LS研総合発表会
2020年度研究成果発表

データ利活用促進に向けたデータ分析に必要なデータを効率的に収集する技法の研究（クラス 1）

【LS研 09分科会】 2021.5.18（火）

Agenda

1. 研究概要と目的
2. 研究対象
3. データ収集技法の検証内容
4. データ収集ガイドライン
5. 結論

当分科会のテーマ

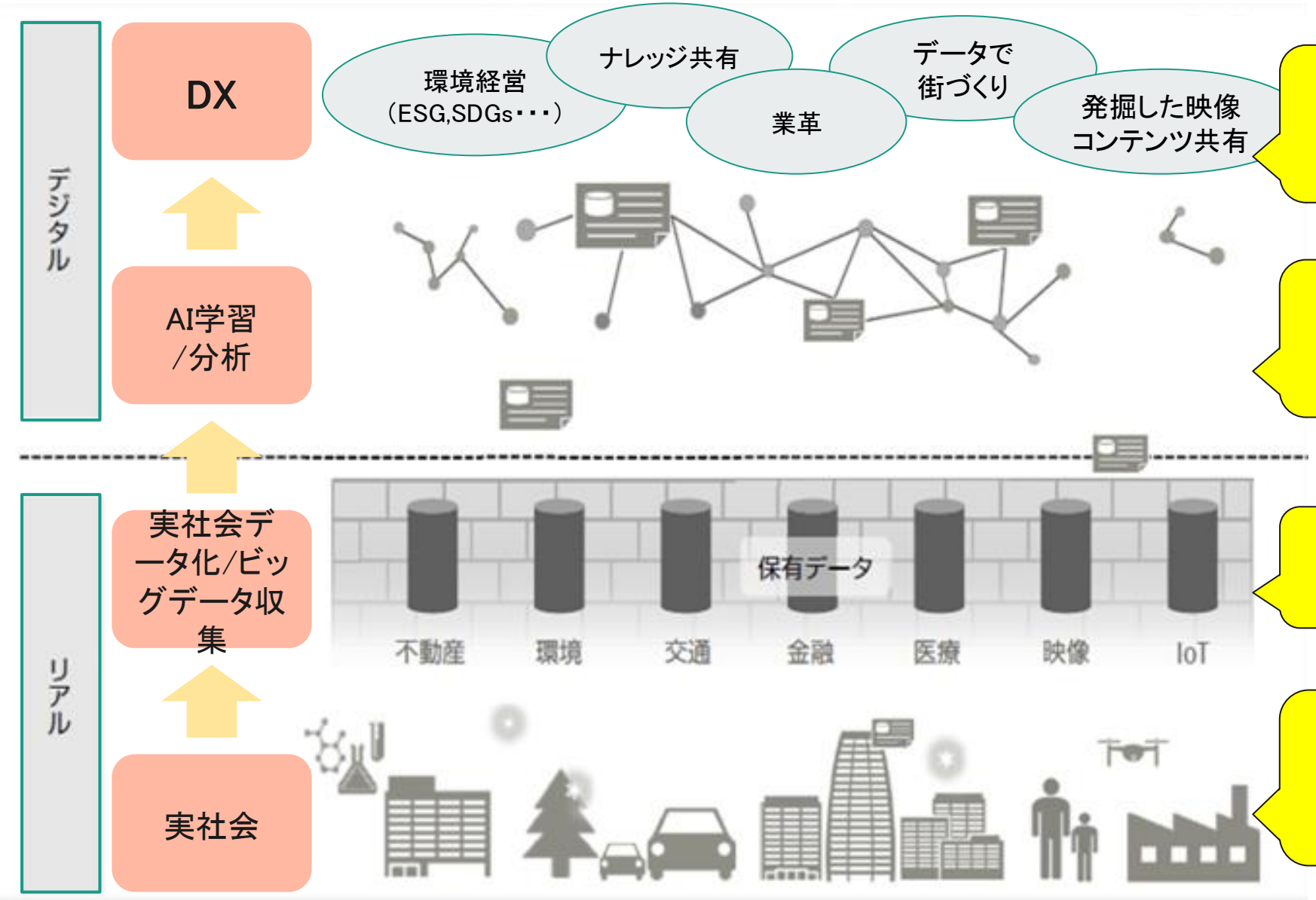
データ利活用促進に向けたデータ分析
に必要なデータを効率的に収集する技法の研究

データ利活用促進に向けた、

データ分析に必要な、

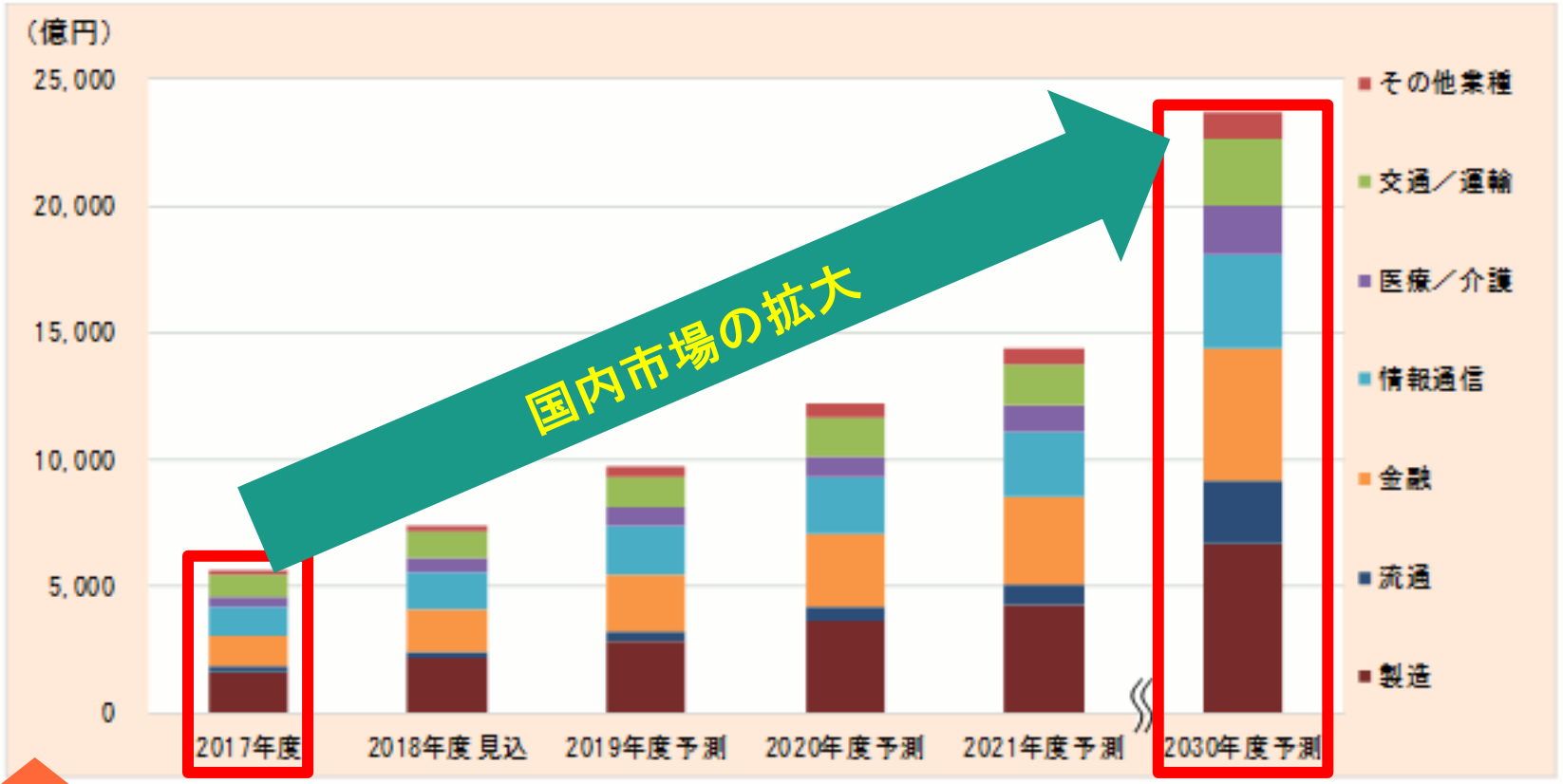
データを効率的に収集する、技法の研究

データ利活用が実現するデジタルトランスフォーメーション



ＤＸの国内市場と注目度

投資金額ベースのDXの業界別の国内市場

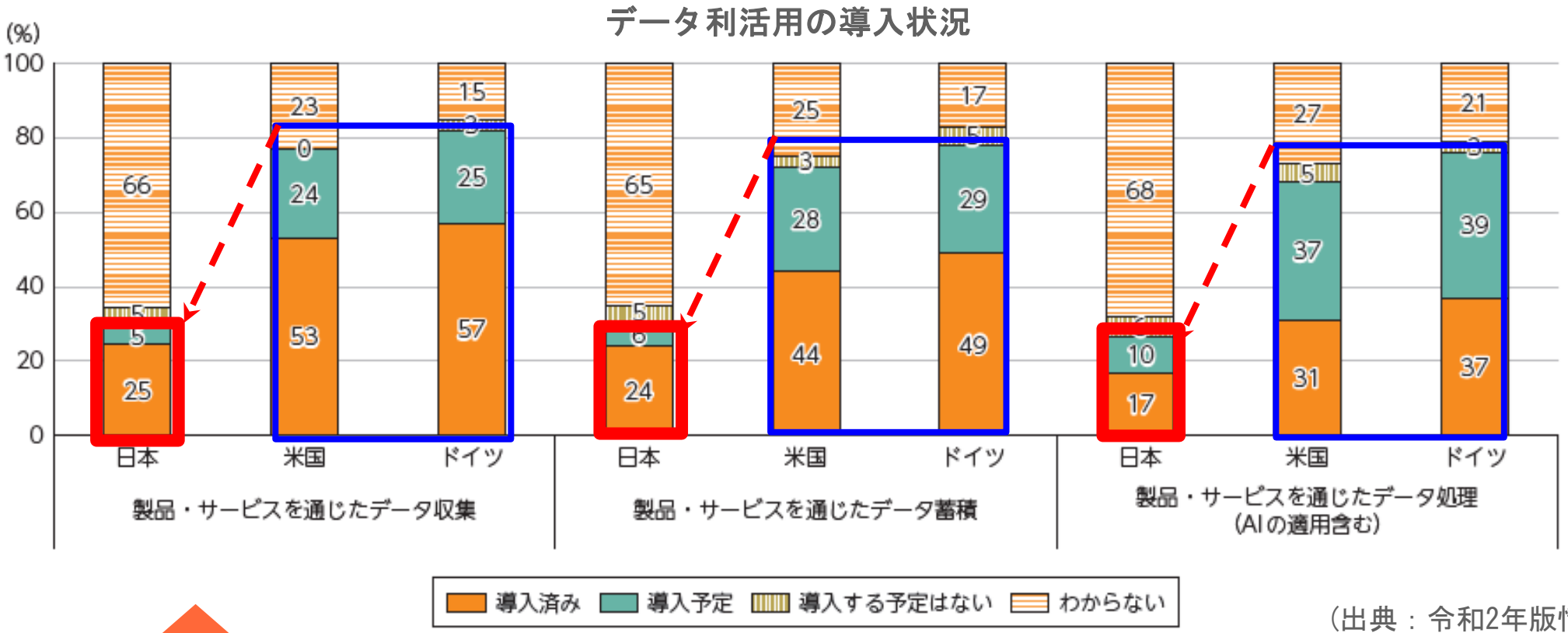


（出典：富士経済グループニュースリリース）

多くの国内企業がDXに取り組む姿勢
または
DXを重要視している

**DXへの取り組みが遅れることは
競争力の低下に繋がる**

日本と海外のデータ利活用状況



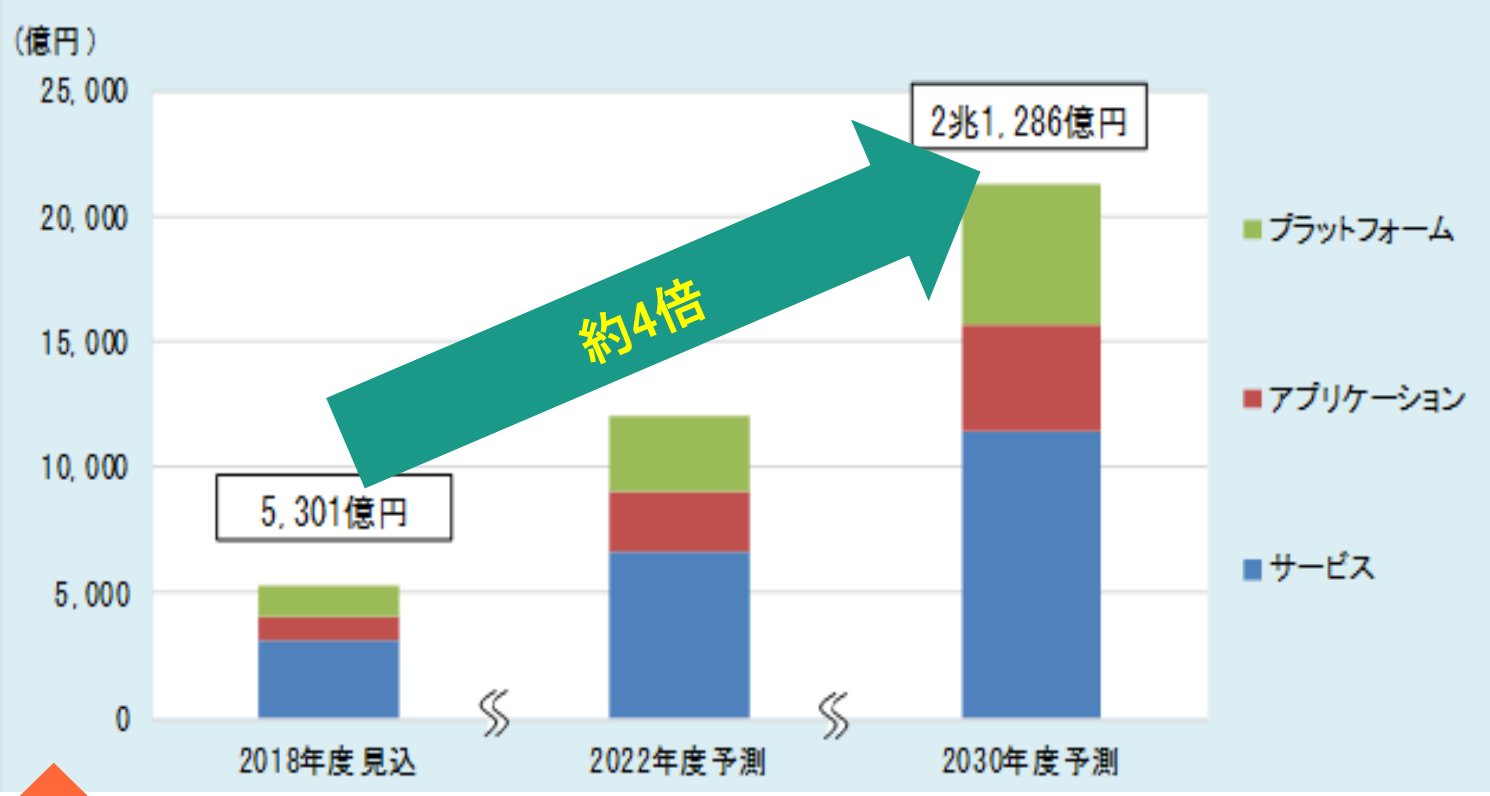
（出典：令和2年版情報通信白書）

日本はデータ利活用の全てのフェーズで
アメリカ･ドイツから遅れている

**データ利活用のスタートラインである
「データ収集」の導入率の向上が先決**

ＡＩ/機械学習に着目する理由

AIの今後の市場予測



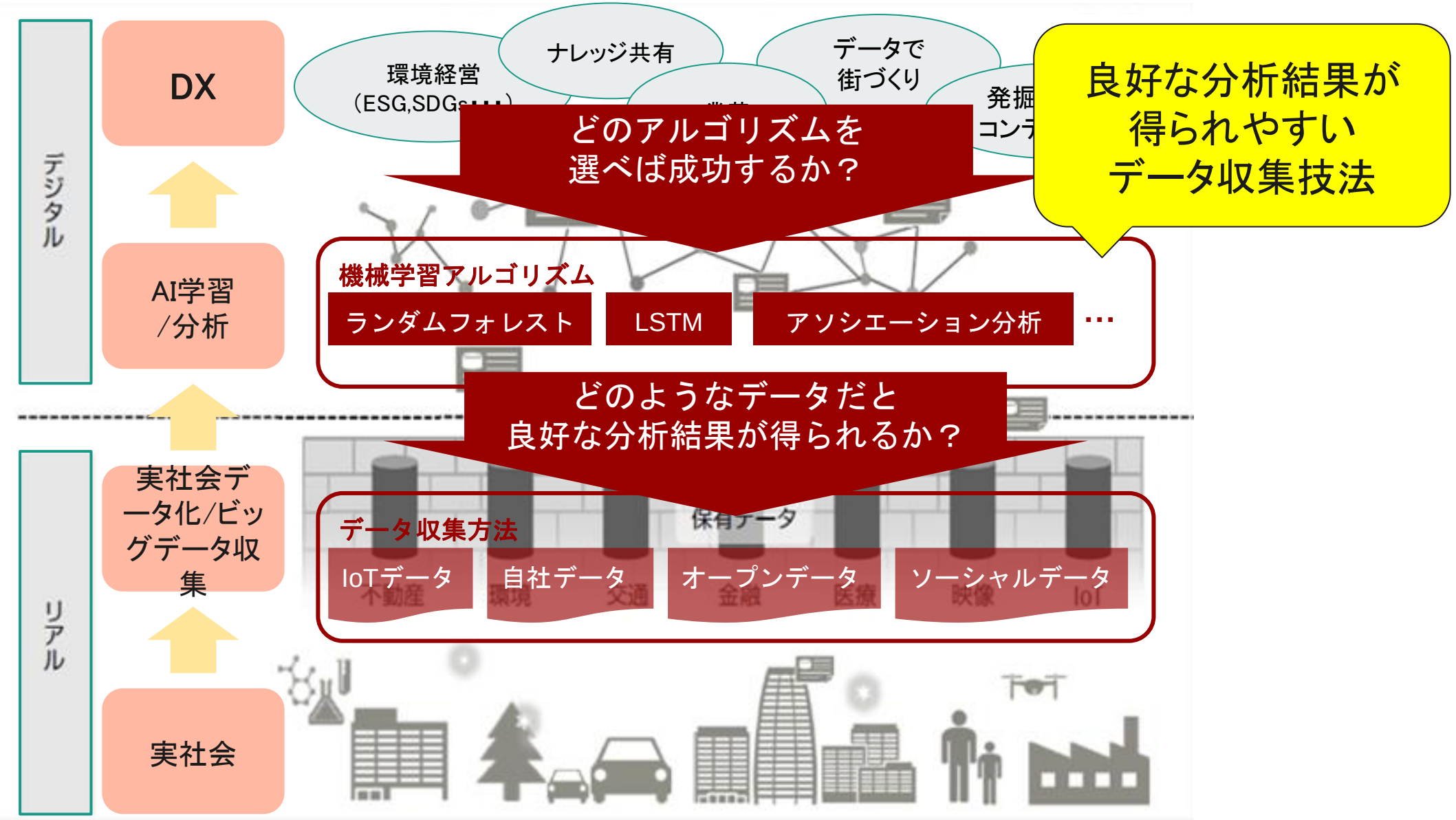
（出典：プレスリリース：『2019 人工知能ビジネス総調査』まとまる（2019/6/7発表 第19039号））

市場規模の拡大、すなわち
AI・機械学習が注目されている

「ＤＸ」とは、データとICT技術で業務に変革をおこし
競争上の優位性を確立すること

大量のデータを扱うICT技術　＝　AI・機械学習

研究の着眼点



研究目的

AI/機械学習において
良好なデータ分析結果が得られるデータ収集技法の提案

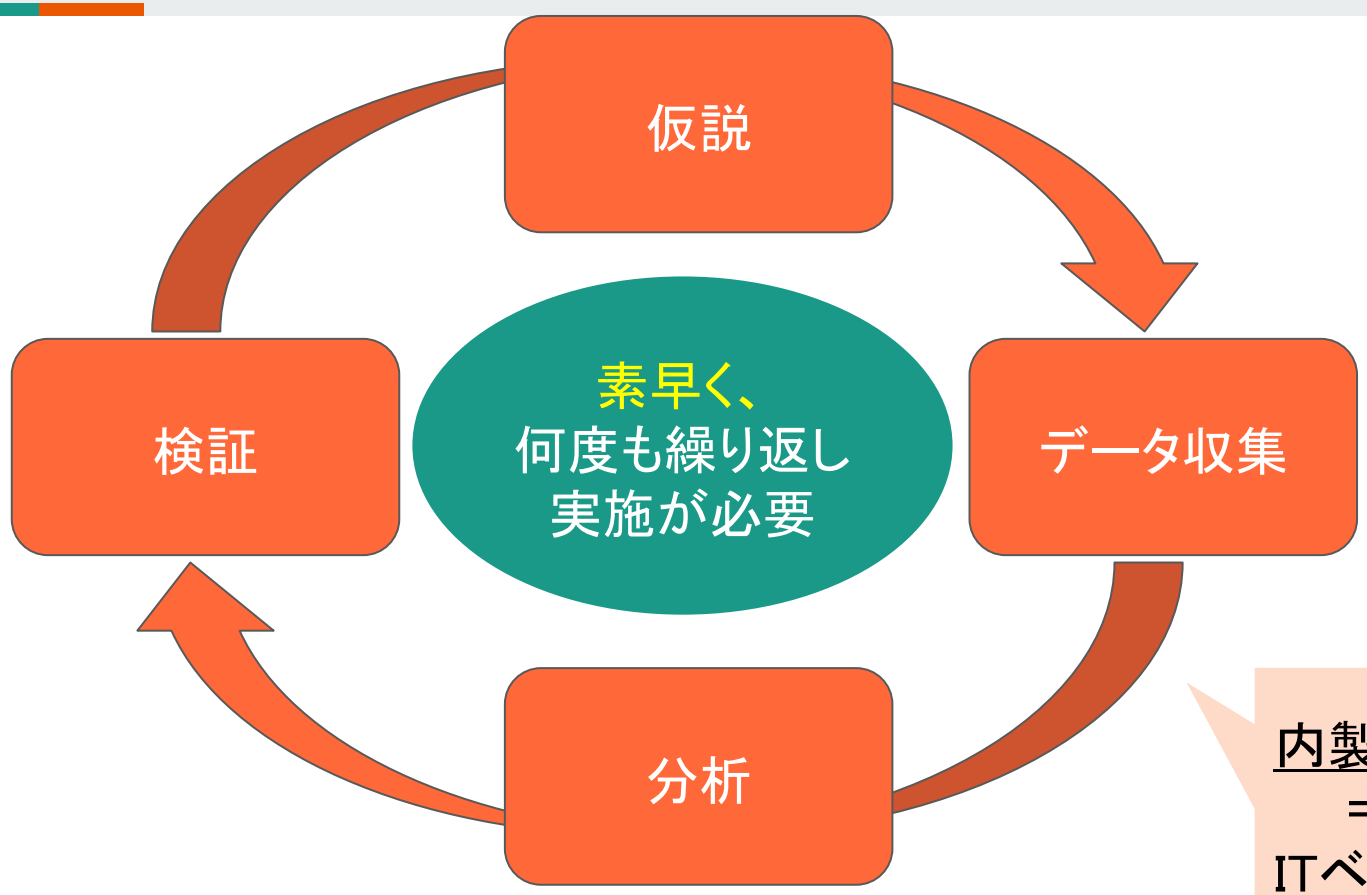
具体的には

・目的に最適な分析アルゴリズムの選定方法の提案

・収集データとデータ分析結果との関係性に着目したデータ収集技法の提案

の2点に注力する。

研究意義



内製の場合
⇒スキルは乏しいが、時間を短縮しやすい
ITベンダー、SIer、データ分析企業の場合
⇒スキルはあるが、時間が必要になりやすい

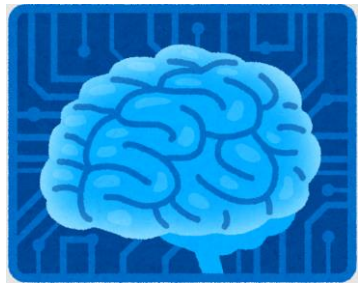
**データ利活用のユーザ自身がデータ分析スキルを持つことで、
データ収集と仮説検証の反復を素早く実施できる**

Agenda

1. 研究概要と目的
2. 研究対象
3. データ収集技法の検証内容
4. データ収集ガイドライン
5. 結論

ＡＩとは

● AI（人工知能）とは…



知的な機械、

特に、

知的なコンピュータプログラムを作る科学と技術

を意味する

ＡＩの活用分野

AI実用化のパターンは大きく、3つに分類される

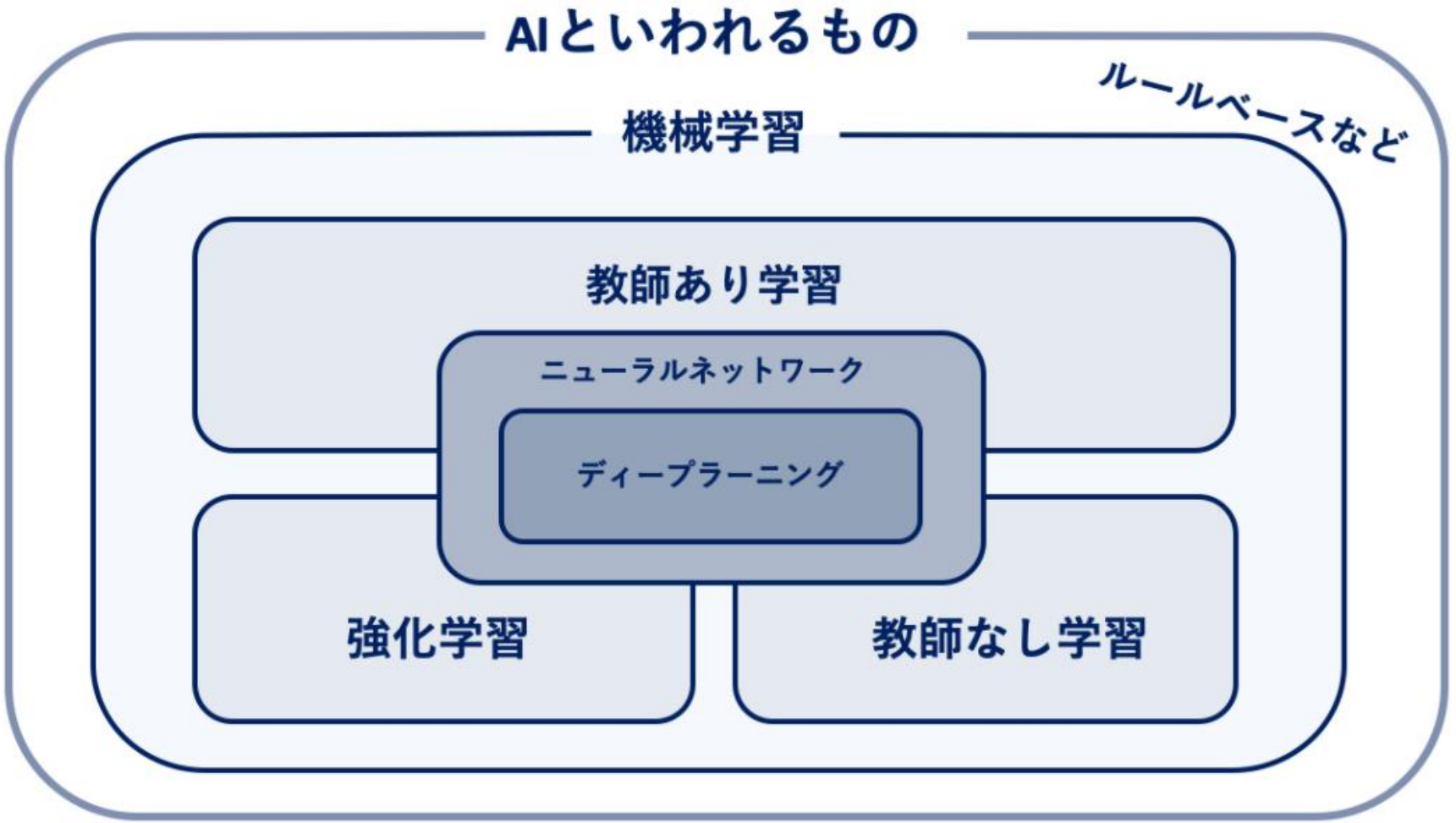
識別 ・ 予測 ・ 実行

識別	予測	実行
音声認識	数値予測	表現生成
画像認識	マッチング	デザイン
動画認識	意図予測	行動最適化
言語解析	ニーズ予測	作業の自動化

人工知能の利用パターン（出典：平成 28 年版情報通信白書）

ＡＩの全体像（構成要素）

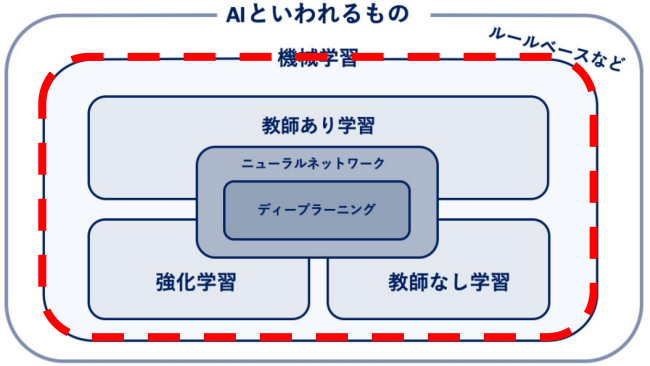
AIという広義の概念の中に、各技術や分野が存在

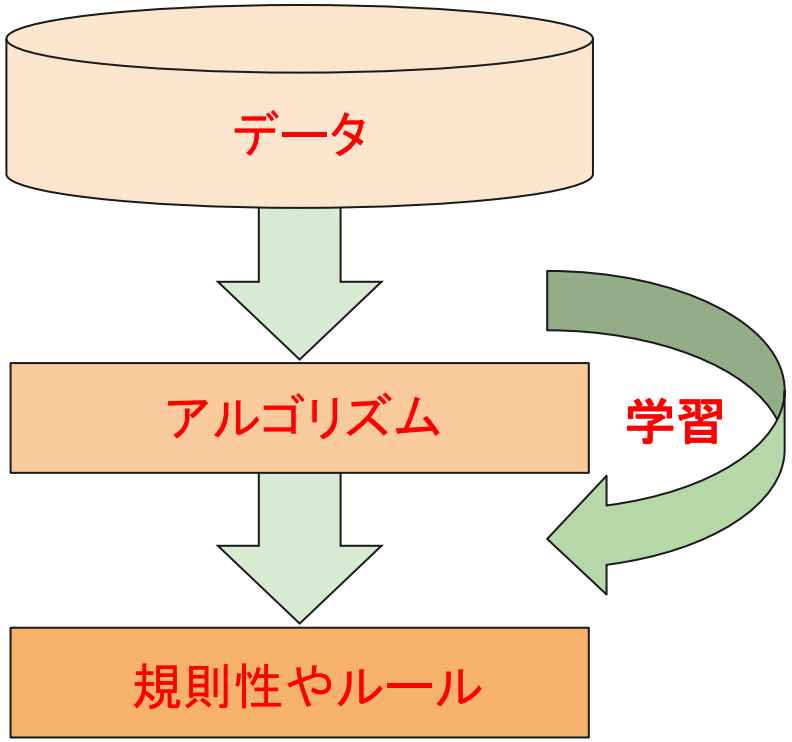


図：機械学習・AI・ディープラーニングの関係性（出典：AINOW）

機械学習とは

機械学習･･･AIという概念に属する技術の1つ



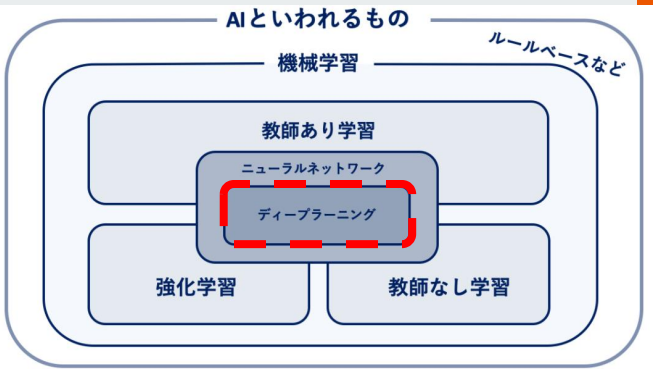


コンピュータが大量のデータを、

様々な分析アルゴリズムを用いて学習し、

そこから規則性やルールを導きだす

ディープラーニングとは①



ディープラーニング･･･機械学習をさらに発展させた技術

■**従来の機械学習**：

データの中のどの要素が結果に影響を及ぼしているのか（特徴量）

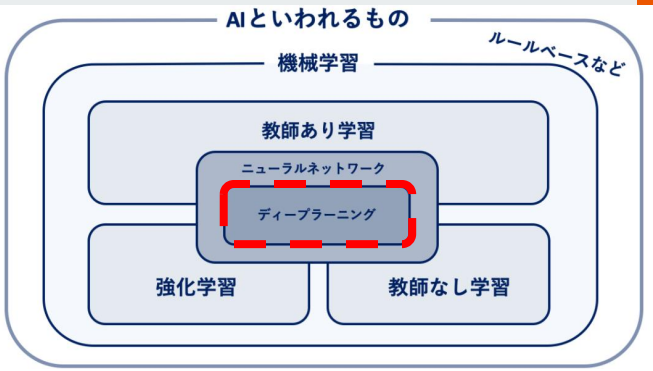
を人が判断、調整する

■**ディープラーニング**：

データの中に存在しているパターンやルールの発見、

特徴量の設定、学習なども機械が自動的に行う

ディープラーニングとは②



●ディープラーニングのメリット…

人が見つけられない特徴を学習できるようになる

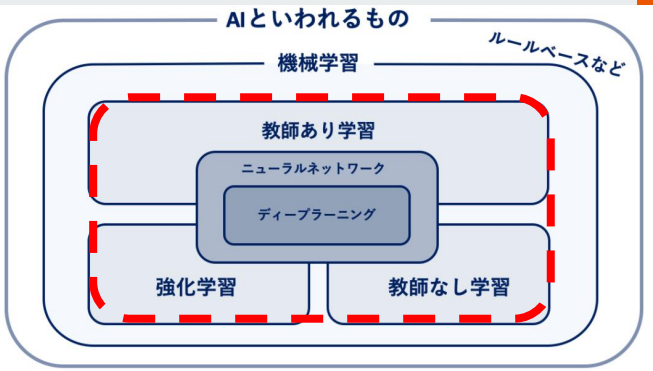
⇒人の認識・判断では限界があった画像認識等の技術が飛躍的に成長

例）犬と猫の判別

犬と猫の違いを人が見れば識別できるが、学習のために特徴を明確に詳細に定義するとなると難しい…

⇒ディープラーニングが人が定義できない特徴まで自動的に抽出してくれる

機械学習の分類

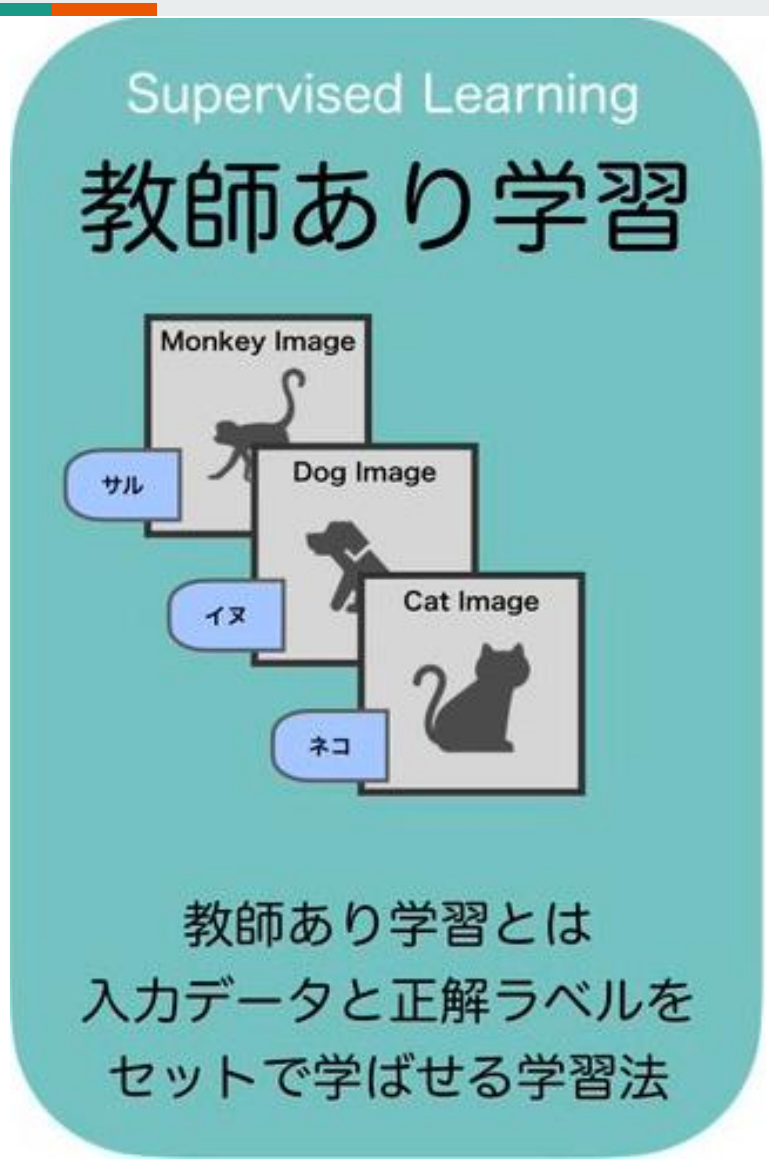


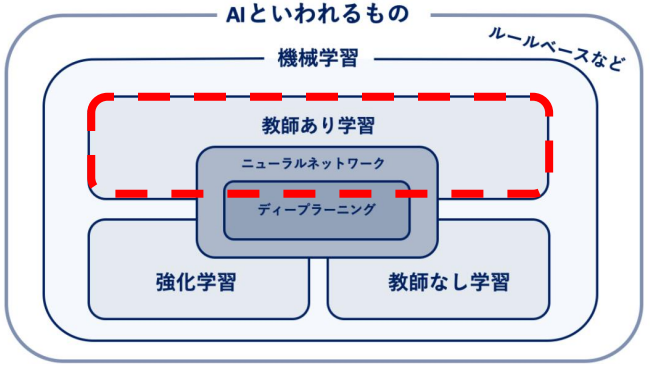
●機械学習の分類

大きく3つに分類される

- 教師あり学習
- 教師なし学習
- 強化学習

教師あり学習





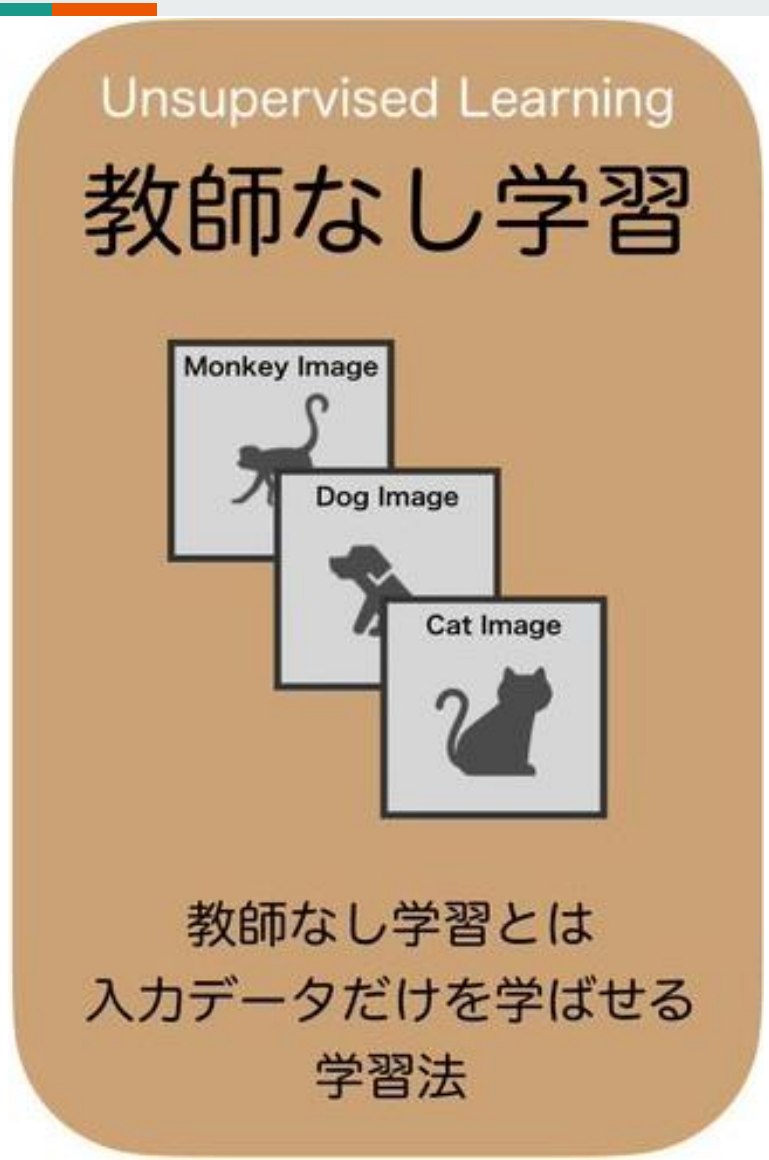
…入力とそれに対する正しい出力が
学習データとしてモデルに与えられ、
モデルは出力を正解に近づけるように学習

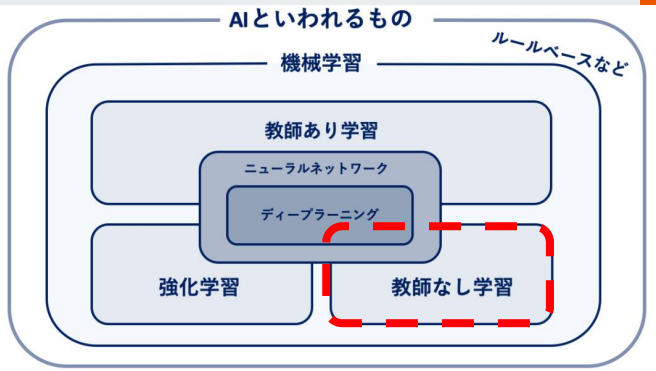
●事例

- 過去の売上から将来の売り上げを予測
- 与えられた動物の画像が何の動物かを識別
- 与えられた英語文章を日本語の文章に翻訳

図：機械学習の学習法（出典：@nekoallllergy.（2020年5月7日）.
https://twitter.com/nekoallllergy/status/1258399783921696768）

教師なし学習



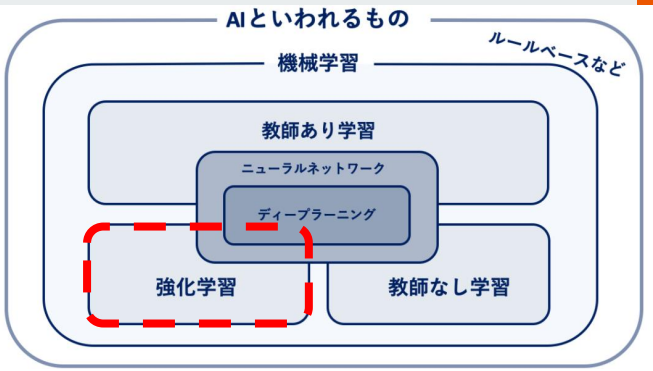


…教師あり学習と違い入力のみが
学習データとしてモデルに与えられ、
モデルは入力の関係性や構造をうまく表現するように学習

●事例
- ECサイトの売り上げデータから、顧客層を分類・認識
- 入力データの各項目間にある関係性の把握
- 不良や異常の検知

図：機械学習の学習法（出典：@nekoallllergy.（2020年5月7日）.
https://twitter.com/nekoallllergy/status/1258399783921696768）

強化学習



…入力のみが学習データとして与えられ、
出力は与えられないが、出力がどれだけ良いかはわかり、
利益を最大化するように自律的に学習

●事例
- 自動車の自動運転
- コンピュータ囲碁プログラム
- 広告配信の最適化

図：機械学習の学習法（出典：@nekoallllergy.（2020年5月7日）.
https://twitter.com/nekoallllergy/status/1258399783921696768）

アルゴリズムの種類と概要

当分科会で調査したアルゴリズムは以下表の通り

	アルゴリズム	
教師あり学習	ランダムフォレスト	決定木
	重回帰	予測
	ラグ特徴量	時系列
	LSTM	時系列（DL）
教師なし学習	アソシエーション分析	関連性
	クラスタリング	分類分け

Agenda

1. 研究概要と目的
2. 研究対象
3. データ収集技法の検証内容
4. データ収集ガイドライン
5. 結論

アプローチ方法

1)自己検証

データと機械アルゴリズムを用いてデータ分析を実施し、
データ収集およびデータ分析のポイントを実証もしくは導き出す

2)実例調査

AI・機械学習を用いた企業のビジネス事例を参考に、
データ収集・データ分析のノウハウを蓄積し
活用パターンと方法を調査し整理

各アルゴリズムにおける調査方法

1)教師あり学習

「自己検証」をベースとして検証

サンプルデータを用意して予測値と正解値の誤差の精度を確認

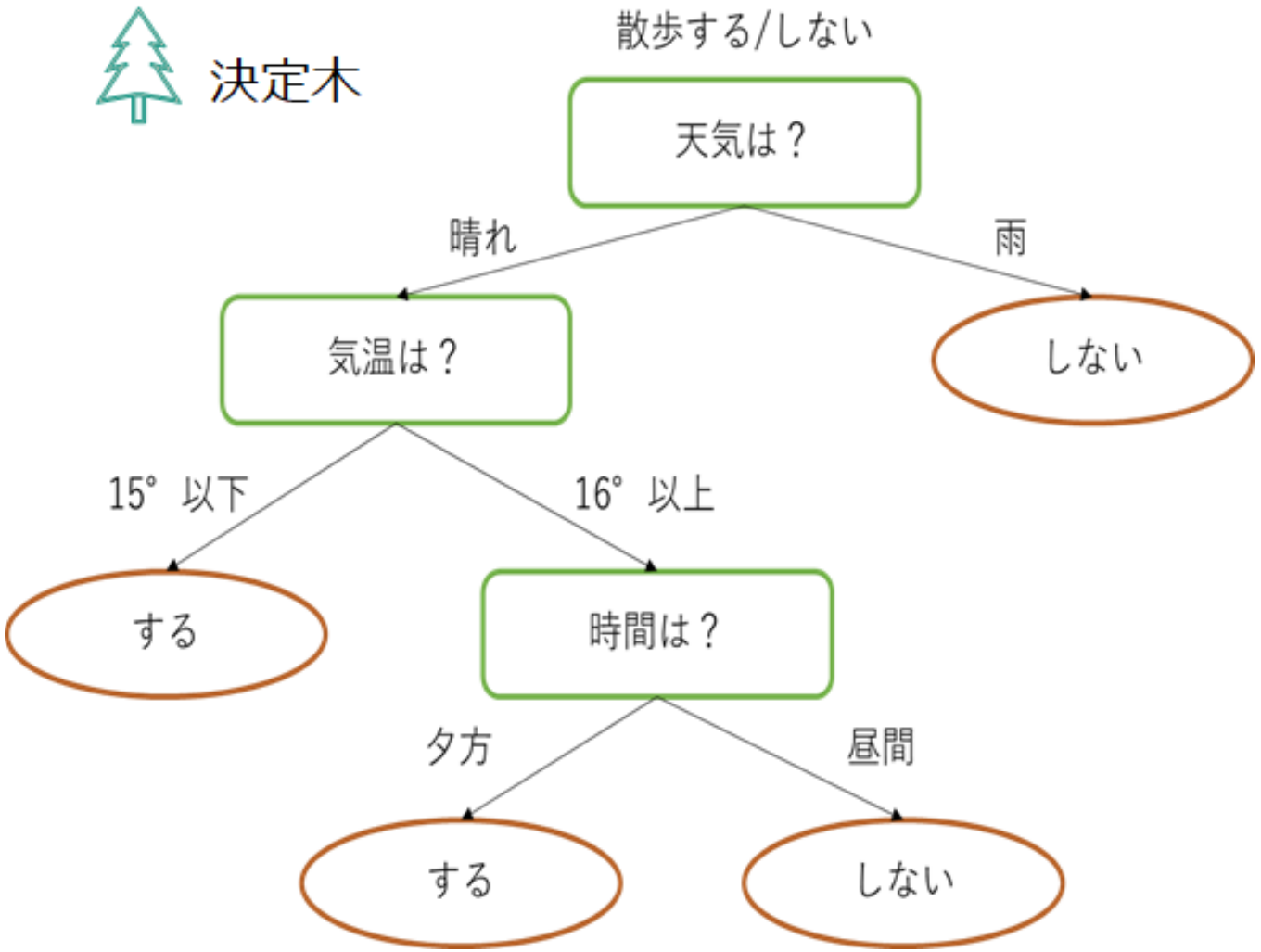
2)教師なし学習

「実例調査」をベースとして検証

アルゴリズムごとに成功事例や失敗事例を収集

利用される分析用途や用意すべきデータセットの情報を整理

ランダムフォレスト



●特徴

- 汎化能力が非常に高く、
  パッケージを用いることで
  容易に実装することが可能

- 決定木を量産、
  統合して結果を出力する

ランダムフォレスト　自己検証（検証方法）

検証方法	Pythonのライブラリであるscikit-learnを使用
データ	職場での欠勤理由データ、タイタニック乗船者の生存データ、 ワイン分類データ、QCMセンサーアルコールデータ、 SPECT画像の診断データ

目的変数と説明変数の相関係数による学習結果への影響を確認

目的変数と説明変数の相関係数が高いデータの方が
分類の正答率は高いという傾向が確認できた

ランダムフォレスト　自己検証（検証結果）

①各データセットの正答率、相関係数の比較

データセット名	正答率	目的変数と説明変数の相関係数の平均値	目的変数と説明変数の相関係数の絶対値の最大値	説明変数同士の相関係数の平均値	クラス数	説明変数の数
職場での欠勤理由データ	**0.416**	0.313	0.55	0.121	28	19
タイタニック乗船者の生存データ	**0.829**	0.213	0.54	0.179	2	7
ワインデータ	0.977	0.493	0.85	0.319	3	10
QCMセンサーアルコールデータ	0.906	0.421	0.71	0.525	5	15
SPECT画像の診断データ	0.657	0.271	0.54	0.298	2	44

②データセットをフィルタした際の正答率、相関係数の比較

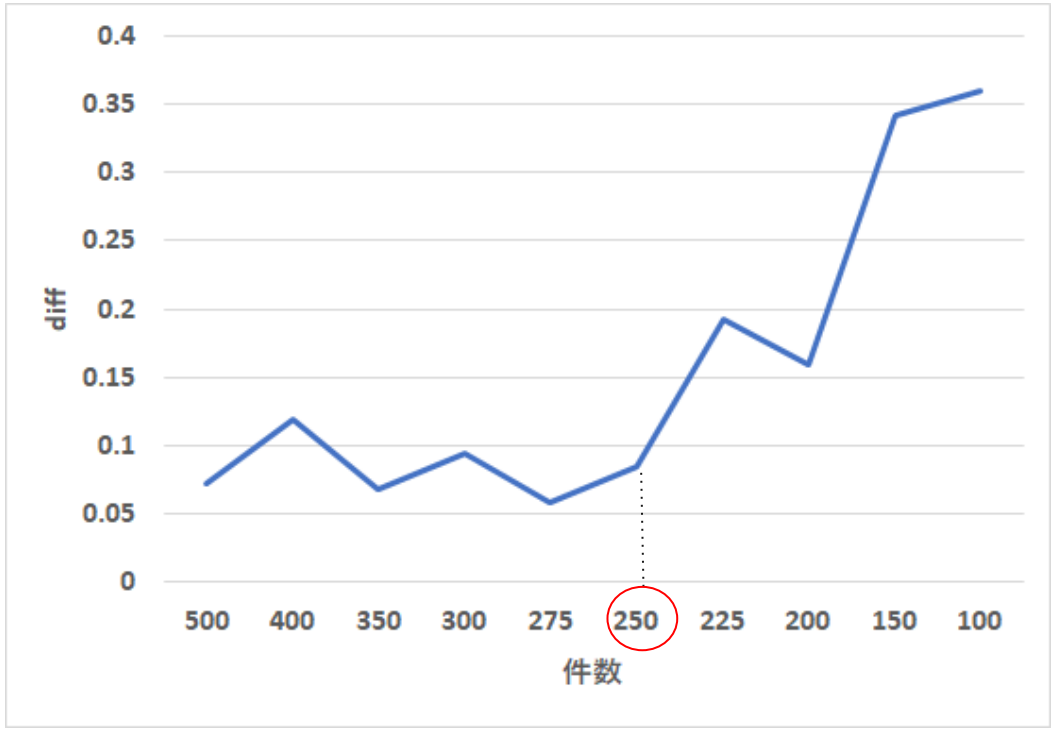
データセット名	正答率	目的変数と説明変数の相関係数の平均値	目的変数と説明変数の相関係数の絶対値の最大値	説明変数同士の相関係数の平均値	クラス数	説明変数の数
職場での欠勤理由データ	**0.308**	0.409	0.55	0.121	28	7
タイタニック乗船者の生存データ	**0.811**	0.327	0.54	0.179	2	4

→ 相関係数の低い説明変数も、正答率に寄与している為、フィルタしない方が良い

ランダムフォレスト　自己検証（検証結果）

③データセットの件数を調整した状態での正答率の振れ幅

職場での欠勤理由データ



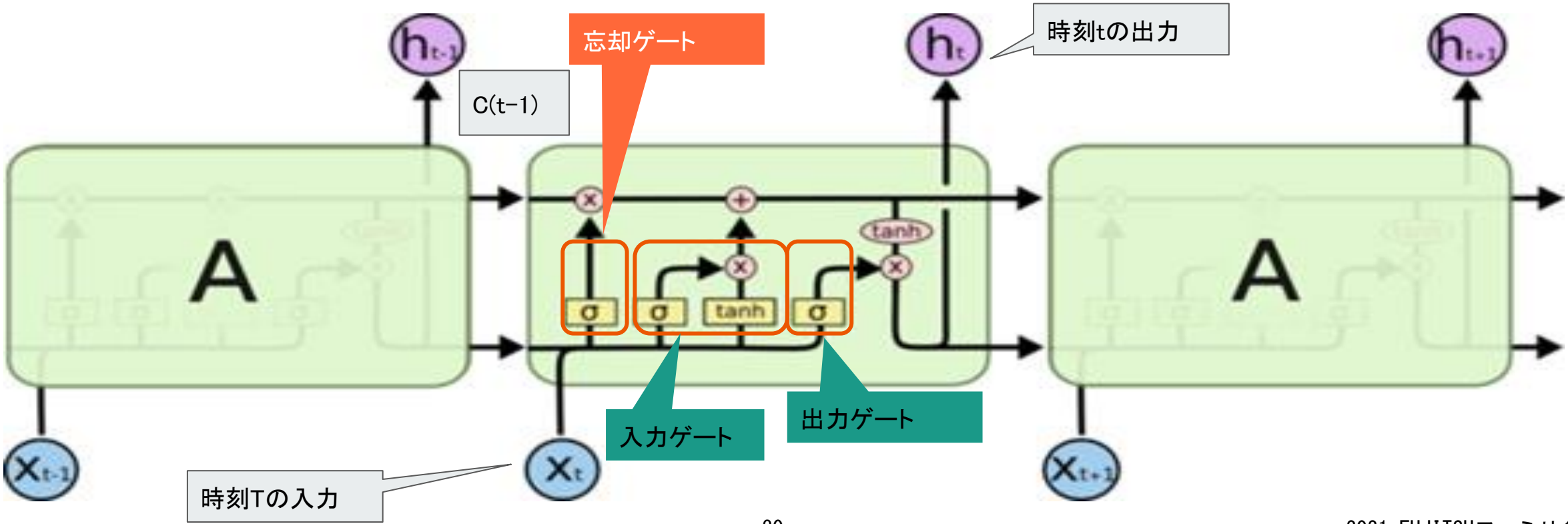
データセット名	件数	抽出データ	正答率	振れ幅
職場での欠勤理由データ	500件	0~500	0.456	0.072
		100~600	0.416	
		200~700	0.488	
	300件	0~300	0.426	0.094
		100~400	0.453	
		200~500	0.52	
		300~600	0.453	
		400~700	0.453	
	100件	0~100	0.56	0.36
		100~200	0.56	
		200~300	0.44	
		300~400	0.48	
		400~500	0.6	
		500~600	0.24	
		600~700	0.44	

データの件数がおおよそ250件以下の場合、正答率の振れ幅が大きくなる傾向を確認

LSTM

・RNN（リカレントニューラルネットワーク）の一種
・CEC（Constant Error Carousel）、入力ゲート、出力ゲート、忘却ゲートを持つ

長期的な依存関係を学習することが可能



ＬＳＴＭ　自己検証（検証方法）

検証方法　Pythonのライブラリであるkerasを使用し、LSTMレイヤーを定義

測定方法　MAE（Mean Absolute Error：平均絶対誤差）

データ　気象庁が公開している2009年～2018年の大阪の1時間毎の気温

パターン　以下の2パターンを検証

1)説明変数の数による精度への影響

過去6時間分、12時間分、24時間分のデータをそれぞれ説明変数とした場合で検証

2)学習データ数の精度への影響

説明変数を過去6時間分の気温データとし、学習用データ数が少ない場合と、多い場合で検証

ＬＳＴＭ　自己検証（結果）

1)説明変数の数による精度への影響

気温データ(時間単位)			
教師データ	2009年1月1日～2018年6月30日		
検証データ	2018年7月1日～2018年12月31日		
比較内容	過去6時間	過去12時間	過去24時間
MAE	0.089175	0.079974	0.068363

2)学習データ数の精度への影響

気温データ(時間単位)		
教師データ	2009年1月1日～2018年6月30日	2018年1月1日～2018年6月30日
検証データ	2018年7月1日～2018年12月31日	
比較内容	教師データ数：多	教師データ数：少
MAE	0.089175	0.116032

ＬＳＴＭ　自己検証（結論）

説明変数の数、学習データが多いほど予測精度が良くなった

特定のデータのみを継続的に取得している場合に有効

アソシエーション分析

データ間の相関関係を発見する際に用いる分析手法

同時購入されている商品の相関関係の分析をバスケット分析という

図：https://boxil.jp/mag/a1565/
アソシエーション分析とは - ツール5選 データの相関関係の見える化より

アソシエーション分析　実例調査（事例）

赤石　雅典、『Pyrhonで儲かるAIをつくる』、日報BP、2020年8月、276-300ページ、978-4296106967

データ：ECサイトの取引履歴データ/商品の購買履歴

1) 商品に対して、①支持度　②確信度　③リフト値を確認する

2) 支持度の閾値を設定して、商品の組み合わせを見比べる

同一商品の色違いや似たような商品が同時に購入されていることがわかった。

Agenda

1. 研究概要と目的
2. 研究対象
3. データ収集技法の検証内容
4. データ収集ガイドライン
5. 結論

データ収集ガイドラインの目的

- 良好なデータ分析結果が得られる
  分析アルゴリズム、ならびにデータ収集技法を提案すること

- AIの利活用分野を整理し、
  データ利活用への参入のハードルを下げること

構成／利用方法

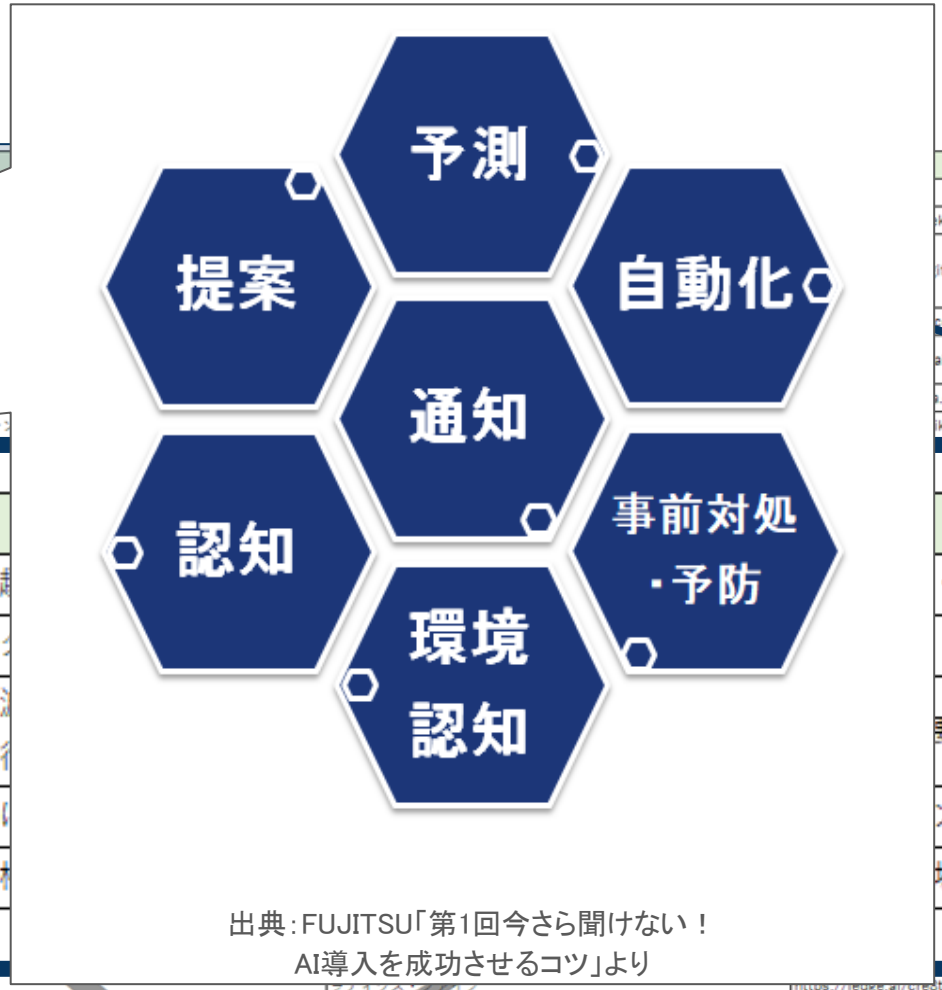
① AI利用分野特定シート

② アルゴリズム選定シート

③ アルゴリズムに適したデータ収集技法シート

①AI利用分野特定シート

No	AI利用分野	事例	データの形式	分析内容
1	予測・提案	マッチングアプリ	年齢・性別・趣	・提案
2	予測	配送ルート予測	スマートメー	
3	予測	タクシー配車予測(交通)	携帯電話の電 タクシーの運	要予測
4	予測	需要予測(アパレル・小売)	ファッション	ンド分析
5	予測	AI農業(生産量調整)	野菜の市場価	場価格



出典:FUJITSU「第1回今さら聞けない！
AI導入を成功させるコツ」より

構成／利用方法

① AI利用分野特定シート

② アルゴリズム選定シート

③ アルゴリズムに適したデータ収集技法シート

②アルゴリズム選定シート

No	分野	ポイント				アルゴリズム	対象シート
		教師データ収集可否	分析対象	メリット	デメリット		
1	提案	否	数値	対象商品が多くても一定時間で計算可能	支持度が低いと計算対象に入らない	アソシエーション分析	アルゴリズムに適したデータ収集技法シート_アソシエーション分析(事例)1
2	提案	否	数値	対象商品が多くても一定時間で計算可能	支持度が低いと計算対象に入らない	アソシエーション分析	アルゴリズムに適したデータ収集技法シート_アソシエーション分析(事例)2
3	提案	否	数値	対象商品が多くても一定時間で計算可能	支持度が低いと計算対象に入らない	アソシエーション分析	アルゴリズムに適したデータ収集技法シート_アソシエーション分析
4	予測	可	数値	非常に汎用性が高いので、様々な分野で使用される、	前処理（データ加工）が必要、	重回帰分析	アルゴリズムに適したデータ収集技法シート_重回帰分析
5	予測	可	数値	時系列データ/連続値データに適する、 分析コストが低い（CPU、処理時間等）	離散データに適さない、 前処理（データ加工）が必要	重回帰分析＋ラグ特徴量	
15	認知	可	数値	説明変数の正規化や標準化が不要 データ量が多くても高速に動作可能	予めクラス分けされた教師データが必要	ランダムフォレスト	アルゴリズムに適したデータ収集技法シート_ランダムフォレスト
16	認知	可	画像	説明変数の正規化や標準化が不要 データ量が多くても高速に動作可能	説明変数は数値データに変換する必要がある 予めクラス分けされた教師データが必要	ランダムフォレスト	アルゴリズムに適したデータ収集技法シート_ランダムフォレスト

構成／利用方法

① AI利用分野特定シート

② アルゴリズム選定シート

③ アルゴリズムに適したデータ収集技法シート

アルゴリズム：重回帰分析

■事例

●検証概要

データセットの説明変数と目的変数の相関性に着目し分析を実施した。また、データセットの件数を調整した場合の影響に関しても分析を実施

●検証方法①

アルゴリズムに重回帰を採用し、データはPythonの機械学習用ライブラリである「scikit-learn」のサンプルデータセットの「ボストンの住宅価格」と「糖尿病の進行状況」を使用した。
説明変数と目的変数の相関性を確認し、全項目を説明変数とした場合、相関性が最も高い3項目のみを説明変数とした場合、
相関性が最も低い3項目のみを説明変数とした場合のMAE（Mean Absolute Error：平均絶対誤差）値による予測精度の比較を行った。
データセットの件数を50件や100件で抽出した状態でMAE値による予測精度の比較も行った。

●検証結果

相関性が高い3項目は相関性が低い3項目より予測精度が良くなることを確認できたが、全項目を説明変数とした場合の予測精度が最も良くなることが見受けられた。
データセットの件数が少ないと用意するデータによってMAE値の幅が大きく異なる傾向が確認できた。（表YYYY）
※比較結果は「説明変数の違いによるMAE値の比較」と「データセットの件数を調整した状態でのMAE値の振れ幅」を参照

●まとめ

データセットの件数が十分であるかを見極める手法の一つとして活用できると考えている。
データセット数の件数を調整してMAE値の振れ幅が大きく変動しない件数を算出することが、データセットの件数が十分であるかを見極める手法の一つとして活用できると考えている。

説明変数の違いによるMAE値の比較

説明変数/データセット	ボストンの住宅価格	糖尿病の進行状況
全項目	3.119781204	45.13216047

構成／利用方法

① AI利用分野特定シート

② アルゴリズム選定シート

③ アルゴリズムに適したデータ収集技法シート

分析アルゴリズム及びデータ集技法の選定
何から取組むべきか分からないという課題の解消

Agenda

1. 研究概要と目的
2. 研究対象
3. データ収集技法の検証内容
4. データ収集ガイドライン
5. 結論

まとめ

研究目的　良好なデータ分析結果が得られるデータ収集技法を提案すること

アルゴリズム活用事例

データ収集・活用事例

→ **データ収集ガイドライン**

研究意義　データ利活用のユーザーである私達自身がAIのスキルを身に付けて、データ利活用を行うことができるようになること

→ **DXの推進**

課題

① データ収集ガイドラインの実ユーザーによる評価

データ収集ガイドラインを実際のユーザーに試用してもらうことができなかった。

② 近年普及が進むアルゴリズムの事例収集

LightGBM等の近年普及が進むアルゴリズムの事例収集が不足した。
今後も新しいアルゴリズムは増えていくため、ガイドラインのアップデートが必要となる。

③ 幅広いデータ事例での検証

分科会内で収集できるデータには限りがある。
さらに幅広いデータを活用した検証が必要となる。

【おまけ】

20年度(コロナ禍)のLS研分科会について

【テレワークの進め方】
- 作業依頼は具体的に内容/範囲を認識合わせしないとズレが生まれやすい
- 特定の人（リーダー）に負荷がかかりやすいため、調整が必要

【各自の感想】
- リモート開催は移動などが無く楽だった。
- 会議中など集中力の維持が難しかった。
- 顔合わせの飲み会などやりたかった。
- 自身のパソコンのみでAI機械学習が実行できる知識を得た。

お礼

ご清聴ありがとうございました！

本研究活動担当幹事で討議への参加や論文の内容について助言を頂きました山内様、LS研事務局及びテクニカルアドバイザーの皆様、分科会開催にあたりご協力いただいた各社関係者の皆様に深く謝辞を申し上げます。

分科会09メンバー　一同