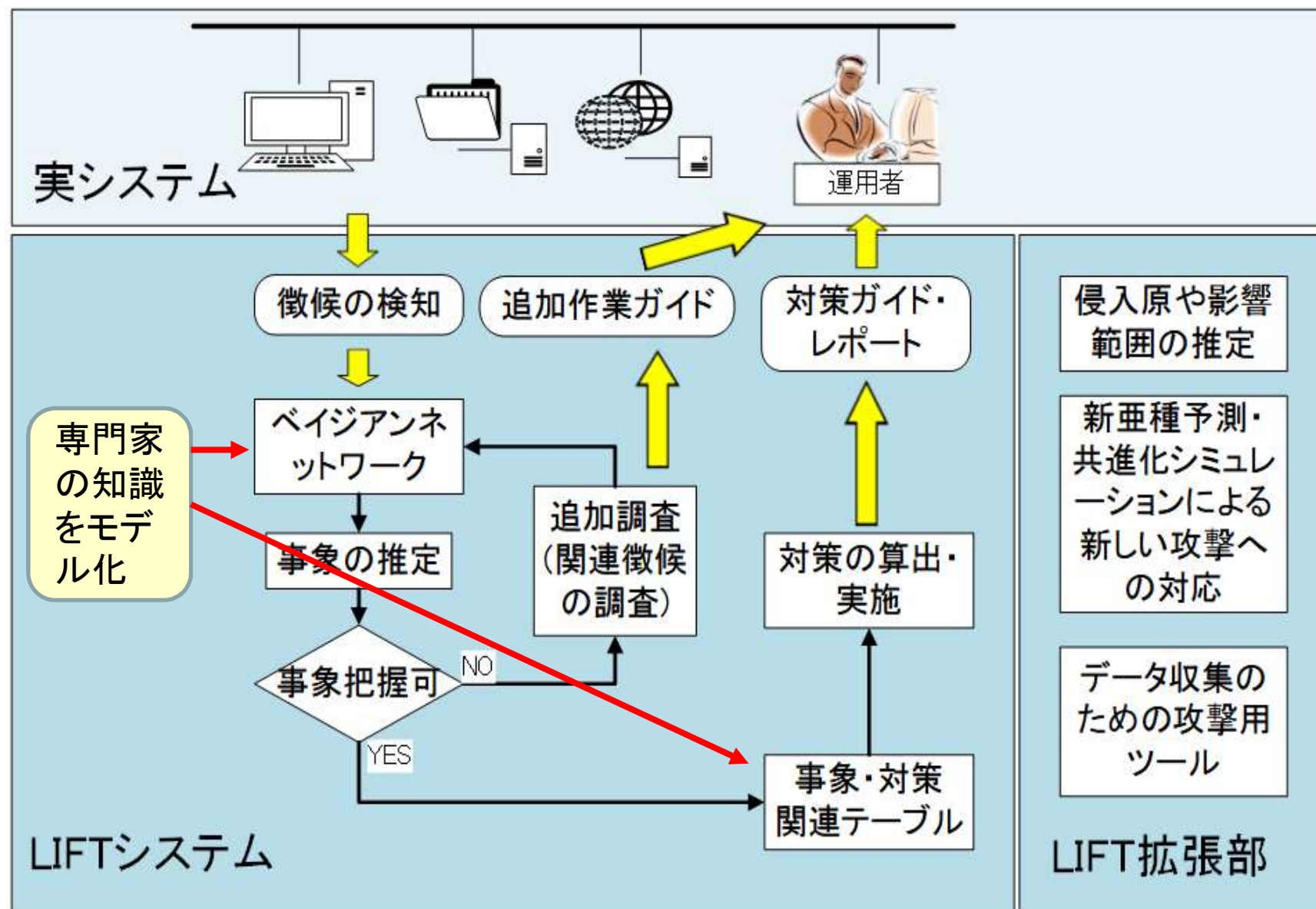
# LIFTプロジェクトの概要



共同開発プロジェクト (リーダー佐々木、上原先生、高倉先生、八槨先生、柿崎先生、日立他) 期間:2013年9月ー2018年3月(第一期)

現状での主な成果:① 方式確立プロト開発 ②原因プロセス発見ソフト製品化



# LIFTシステムの概要



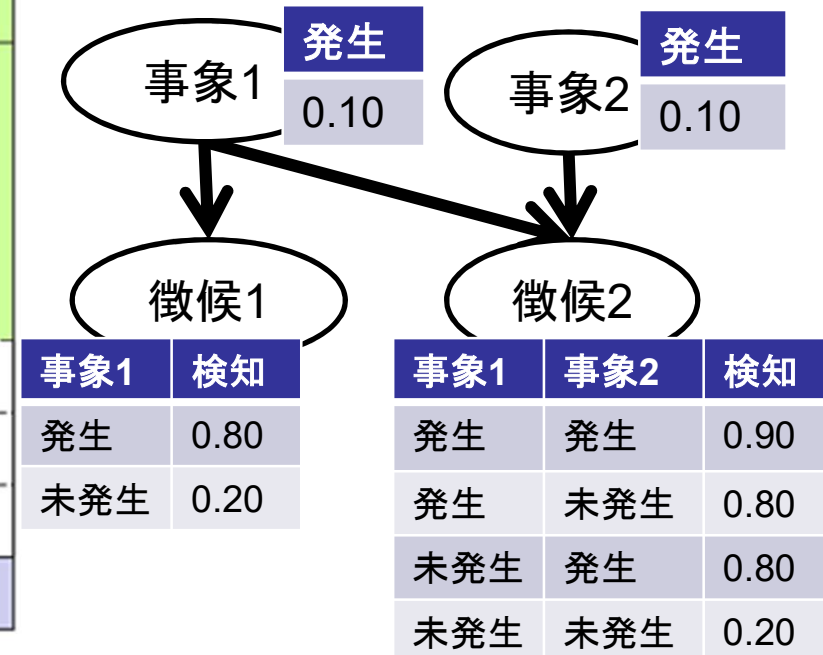
# ベイジアンネットを利用する理由

- ベイジアンネット: AIの一種。事前主観確率とデータ観測利用
- テーブルよりベイジアンネットの方が良い結果が期待できる
  - 検知されていないという情報を利用できる
  - 徴候が検知されたとき関連したものは上昇、関連していないものが減少するという仕組みが入っている

＜テーブル＞

	攻撃事象	徴候	プロキシ					
			立ち上がり	不自然なプロセスの	信	プロキシを経由しない通	443以外のCONNECT メソッドを利用した通信	長時間のセッション
基盤構築 フェーズ	端末が不正プログラムを起動	0.3						
	 C&Cサーバへ接続	0.6		0.6	0.6	0.4		
	 必要な機能のダウンロード	0.4		0.4		0.3		
	端末の情報入手	0.5				0.2	0.4	

＜ベイジアンネット＞



# 事象推定時におけるLIFTの画面

LIFT 設定 ▾

2018/12/15 23:39:55  
マルウェアがC&Cサーバとの通信を行う  
[詳細](#)

2018/12/15 23:39:55  
通常とは異なるUser agentによる外部への通信  
[詳細](#)

2018/12/15 23:39:55  
ブラックリストに登録されているドメインへのアクセス  
[詳細](#)

**兆候の検知**

TODOリスト 徴候一覧 **事象一覧** 対策一覧 ネットワーク  
[更新](#)

事象ID	事象名	推定状態
1	マルウェアが添付されたメールが届く	なし
2	社員がメールに添付された不正プログラムを起動する	なし
3	マルウェアがC&Cサーバとの通信を行う	推定済み
4	必要な機能のダウンロード	なし
5	攻撃基盤の端末の中の情報入手する	なし
6	攻撃者が攻撃基盤から内部ネットワークを探索する	なし
7	攻撃者が攻撃基盤から他の端末へ侵入し、攻撃基盤を増やす	なし
8	攻撃者が攻撃基盤からサーバへ侵入する	なし
9	機密情報の送信	なし
10	端末の破壊	なし

**事象の推定**

既存の攻撃に対しては正しく対応  
(十分性については確信が持てない)

Ryoichi Sasaki et al. "Development and Evaluation of Intelligent Network Forensic System LIFT Using Bayesian Network for Targeted Attack Detection and Prevention" International Journal of Cyber-Security and Digital Forensics (IJCSDF) 7(4): pp344-353, 2018 (to appear)

# 目次

---

1. AIとセキュリティに関する4つの観点
2. AIを利用した攻撃: Attack using AI
3. AI自身による攻撃: Attack by AI
4. AIに対する攻撃: Attack to AI
5. AIを利用したセキュリティ対策: Measure using AI
6. おわりに



# サイバーセキュリティとAI の関連に関する考察

(1) Cybersecurity対策に関しても、データが十分あり、AI技術の1つである機械学習が使えるような対象は現在でも有効。

ただし、動的変化に対応が必要。

(2) いずれにしても、サイバーセキュリティのAI応用は今後も重要な研究分野

(3) AI応用では、機械学習だけではなく集合知やIA(Intelligence Amplifier)、ベイジアンネットなどの技術も大切に。





