資料1

AIへのデータ利用の状況

平成31年3月13日

事　務　局

日本企業のAIへのデータ利用について考えられる方向性（案）

インターネット業界を代表する海外のデジタル・プラットフォーマーが大量のデータを保有・活用する中で、日本企業がAIにおいてどのようにデータを利用していくのかについては、次の3つの方向性が考えられる。

①　IoTを活かしてリアルデータを最大限収集・活用すること 【P4～5】

②　各社の戦略や公共性を踏まえてオープン化・連携することでデータ量を補完すること 【P6】

③　（直接的な教師データが不要な）「教師なし学習」や「強化学習」に力点を置く等によりデータ不足を補完すること 【P7～12】

AI（機械学習）における学習用データの量と質の重要性

- **一般的に、AIによる分析精度を向上させるには、学習用データの「量」と「質」が重要とされている。**
- **データの「質」については、不明瞭なデータや偏ったデータで学習しても、十分な精度を出すことは難しい。**
- **データの「量」の重要性については、一例ではあるものの、実験により確かめられている。**

学習用データが増えた場合の精度の違いを実験した結果、精度は概ね学習用データ数のlogスケールに比例して向上。

1．学習用データとして、Googleが保有する3億枚のラベル付き画像を活用し、学習に用いるデータ数を変えて学習（モデルを作成）
2．学習後のモデルでオブジェクト検出（object detection）を行った結果を比較（本番用データには2種類のデータセットを活用）

縦軸は平均精度（Average Precision）であり、画像中に存在する物体をモデルが見つけることができた割合、画像中に存在しない物体をモデルは無いとみなすことができた割合等を基に算出。

精度が高い

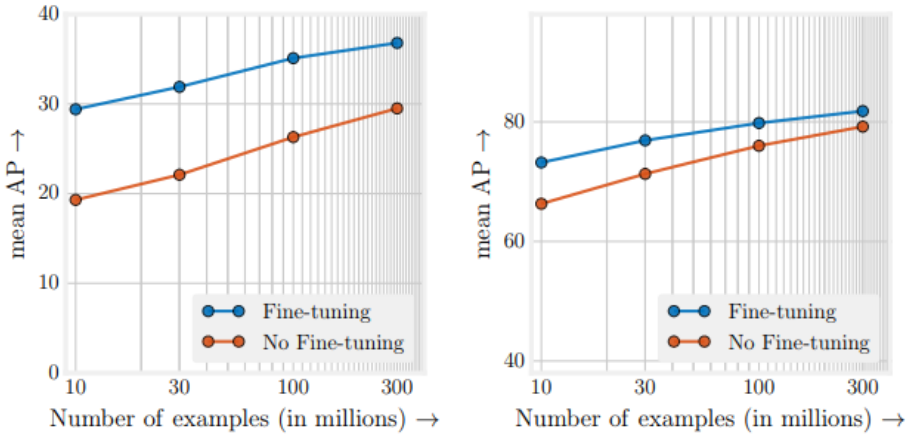


Figure 4. Object detection performance when initial checkpoints are pre-trained on different subsets of JFT-300M from scratch. x-axis is the data size in log-scale, y-axis is the detection performance in mAP@[.5,.95] on COCO minival* (left), and in mAP@.5 on PASCAL VOC 2007 test (right).

Fine-tuning（ファインチューニング）：既存のモデルの一部を再利用して、新しいモデルを構築する手法。

学習用データの量が多い

出典：Chen Sun, Abhinav Shrivastava, Saurabh Singh and Abhinav Gupta ”Revisiting Unreasonable Effectiveness of Data in Deep Learning Era”, 2017.
https://arxiv.org/abs/1707.02968

海外のデジタル・プラットフォーマーの保有するデータの状況

● **インターネット業界を代表する海外のデジタル・プラットフォーマーは、サービス利用者から膨大な量のデータを収集・蓄積・活用し、ビジネスを展開している。**

Google

- Googleでの検索回数は、1日に55億回（1分で380万回）
- YouTubeでは、1日に65年間分（1分で400時間）の映像がアップロード
- Googleフォトでは、1年で13.7ペタバイトの画像がアップロード
- Googleが保有するデータセンターのストレージ容量は10～15エクサバイトという試算も存在

amazon

- Amazonでは、3億人を超える顧客情報を保有し、1年で50億個以上※1の商品を販売

※1　プライム会員（約1億人）のみを対象とした数値

facebook

- Facebookでは、1日に3億5000万枚の画像がアップロードされ、1日に4ペタバイトのデータが生成

Alibaba.com
Global trade starts here.™

- Alibabaでは、6億人を超える顧客情報を保有し、1日で10億個※2の商品を販売

※2 「独身の日」（11月11日）の特別セール

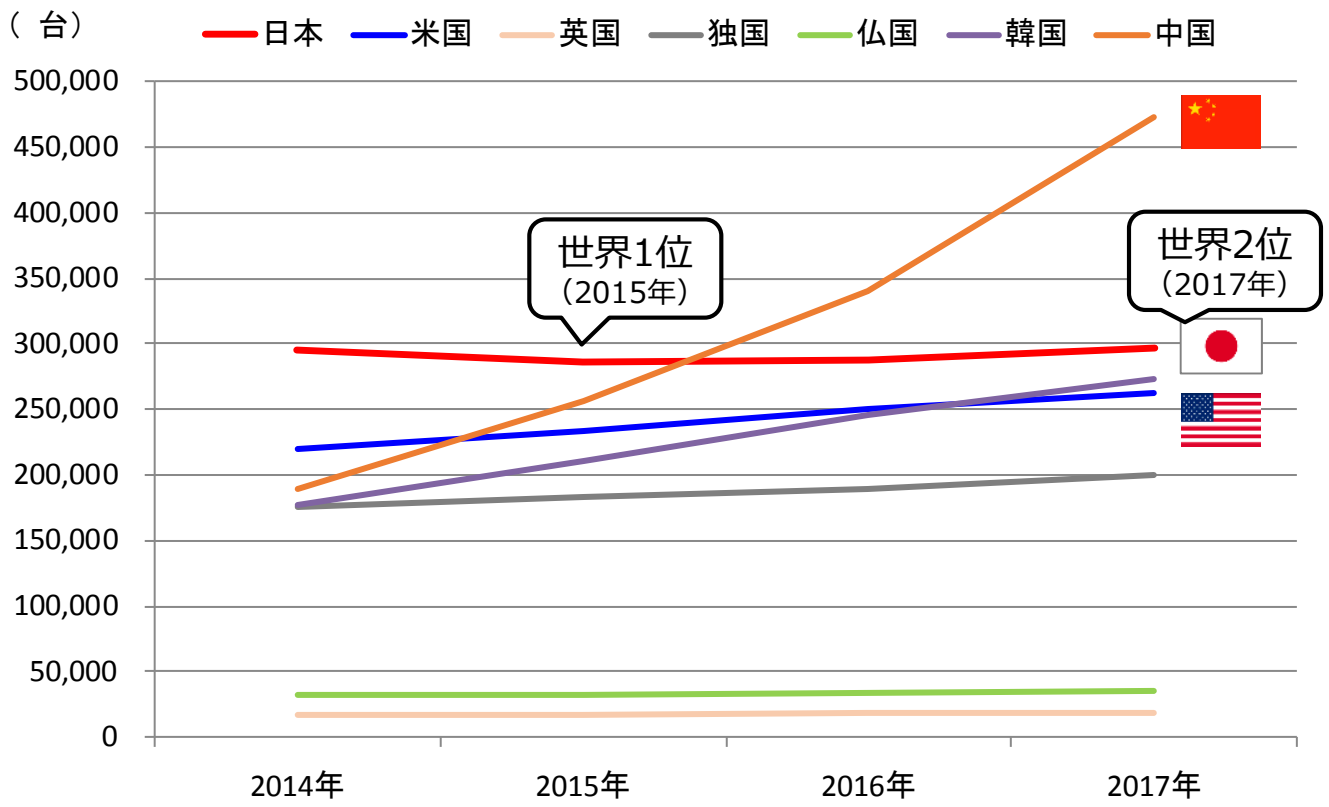
×1000	×1000	×1000	×1000	×1000
KB(キロバイト) →	MB(メガバイト) →	GB(ギガバイト) →	TB(テラバイト) →	PB(ペタバイト) →	EB(エクサバイト)

出典：公開資料より事務局作成

リアルデータに関する日本の可能性

● リアルデータについては、製造現場や医療・ヘルスケア分野を中心に、日本が強みを活かしてデータを収集・蓄積・活用していくことで、AI時代の主役になれる可能性があるのではないか。

産業用ロボットの稼働台数



出典：日本ロボット工業会「世界の産業用ロボット稼働台数推定」を基に事務局作成
※　対象はマニピュレーティングロボットのみ

医療・介護データ

名称	データの規模
NDB レセプト情報・特定健診等情報データベース	■医療レセプト • 1億2000万人 • 約148.1億件(2018年3月) ■特定健診データ • 2400万人 • 約2.3億件(2018年3月)
介護DB	■介護レセプト • 約8.6億件(2018年3月) ■要介護認定情報 • 約5千万件(2018年3月)
NIS Data	■入院患者データ • 700万人
CMS Data	■レセプトデータ • 5000万人 （2012年、健康保険加入者）

出典：厚生労働省「第74回社会保障審議会介護保険部会資料」、松居宏樹「医療ビッグデータ利用の現状と課題」を基に事務局作成

（参考）GAFAによるリアルデータ収集の動向

● GAFAがリアルデータ収集を進めるために進出している産業を、日本の重点産業分野（「未来投資戦略 2018」に記載されているデジタル戦略を進める重点産業の主要分野）別に整理すると、進出数によって3領域に分類される。

	日本の重点産業分野	Google	Apple	facebook	amazon
GAFAの全てが進出済でありGAFA間競争も激戦となっているレッドオーシャン領域（5分野）	医療	○	○	○	○
	介護	○	○	○	○
	AI次世代家電	○	○	○	○
	デジタル・ガバメント	○	○	○	○
	中小企業の生産性革命	○	○	○	○
Facebook社を除く3社が進出しておりレッドオーシャンになりつつある領域（7分野）	自動運転・公共交通のスマート化	○	○		○
	健康	○	○		○
	スマートバイオ	○	○		○
	エネルギー転換・脱炭素化	○	○		○
	Fintech／キャッシュレス	○	○		○
	インフラ管理の高度化	○	○		○
	スマート農林水産業	○	○		○
0～2社しか進出していないブルーオーシャン領域（8分野）	宇宙	売却, △	△	○	○
	AI次世代ロボット	売却, 解散	○		○
	航空機	○			○
	PPP／PFI手法の導入加速	○			○
	スマートシティ	○	△	△	△
	スマートマテリアル	○	△		
	観光・スポーツ・文化芸術	△	△	△	△
	サプライチェーン（製造、卸売、小売）	△	△		△

出典：大和総研「GAFA が異業種分野への進出を加速」を基に事務局作成

データの公開・共有による学習用データの増強

● **個社が保有するデータではデータ量・種類に限界があるため、企業・業種を超えた多様なデータ共有により、社会課題の解決やイノベーションの創出を目指す取組が活発化している。**

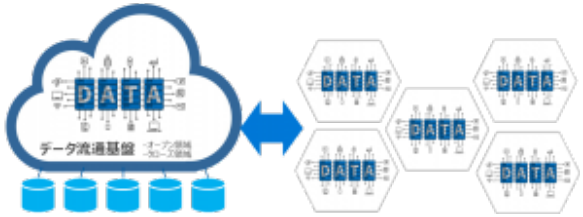
AIデータ活用コンソーシアム

AIの研究と利活用を推進するためにはデータ（特に自然言語や画像などの日本固有のデータ）が重要であり、日本におけるAIの研究と利活用をより一層加速させるべく、以下のような活動を進める。

- 日本固有の自然言語、画像をはじめとする開かれたデータの流通の場の提供
- データ流通基盤の社会、企業における実装および活用の促進
- AIによるオープンイノベーションを通じて社会課題の解決を促進

等

◆発足日：2019年3月6日（水）

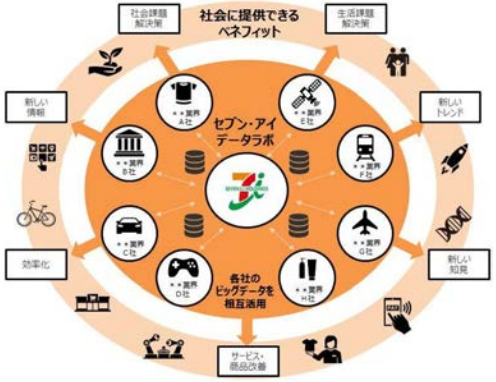


出典：AIデータ活用コンソーシアム
http://www.aidatacon.com/2019/03/06/established/

セブン&アイ・データラボ

幅広い業界の参加企業それぞれが保有する豊富な統計データから得られる知見を相互活用し、そこから生じる新たな知見によって生活課題や社会課題の解決を目指す。

◆開 始 日：2018年6月1日（金）
◆参加企業：ANA ホールディングス株式会社、株式会社 NTTドコモ、株式会社ディー・エヌ・エー、東京急行電鉄株式会社、東京電力エナジーパートナー株式会社、株式会社三井住友フィナンシャルグループ、三井物産株式会社 等 10社（当初）

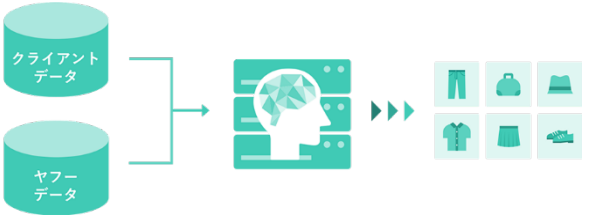


出典：セブン&アイ・ホールディングス
https://www.7andi.com/library/dbps_data/_material_/_files/000/000/003/005/2018060101.pdf

ヤフー「データフォレスト構想」

ヤフーのデータから得られる様々なインサイトを、ヤフー社内で活用するだけでなく、ヤフーと企業、ヤフーと自治体、また、企業間や自治体間など、参画するプレイヤーがデータを相互利活用することで、それぞれが成長し、さらに多くのデータが集まるエコシステムを目指す。

◆開 始 日：2018年2月8日（木）
◆参加企業：約20社の企業と商品開発、需要予測などのテーマで実証実験を進めている。

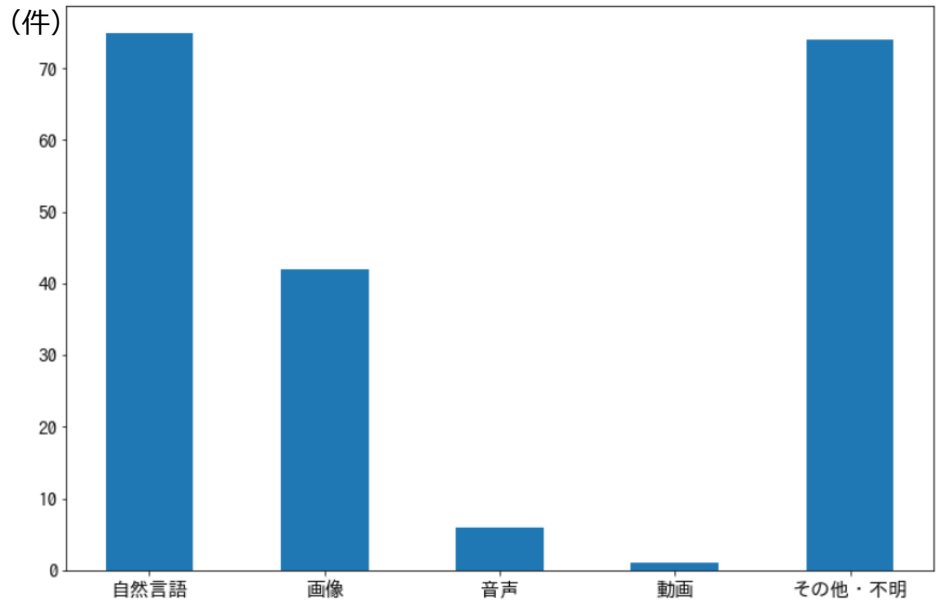


出典：ヤフー
https://dataforest.yahoo.co.jp/

日本のAIにおける学習用データの状況

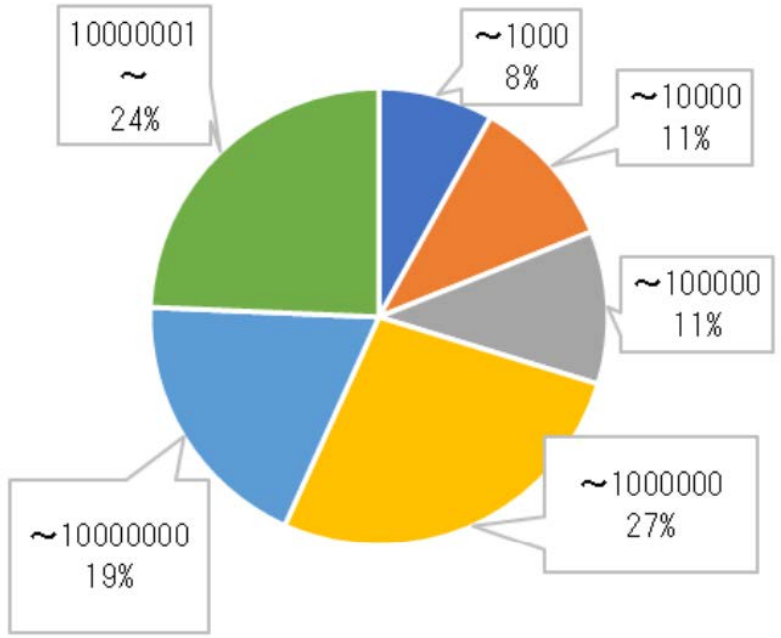
- 限定的な事例をベースとしているものの、日本において、AIを活用した商品サービスで活用されているデータは自然言語（テキスト）や画像が多く、学習用データの量については、10,000件以下の事例が約2割を占めているとの調査結果がある。
- AI時代には、一般にデータを大量に持つプレイヤーが優位と考えられているものの、ニッチなフィールドから収集されたスモールデータでビジネス展開していくストーリーも考えられるのではないか。

AI関連商品サービスに使われるデータの種別



※ 事例収集期間・方法： 2016年1月1日～2017年8月31日の間にPR TIMESにおいて配信されたニュースリリースの中で人工知能でヒットするものを収集。収集件数は183件。

AI関連商品サービスで適用されるデータ量



※ データ量が分かった37事例のみで集計。

出典：Innovation Nippon「人工知能と日本2017」
http://www.glocom.ac.jp/news/3246

AIにおける学習方法と用途

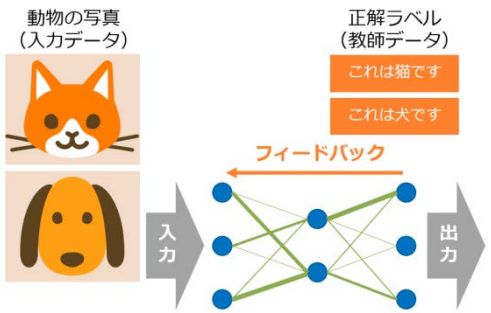
- **AIにおける学習方法には、大別して「教師あり学習」、「教師なし学習」、「強化学習」の3つがあり、それぞれに適した用途がある。**
- **必ずしも単一の学習方法を用いる必要はなく、複数の学習方法を組み合わせて学習させることもある。**

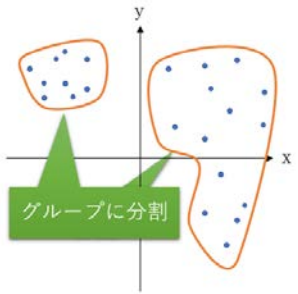
AIにおける代表的な学習方法

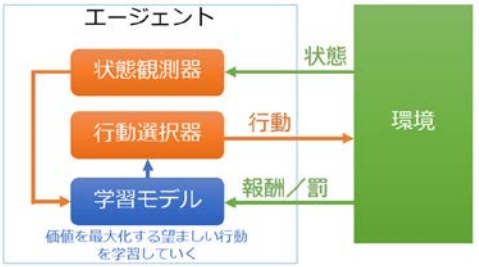
名称	内容
教師あり学習	• 教師あり学習では、正解データ（目的変数）を含むデータセットを入力として利用する。 • 目的変数を除く残りのデータ（説明変数）から得られる出力結果に着目する。その値ができるだけ正解に近くなるような特徴量を探し出して、モデルを作成する。 • 最後に、正解データを持たないデータセット（新規データ）に作成したモデルを適用し、予測結果を得る。
教師なし学習	• 教師なし学習では、正解データ（目的変数）を含まず、説明変数だけのデータセットを入力値として用いる。 • データセット全体から特徴量を抽出してモデルを作成する。 • 最後に、出力結果はただの数値でしかないため、ラベルを付けて意味のある情報にする。
強化学習	• 強化学習では、与えられた環境下でエージェントが最大の勝ちを得るための行動を学習する。 • 例えば、コンピュータゲームを考えると、ゲームを行うプレイヤ（エージェント）は現在の打ち手（状態）からゲームをクリア（報酬）するために、どのような打ち手（行動）を取ればよいかを学習する。

出典：インターフェース（2017年12月）等より事務局作成

具体的なイメージ







出典：Deep Insider

用途の例

- 売上予測、需要予測、株価予測
- 不正検知、故障検知
- 画像分類
- 顧客維持

- レコメンド
- 顧客セグメンテーション
- ターゲットマーケティング

- ゲーム
- 広告最適化
- 自動運転車
- ロボット制御

各学習方法において必要となるデータ

● 「教師あり学習」においてはラベル付きデータが必要である等、各学習方法で必要となるデータの質・量は異なる。

学習方法と必要となるデータ

学習方法	必要となるデータ
教師あり学習	ラベル付きデータ
半教師あり学習	少 ラベル付きデータ ／ 多 ラベルなしデータ
教師なし学習	ラベルなしデータ
強化学習	自ら作成したデータ

パーセプトロン、アソシアトロン、サポート・ベクタ・マシン、決定木、k近傍法、線形回帰、折りたたみニューラル・ネットワーク、再帰型ニューラル・ネットワーク

自己組織化マップ、ボルツマン・マシン、主成分分析、クラスタ分析、オート・エンコーダ、敵対的生成ネットワーク

Qラーニング、Deep Q Network、Sarsa

画像認識や音声認識でよく用いられる折りたたみニューラルネットワーク（CNN：Convolutional Neural Network）、自然言語処理で用いられる再帰型ニューラル・ネットワーク（RNN：Recurrent Neural Network）といったアルゴリズムがある。

少量のラベル付きデータを活用することで、「教師なし学習」よりも効率的に学習することができる。

学習データのノイズ除去等を行う際に活用される「オート・エンコーダ」、生成器と識別器が競いながら学習する「敵対的生成ネットワーク」といったアルゴリズムがある。

与えられた正解の出力をそのまま学習すれば良いわけではなく、より広い意味での「価値」を最大化する行動を学習する。つまり、その瞬間ではなく、長期的な視点で学習するため、何回も試行錯誤が必要になる。

学習用データ作成の流れ

学習用データの作成には、一定の作業が必要

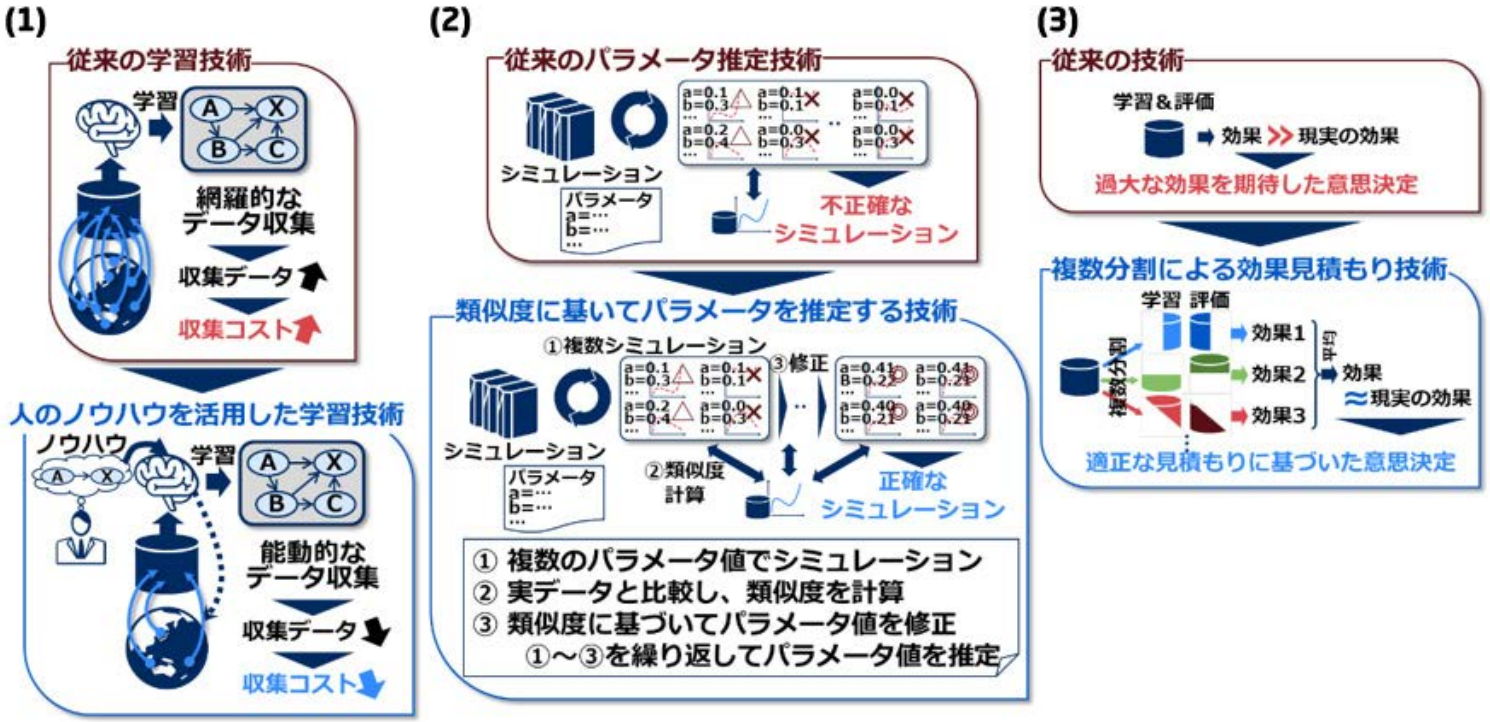
データ → クレンジング（ノイズ除去等） → ラベルなしデータ → ラベル（正解データ）付与 → ラベル付きデータ

出典：各種資料等より事務局作成

（参考）学習用データの不足を補うための技術の例①

■学習効率の高いデータを能動的に収集して学習し、推定精度を向上させる技術

1. 人のノウハウを取り入れて、学習効率の高いデータを能動的に収集して学習する技術
2. 収集したデータをもとに、実世界の事象の複数のシミュレーション結果の類似度に基づいてパラメータの修正を自動で繰り返し、正しいパラメータを推定する技術
3. AIの分析結果に基づく意思決定時に、収集データを学習用と効果評価用に分割した複数パターンで効果を見積もり、少数データの偏りに影響されにくい意思決定を可能にする技術

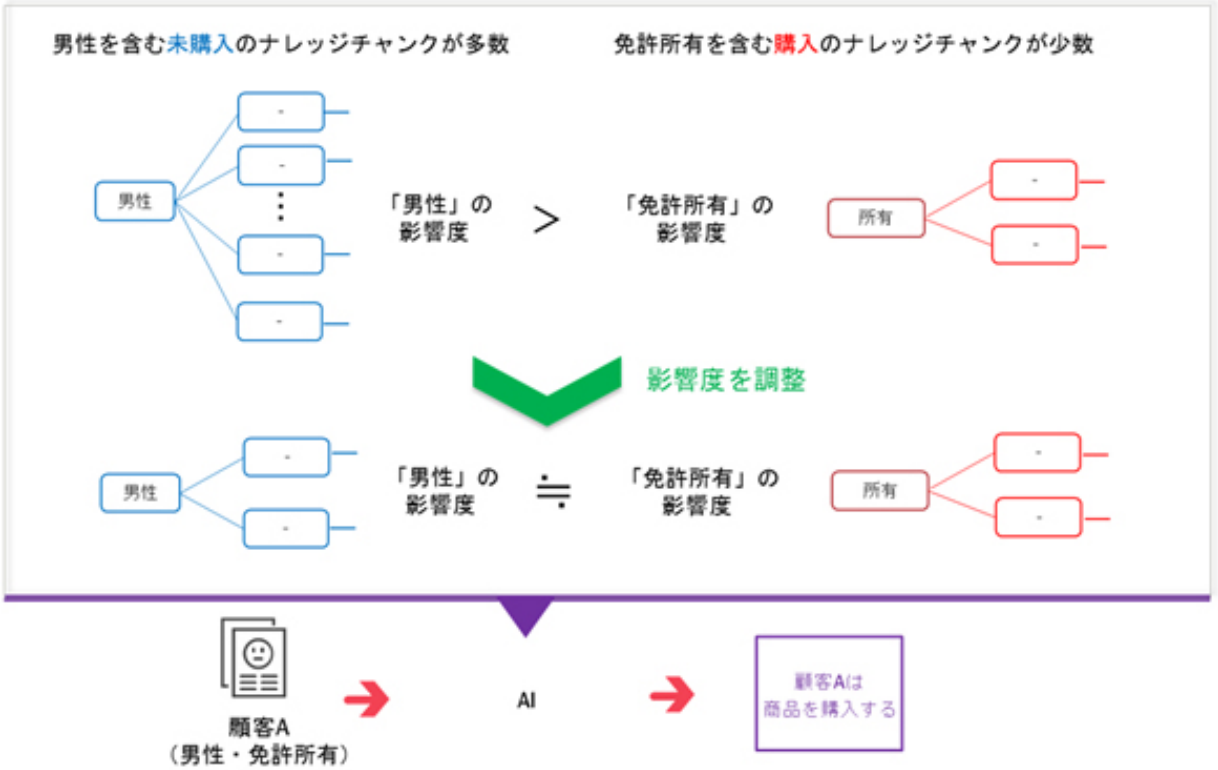


出典：NEC
https://jpn.nec.com/press/201807/20180710_03.html

（参考）学習用データの不足を補うための技術の例②

■学習用データの量が少ない場合や偏りがある場合にも、高精度な判断を可能とする技術

1. データの項目どうしをすべて組み合わせ、その大量の組合せを仮説として重要度の高いものを選別
2. 仮説を構成する項目の重複関係に基づいてそれぞれの影響度を制御することで、どの仮説に対しても均等に学習することができ、データに偏りがある場合でも従来よりも高精度な判断を下すことが可能。また、仮説は論理的な表現で記述されているため、人間にも判断理由を理解することも可能。
3. 本技術により、判断したい対象のデータが少ない医療やマーケティングなどの現場でもAIを活用できるようになる。



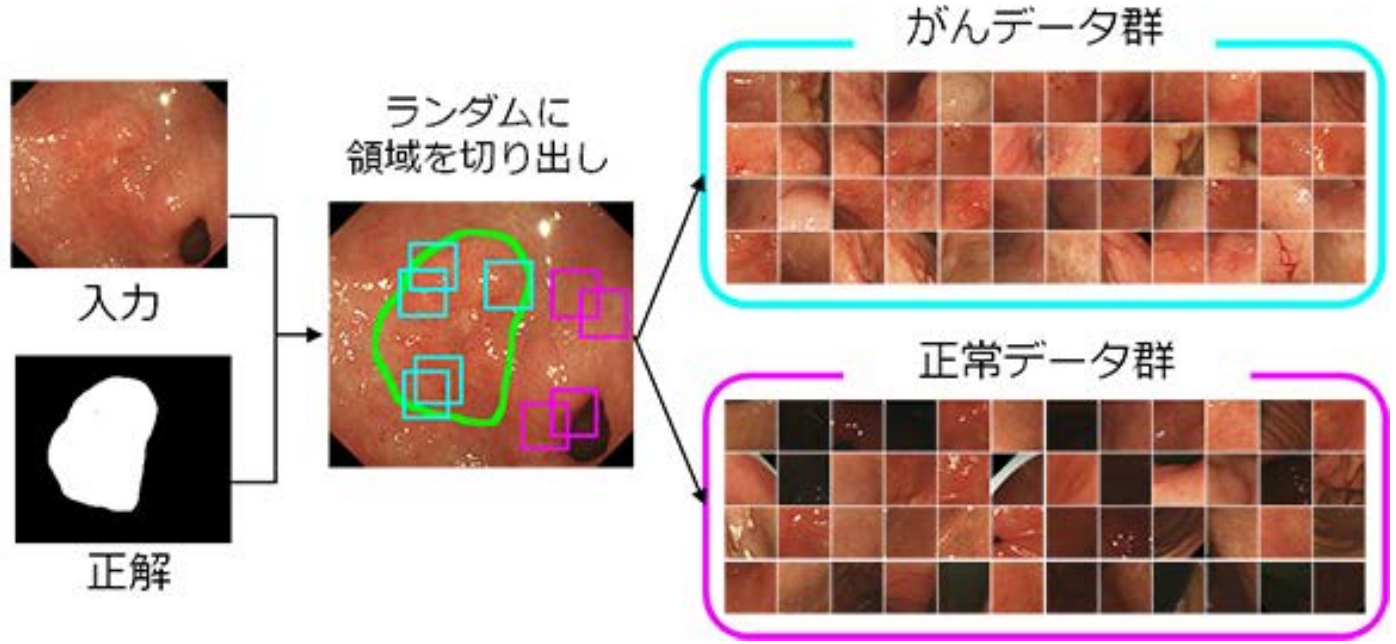
出典：富士通研究所
http://pr.fujitsu.com/jp/news/2018/09/19-2.html

（参考）学習用データの不足を補うための技術の例③

■少数の正解データにより構築されたAIによる早期胃がんの高精度な自動検出法

ディープラーニングを画像中の物体検出へ応用する場合、一般には数十～数百万枚の正解画像が学習用データとして必要であるが、早期胃がんの場合、良質の正解画像を大量に収集することは困難であり、以下のアプローチによりこの問題を解決。

1. 早期胃がんの領域を正解として与えた正解画像と正常画像の計約200枚を学習用データとして用意
2. 「がんの部分」と「正常の部分」を確実に含む領域をランダムにそれぞれ約1万枚切り出し、合わせて約2万枚の画像を取得
3. データ拡張技術（拡大縮小、反転、回転シフト、色変換など）を利用して画像を約36万枚まで拡大
4. その画像を学習させた結果、陽性的中率（コンピュータが「がん」と判断した画像中、実際に「がん」であった割合）は93.4%、陰性的中率（コンピュータが「正常」と判断した画像中、実際に「正常」であった割合）は83.6%と高い確率となった。



出典：理化学研究所
http://www.riken.jp/pr/press/2018/20180721_2/

AIにおけるデータ活用プロセス

- **AIの開発・利用に当たっては、「学習用データ」と「本番用データ」を（必要な場合加工した上で）活用する。**

基本的に
本番用データ
とは別のデータ

①学習用データの収集
- 社内蓄積データ（既存）の活用
- 社内で新たにデータを収集して活用
- 外部データ（有料/無料）の活用

②学習用データの加工（前処理）
- クレンジング（ノイズ除去等）
- データ拡張（左右反転等によりデータを増やす）

③モデルの作成（AI）
- 自社開発（すべて）
- 自社開発（一部は既存のフレームワーク等を活用）
- 委託開発
- サービスを利用

開発はせず、提供されているサービス/アプリケーションを利用する場合は①～③を省略できる。

本番用データを活用した分析（④～⑥）を通じてAI（モデル）を改善する。

④本番用データの収集
- 社内蓄積データ（既存）の活用
- 社内で新たにデータを収集して活用
- 外部データ（有料/無料）の活用

⑤本番用データ加工（前処理）
- クレンジング（ノイズ除去等）

⑥AIによる分析
- 分析の実行（③のモデル）

⑦結果の活用
- 業務等の自動化・効率化
- 既存業務・サービスの高度化
- 新サービスの開発

出典：公開資料等より事務局作成

（参考）学習用データ①

- **AIにデータは必要不可欠であるが、研究開発等を目的としたデータ・セット（データ集）が無料で公開されており、これらを学習用データとして活用してAIモデルを作成することも可能となっている。**

カテゴリ	データ提供者名	データ名	データ内容
動画	Google	YouTube-8M Dataset	• 700万件以上（45万時間）の動画に関する情報を公開。
	Google	YouTube-BoundingBoxes Dataset	• 24万件の動画を公開。
	Deepmind	Kinetics	• 30万件のYouTube動画に、400種類に分類された人間のアクションがラベリングされたデータを公開。
	University of Central Florida	UCF101 - Action Recognition Data Set	• 人間の行動101件が分類されたラベリングされた約13,000の動画を公開。
	Google	AVA	• 動画の中に人間の複数の行動に関するラベリングが付与されている動画を公開。
	twentybn	20BN-JESTER DATASET V1	• ハンドジェスチャーのラベルが付与された動画データセットを約15万件公開（twentybnはドイツベースのベンチャ）
	MIT/IBM	Moments in Time Dataset	• 3秒の動画にアクションラベル（ラベル数は約340件）が付与されており、公開件数は100万件。
画像	Yann LeCun氏他	MNIST	• 手書き文字の数字「0～9」に正解ラベルが付与されたデータセットを公開。
	University of Tronto	CIFAR-10	• 10種のラベリングが付与された、6万件の画像を公開。
	Zalando	Fashion-MNIST	• ファッションに関する画像6万件を公開（テストデータも1万件公開）。
	Computer Vision Laboratory	Food 101	• 101のラベルが付与された、約10万件の食品画像を公開。
	University of Washington	MegaFace	• 顔認識アルゴリズムのコンテストを実施し、67.2万人分、470万枚の画像を公開。

出典：codexa「【24個掲載】機械学習で使えるデータセット一挙勢揃い！」、
他公開情報より事務局作成

（参考）学習用データ②

カテゴリ	データ提供者名	データ名	データ内容
画像	The Chinese University of Hong Kong	CelebA Dataset	• 40のラベルが付与された、20万人以上の世界中の有名人の顔の画像を公開。
	United States Department of Defense	The FERET Database	• 1,199名を異なる角度で撮影した画像を約11,000公開。
	Qiong Cao氏他	VGGFace2 Dataset	• 9,131名分の331万に及ぶ顔のデータを公開。
	NIH	NIH Chest X-ray Dataset of 14 Common Thorax Disease Categories	• 14の胸部疾患に分類分けされた3万人の肺のレントゲン写真11万件のデータを公開。
その他	Amazon	Public Data Sets	• 地理空間データ（衛星画像）、環境データ（気象画像）、ゲノム、Webデータ等複数データを公開。
	Microsoft	Azure ML datasets	• Azure ML（クラウドでAI機能を提供するサービス）で利用可能なデータセットを公開。
	DataMarket	DataMarket	• 為替レート、人口推移、魚の漁獲量等の時系列のデータセットを公開。

出典：codexa「【24個掲載】機械学習で使えるデータセット一挙勢揃い！」、
他公開情報より事務局作成