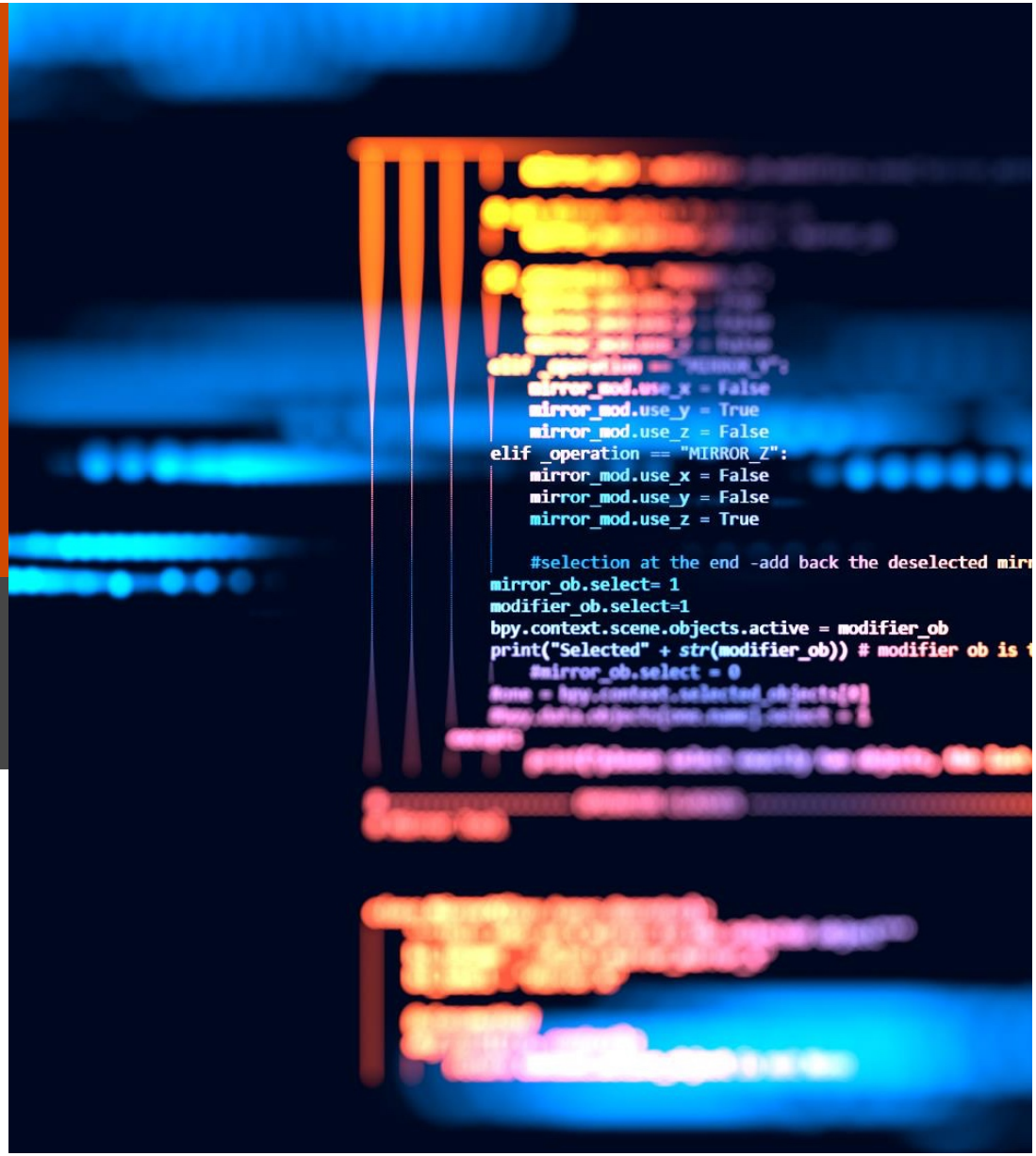
スマートサイバー

AI活用時代の
サイバーリスク管理

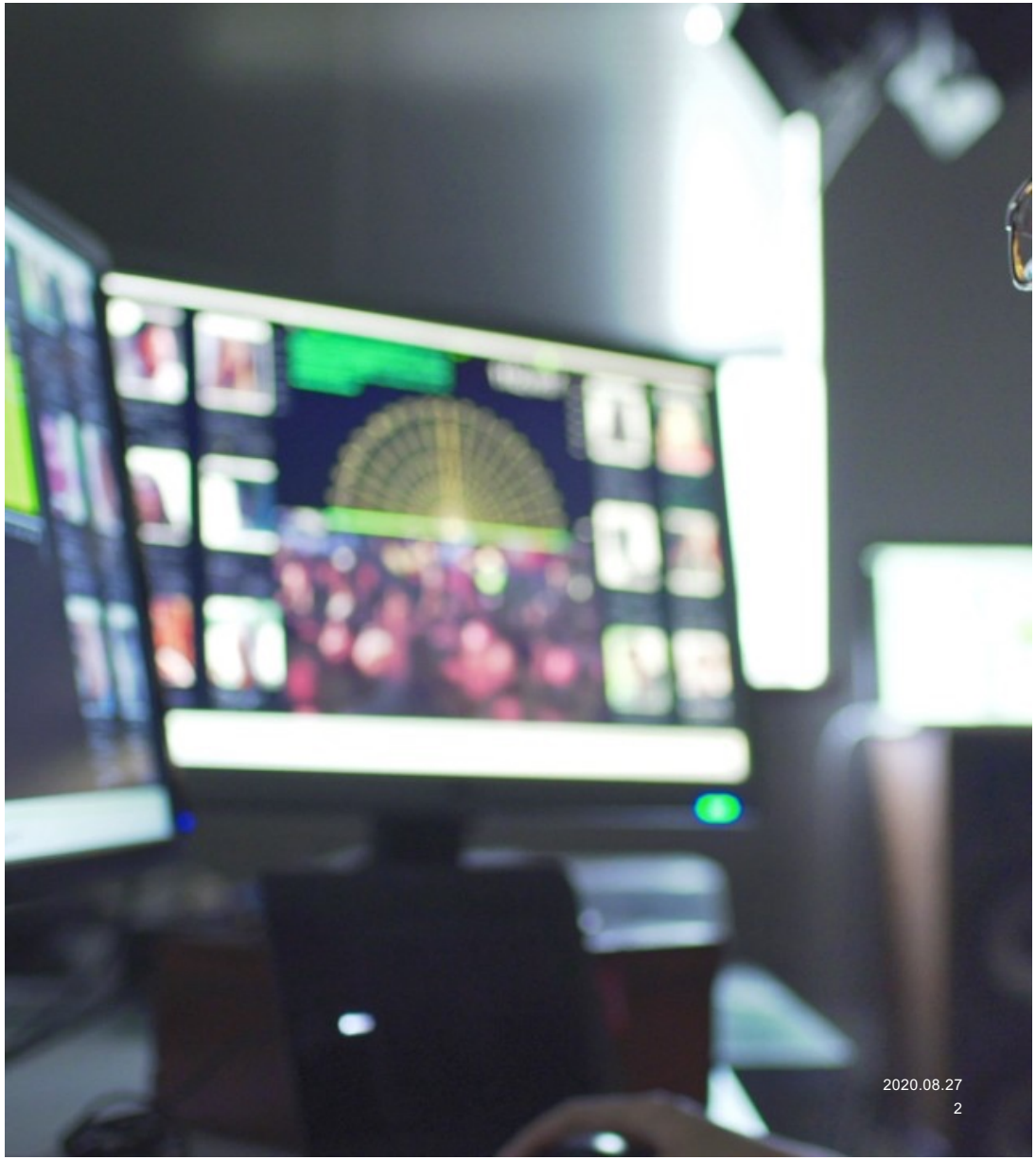
PwCコンサルティング合同会社 パートナー　丸山 満彦

pwc



“

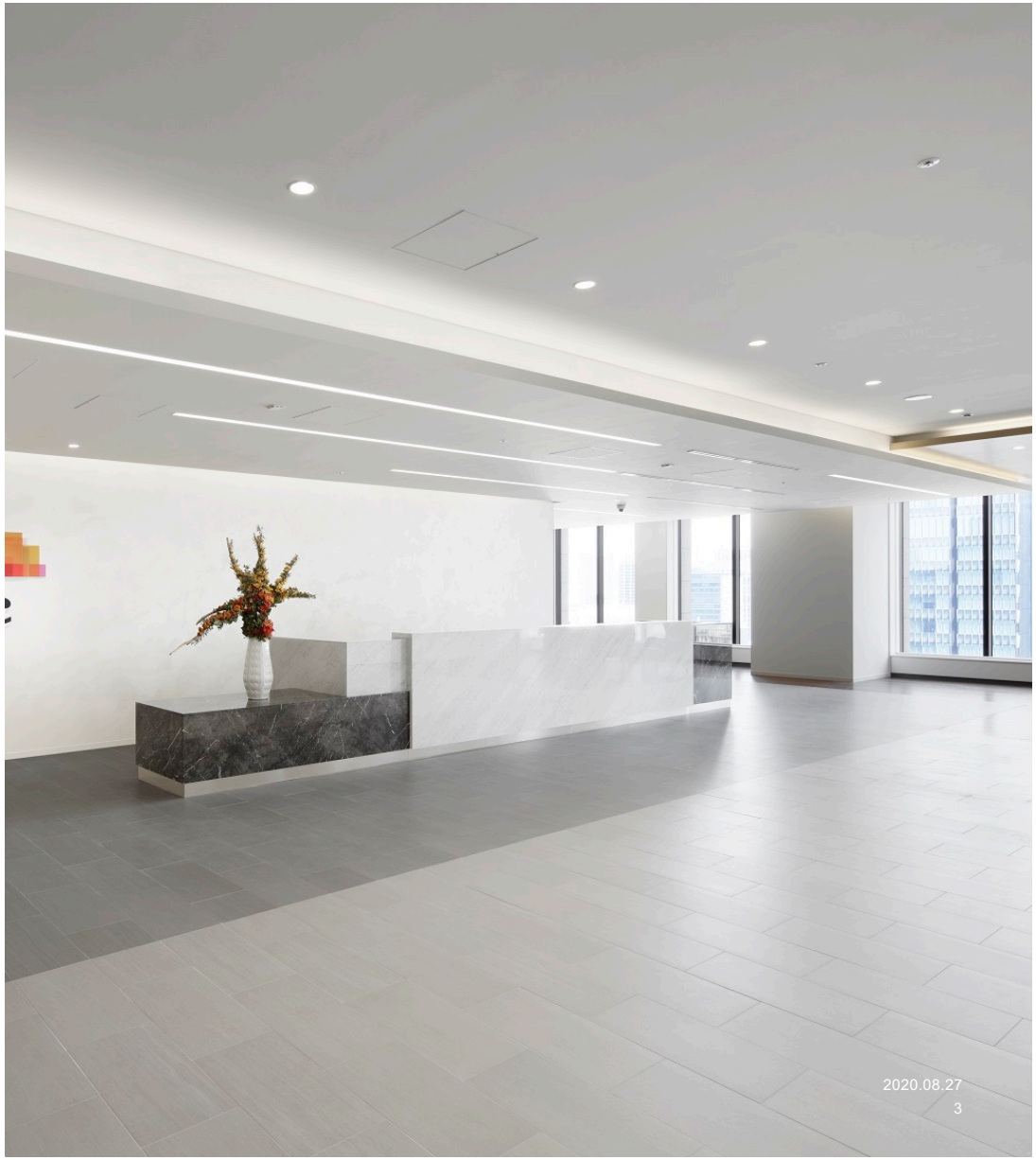
機械学習、深層学習をはじめとするいわゆる人工知能技術（AI）の社会での実装が進んできています。サイバーリスクの防御の面でも機械学習、深層学習を活用したサイバー防御製品やサービスが広がってきています。サイバーリスク管理にAIがどのように活用できるのか、人間とのかかわりはどうすべきか、そしてAIを活用したサイバー攻撃、AIに対するサイバー攻撃といったことにも触れていきながら、これからの課題を考えていきたいと思います。



スマートサイバー

AI活用時代のサイバーリスク管理

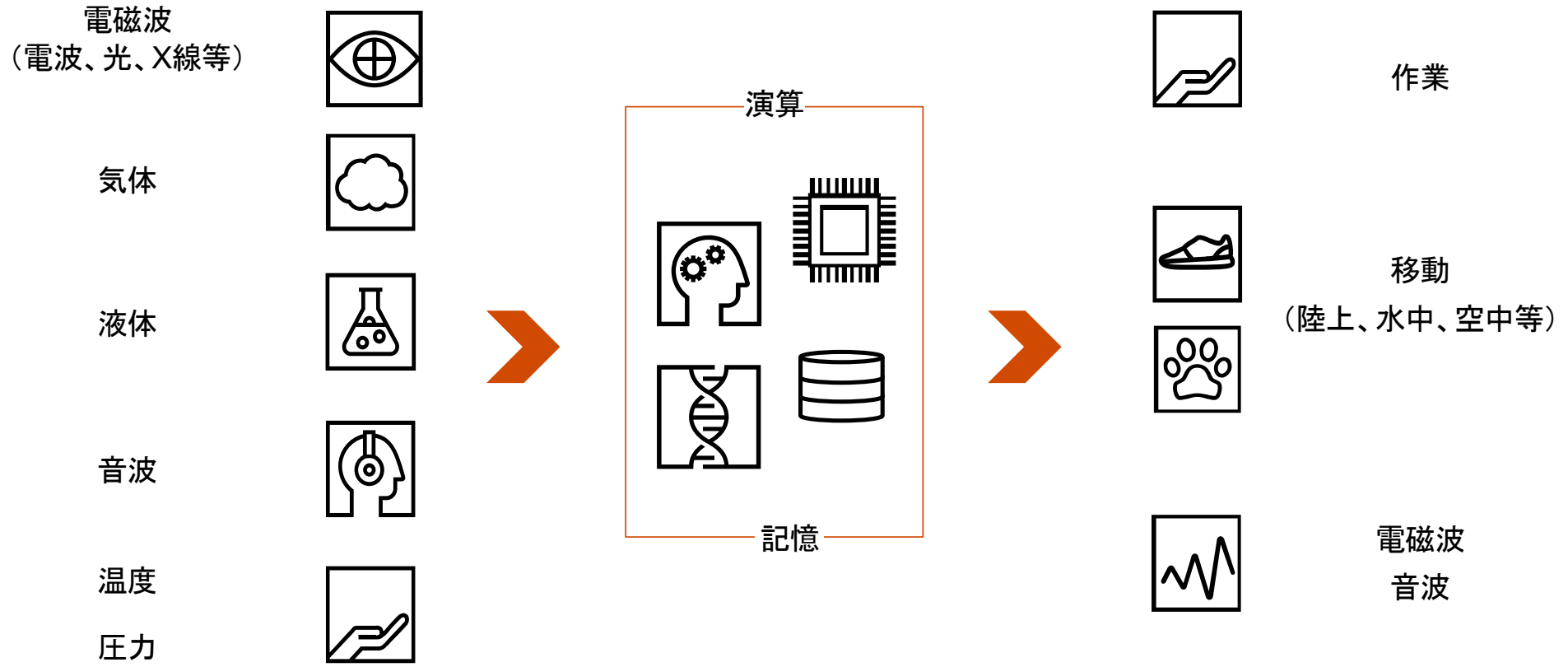
1. AIとサイバーリスク
2. 機械学習を活用したサイバー防御
3. 機械学習を活用したサイバー攻撃
4. 機械学習に対する攻撃
5. AI活用時代のサイバーリスク管理の課題



0

AIについて“ちょっと”
考えておこう

動物とAI付きロボット



AIは最後に人間になれるのか？

アトム

アァ・・・・・僕にはわからない

ぼくにはちっともわらない!!

音楽ってただ音を順序よくならべたものとしか、感じないんだい

絵を見たって音楽を聞いたってアァよかったなァよかったなァって思ったことがないの

人間の友達

君は泣いたり怒ったりできるのに

うん　でもそれだけのことしかできないんだ

人間のようにきれいだなァよかったなァと感じないんだよ

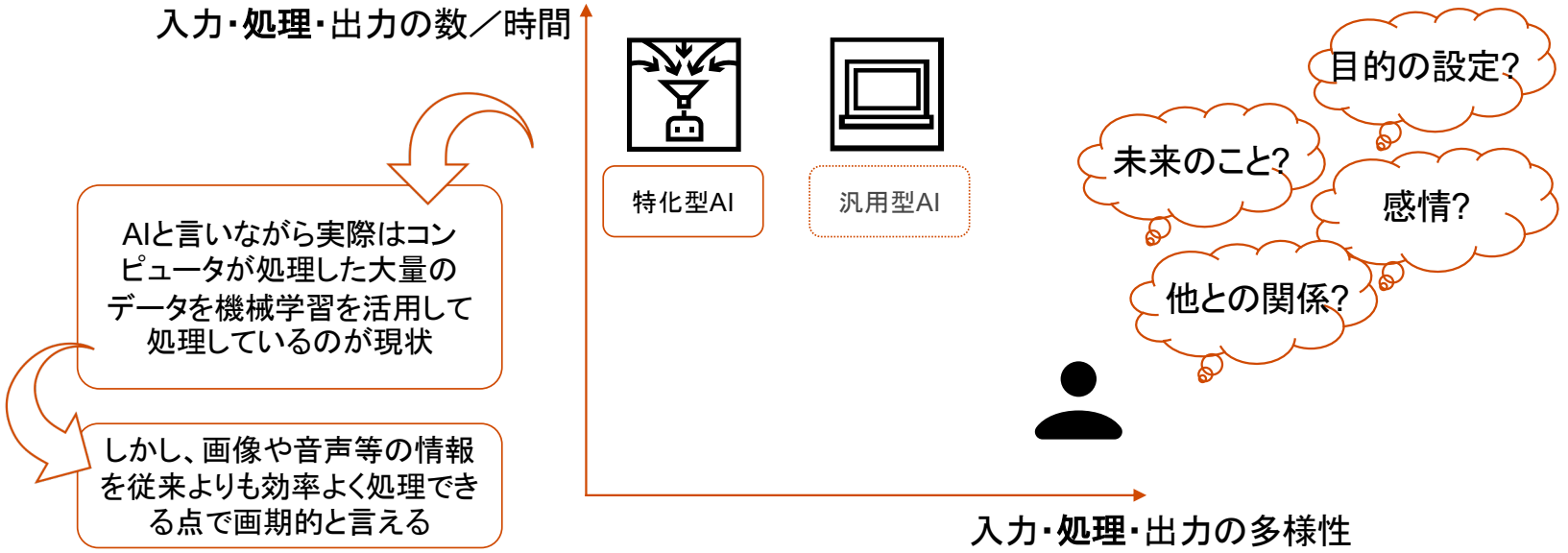
鉄腕アトム　第2巻
「アルプスの決闘の巻」
手塚治虫

AIは何が得意で何が不得意か

今のAIは人間からはほど遠く、限りなく従来のコンピュータシステムに近い

人間は多様な情報を総合的に判断し、多様な形態でアウトプットすることが得意

従来のコンピュータは特定の情報を蓄積し、大量反復的に利用し、特定の形態でアウトプットすることが得意



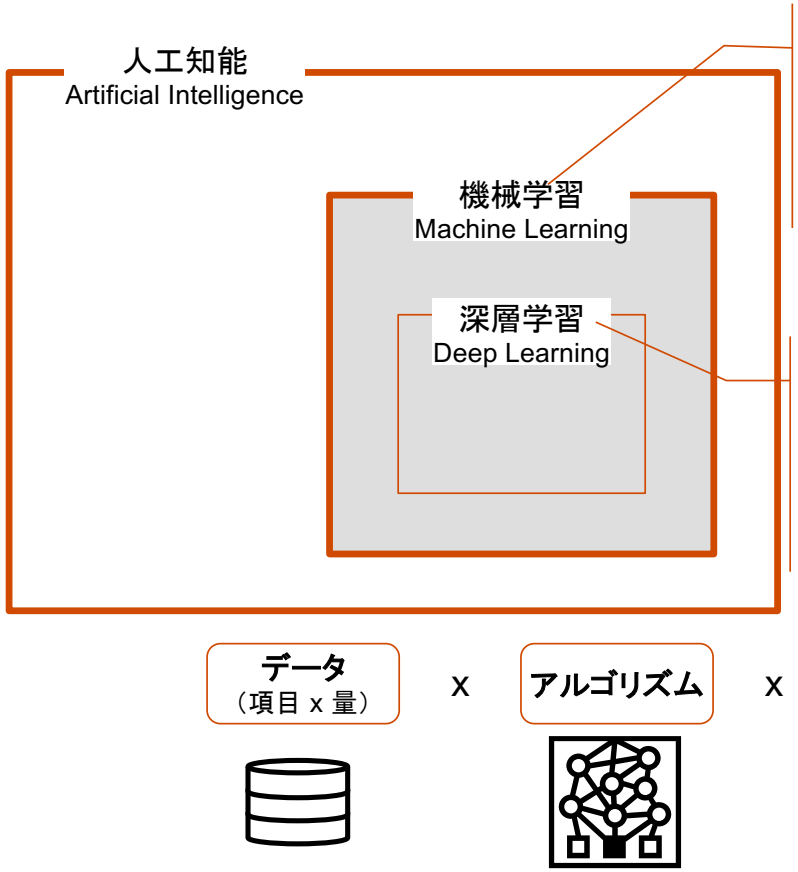
第三次AIブームと言ってもアトムにはほど遠い

これ以降、ここで取り扱うAIは機械学習(ML)に限ることにします

人工的につくった知的な振る舞いをするもの（システム）である
溝口 理一郎
（北陸先端科学技術大学院大学 教授）

人工的につくられた人間のような知能、ないしはそれをつくる技術
松尾 豊
（東京大学大学院工学系研究科 教授）

defined as a system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation
Kaplan, Andreas; Haenlein, Michael (1 January 2019). "Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence". *Business Horizons*. **62** (1)

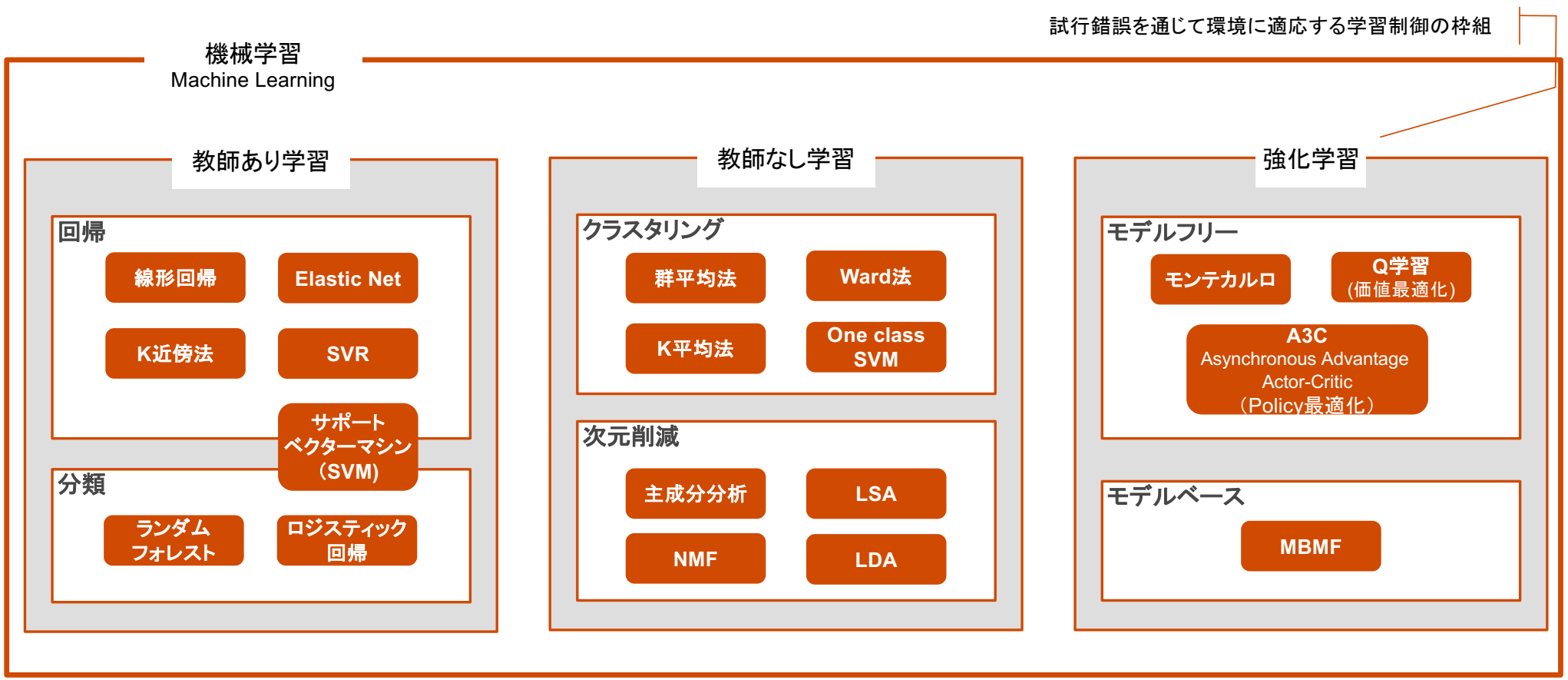


明示的にプログラムしなくても学習する能力をコンピュータに与える研究分野
“Field of study that gives computers the ability to learn without being explicitly programmed”
Arthur Samuel
（コンピューターサイエンティスト）

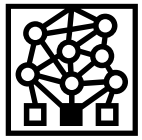
多層のニューラルネットワークを活用し、物事の特徴を抽出する技術。

データに存在しているパターンや規則の発見、特徴量の設定、学習などを機械自身が自動的に行う。

（参考）機械学習の分類例


試行錯誤を通じて環境に適応する学習制御の枠組
機械学習
Machine Learning
教師あり学習
回帰
線形回帰
Elastic Net
K近傍法
SVR
サポート
ベクターマシン
（SVM）
分類
ランダム
フォレスト
ロジスティック
回帰
教師なし学習
クラスタリング
群平均法
Ward法
K平均法
One class
SVM
次元削減
主成分分析
LSA
NMF
LDA
強化学習
モデルフリー
モンテカルロ
Q学習
(価値最適化)
A3C
Asynchronous Advantage
Actor-Critic
（Policy最適化）
モデルベース
MBMF

（参考）機械学習の分類例



タスク

回帰（または予測）
- 前の値に基づいて次の値を予測するタスク

分類
- 物事を異なるカテゴリーに分類するタスク

クラスタリング
- 類似性によってグループ化するタスク

相関ルール学習（または推奨）
- 以前の経験に基づいて何かを推奨するタスク

次元削減
- 一般的で最も重要な関数を検索するタスク

生成モデル
- 以前の知識に基づいて何かを作成するタスク

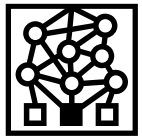
タスクを解決する方法

過去の傾向
- 教師あり学習（タスク駆動アプローチ）
- アンサンブル学習

今のトレンド
- 教師なし学習（データ駆動型アプローチ）
  通常、データの異常を見つけることを目的とする
- 半教師あり学習
  教師ありアプローチと教師なしアプローチの両方の利点を組み合わせた手法

将来の傾向
- 強化学習（環境駆動型アプローチ）
- アクティブラーニング

（参考）深層学習のアルゴリズム例

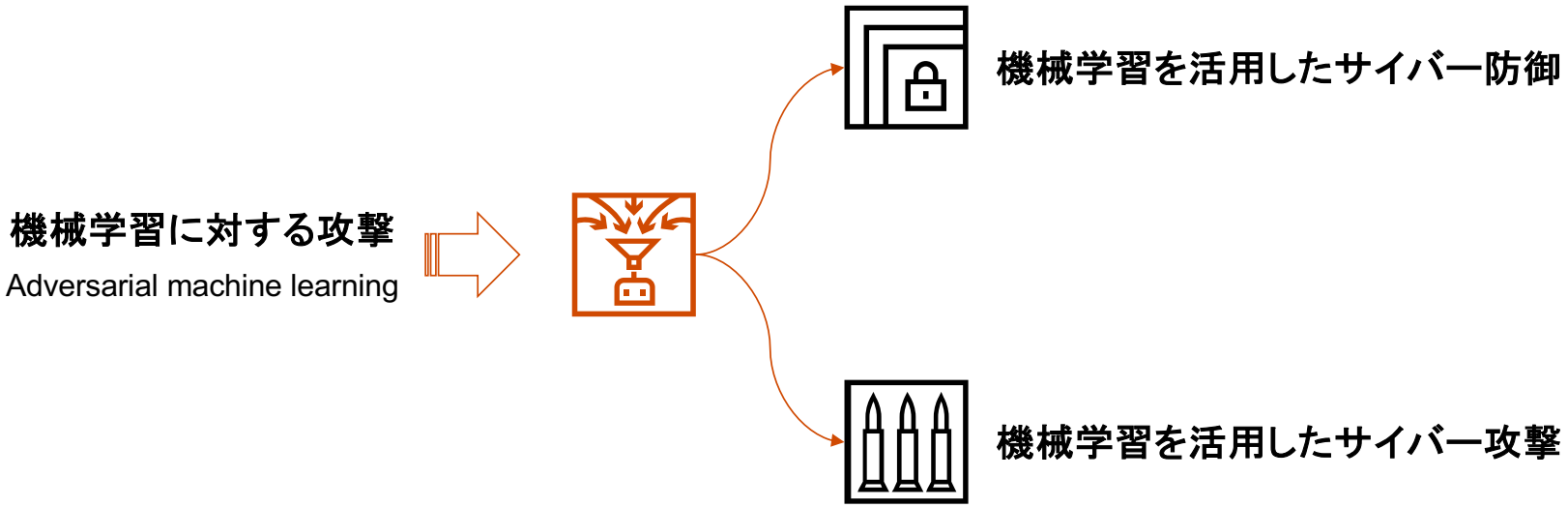


回帰	**再帰型ニューラルネットワーク（RNN)**	音声データ、テキストデータ
	…	
分類	**畳み込みニューラルネットワーク（CNN)**	画像データ
	…	
クラスタリング	自己組織化写像	
	…	
相関ルール学習	deep belief network (DBN)	
	…	
生成モデル	敵対的生成ネットワーク	
	…	

1

機械学習と
サイバーリスク

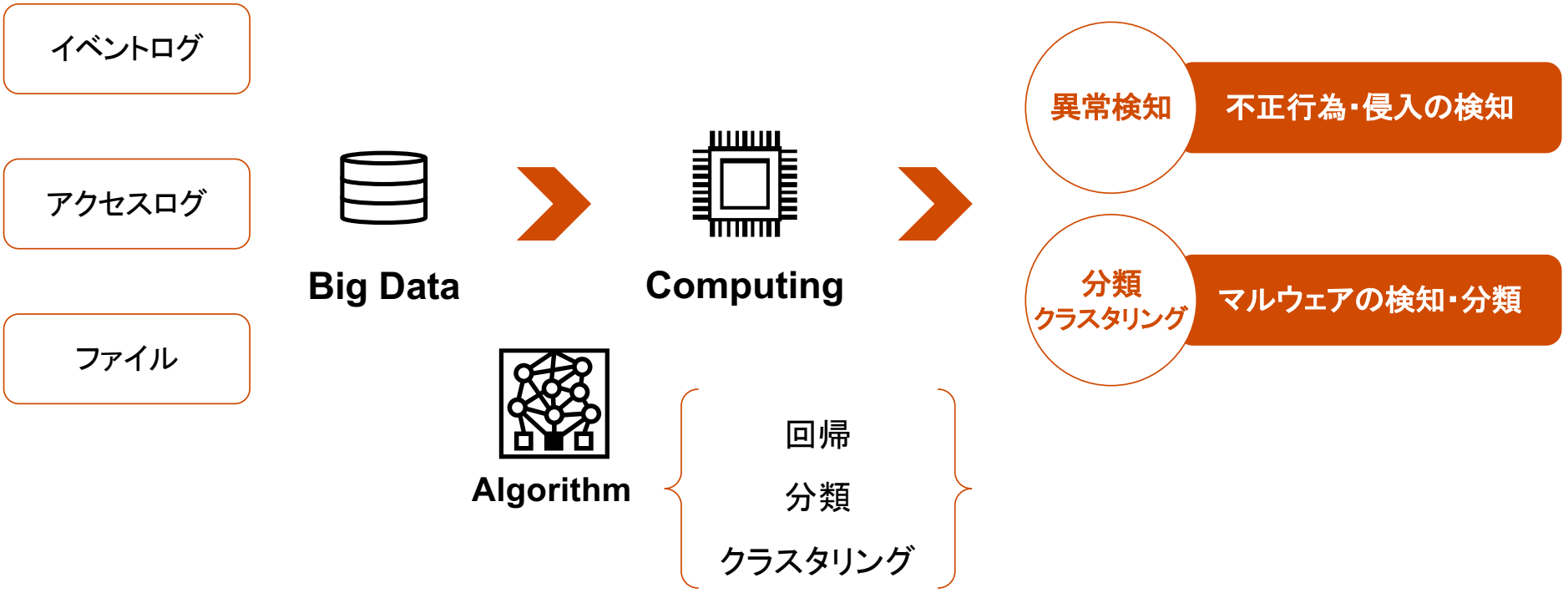
AIとサイバーリスクの接点



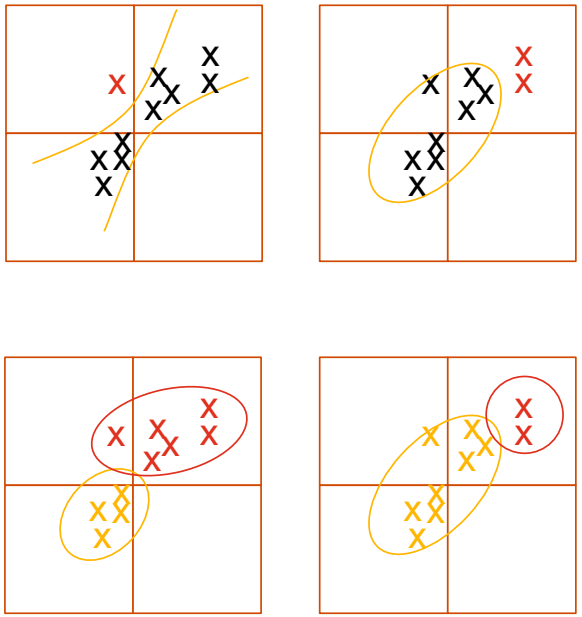
2

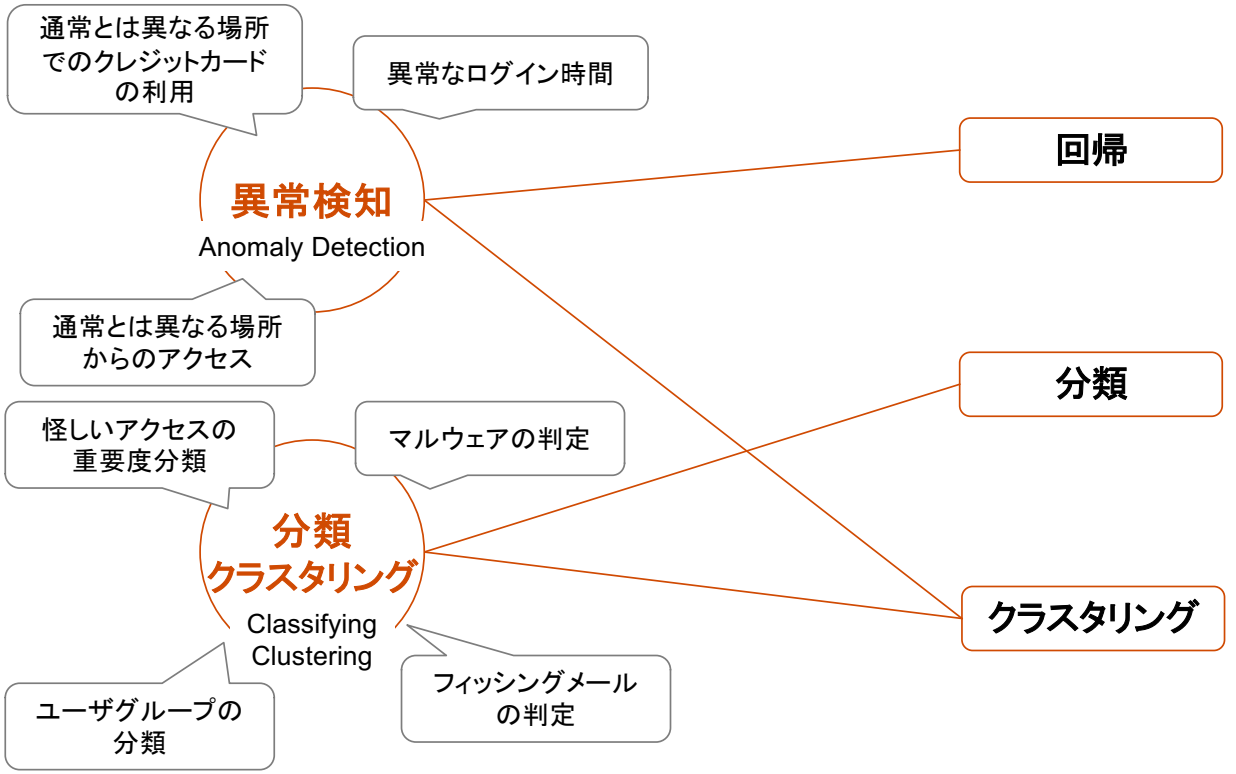
機械学習を活用したサイバー防御

サイバー防御において機械学習が活用しそうな分野例



防御における機械学習の利用イメージ





機械学習はどの対策で有効に機能しそうか？

異常なユーザアクションの検出

マルウェアの検出

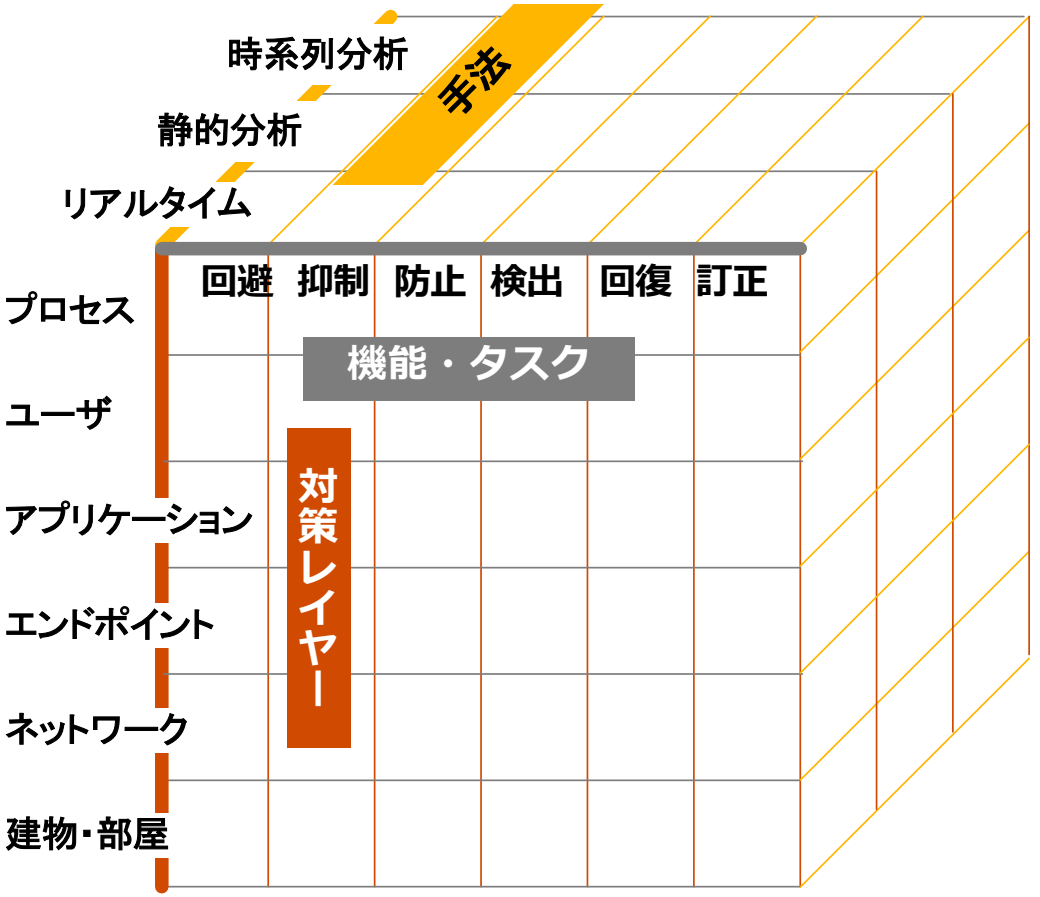
詐欺メール等の検出

異常なユーザアクションの検出

悪意あるスクリプト検出

マルウェアの検出

悪意あるURL検出



（参考）各レイヤーにおける機械学習を活用した対策例

	教師あり学習		教師なし学習
	回帰	分類	クラスタリング
プロセス動作	次のユーザアクションを予測し、クレジットカード不正などの外れ値を検出する	既知のタイプの不正を検出する	ビジネスプロセスを比較し、外れ値を検出する
ユーザ行為	ユーザアクションの異常を検出する （例：異常な時間のログイン）	ピアグループ分析のために異なるユーザをグループに分類する	ユーザをグルーピングし、外れ値を検出する
アプリケーション	HTTPリクエストの異常を検出する （例：XXE攻撃、SSRF攻撃、認証バイパス等）	インジェクションなどの既知のタイプの攻撃を検出する （例：SQLi、XSS、RCE等）	DDOS 攻撃や大量搾取を検出するためのユーザアクティビティのクラスタリング
エンドポイント	実行可能なプロセスの次のシステムコールを予測し、実際のプロセスと比較し、外れたプロセスを検出する	スキャンやスプーフィングのようなネットワーク攻撃の異なるクラスを識別する	安全な電子メールゲートウェイ上でのマルウェア保護（例：合法的なファイルの添付ファイルを外れたものから分離する）
ネットワーク	ネットワークパケットのパラメータを予測し、通常から外れたパケットを検出する	スキャンやスプーフィングなど、さまざまなクラスのネットワーク攻撃を識別する	フォレンジック分析

2018.10.04 Machine Learning for Cybersecurity 101 by Alexander Polyakov
https://towardsdatascience.com/machine-learning-for-cybersecurity-101-7822b802790b
を基に作成

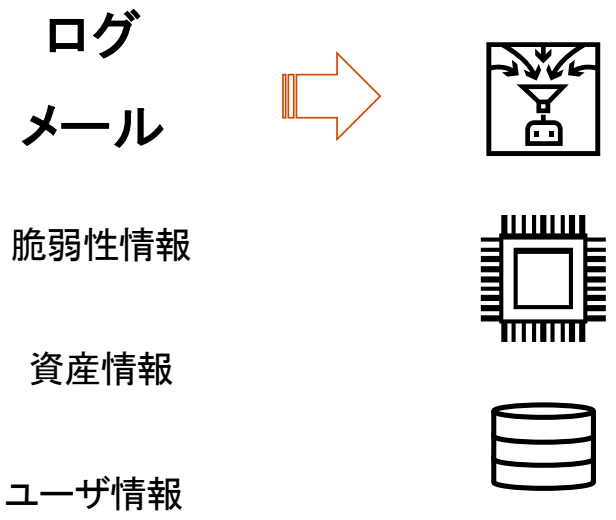
Cyber Security Flameworkで機械学習が活用できそうな部分例

異常とイベント（DE.AE）:
異常な活動は、検知されており、イベントがもたらす潜在的な影響が、把握されている。

DE.AE-1:
ネットワーク運用のベースラインと、ユーザとシステムで期待されるデータフローが、定められ、管理されている。

DE.AE-3:
イベントデータは、複数の情報源やセンサーから収集され、相互分析されている。

DE.AE-5:
インシデント警告の閾値が、定められている。



防御における機械学習の活用

機械学習を利用した防御により、効率的・効果的な防御が可能となる

3

機械学習を活用した
サイバー攻撃

サイバー攻撃において機械学習が活用しそうな分野例

SNS

漏洩した
ID/Password
リスト

メール文書等

Big Data

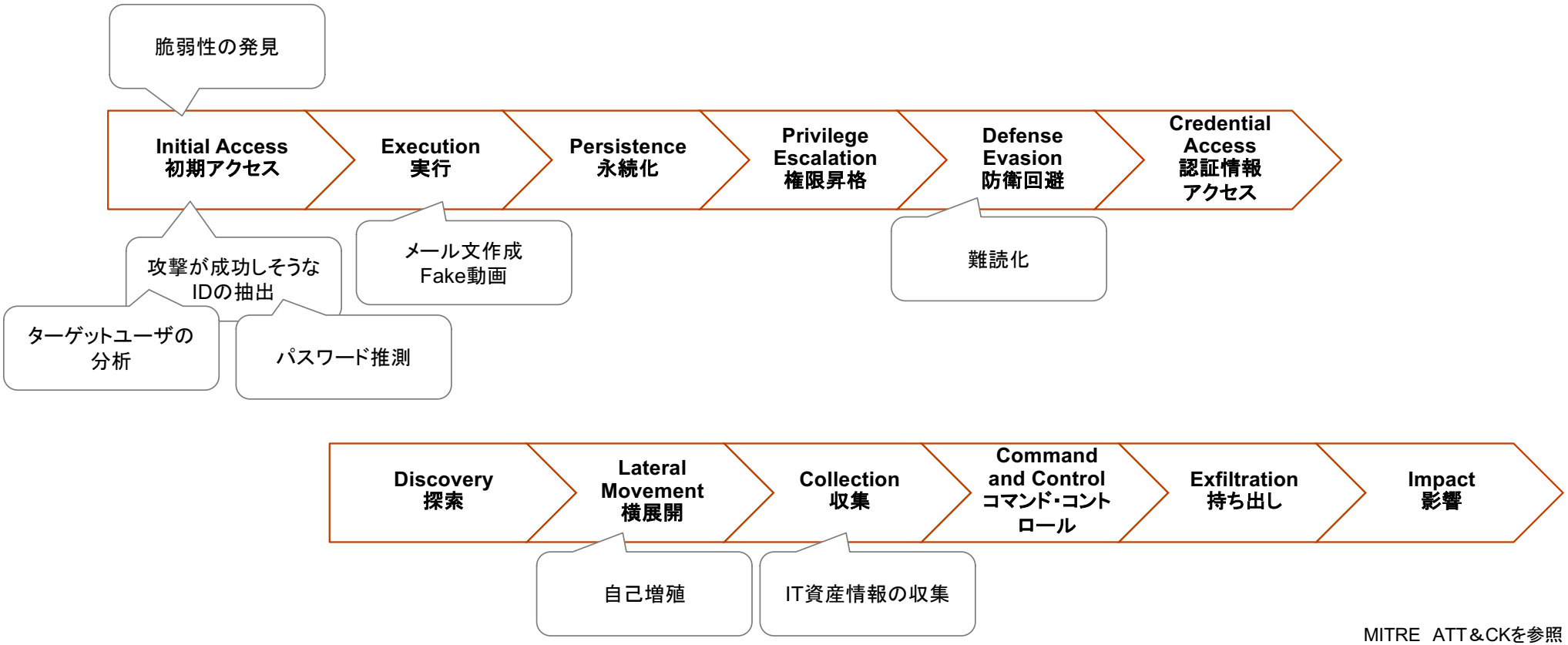
Computing

Algorithm

- 回帰
- 分類
- クラスタリング
- 相関ルールの学習
- 生成モデル

- 情報収集
- なりすまし
- 不正アクセス
- 攻撃

攻撃者はどこで機械学習を活用できるか？



MITRE ATT＆CKを参照

機械学習による攻撃はこれから洗練されていく？

情報収集

情報収集

なりすまし

不正アクセス

攻撃

- スパムメールのリンクをクリックしそうなID群の分離
  - メール文面の分析から興味の対象を理解する
- お金を支払いそうなID群の分離
  - SNSに上げている写真を分析し経済力を理解する
- 交友関係等のマップの作成
  - 信頼の厚い友人等の把握とその人とのメールのやりとりを理解する
- 仮想化ネットワークの設定の推測等

Social Mapper

機械学習による攻撃はこれから洗練されていく？

なりすまし

情報収集

なりすまし

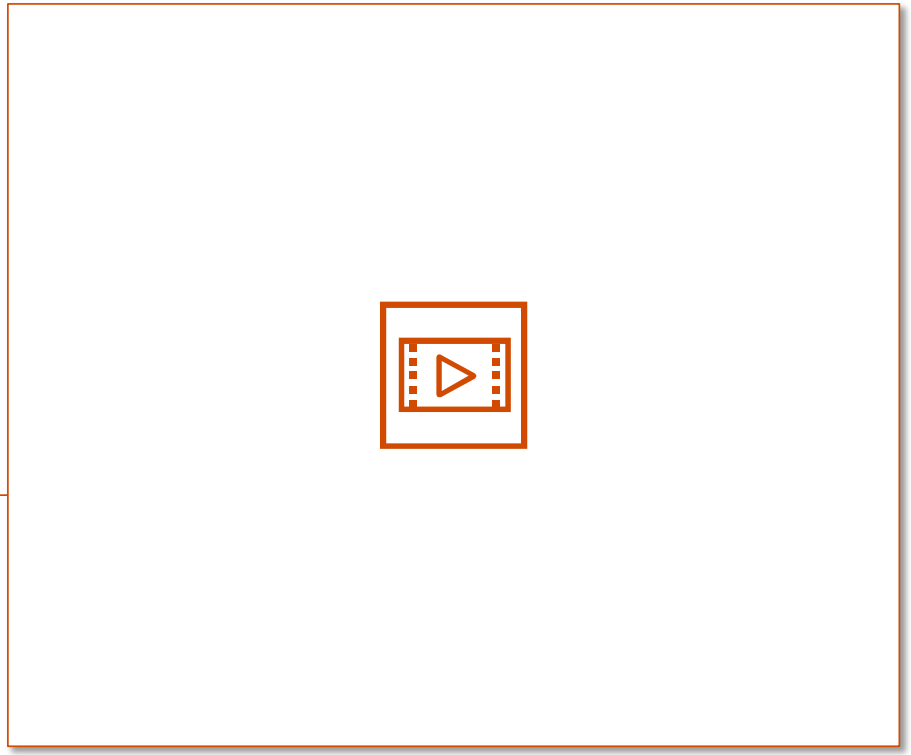
不正アクセス

攻撃

スパムメールの作成

偽造メールの作成

Fake動画の作成



https://www.bbc.com/news/av/technology-40598465

機械学習による攻撃はこれから洗練されていく？

不正アクセス

- 情報収集
- なりすまし
- **不正アクセス**
  - CAPTCHAバイパスのための機械学習
  - パスワードブルートフォースのための機械学習
- 攻撃

機械学習による攻撃はこれから洗練されていく？

攻撃

情報収集

なりすまし

不正アクセス

攻撃

State
一般的なカテゴリの特徴からなる2350次元の特徴ベクトルを使用
- PE ヘッダのメタデータ
- セクションのメタデータ：セクション名、サイズ、特性
- テーブルのメタデータのインポートとエクスポート
- 人間が読める文字列の数 (ファイルパス、URL、レジストリキー名など)
- バイトヒストグラム
- 2D バイトエントロピーヒストグラム

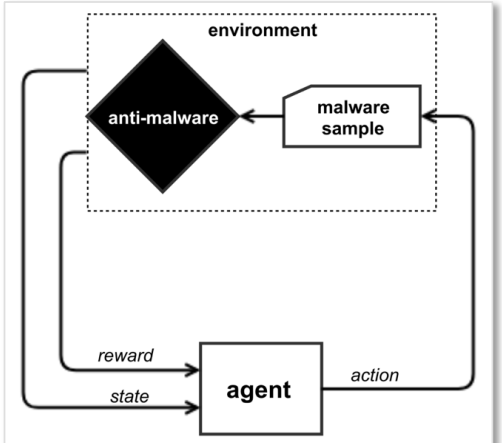


Figure 1: Markov decision process formulation of the malware evasion reinforcement learning problem.

Action
- 未使用のIATに関数を追加
- 既存セクション名の操作
- 新規（未使用）セクションの作成
- セクションの最後の余分なスペースにバイト列を追加
- 元のエントリポイントにジャンプするだけの新しいエントリポイントの作成
- 署名者情報の削除
- デバッグ情報の操作
- バイナリをパックまたはアンパック
- PEヘッダのチェックサムの変更
- PEファイルの最後にバイト列を追加

脆弱性発見のための機械学習

マルウェア作成のための機械学習

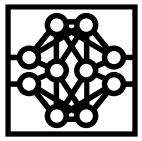
偽オンラインレビュー等（crowd turfing）のための機械学習

OpenAI Gymのマルウェア環境

Learning to Evade Static PE Machine Learning Malware Models via Reinforcement Learning
Hyrum S. Anderson, Anant Kharkar, Bobby Filar, David Evans, Phil Roth

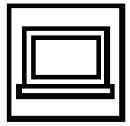
攻撃における機械学習の活用

攻撃者を支援するツールとして機械学習の活用が進むだろう



ターゲットに関する情報が
ネット上に分散して大量に
存在している

機械学習が手軽に利用
できる環境が揃ってきた



機械学習の活用が進むだろう

情報収集

なりすまし

不正アクセス

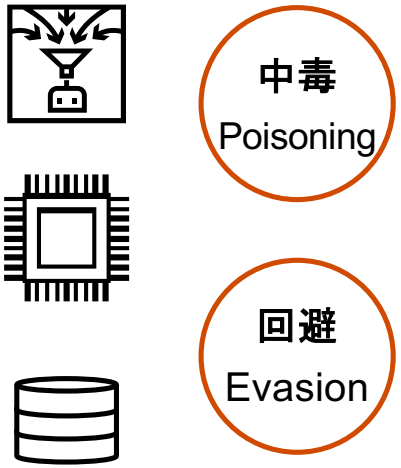
攻撃

4

機械学習に対する攻撃

機械学習に対する攻撃（敵対的機械学習 Adversarial Machine Learning）

機械学習が人間を真似たものであるので、教え方一つで変わる



中毒
Poisoning

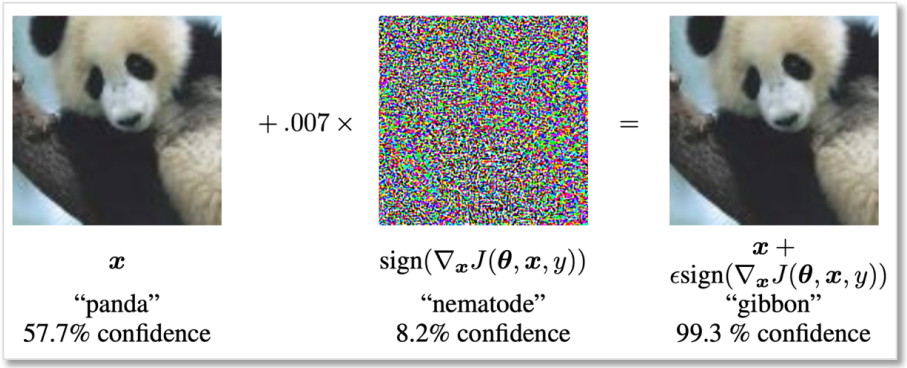
学習入力データに細工する

回避
Evasion

学習結果を回避させる入力をする

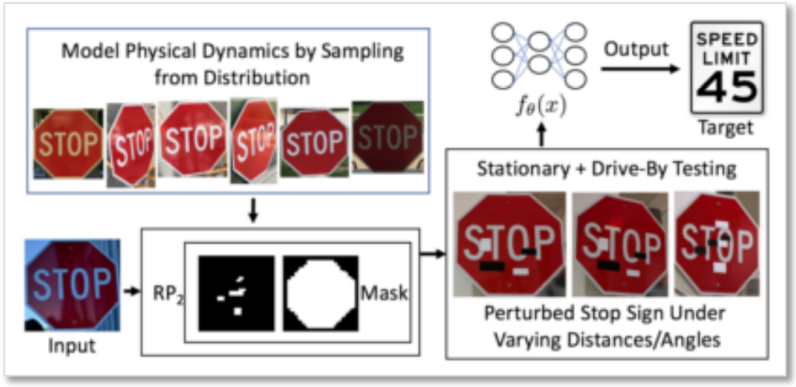
敵対的学習（回避）の例

機械学習は内容を理解しているわけではないので例えばデータエ夫すれば騙せる



[Szegedy et al.: Intriguing properties of neural networks. ICLR2014.]

Explaining and Harnessing Adversarial Examples
Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy
https://arxiv.org/abs/1412.6572



[Evtimov et al.: Robust Physical-World Attacks on Machine Learning Models. 2017]

Robust Physical-World Attacks on Deep Learning Models
Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao,
Atul Prakash, Tadayoshi Kohno, Dawn Song
https://arxiv.org/abs/1707.08945

敵対的攻撃の分類例

トレーニング中はデータを汚染する攻撃、本番中は誤判断させる攻撃が考えられる

データやアルゴリズムに関する情報を入手する

ステージ ＼ 目標	スパイ活動 Espionage	妨害 Sabotage	不正 Fraud
中毒 Poisoning：トレーニング中	中毒による推論	中毒 トロイの木馬 バックドア	中毒
回避 Evasion：本番運用中	推論攻撃	回避（過検知） 敵対的再プログラミング	回避（検知もれ）

異なる目的に利用できるように変える

How to attack Machine Learning (Evasion, Poisoning, Inference, Trojans, Backdoors)
Alexander Polyakov
Aug 6, 2019
https://towardsdatascience.com/how-to-attack-machine-learning-evasion-poisoning-inference-trojans-backdoors-a7cb5832595c
を基に作成

ホワイトボックス、グレーボックス、ブラックボックス

ブラックボックスを維持できれば本番運用中の攻撃はかなり防げそうです

容易 ← **攻撃の容易性** → **困難**

ホワイトボックス	グレーボックス	ブラックボックス
攻撃者は、 データセット ニューラルネットワークのタイプ、 構造、レイヤー数、 トレーニングされた全ての重み を含む、 システムに関する すべてが知られている	攻撃者は、 データセット、 ニューラルネットワークのタイプ、 構造、レイヤー数等 の詳細を知っている	攻撃者は、 システムに関する情報は 何も知らない

機械学習に対する攻撃

機械学習の利用が進めば進むほど、脅威は大きくなる可能性があります

学習 → 柔軟なシステム → 変化させられるという攻撃を受ける

攻撃を受けた結果かどうか
人間が**判断するのが**
難しい場合がある

機械学習結果への
依存が増加

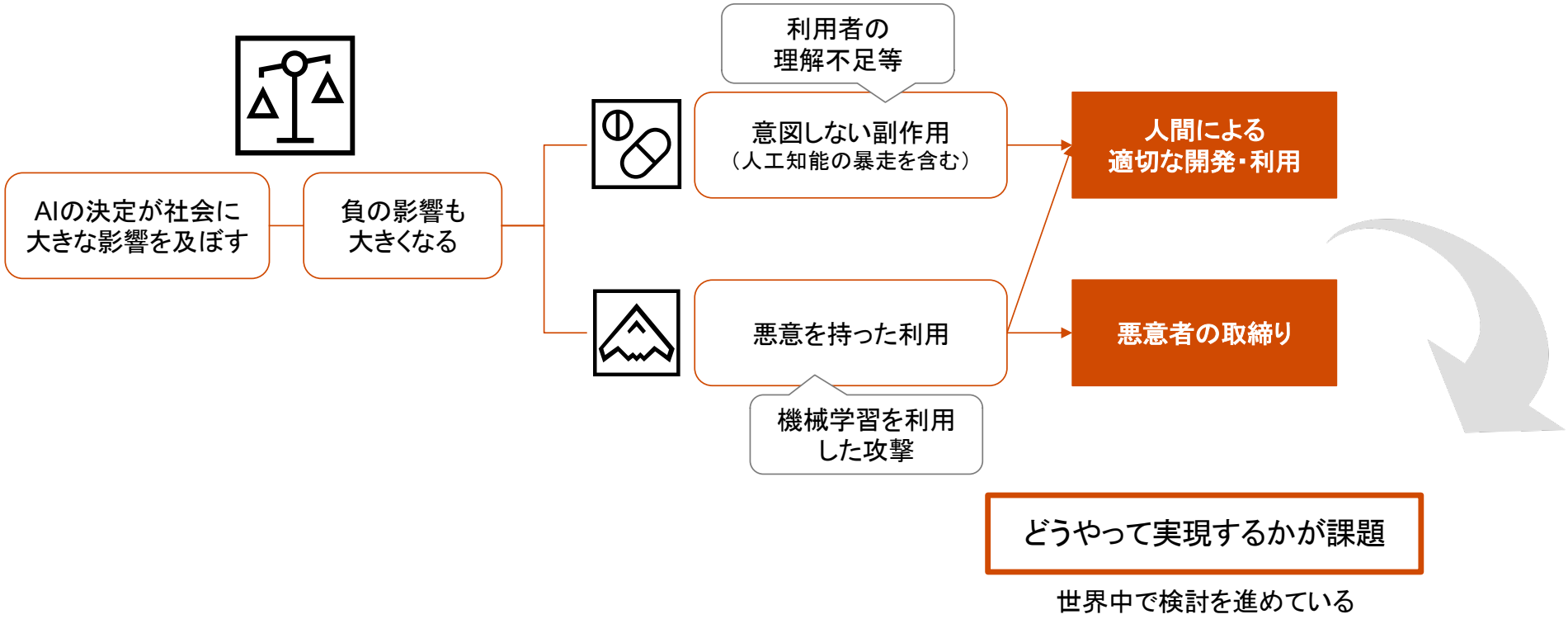
脅威は大きくなるだろう

5

AI活用時代のサイバーリスク管理の課題

AI活用自体の課題

ガバナンスと倫理問題はある。どうやって実現するかが課題



機械学習活用におけるサイバーリスク管理の課題
（特に防御面での活用に関して）

機械学習は機能、ツールであり特定の問題を解決する上で役割を果たすが、すべての問題が機械学習で解決できるわけではない

機械学習は、さまざまなユースケースで常に良い結果をもたらすという考え方はまったくの誤り

機械学習は、他の選択肢と比較してリソースを大幅に消費することが多い

機械学習への過度の依存は、誤った安心感を生み出す可能性がある

機械学習の保護を適切にしなければ、攻撃され、誤った判断をし続けることになる

スマートサイバー

AI活用時代のサイバーリスク管理

1. AIとサイバーリスク
2. 機械学習を活用したサイバー防御
3. 機械学習を活用したサイバー攻撃
4. 機械学習に対する攻撃
5. AI活用時代のサイバーリスク管理の課題

Question

Thank you

www.pwc.com/jp