

日本のサイバーセキュリティを「連携」「学び」「創造」

パネルディスカッション 「AIセキュリティ」 その脅威と対策を考える

JNSA NSF 2020
IoTセキュリティWG
2020/1/21

パネルディスカッション

～「AIセキュリティ」その脅威と対策を考える～

「システム開発の50%が機械学習」といわれる昨今、機械学習を用いたシステムの品質確保に注目が集まっております。一方で論文ベースながら「機械学習の脆弱性を狙った攻撃」も数多く取り上げられており、今後は「**機械学習におけるセキュリティ**」が注目を浴びると考えられます。

海外論文等で取り上げられている「機械学習に特有なセキュリティ上の脅威とその対策」を紹介し、その有効性、今後の脅威、課題などを様々な視点から議論します。

パネリスト紹介

石川 冬樹先生



国立情報学研究所 アーキテクチャ科学研究系 准教授
先端ソフトウェア工学・国際研究センター 副センター長
機械学習工学研究会（MLSE）主査
AIプロダクト品質保証コンソーシアム 副運営委員長

パネリスト紹介

福田 尚弘



パナソニック株式会社 LS社 技術本部 イノベーションセンター
JNSA IoTセキュリティWGメンバー

パネリスト紹介

松岡 正人



日本シノプシス合同会社ソフトウェア
インテグリティ グループ
シニアプロダクトマーケティングマネージャー
IoTセキュリティWGリーダー

何故AIセキュリティか？

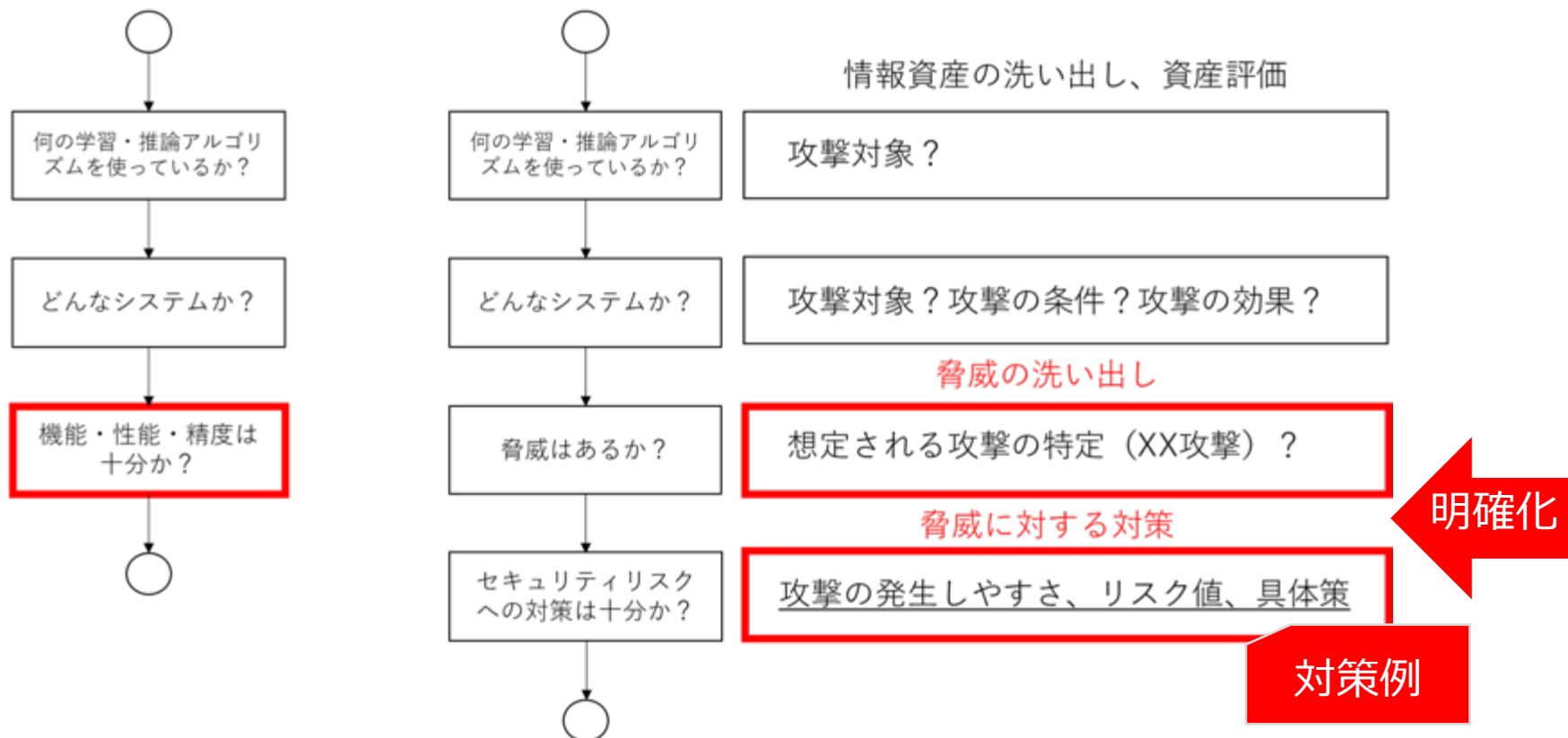
パネルディスカッション
「AIセキュリティ」
その脅威と対策を考える

JNSA IoTセキュリティWG

～AIセキュリティ調査の動機～

品質：システム (x)

セキュリティ：システム (x)



機械学習の脅威と対策を明確化

機械学習システムの基礎知識

パネルディスカッション
「AIセキュリティ」
その脅威と対策を考える

機械学習システムの基礎知識

「ML(機械学習)とは」

“学習分野であり、明示的にプログラムしなくてもコンピュータに学習機能を提供する。”

アーサー・サミュエル、1959

“Pで測定されるTでのパフォーマンスが経験Eで向上する場合、コンピュータプログラムは、あるタスクTおよびいくつかのパフォーマンス測定Pに関する経験Eから学習する”

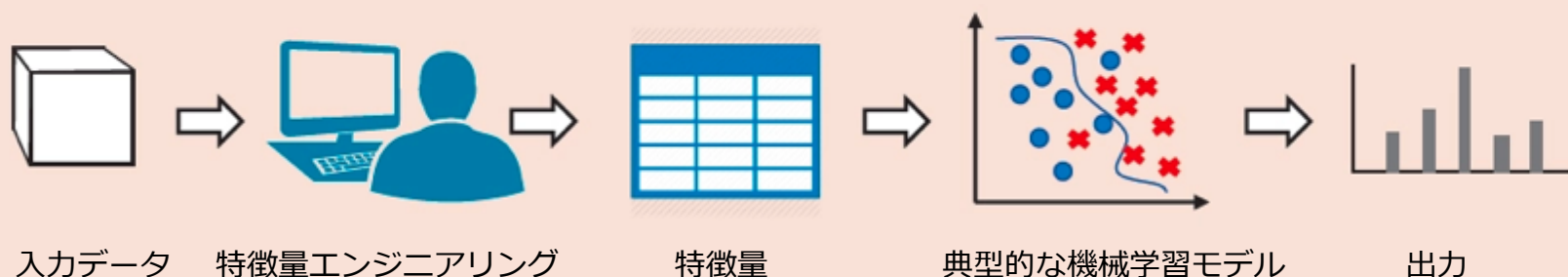
トム・ミッチェル、1997

*Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition
<https://learning.oreilly.com/library/view/hands-on-machine-learning/9781492032632/>

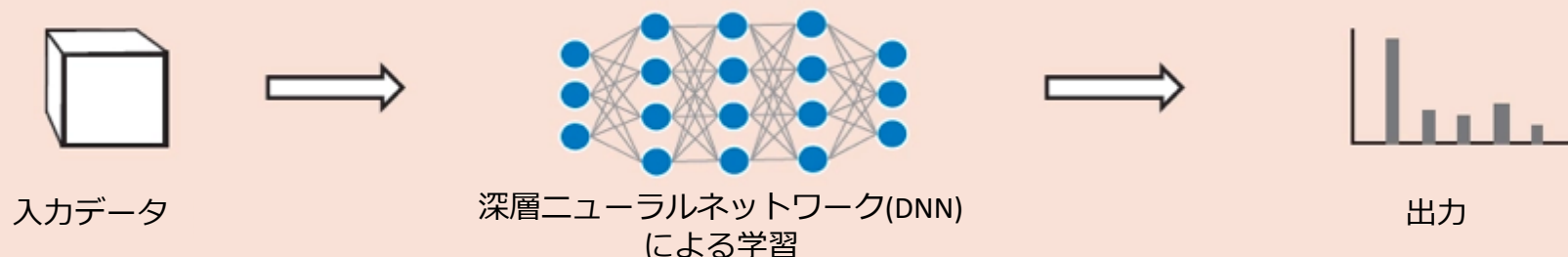
機械学習システムの基礎知識

「ML(機械学習)の仕組み」

典型的な機械学習



深層学習



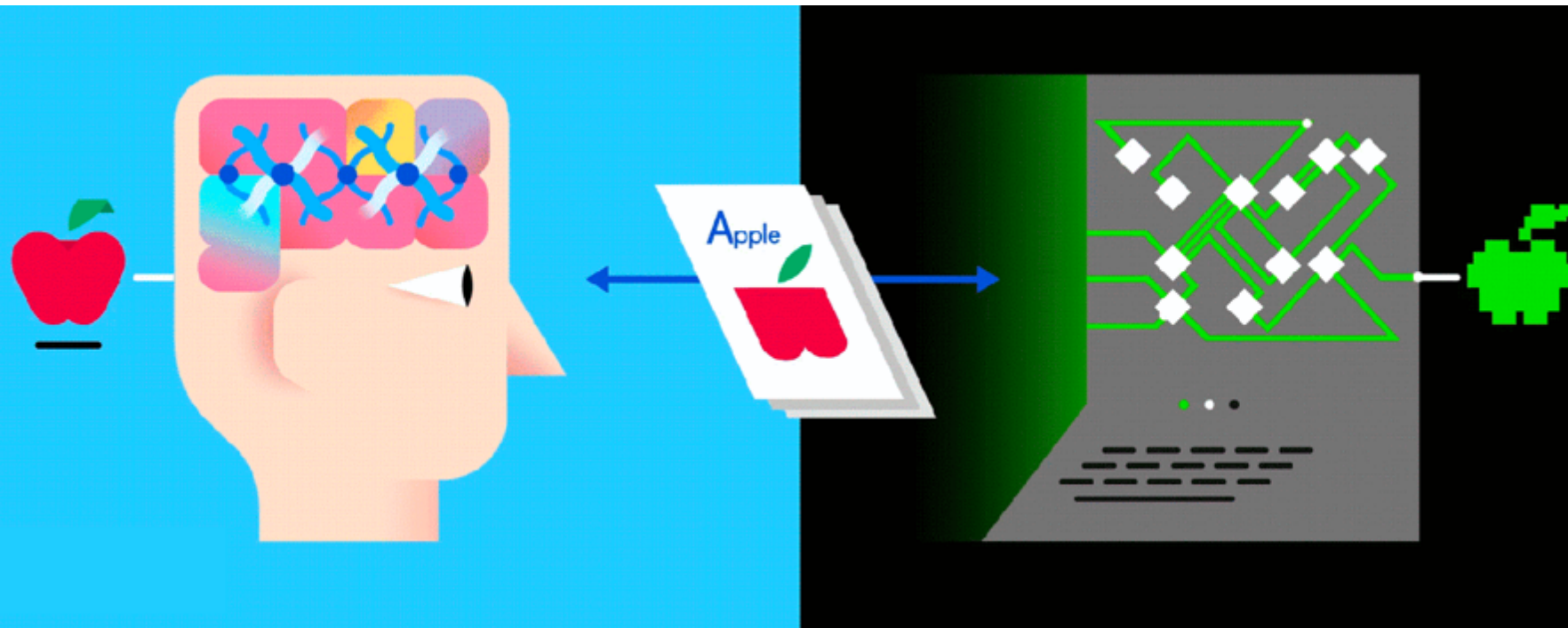
*ACM: Techniques for Interpretable Machine Learning

<https://cacm.acm.org/magazines/2020/1/241703-techniques-for-interpretable-machine-learning/fulltext>

機械学習システムの基礎知識

「人間との違い」

人工ニューロンは非常に多くの層で構成できるが単方向のみ、人間のニューロンは複数方向で接続

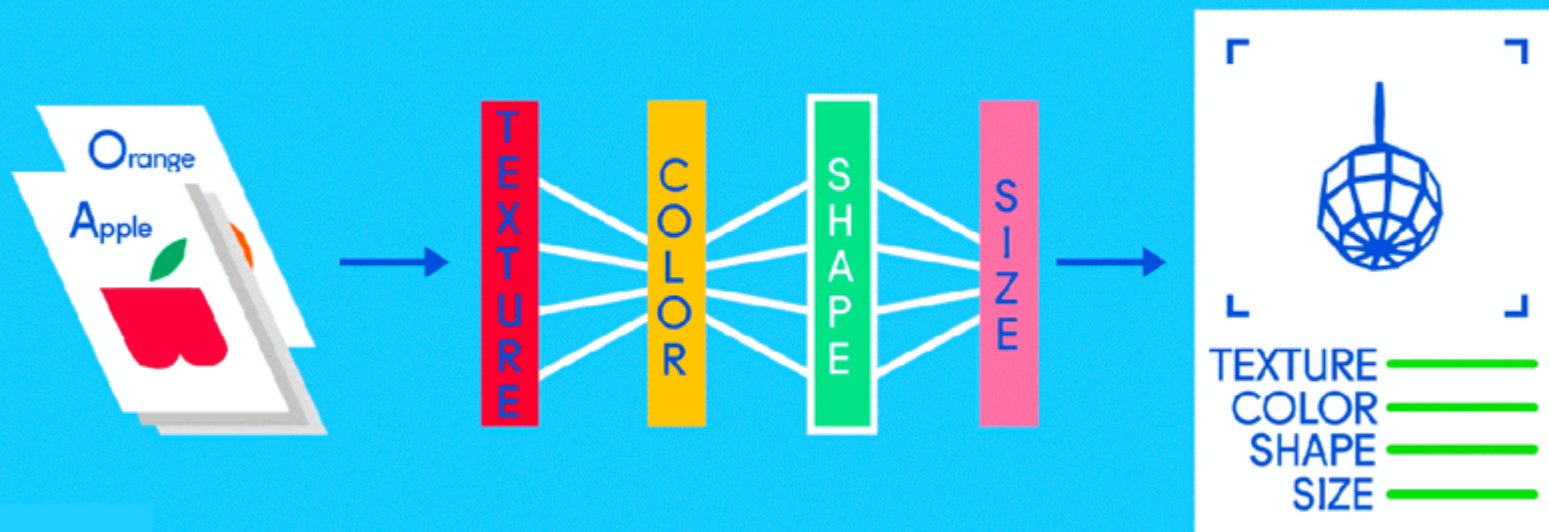


*TechCrunch: 数学知識もいらないゼロからのニューラルネットワーク入門
<https://jp.techcrunch.com/2017/04/15/20170413neural-networks-made-easy/>

機械学習システムの基礎知識

「学習」

正解と不正解のデータを与え、画像の特徴を認識して判断させる

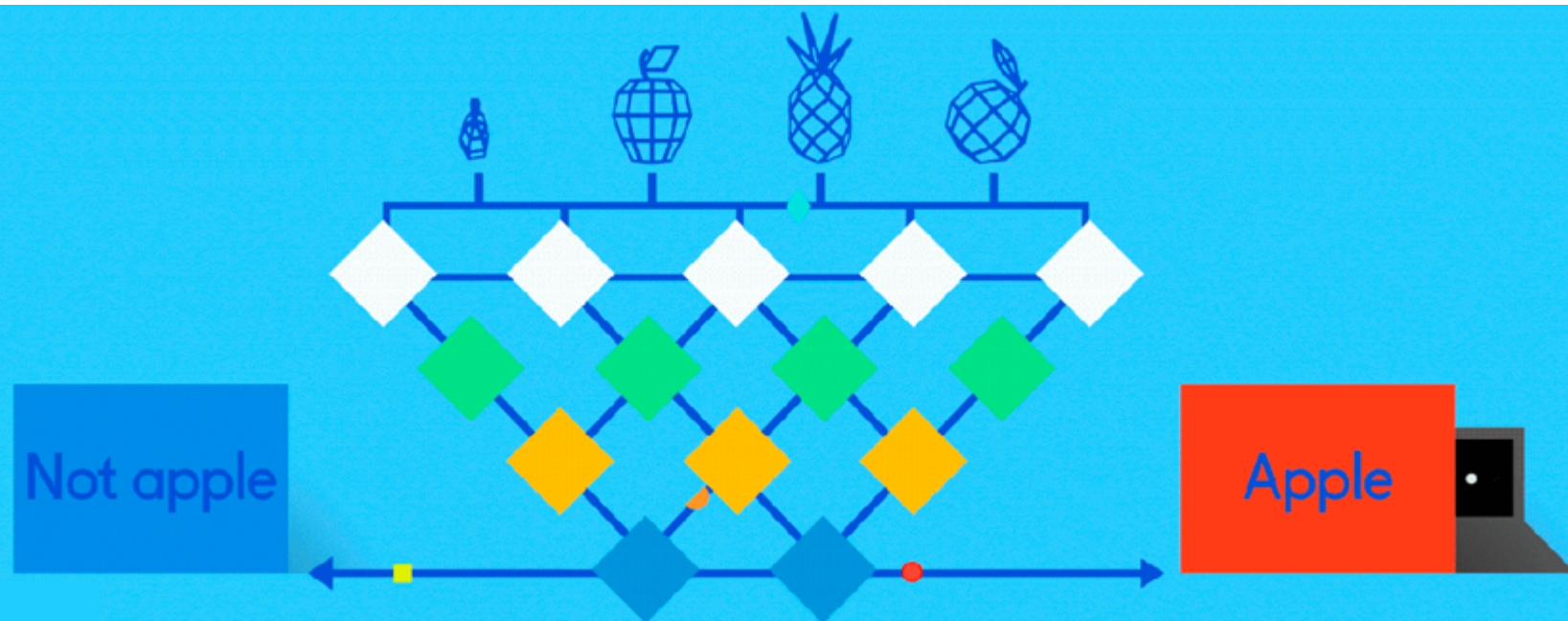


*TechCrunch: 数学知識もいらないゼロからのニューラルネットワーク入門
<https://jp.techcrunch.com/2017/04/15/20170413neural-networks-made-easy/>

機械学習システムの基礎知識

「NNの評価」

精度を向上させるにはニューロンに修正（重み）を加え、学習を繰り返す他ない



*TechCrunch:数学知識もいらないゼロからのニューラルネットワーク入門
<https://jp.techcrunch.com/2017/04/15/20170413neural-networks-made-easy/>

機械学習システムの基礎知識

用語の整理

- ANN：人工ニューラルネットワーク
 - 人為的に作成されたニューラルネットワーク(NN)の総称
- MLP：多層パーセプトロン
 - 入力、出力の各層との間にひとつの隠れ層を含むNN
- DNN：深層ニューラルネットワーク
 - 複数層で構成されるMLP
- CNN：畳み込みニューラルネットワーク
 - 格子状に特徴をつかむ構造のNN
- RNN：再帰型ニューラルネットワーク
 - NNの出力を他のNNの入力に使う再帰的構造のもの

機械学習システムの基礎知識

MLの課題

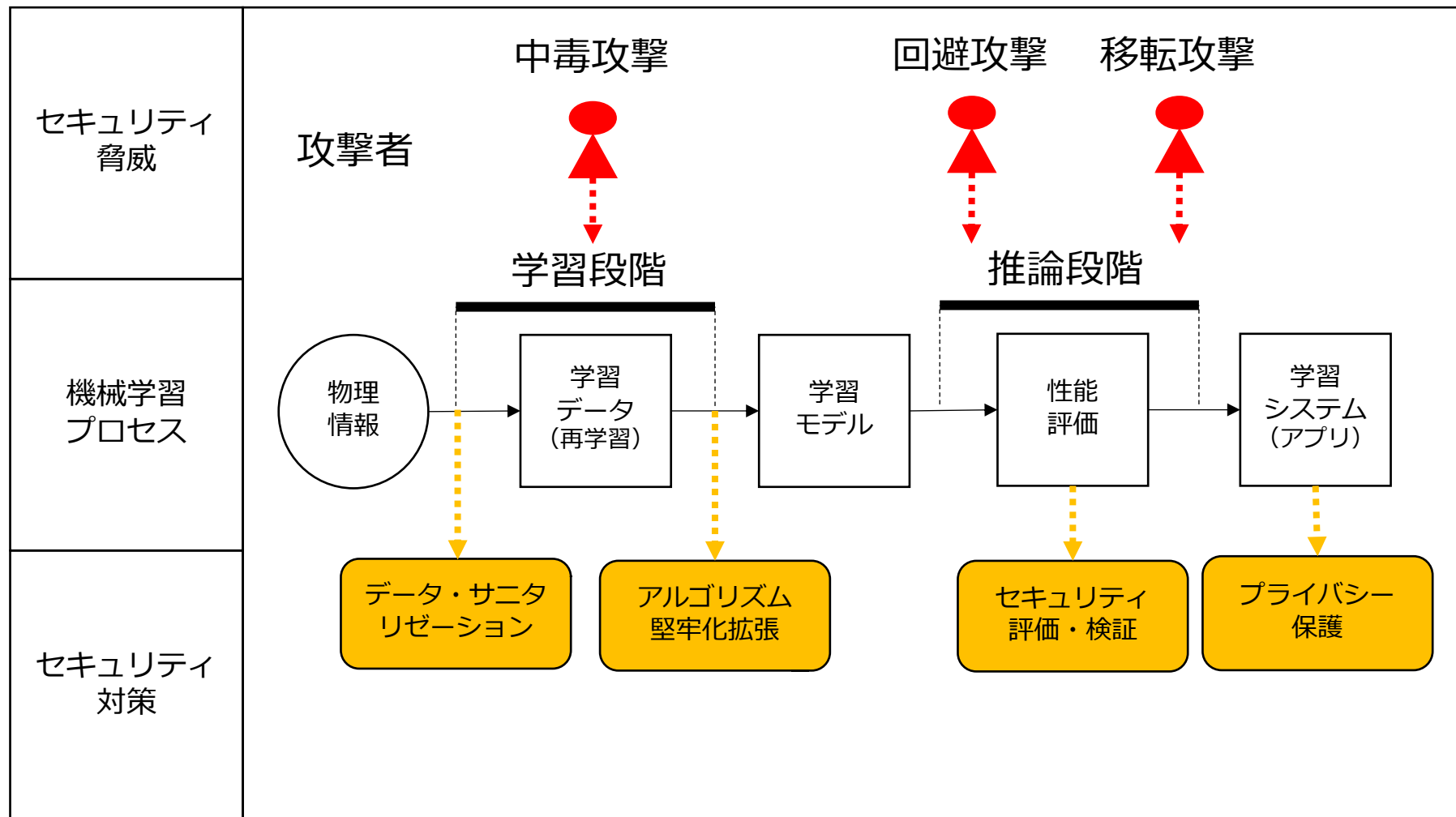
ブラックボックス化された複雑なモデルによって動作するため、予期せぬ結果を招く場合がある

- スпамフィルターをフィッシングメールがすり抜ける
- アンチウィルスソフトによるマルウェアの誤検知
- 自動翻訳における誤訳
- 自動運転車の誤認識による事故など

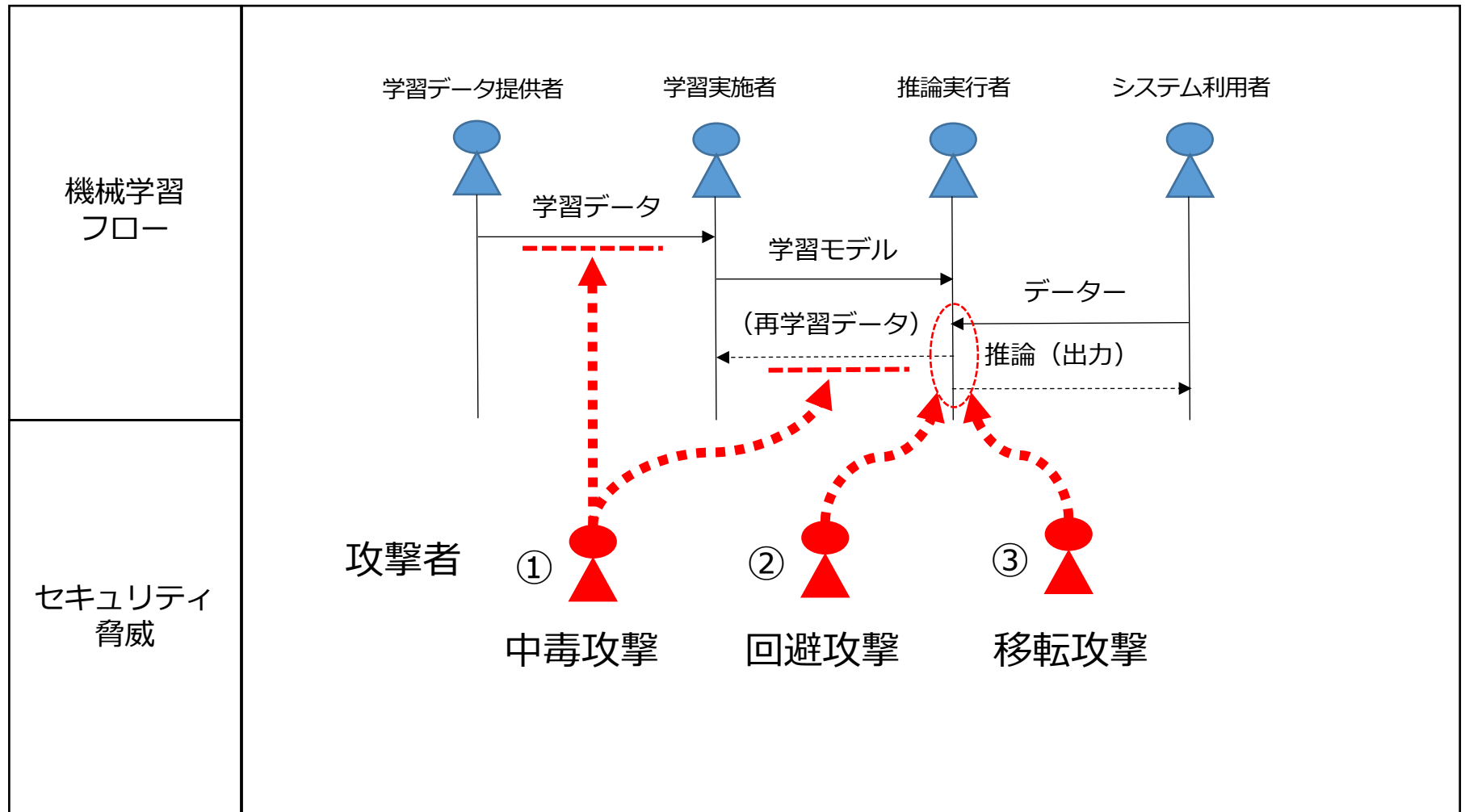
機械学習特有の脅威／攻撃

パネルディスカッション
「AIセキュリティ」
その脅威と対策を考える

機械学習と脅威



機械学習と脅威（フロー）

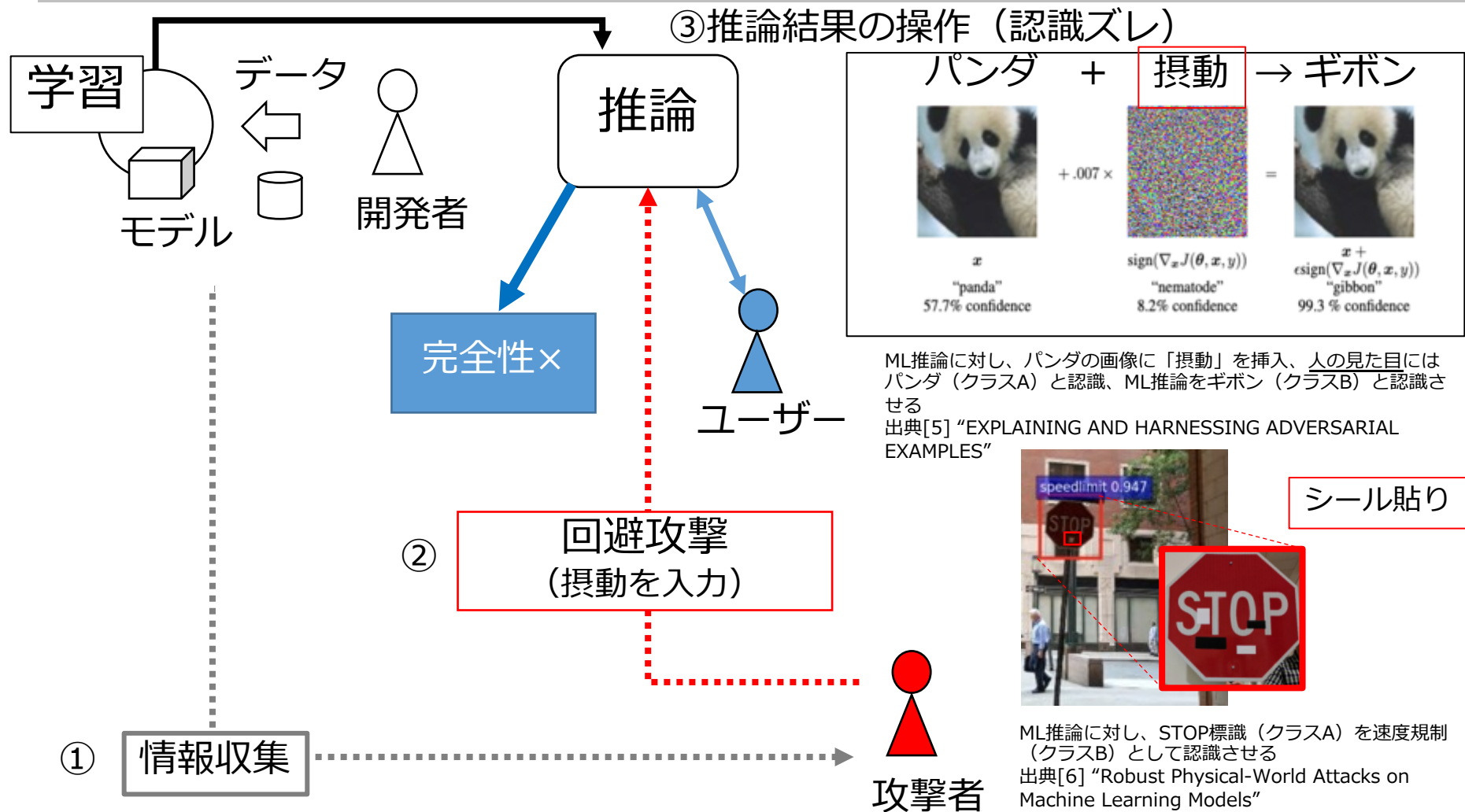


機械学習への攻撃

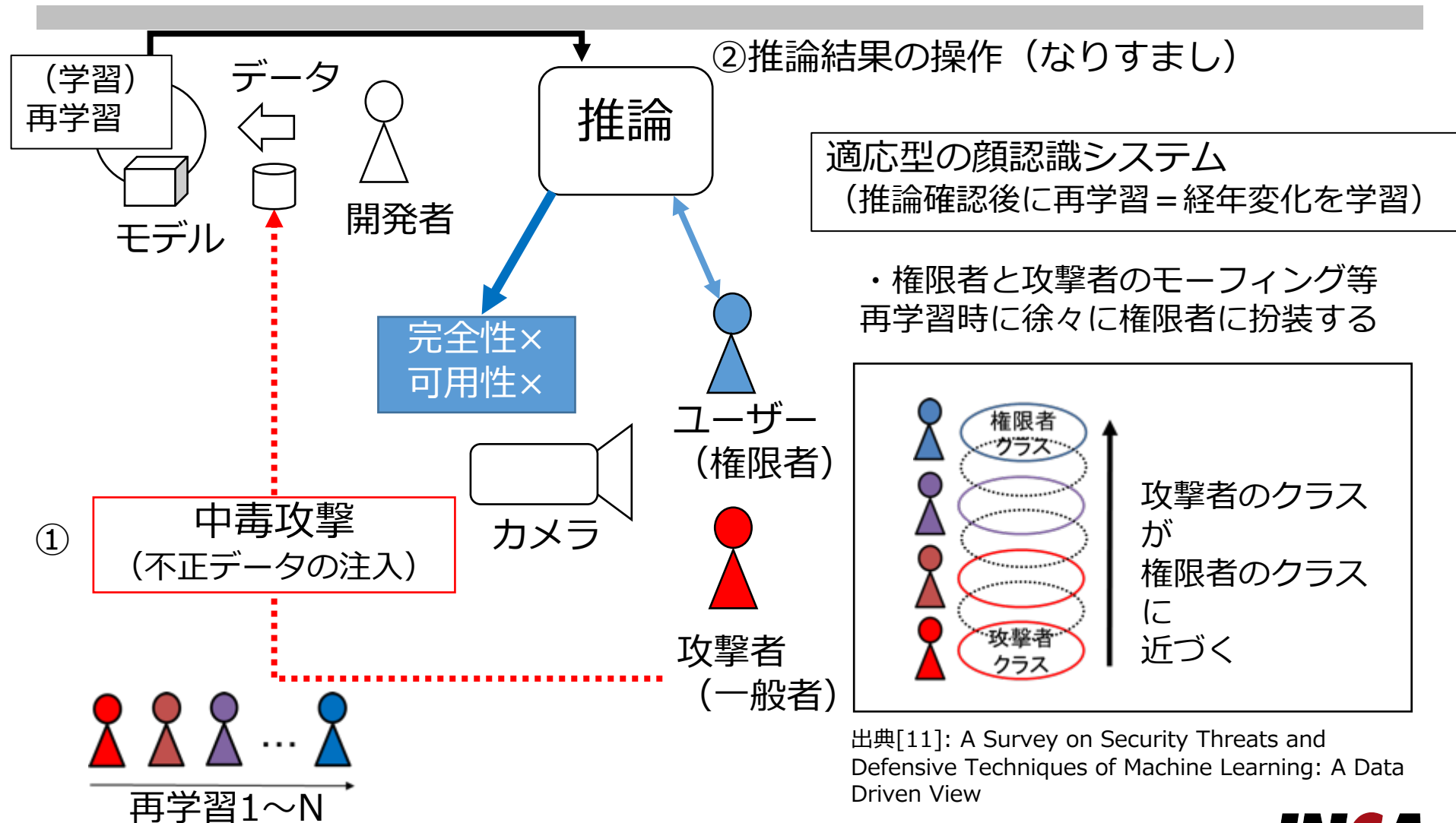
攻撃（脅威）	サブ分類	内容
回避攻撃 (Evasion Attacks)	・透明化攻撃 (Stealth) ・なりすまし攻撃 (Impersonate)	人間には認識できない「 <u>摂動</u> を含んだデータ入力」により、 <u>人間と機械学習の推論エンジンとで異なる認識を起こす</u> 攻撃（画像、音声、文字等）
中毒攻撃 (Poisoning Attacks)	・可用性攻撃 (Availability) ・バックドア攻撃 (Backdoor)	学習データへの「 <u>不正データの入力</u> （注入）」により、 <u>学習モデルの境界を何らかの方法によりシフト</u> する攻撃 （機械学習のモデル境界を <u>大量の不良データの注入</u> により使用不能とする可用性攻撃と、 <u>少量の洗練されたデータ注入</u> によりバックドアを生成する <u>バックドア攻撃</u> がある）
移転攻撃 (Inversion Attacks)	・プライバシー攻撃 (Privacy) ・メンバーシップ推論攻撃 (Membership)	機械学習の「 <u>推論エンジンへのデータ入出力または反応</u> 」によって、元データなどの <u>機密情報</u> （またはモデル自体）を <u>抽出</u> する攻撃 （メンバーシップ推論攻撃では、敵対者が手元データが相手のデータセットに含まれているかを探る攻撃）

※違う名称、分類もあるが、ここでは3つの攻撃まとめた

回避攻撃(Evasion Attack)

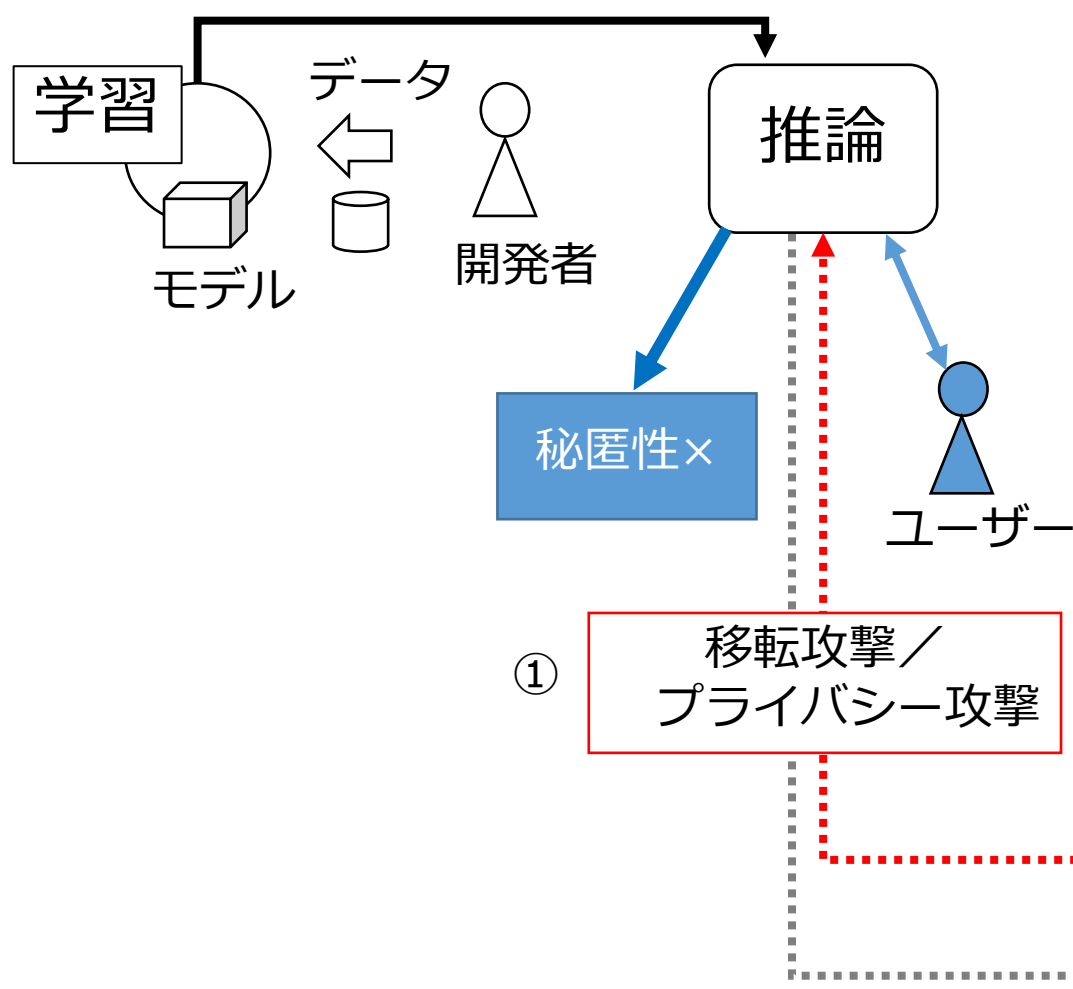


中毒攻撃(Poisoning Attack)



出典[11]: A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View

移転攻撃(Inversion Attack) プライバシー攻撃(Privacy Attack)



ML推論にアクセスし、学習に使ったデータ（顔画像）を類推



攻撃によって類推した新モデルを使って復元した画像



旧モデルの訓練に使った画像

出典[7] “Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures”

回避攻撃の分類

攻撃パターン	概要	備考
勾配ベース	- モデルの勾配へのアクセスを必要とするホワイトボックス攻撃の一種。攻撃者は、モデルの勾配の詳細な理解により、攻撃を数学的に最適化する。「強化モデル」を攻撃する最適な手法でもある。攻撃者がモデルの勾配にアクセスできる場合、攻撃者は常にモデルを欺く敵対的な新しいセットを作成できる。 (隠蔽によるセキュリティ以外の防御はかなり困難)	
信頼スコア	出力された分類信頼度を使用してモデルの勾配を推定し、上記の勾配ベースの攻撃と同様のスマート最適化を実行する。(攻撃者がモデルについて何も知らなくてもよい、ブラックボックス攻撃の一種)	
ハードラベル	- モデルによって出力されたラベル (cat、dog、hotdog等) のみに依存し、信頼スコアを必要としない。攻撃は弱い、より現実的な攻撃。 <u>(回避攻撃で最も強力な攻撃は、境界攻撃)</u>	境界攻撃
代理モデル	追加の手順が必要なことを除いて、勾配ベースの攻撃と非常に似ている。敵がモデルの内部にアクセスできないが、ホワイトボックス攻撃を仕掛けたい場合、最初にターゲットのモデルを自分のマシンで再構築しようとすることができる。	<u>リバースエンジニアリング</u>
		<u>コピー作成</u>
		敵対的な例を生成
ブルートフォース	最適化をまったく使用せずに敵対的な例を生成し、代わりに左記の単純な攻撃を行う。	画像をランダムに回転/平行移動
		一般的な摂動の適用
		ガウスノイズを追加

中毒攻撃の分類

攻撃パターン	概要	備考
ロジック変造	最も危険攻撃。 <u>攻撃者がアルゴリズムと学習方法を変更できる場合、ロジックの変造が起こせる。</u> （攻撃者は必要なロジックを簡単にエンコードできる。）	学習を信頼できない誰かに任せた場合、HTTPなど非暗号通信を傍受され盗聴・改ざんされた場合、モデルファイルを改ざんされた場合など
データ変更	攻撃者はアルゴリズムにアクセスできないが、訓練データを変更、追加、削除できる。（ラベルの操作または洗練された攻撃者によるデータ入力操作）	ラベルの操作（バックドア攻撃）
		ラベルの操作（可用性攻撃）
		データ入力操作（分類境界シフト）
		データ入力操作（クラスター距離変更）
データ挿入	データ変更であるが「追加」に限定。攻撃者が新しいデータを訓練プールに注入できる場合	データ入力操作（透明透かし追加）
		トラフィックに摂動を挿入
転移学習	最も弱い攻撃。2回目のトレーニングを行うにつれて、元のメモリ（毒を含む）が希釈されるため。	

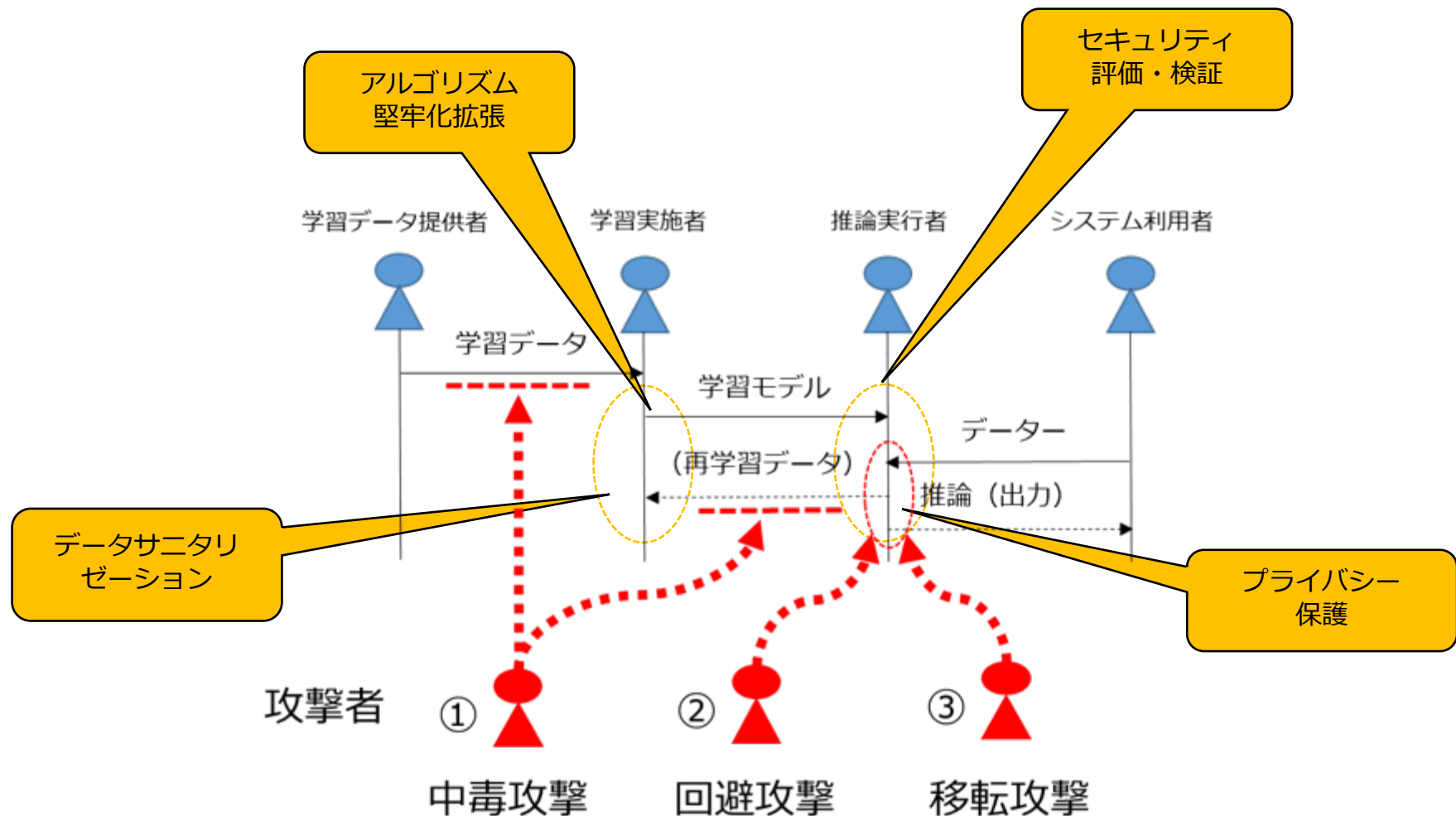
移転攻撃／プライバシー攻撃の分類

攻撃パターン	概要	詳細
データ	攻撃者が手元にデータポイントを持っており、元のデータセットに属しているかどうかを知りたい場合 例) 名前が機密性の高い医療リストに含まれているかどうかを知りたい場合	元のターゲットモデルに基づいて構築された「シャドー」モデルを使用。 患者が退院データセットに「入っている」などを判断する手法
モデル	ブラックボックスアクセス権を持つが、MLモデルのパラメーターまたはトレーニングデータに関する予備知識がない攻撃者が、モデルの機能を複製（モデル抽出）を実行。 （入出力ペアを観察してモデルパラメーターを抽出する。 ただし、信頼スコアへのアクセスに依存）	APIにいくつかのクエリを送信するだけでモデル全体をダウンロード モデルの勾配にアクセスして作成された回避攻撃 （ホワイトボックス回避攻撃）

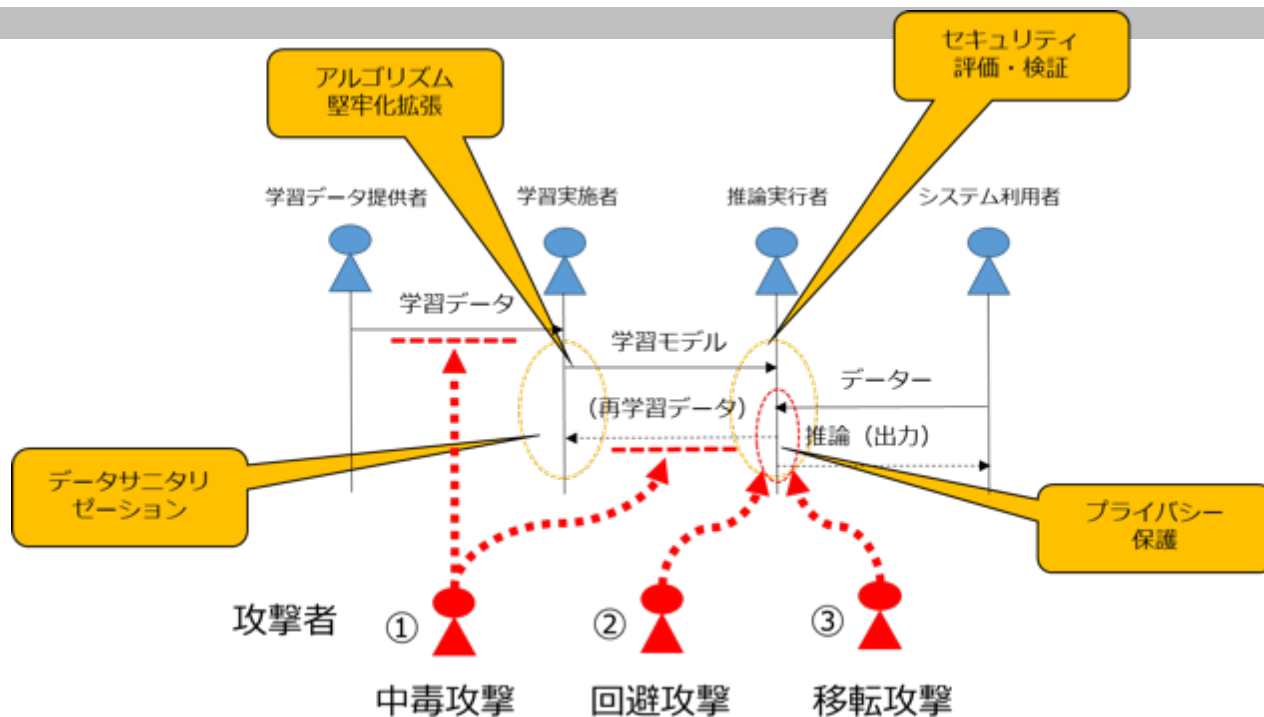
機械学習でのセキュリティ対策

パネルディスカッション
「AIセキュリティ」
その脅威と対策を考える

機械学習での脅威とセキュリティ対策



機械学習での脅威と対策（詳細）



対中毒攻撃

異常値検出／モデル精度に与える影響分析

- ・入力を意図的に摂動 (STRIP)
- ・境界シフトの警告 (LOOP)
- ・回帰 (TRIM)

対回避攻撃

全網羅／経験的防御

- ・摂動攻撃パターンの網羅 (困難)
- ・敵対的訓練 (中毒攻撃の副作用)
- ・勾配マスキング
- ・入力変更 (ノイズ除去)
- ・検出技術
- ・NULLクラス

対移転攻撃

API強化／サニタイズ／モデル強化／検出／差分プライバシー

- ・API非公開
- ・漏らしたくないデータの除去
- ・モデル選択／フィット制御／知識管理
- ・攻撃パターンの異常性判定
- ・ユーザー入力のランダム化

機械学習への対策（回避攻撃）

攻撃	対策	概要	備考
回避攻撃	全網羅	摂動を含むパターンの全網羅	現実には不可能
	経験的防御	敵対的訓練	中毒攻撃をもたらす危険性あり
		勾配マスキング	資産移転が可能とされ効果少ない
		入力変更	ノイズ除去等
		検出	畳み込みフィルターなど
		追加クラス（NULLクラス）	

機械学習への対策（中毒攻撃）

攻撃	対策	概要	備考
中毒攻撃	異常検出	異なるデータの注入と検出	中毒攻撃の特徴を検出
	モデルの精度に影響する内容の影響度分析	入力を意図的に摂動	STRIP
		境界シフトの警告	LOOP
		回帰	TRIM

機械学習への対策（移転攻撃）

赤字：AI特有でない対策

攻撃	対策	概要	備考
移転攻撃	API強化	APIを非公開	アクセス制限
		信頼スコアでなくハードラベルのみ公開	匿名化（属性削除）
		信頼性スコア公開の場合は {高、中} など抽象化	匿名化（属性情報の一般化）
	データサニタイズ	漏らしたくないデータ除外（クレジットカードなど）	匿名化（属性削除、攪乱ノイズ付加）
	モデル強化	モデル選択（ベイジアンは決定木より堅牢など）	
		フィット制御（オーバーフィットからモデルよりデータ抽出・正規化）	
		知識管理（モデルを少数共有に制限）	アクセス制限
	検出	クエリーパターンを一般ユーザーと比較、異常性を判定	
	差分プライバシー	ユーザー入力のランダム化	匿名化（攪乱、疑似データ生成）
		元データのランダム化	
		モデルのパラメーターの摂動	
		損失関数の摂動	匿名化（攪乱）
		推測中の出力の摂動	
	プライバシー保護ELM	多入力依頼計算型プライバシー保護機械学習	加法準同型暗号 PP-ELM: Privacy Preserving Extreme Learning Machine

その他（対策関連）

パネルディスカッション
「AIセキュリティ」
その脅威と対策を考える

その他： {ホワイト、ブラック、グレー} ボックス

ターゲットの情報	内容
ホワイトボックス	ターゲットの基礎となるデータ分布、モデルのアーキテクチャー、使用される最適化アルゴリズム、重み、バイアスなどがわかっている
ブラックボックス	ターゲットの情報を何も知らない
グレーボックス	上記の中間 (例：攻撃者はターゲットのモデルを類推できるがデータはわからない、など)
ノーボックス	代理モデル（攻撃者は標的のMLシステムの理解に基づいてモデルを再構築） (例：ニューラルネットワークなどで構築)

その他：制限（攻撃を制限する方法）

対象	内容
画像	画像では、摂動空間を「距離」メトリックで制限するのが一般的
マルウェア	マルウェアでは、攻撃者は特定の場所／特定の方法でのみファイルを攻撃できる
物理デバイス	物理デバイス（衛星、車、ドローン、監視カメラ）に展開されたシステムでは、攻撃者は物理ドメインの入力の変更に制限される場合がある
訓練	訓練時の攻撃では、攻撃者は以下の点を必要とする。 a) システムが新しいデータに基づいて継続的に再訓練を行う （そうでなければ、悪意のデータを注入できない） b) システムが外部ソースからデータを取り込み、何らかのループを形成して承認する
プライバシー	プライバシー攻撃では、一般に攻撃者はクエリ制限のないパブリックエンドポイントを必要とし、信頼スコア（95%など）の出力応答を期待する （信頼スコアは限定的。ほとんどのウイルス対策製品は、ファイルが「悪意のある」または「良性」であると判断した場合に、詳細な情報を提供しない）

その他：{誰、何故、如何} に？

誰（WHO）	何故（WHY）	備考
スマートな研究者／技術者	フーリングアラウンド	（弄ってみるレベル）
賞金稼ぎプログラムに参加する ホワイハット	賞金稼ぎ	
ペンテストを行うレッドチーム	ペンテスト	
商用MLシステムを攻撃するハクティビスト	主張の証明	
商用MLシステムを攻撃するブラックハット	金銭的報酬	（エクスプロイト展開、ダーク ネット販売）
商用MLシステムを攻撃する 組織的ブラックハットグループ	各々（金銭、証明、政治的目的達成）	（Anonymous、The Shadow Brokers、Legion of Doom）
国家が後援する組織	サイバー戦争の目的達成	

如何に（HOW）（軸）	備考
タイミング	トレーニング、推論
機能	ホワイハット、ブラック、グレー・ボックス
制限	摂動空間（ズレ幅）、機能性、領域、 再トレーニング頻度
代替案	

出典：[2] “Calypso AI、Ilja Moisejevs氏”の機械学習に関するセキュリティブログ（Toward Data Science）

まとめ／仮説

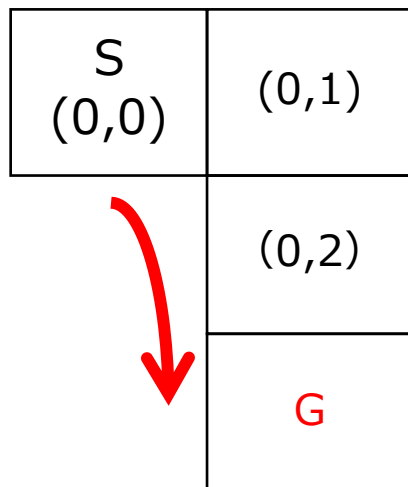
- 機械学習の攻撃（脅威）を分類
 1. {回避、中毒、移転} 攻撃に分類
 2. 各攻撃に様々な手段がある
 3. 攻撃のしやすさは {回避> 中毒、移転} 攻撃の順
- 脅威に対応するセキュリティ対策
 1. {回避> 中毒、移転} 攻撃の順で守りにくい
 2. {ホワイト>ブラック} ボックスの順で守りにくい
 3. 移転攻撃はAIによらないセキュリティで保護
(差分プライバシー)
- 可能な限り攻撃を制限する
- 攻撃者が誰かを見極めて対策を強化

機械学習を使った攻撃

パネルディスカッション
「AIセキュリティ」
その脅威と対策を考える

強化学習について（Q-学習）

Q-学習：報酬付きの状態遷移表



状態	行動	報酬
(0,0)	上	-1
	右	50
	下	-1
	左	-1
(0,1)	上	-1
	右	-1
	下	50
	左	50
(0,2)	上	50
	右	-1
	下	100
	左	-1

3状態×4行動 = 12パターン

強化学習について (Q-学習)



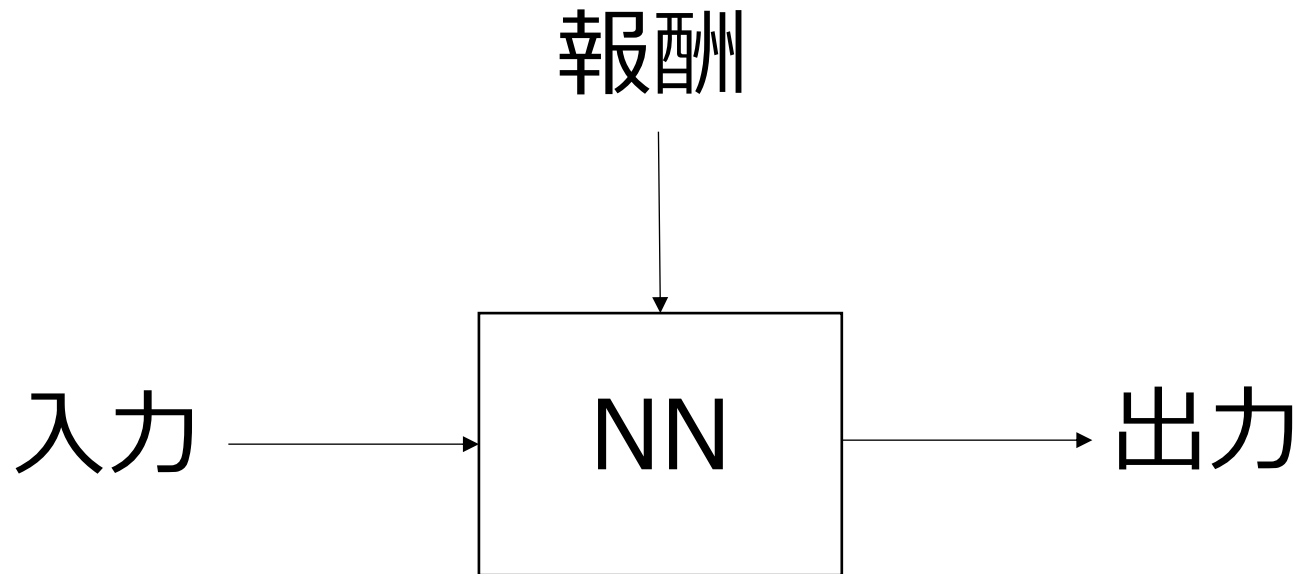
チェス (8x8)

状態	行動	報酬
???	???	???
	???	???
	???	???
	???	???

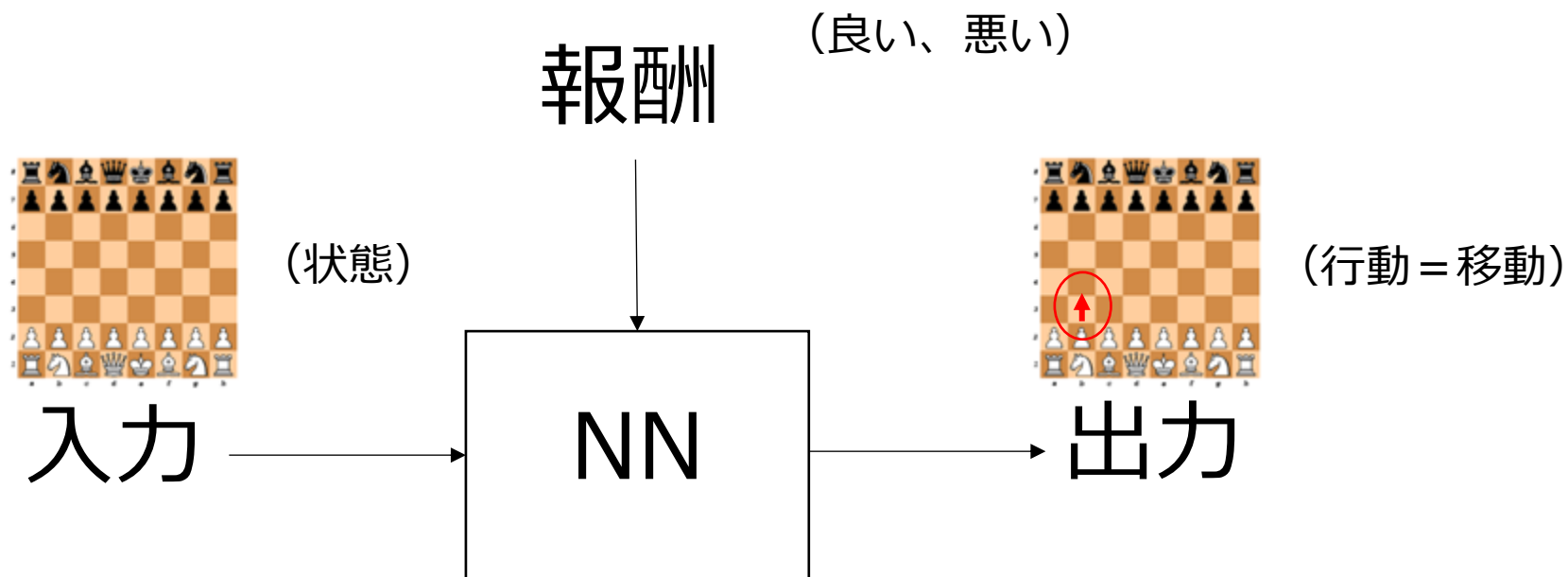
$\sim 10^{47}$

規模が大きいと機能しない (次元の呪い)

強化学習（ニューラルネットワーク）

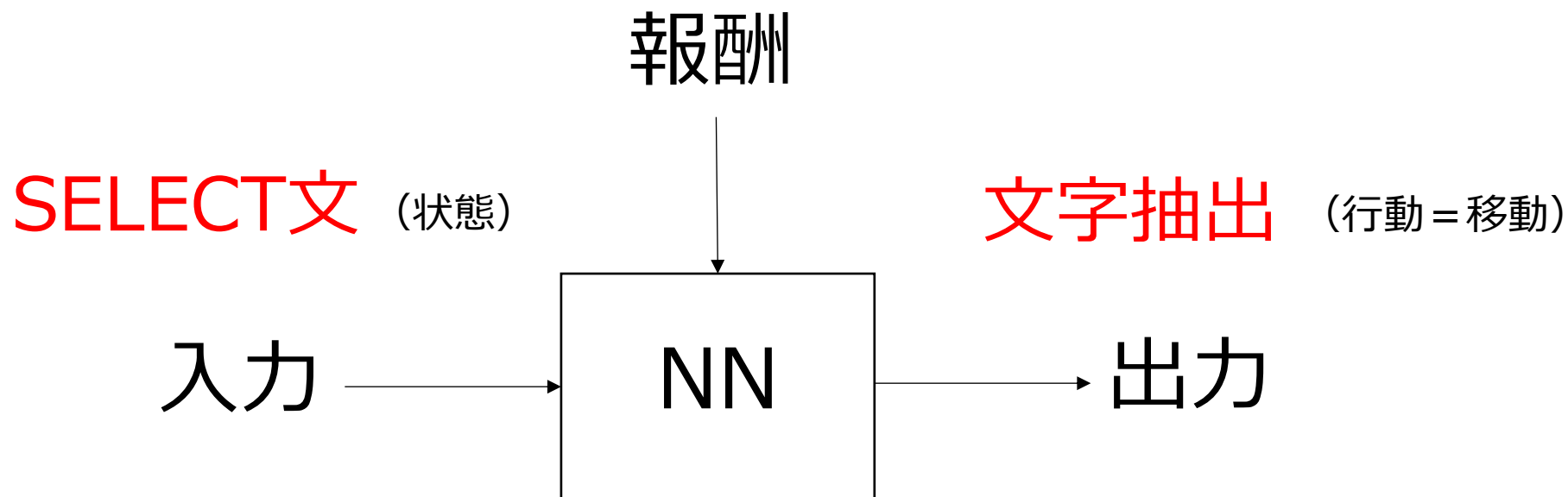


強化学習（ニューラルネットワーク）



強化学習（SQL攻撃）

HTTP: (200, 500)



出典[8] DEFCON 25 2017 Weaponizing Machine Learning

ファジング (Fuzzing)

予測不可能な「入力データ」を与えることで意図的に例外を発生させ、その例外の挙動を確認する手法

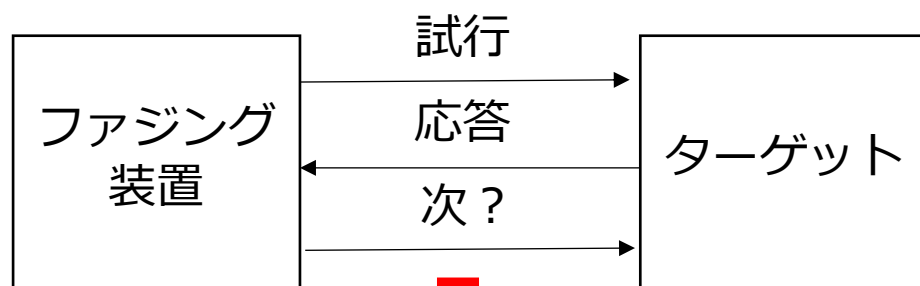
手法	概要
テストケースの事前生成	問題を生じる可能性が高いもの（0、1、その他境界値周辺の値など）を事前に生成（最低品質の確認）
ランダム	テストデータをランダムに生成
突然変異、ブルートフォース	入力データを連続的に変化させてテスト（時間かかる） （突然変異は指定値からの変動、ブルートフォースは総当たりで入力データを連続的に変化させる）
自動プロトコル生成	入力データのプロトコル仕様から、静的に決定する部分と動的に決定する部分を切り分け、動的部分に対して変動の自動化を図る（静的部分と動的部分が記された定義ファイルの作成が必要）

既存のファジングは自動化に対応しているが設定は経験値によるところが大きい

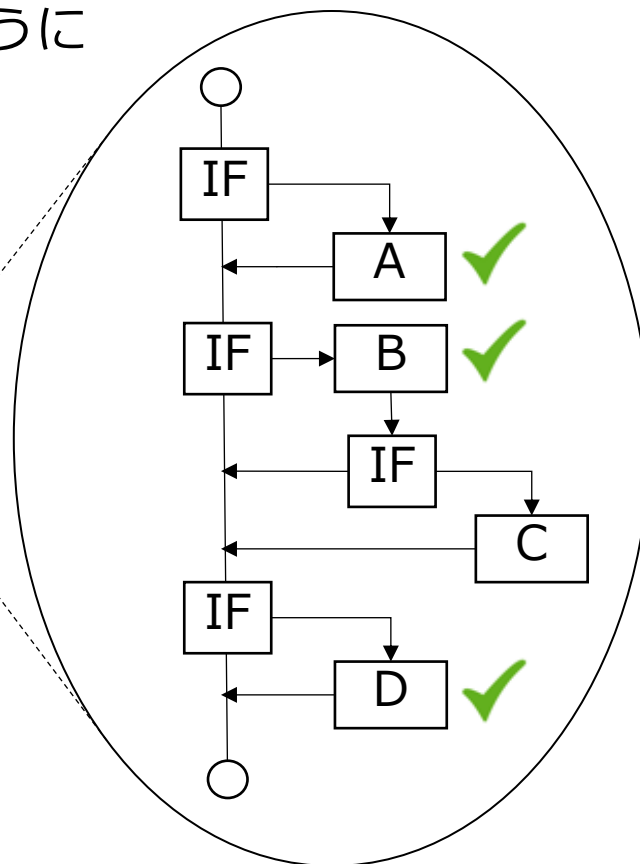
ファジングの課題

既存のファジングの課題 = 次に入力（ファズ）すべき値を何にするか？

- ① ランダムに頼る
- ② コード網羅率を最大（図A～D）にするように設定を調節



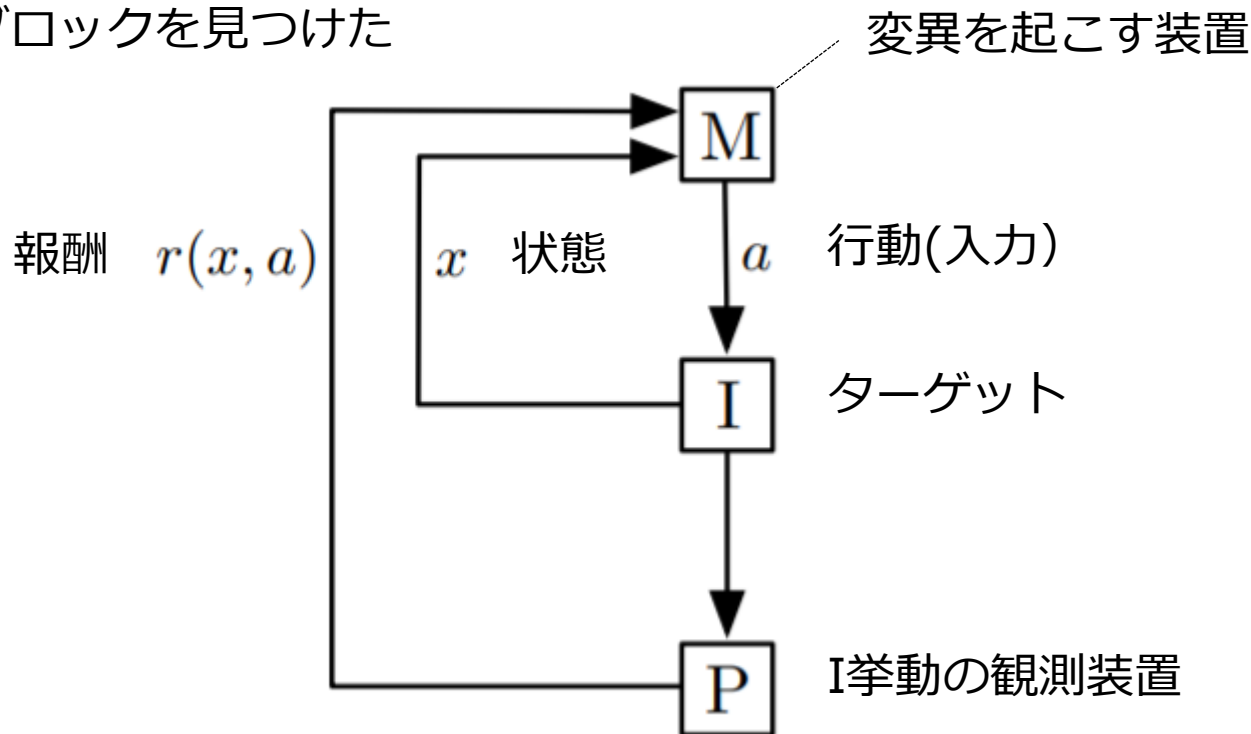
③ 「次」の値を決めるのに
Q-学習を用いる



機械学習を用いたファuzzingの例

ポイントは報酬をどう作るか？

- ①新しい処理ブロックを見つけた
- ②実行パス長
- ③実行時間
- など



出典[10] "Deep Reinforcement Fuzzing"

まとめ／仮説

- データの次元数が大きくなり過ぎると機能しない
- アルゴリズム（数式等）の方が効率的な場合もある
- ファジングでは報酬をどうするかが難しい
- 機械学習のファジング適応の論文では、
ランダム等のファジングに比べ
高々 {7%、30%} 向上

機械学習のユースケースと攻撃

パネルディスカッション
「AIセキュリティ」
その脅威と対策を考える

機械学習のユースケースと脅威・対策

ユースケース	攻撃	脅威			対策			
		回避攻撃	中毒攻撃	移転攻撃	アルゴリズム 堅牢化拡張	データ サンタリゼーション	プライバシー保護 (差分プライバシー)	セキュリティ評価・検証 (AT)よりなご
自動運転（運輸）	標識詐称（見た目にわからない標識）	○			○	○		○
スパムフィルター （サイバーセキュリティ）	マルウェア検知をバイパス （対策者有利、フィルターの見逃ししかない）	○				○		○
顔認識（小売り）	- 顔認識をバイパス - なりすまし（サンプルの周辺の入力をランダムサンプリングし、敵対的な例を見つける＝カーネル法） - 顔データ窃盗：オラクル攻撃（→ビヘイビア検知）	○	○		○	○	○	○
音声コマンド （スマートホーム）	音声コマンド偽装	○			○	○		○
Web評判（メディア）	映画・レビュー、ホテル評判をだます 政治誘導（ディープフェイク）	－	－	－	－	－	－	－
マネー（金融）	トリック、不正検知をだます	－	－	－	－	－	－	－

パネルディスカッション

パネルディスカッション
「AIセキュリティ」
その脅威と対策を考える

議題（AIセキュリティについて）

1. AIでのセキュリティ対策（～近い将来、～課題）
2. AIによる攻撃（～近い将来、～課題）
3. AI脅威論とセキュリティについて
（10年後、セキュリティ対策・運用の現場はどうなるか？）

参考文献

- [1] A Survey on Security Threats and Defensive Techniques (IEEEAcess、 2018)
https://www.researchgate.net/publication/323154427_A_Survey_on_Security_Threats_and_Defensive_Techniques_of_Machine_Learning_A_Data_Driven_View
- [2] “Calypso AI、 Ilja Moisejevs氏”の機械学習に関するセキュリティブログ (Toward Data Science)
<https://towardsdatascience.com/@iljamoisejevs>
<https://aibusiness.com/ai-researchers-develop-backdoor-security-threat/>
- [3] JPCERT/CC IoTセキュリティのためのチェックリスト (2018)
<https://www.jpCERT.or.jp/tips/2016/wr162501.html>
- [4] JNSAコンシューマー向けIoTセキュリティガイド (2015)
https://www.jnsa.org/seminar/2018/0226/data/3_koshiishi.pdf
- [5] EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES, Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy Google Inc., conference paper at ICLR 2015
<https://arxiv.org/abs/1412.6572>
- [6] Evtimov, Ivan, et al. “Robust Physical-World Attacks on Machine Learning Models” Cornell University, 7 Aug. 2017
<https://arxiv.org/abs/1707.08945>
- [7] Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, Matt Fredrikson Carnegie Mellon University, Somesh Jha University of Wisconsin–Madison, Thomas Ristenpart Cornell Tech, 2015
<https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf>
- [8] DEFCON 25 2017 Weaponizing Machine Learning Petro, Morris Stream 30 July 2017
<https://www.machinelearning.ai/machine-learning/defcon-25-2017-weaponizing-machine-learning-petro-morris-stream-30july2017/>
- [9] Automated Software Vulnerability Testing Using In-Depth Training Methods, Alexandr Kuznetsov, Oleksiy Shapoval, Kyrylo Chernov, Yehor Yeromin, Mariia Popova, Olga Syniavska V. N. Karazin Kharkiv National University, Svobody sq., 4, Kharkiv, 61022, Ukraine
<http://ceur-ws.org/Vol-2353/paper18.pdf>
- [10] Deep Reinforcement Fuzzing, Konstantin Bottinger, Patrice Godefroid, Rishabh Singh Fraunhofer AISEC, Microsoft Research, 2018
<https://arxiv.org/pdf/1801.04589.pdf>
- [11] ニューラルネットワークの基礎解説：仕組みや機械学習・ディープラーニングとの関係は, ビジネス+IT, 2017/03/15
<https://www.sbbbit.jp/article/cont1/33345>