

# [ Data-Centric / HITL ]を用いた AI開発の実践例

株式会社APTO  
高品 良



## 会社概要

AI開発の80%以上を占めるデータの収集、  
アノテーションに特化したサービスを提供。

## サービス

- 1) アノテーションツール提供
- 2) アノテーション受託/AI開発受託
- 3) データセット販売事業
- 4) ノーコード/ローコードプラットフォーム提供

## 会社情報

会社名 株式会社APTO  
設立 2020年1月20日  
代表者 高品 良  
拠点 東京都渋谷区神南1-5-14 三船ビル403



## - Investors



## - Programs & Award



## - Media Coverage



## - Collaboration

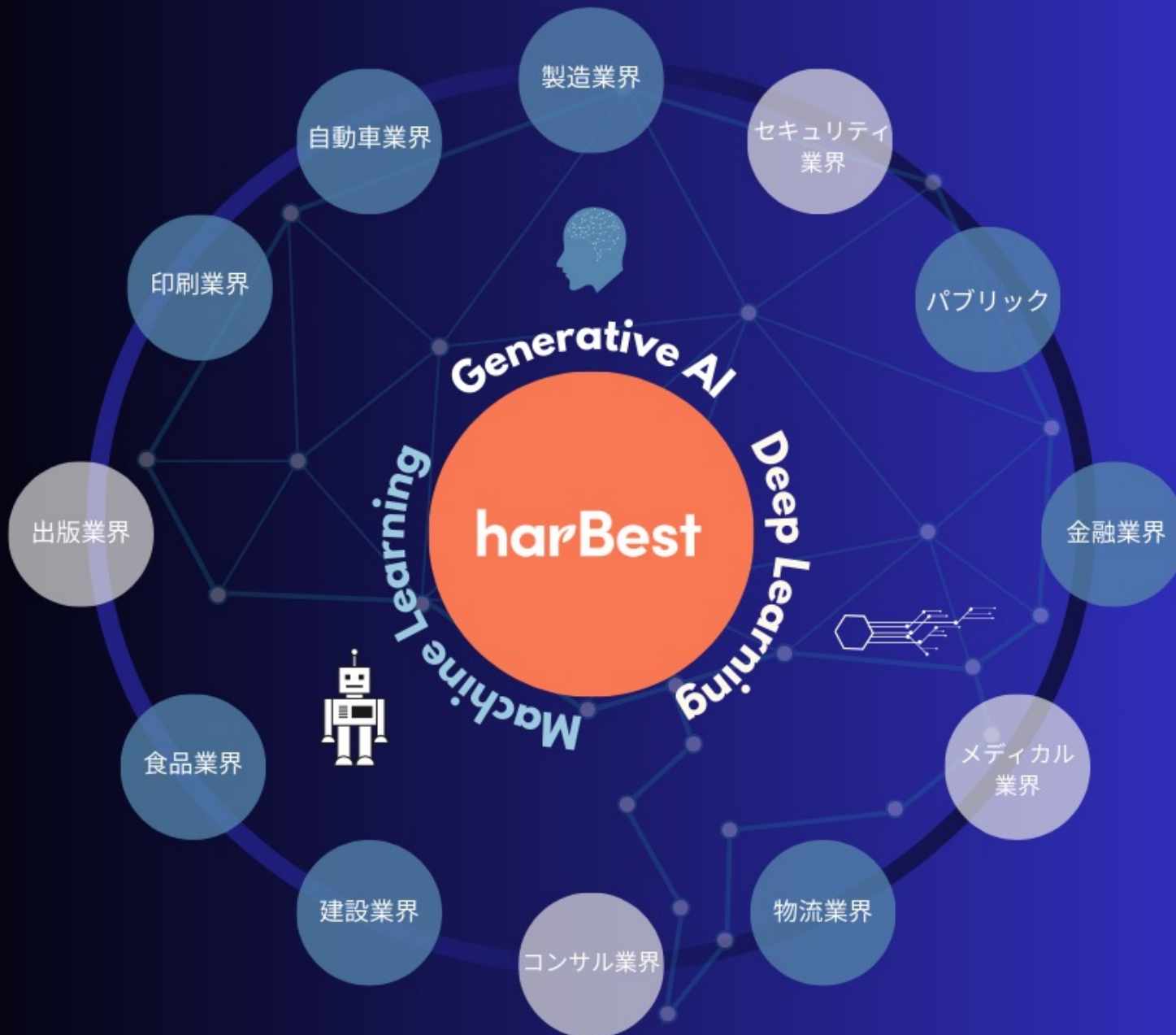


## AI Solution

日本を代表する企業に対して  
AI開発の課題を解決し、  
日本をAI先進国へ

## HITL System

全国にいるクラウドワーカー/  
エキスパートワーカーによって  
リアルタイムでデータ収集、  
アノテーションが可能





株式会社APTO 代表取締役CEO

## 高品 良

### <学歴>

大学で経営工学を専攻（プログラミングに触れる）

### <職歴>

2013年4月：インフラ/バックエンドエンジニア（大規模機関係システム開発）

2015年10月：IT案件にエンジニアとして参画（ITコンサル会社）

2017年11月：VRコンテンツ制作会社を設立  
（VR物件、エンタメコンテンツ制作）

2020年1月：APTOを創業

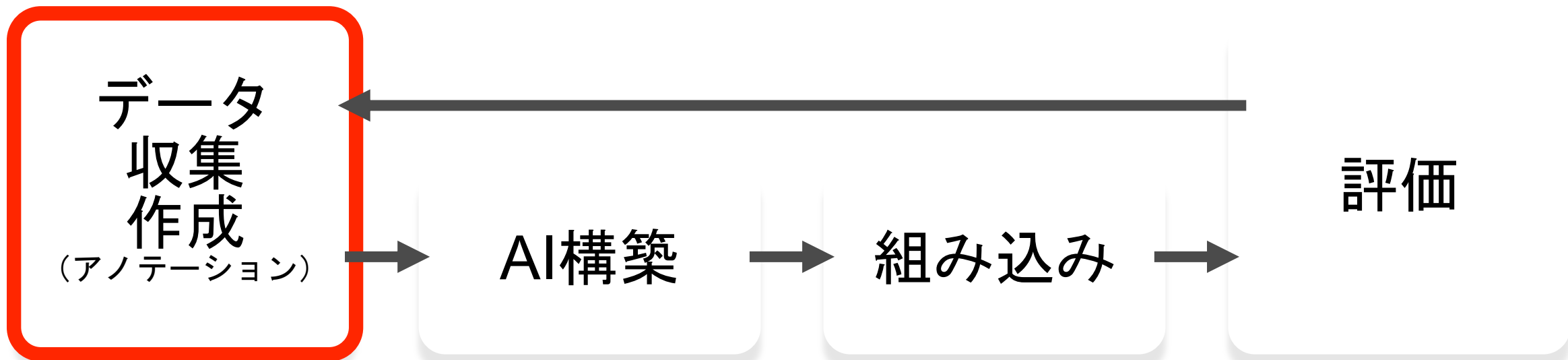
AI開発に興味をもち、自動炎上検知サービスを開発しようとしたが、精度があがらず頓挫（データ不足が課題）

AI開発にはデータが非常に重要だと気づき、データに特化したAPTOを創業

2020年12月：プロダクト開発を本格化（シードの資金調達を完了）

2021年10月：harBest（データ収集、データ作成プラットフォーム）をローンチ





AI開発の

80%

出典 : MLOps: From Model-centric to Data-centric AI  
著者 : Andrew Ng

AI開発・データ活用をしたいけど・・・

自社にデータあるけど  
とても整理  
しきれない・・・



作業をしたいが人員が  
足りない・・・



肝心なデータ  
がない・・・



意外な現状！！

- ・ 高給AIエンジニアがアノテーション作業
- ・ 専門知識不要な作業も多い現実

# harBest



# harBest







- 1) アノテーションチーム管理
- 2) アノテーション品質管理
- 3) メジャーなデータフォーマットへの対応

月額3万円～

## ■ダッシュボード

ダッシュボード/プロジェクト一覧 > プロジェクト一覧

### プロジェクト詳細

プロジェクトを編集 事前テストを編集 データをダウンロード

プロジェクトサマリー アノテーションデータ一覧

データセットアップロード待ち アップロード中 開始待ち 実施中 ポイント付身確認中 完了

#### プロジェクトサマリー

ねちこやんを認識するアノテーション ねちこやんを認識するアノテーション

作成日 2023/8/23 猫を認識 API Keyを発行

猫の一般概念をねこ、ねこやん、ぬっこなどに拡張するためのラベリング作業をおこなってまいります。ねこやん、ねこちゃんなどはエラーパターンとなります。

#### アノテーションの条件

作業タイプ: 画像アノテーション  
ワーカー: クラウドワーカー 男性 30-35歳  
ラベル: 猫 犬 鳥  
1セット毎の作業回数: 5件  
重複させる作業回数: 5回  
作業者: 3名 [メンバーを見る](#)

#### アノテーション作業の進捗

2023/8/30 プロジェクト期限

完了率: 100.0%  
完了件数: 10,000 / 10,000  
重複完了件数: 10,000 / 10,000  
予算: 5,000 pt

## ■アノテーションツール

harBest

HOME  
プロジェクト

### プロジェクト一覧

#	プロジェクト名	プロジェクトID	実施期限	登録日	詳細
1	対象物を四角形で囲む作業	01GP5W410MVPQ0MT03EQPQZ3	2024-02-24	2023-01-07	<a href="#">開始する</a>
2	対象物を多角形で囲む作業	01GP5W65DQP6A664QF09Z8ACC9	2024-02-23	2023-01-07	<a href="#">開始する</a>
3	対象物を塗りつぶす作業	01GWNW1WX3452ETKWDS0NS1WEN	2024-05-04	2023-03-29	<a href="#">開始する</a>

2023 株式会社APTO

利用規約 プライバシーポリシー 特定商取引に基づく表示

案件をAPTOがコントロールし、クラウドワーカーに作業依頼を行ないます。  
データ品質、機密情報、作業者への報酬支払いは当社が責任をもって管理致します。

※プロジェクト管理主体者はAPTO



## LLM開発に必要な権利クリアなデータセットを提供

No.	Information	Question	Answer harBest	Category
1	抗生物質感受性試験は、特定の微生物が特定の抗生物質に対して感受性を持っているかどうかを決定する実験的な方法です。この試験は、最も適した抗生物質の治療を選定するために使用されます。通常、このテストは細菌の株を抗生物質に曝露し、その成長を抑制するかどうかを観察	抗生物質感受性試験の主な目的は何ですか？	最も適した抗生物質の治療を選定する目的です。	医学・微生物学
2	経皮毒とは、皮膚を通じて体内に取り込まれる毒物のことです。これは農薬や一部の化学物質、薬品が該当します。皮膚が薄い部位や傷口に接触すると、経皮毒の吸収率が高まる可能性があります。このような毒物は、皮膚を通して血流に直接入ることができ、体内で様々な影響を及ぼ	経皮毒の吸収率が高まる可能性がある皮膚の状態は何ですか？	経皮毒の吸収率が高まる可能性がある皮膚の状態は、皮膚が薄い部位や傷口です。	医学・毒学
3	市場セグメンテーションはマーケティング戦略の一環で、市場をより小さな類似のグループまたはセグメントに分割するプロセスを指します。このアプローチにより企業は特定のセグメントに特化した製品やサービスを提供し、顧客満足度を高めることができます。また、市場セグメンテーションは広告やプロモーションの効果を高めるのにも使用されます	市場セグメンテーションで利用されるセグメンテーションの基準をいくつか挙げてください。	市場セグメンテーションで利用されるセグメンテーションの基準は、地理的、人口統計学的、心理的、行動的な要因です。	マーケティング
4	土地利用計画は、土地の資源を効率的かつ持続可能な方法で管理と整備を行うための総合的なプランニングです。これには、住宅、商業、産業、農業、自然保護区など、多様な用途に対応する必要があります。土地利用計画は、法的規制、地域社会のニーズ、環境保全など多くの要素を	土地利用計画が考慮すべき多くの要素には何が含まれていますか？	法的規制、地域社会のニーズ、環境保全が含まれています。	都市計画・環境科学
5	連立方程式は、複数の未知数を含む複数の方程式からなる問題であり、それらの方程式がすべて満たされる未知数の値を見つけ出すことが目的です。解法にはグラフ法、代数法、行列法などがあり、数学の他、物理学や工学、経済学でも用いられます。連立方程式の理論は高校教育でも	連立方程式を解くための方法にはどのようなものがありますか？	グラフ法、代数法、行列法などです。	数学
6	ホームオスタシスは、生物の体内環境が一定の範囲内で安定して維持されるメカニズムを指します。これには、体温、pH、塩分濃度、血糖値などが含まれます。ホームオスタシスは、ホルモン分泌、神経伝達、代	ホームオスタシスが維持する体内環境の要素には何が含まれますか？	ホームオスタシスが維持する体内環境の要素には体温、pH、塩分濃度、血糖値が含まれる。	生物学
7	情報倫理は、情報技術と人間活動が交錯する領域での倫理的問題を研究する学問です。これには、プライバシー保護、著作権、データセキュリティなどが含まれます。情報倫理は、急速なテクノロジーの進展によって新たな課題が常に出現するため、非常に動的なフィールドです。適切	情報倫理が研究する倫理的問題にはどのようなものがありますか？	情報倫理が研究する倫理的問題にはプライバシー保護、著作権、データセキュリティなどが含まれます	情報科学・倫理学
8	生化学的酸素要求量 (BOD) は、ある水体に含まれる有機物質の量を示す指標として利用されます。有機物質が分解する過程で酸素が消費されるため、BODが高いということはその水体に多くの有機物質が存在していることを意味します。BODは、一般に5日間のインキュベーション期間を用いて測定され、単位はmg/Lとして示されます。水質汚染の	BODが高いということは何を意味し、その値はどのようにして測定されますか？	その水体に多くの有機物質が存在しており、単位はmg/Lとして示されます。	環境科学



## 現場で収集困難な異常データのバーチャル作成

### ■鉄板のサビを仮想的に再現



バーチャル上で作成されたデータを、学習用データ、評価用データとして利用することで精度向上を実現

## 製造業/メーカー

**RICOH**

**muRata**  
INNOVATOR IN ELECTRONICS

**Panasonic**

**MITSUBISHI ELECTRIC**

**FUJITSU**

**Canon**

## 情報・ITサービス

**SCSK**

**RevComm**

al+

**LIGHTBLUE TECHNOLOGY**

**mcs**  
マイクロコントロールシステムズ株式会社

**AMBL**

## 生成系AI

**stability.ai**

**RIMEN 理化学研究所**

## 商社

**RYODEN**

**HARADA**

## 学術機関

**東京大学**  
THE UNIVERSITY OF TOKYO



## 警備

**CSP** セントラル警備保障

## 化粧品

**hoyu**

## 金融

**ORIX**

## EC

**PROTO**

## 印刷

**TOPPAN FORMS**

## 小売

あなたと、コンビニ、  
**FamilyMart**

## マーケティング

**CCC**  
MARKETING

## 食品

**Aj**  
AJINOMOTO

## ハウスメーカー

**SEKISUI HOUSE**

## 衛星

**Solafune**

## 農業

**AGRIST**

## 植物

**Green Snap**

## 鉄道

**JR**  
鉄道総研

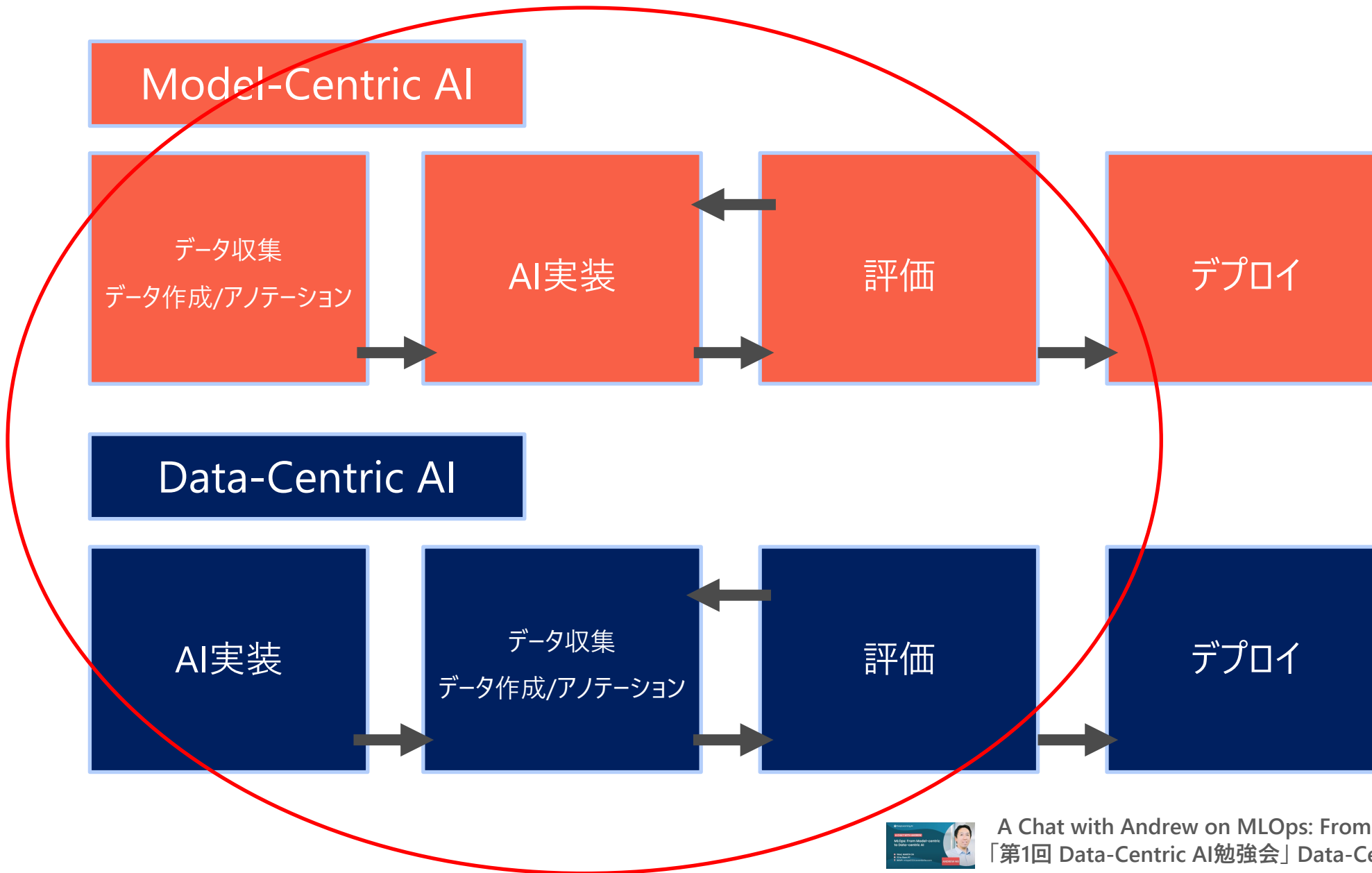
## Model-Centric AI

- ・データ収集を行い、データに含まれるノイズに耐えられるようなモデルを開発する
- ・データを固定し、「コード/モデル」を繰り返し改良していく

## Data-Centric AI

- ・データの一貫性を最重要として、アノテーションツールなどを使って、データの品質を改善する
- ・コード/モデルを固定し、元となる「データ」を繰り返し改良していく







# Model-Centric vs. Data-Centric

	鉄製品の欠陥検査	ソーラーパネルの欠陥検査	表面検査
Baseline	76.2%	75.68%	85.05%
Model-Centric	76.2% (+0%)	75.72% (+0.04%)	85.05% (+0%)
Data-Centric	<b>93.1% (+16.9%)</b>	<b>78.74% (+3.06%)</b>	<b>85.45% (+0.4%)</b>



・外観検査プロジェクトにおいて、ベースライン方式をどれだけ改善できるか「Model-Centric」と「Data-Centric」を比較

・Data-Centricでは大きな改善に成功。また、別の実験では新規でデータを追加するよりもノイズデータを削除した方が大幅に精度改善につながったとする報告も。



A Chat with Andrew on MLOps: From Model-centric to Data-centric AI  
「第1回 Data-Centric AI勉強会」Data-Centric AI Community, YouTube

# Data-Centric Approach



アノテーター-A



アノテーター-B



アノテーター-C



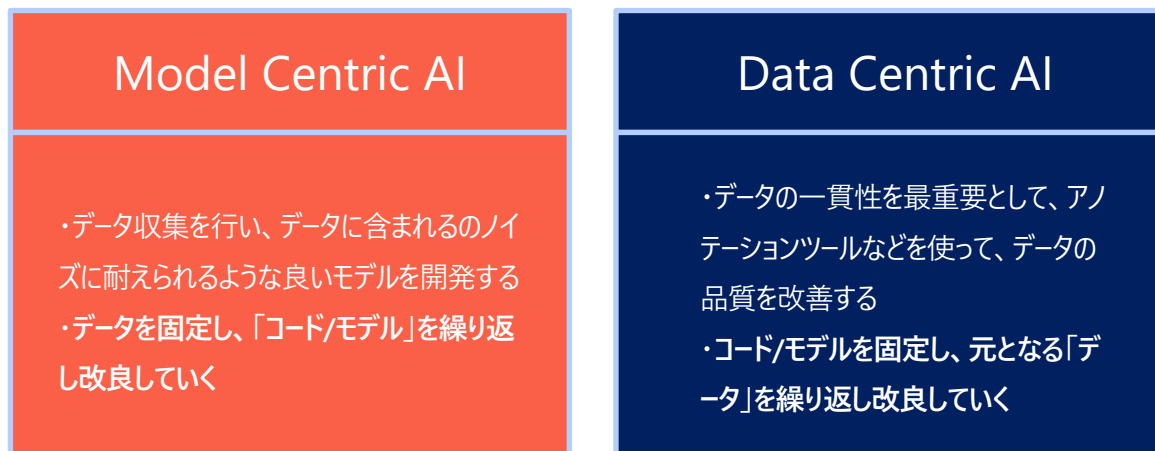
- ・アノテーターによるラベルの偏りを限りなく減らした、アノテーション規則が一貫したデータセットの構築を目指す
- ・一貫性の欠如など、データ上の問題の発見、解決をシステマティックに行う必要がある

## Big Data → Good Data

- ・定義が一貫している（曖昧性のないラベルを定義）
- ・重要なケースをカバーしている（レアケースなど入力を十分にカバー）
- ・本番データからのタイムリーなフィードバック（データ分布がデータドリフトやコンセプトドリフトをカバーしている）
- ・サイズが適切である



## Summary (Prerequisite Concepts)



・MLOpsの最重要タスクは、機械学習プロジェクトのライフサイクル全体を通じて、高品質なデータを保証すること

・今後はData-CentricなAI開発を継続的に、効率的に、システムティックに実現するためのデータ作成/アノテーションツールが重要になる





## Data-Centric + HITL

### Data Centric AI

- ・データの一貫性を最重要として、アノテーションツールなどを使って、データの品質を改善する
- ・コード/モデルを固定し、元となる「データ」を繰り返し改良していく

・MLOpsの最重要タスクは、機械学習プロジェクトのライフサイクル全体を通じて、高品質なデータを保証すること

・今後はData-CentricなAI開発を継続的に、効率的に、システムティックに実現するためのデータ作成/アノテーションツールが重要になる



「じゃあ、どうやって高品質なデータを一定量収集するの？」

「収集したはいいけど、どうやって高品質なラベリング/アノテーションをすればいいの？」

## [ 人 / Human-in-the-Loop ]

## Human-in-the-Loopとは

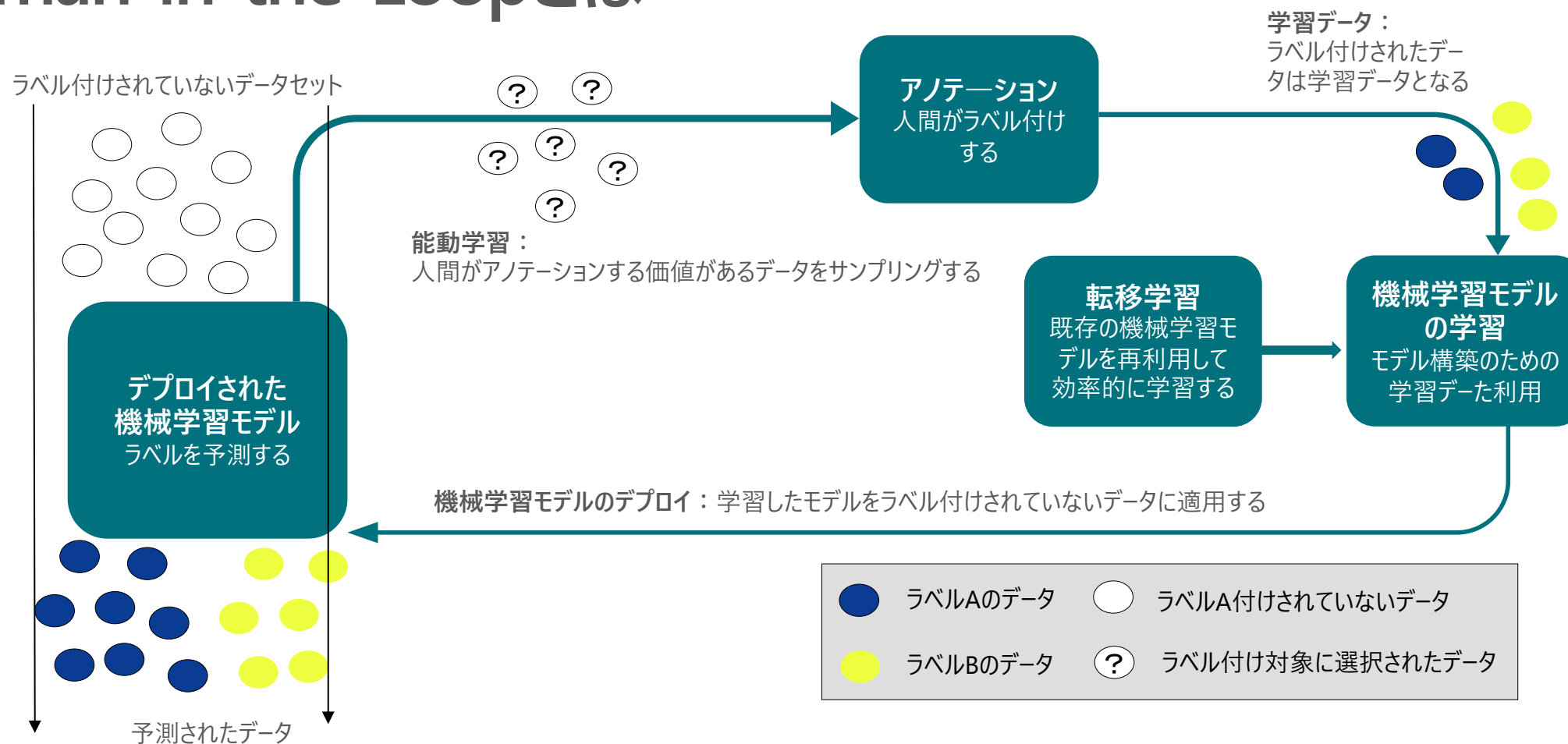
機械学習モデルと人間が相互補完しながら動作するシステムを意味します。機械学習を利用するアプリケーションにおいて、人間と機械の知能を融合するための一連の戦略であり...

早い話が、

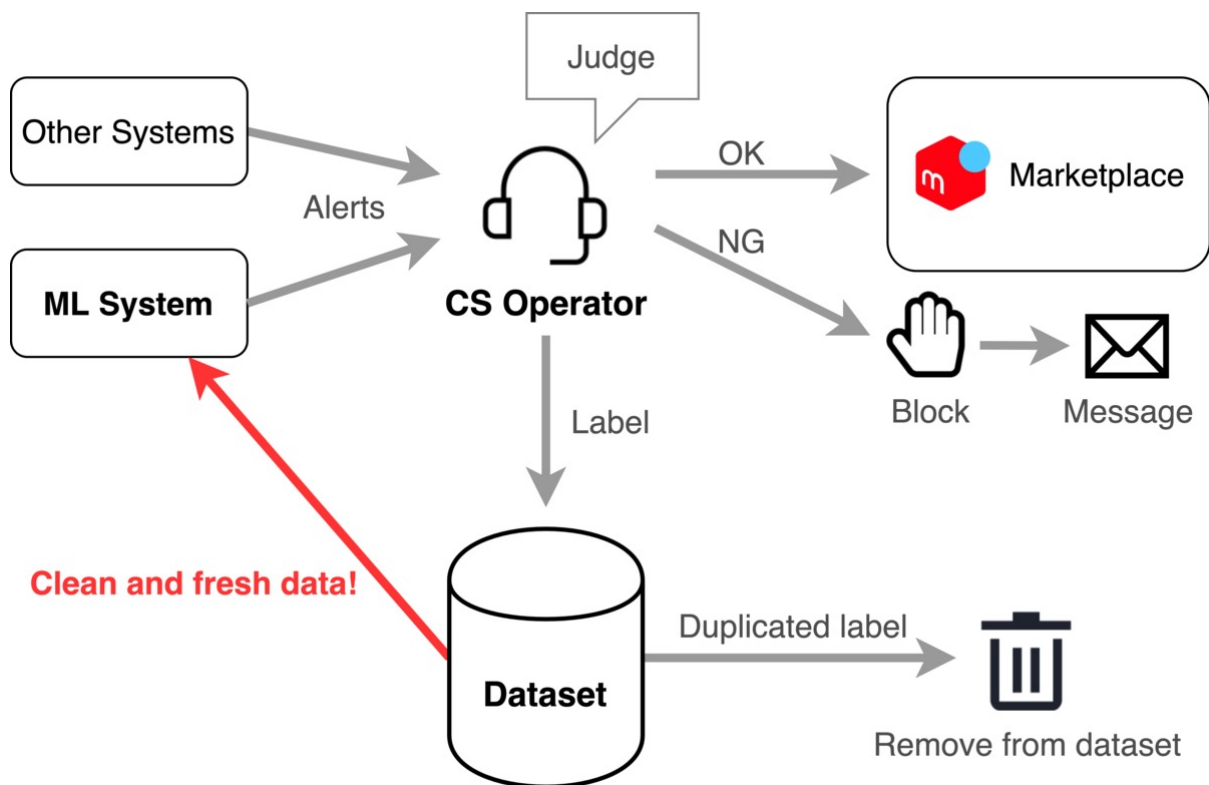
- モデルの精度向上
- 機械学習モデルが目標とする精度に到達するまでの時間を短縮する
- 人間と機械の知能を融合し、モデルの精度を最大化する
- 機械学習を用いて人間の作業効率を上げる

以上を目的とした戦略のこと

## Human-in-the-Loopとは



# メルカリ × Human-in-the-Loop



## 違反出品検知システムにおけるHITL

### <前提>

- ・ 違法出品検知はアルゴリズムは基本的に劣化
- ・ ポリシー変更は正解ラベル定義の変更と同義 (=教師データ変更が必要)

### <解決策>

メルカリはシステムによって検知された違反商品はCSが必ずすべて確認する運用をとっている

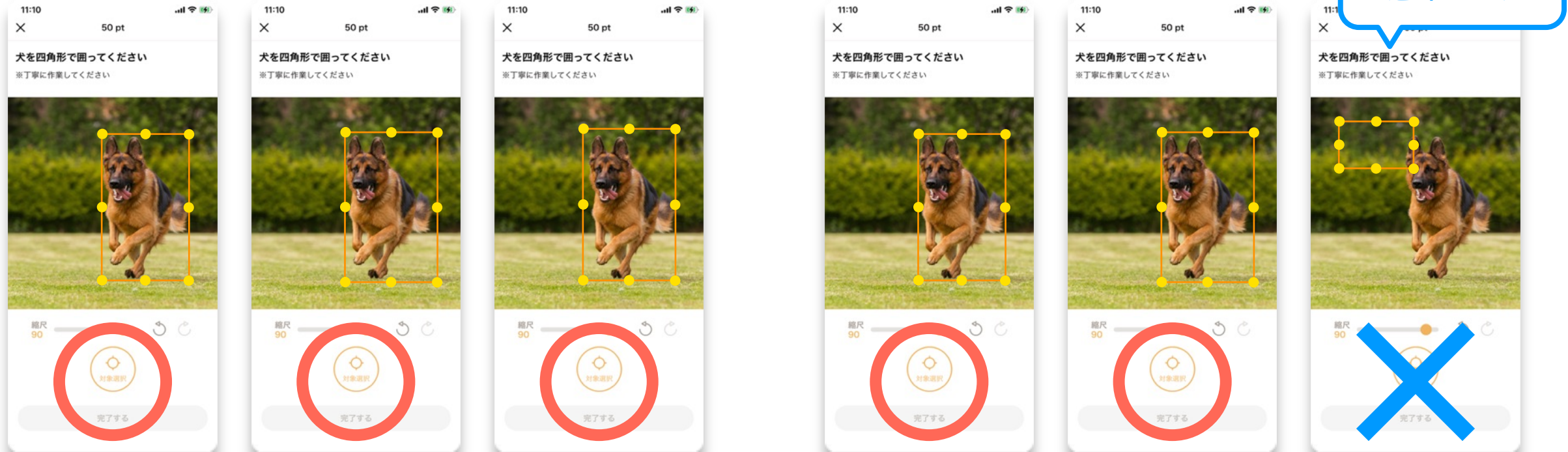
蓄積された正確で新しいデータを用いてモデルの再学習を行うことで機械学習システムの精度が保っています。

ルールベースシステムがCSオペレーターに候補を送信することでMLモデルへのラベルを新たに蓄積し、このラベルからMLモデルが新しい違反を学習することが可能

ポリシーの変更に伴い、無効となる教師データがどうしても存在するが、予めうまく切り分けられるログを残すことで急なポリシーの変更に対して無効となったデータを、最小限の大きさとデータセットから除外することができる

# harBestの特徴①

## 奇数のデータで比較する自動チェック機能



信頼度：高評価

信頼度：低評価

▶ 信頼度が高い良質のデータを集めることが可能



# harBest

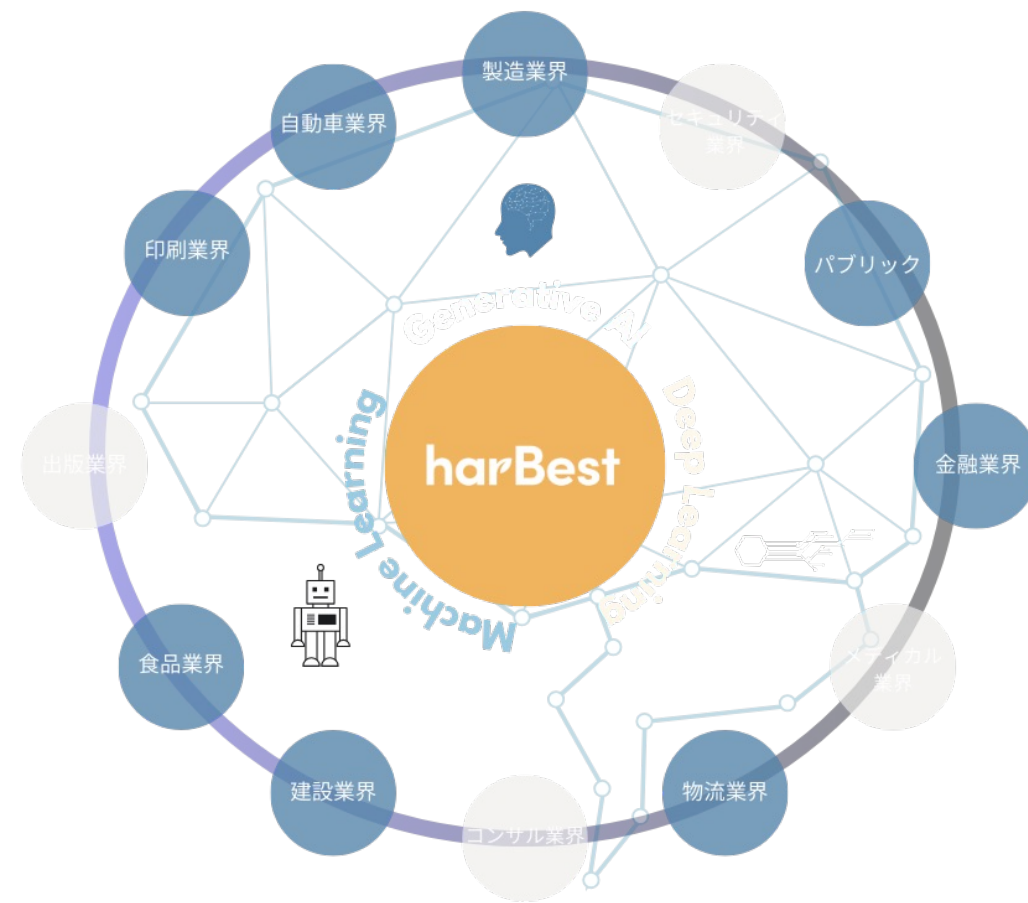
案件をAPTOがコントロールし、クラウドワーカーに作業依頼を行ないます。  
データ品質、機密情報、作業者への報酬支払いは当社が責任をもって管理致します。



- ・イントロダクション（3分）
- ・前提となる考え方 – Data-Centric / Model-Centric（3分）
- ・Human-in-the-Loop (HITL)（3分）
- ・各業界における Data-Centric AI開発事例（10分）
- ・アノテーションプラットフォーム「harBest」が目指す社会（1分）



出典：MLOps: From Model-centric to Data-centric AI (Andrew Ng)



## Data-Centric × Human-in-the-Loop × 農業 × harBest



課題：ピーマン自動収穫機の精度向上

2019年から「農作物の自動収穫機」の開発に着手

- 収穫機に付帯するAIカメラが、収穫可能かどうかを判断

実際の依頼内容

- 機器に搭載されているカメラで動画/画像を撮影、手作業で切り出し、分類、アノテーションを実施  
アノテーション要件

- ①取れるピーマン、②枝が刈り取りの邪魔しているピーマン、③葉っぱに隠れているピーマン、④大きさが適正でないもの、⑤その他の例外に分類・ラベリング

Point:

収集: どうやってカテゴリごとに適切かつ大量のピーマン画像を収集するか

Labeling: 教師データに量、質ともに高精度なアプローチが求められる

Firtering: 枝を欠損することなく、適当な成長度のピーマンを検出

結果: 10%~精度向上に成功。また、枝を切ってしまうことによって起きていた生産量にも寄与。





## Data-Centric × Human-in-the-Loop × 製造業 × harBest



課題：実装ラインの稼働率向上 / 人的な監視・管理コストの削減

高額な機械、実装ラインの稼働率が低減していた

- 機械1台ごとに「シグナルライト」を設置
- 稼働状況を8パターンに分けて分類、アノテーション

Point:

収集: 工場内に設置したカメラの動画から画像を切り出し

Labeling: 稼働パターンに沿って、8種類のラベルを想定

Firtering: 曖昧な信号、確認難度の高いシチュエーションも想定

結果：15%～30%前後認識率の向上に成功。各ラインの稼働率も大幅に伸び、生産の最適化を実現。



# Data-Centric × Human-in-the-Loop × インフラ × harBest





## Data-Centric × Human-in-the-Loop × 食品メーカー

ITmedia NEWS > 企業・業界動向 > 鶏肉加工にもAI自動化の波 リサイクル部位の7割を...

### 鶏肉加工にもAI自動化の波 リサイクル部位の7割を唐揚げ用に——ニチレイの食品ロス削減術

🕒 2021年10月26日 12時23分 公開

[吉川大貴, ITmedia]

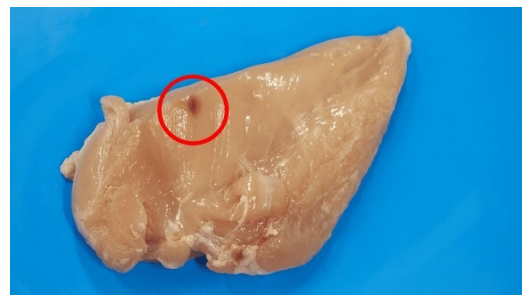


PR 業務改善のプロに聞く 「生産性向上」に必要な視点とアクション

PR 面倒な「固定IP問題」があっさり解決 リモートワークや開発環境アクセスにも

冷凍食品の製造販売を手掛けるニチレイフーズ。今、同社が鶏肉を出荷する上で欠かせない存在になっているものがある。AIだ。

同社ではこれまで、唐揚げなどの材料となる鶏肉を検査し、血合いを取り除く工程を、工場員の目視と手作業に任せていた。しかし従来の手法では端材が増えやすくフードロスにつながる他、工場員の負担にもなることから、画像認識AIを活用して自動化。食用にできず飼料や肥料にリサイクルしていた鶏肉を7割減らせたという。



鶏肉の血合い（ニチレイフーズの公式サイトから引用）

課題：機器による選別後に、人手や目視による検品が必要でコストを圧迫。

2016年から「原料の選別に用いる画像認識AI」の開発に着手

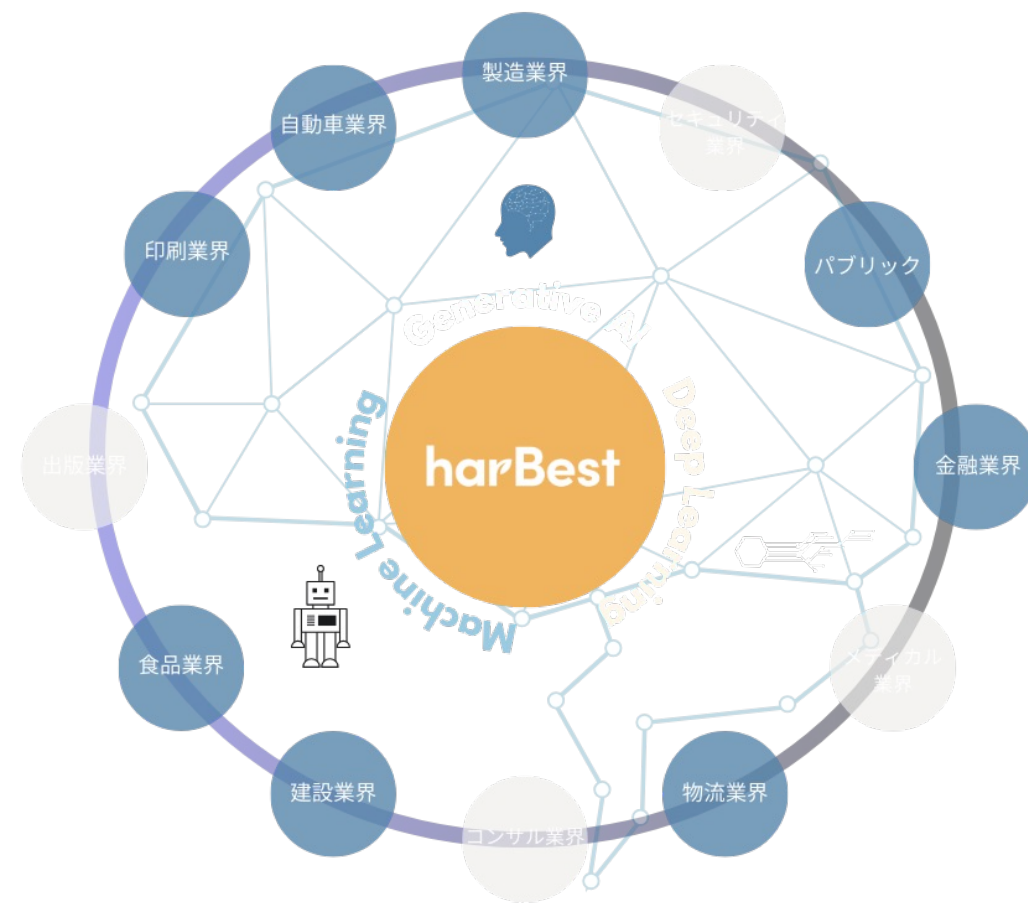
- 工程を分解 ①原料の目視確認 ②血合いの有無・位置確認 ③除去
- ①'ベルトコンベヤーに流れてくる鶏肉を、目視ではなくカメラで撮影 夾雑物(きょうざつぶつ)箇所を独自の撮影技術で強調
- ②'カメラで撮影した画像と、ベルトコンベヤーがどれだけ動いたか計測できる「エンコーダー」で取得した座標情報を基に、血合いの有無や位置を検出
- ③'独自技術で開発した機器により除去（詳細非公開）

結果：食用にできず飼肥料にリサイクルしていた鶏肉を7割減らすことに成功  
また処理速度が約4倍に向上し、除去率も1.5倍に。

- ・イントロダクション（3分）
- ・前提となる考え方 – Data-Centric / Model-Centric（3分）
- ・Human-in-the-Loop (HITL)（3分）
- ・各業界における Data-Centric AI開発事例（10分）
- ・アノテーションプラットフォーム「harBest」が目指す社会（1分）

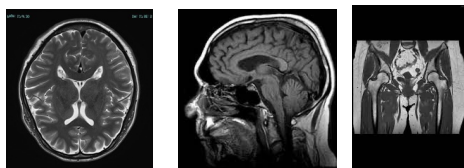


出典：MLOps: From Model-centric to Data-centric AI (Andrew Ng)



## harBest はデータ提供No.1 プラットフォームへ

希少データの収集/アノテーション



①データセット提供  
MRI画像/CT画像



⑥手数料支払



④データセット販売



⑤データセット購入料金

製薬会社 / AI企業



収集困難な医療用データセットを提供  
(日刊工業新聞様)



株式会社APITO、順天堂大学AIインキュベーションファーム2023年第1回JASTARプロジェクト発表

順天堂大学AIインキュベーションファーム2023年第1回JASTARプロジェクトに発表されました

APITO Inc. 2023年12月18日 18時27分

AI開発データ提供、データ提供を加速、急ぎ、迅速に医療用データプラットフォーム「harBest」を開発する。順天堂大学AIインキュベーションファーム2023年第1回JASTARプロジェクトに発表されました。本プログラムは順天堂大学に属し、医療現場に活用するデータ提供を加速することを目的としています。

**JASTAR** **APITO**

医療用データの活用でAI医療を加速

2023年度「AIスタートアップ実証プロジェクト」(JASTAR)において、株式会社APITOプロジェクトに発表されました。

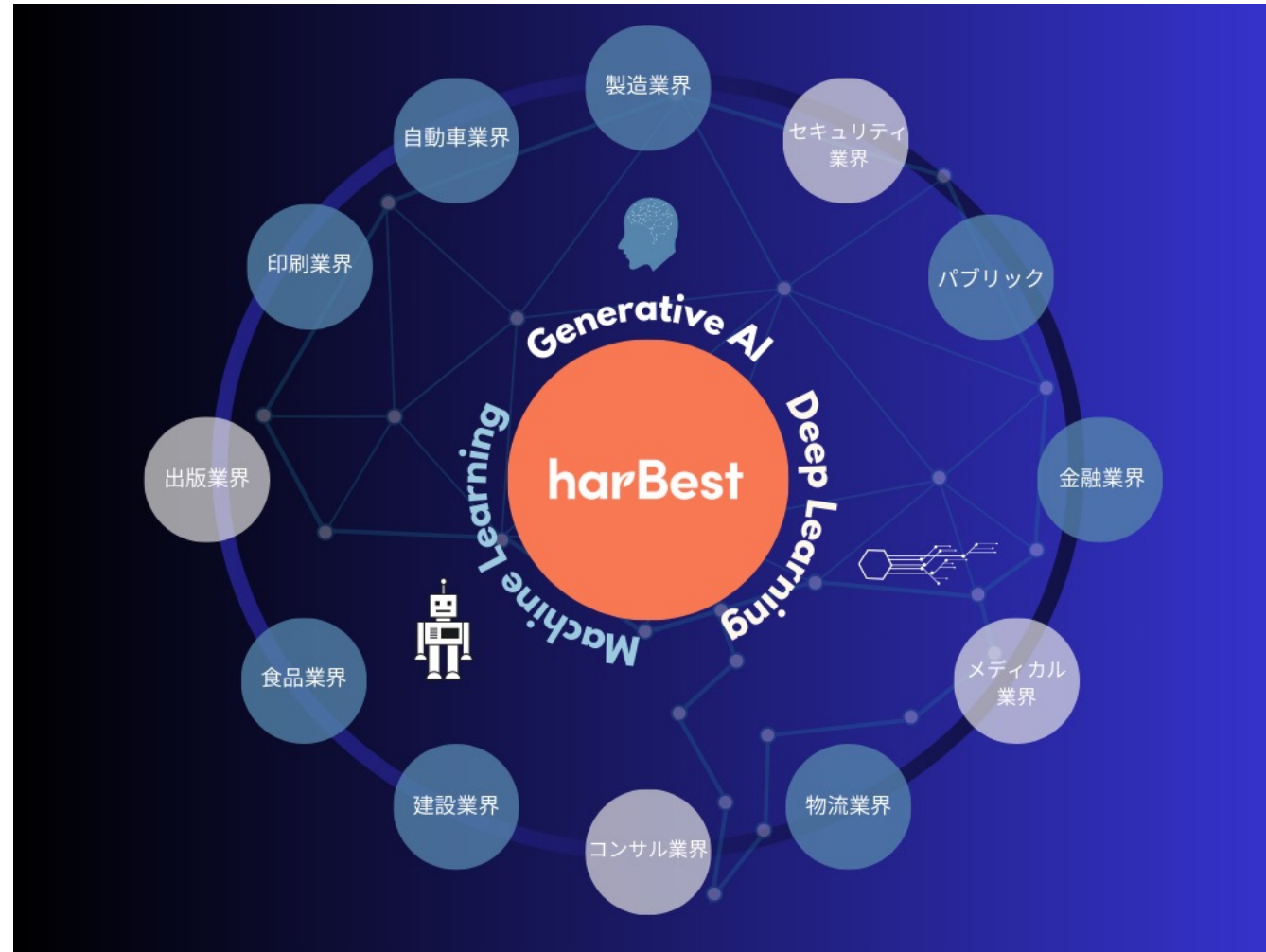
JASTARについて

現場で収集困難な異常データのバーチャル作成



LLM開発に必要な権利クリアなデータセットを提供

No.	Information	Question	Answer HarBest	Category
1	抗生物質感受性試験は、特定の微生物が特定の抗生物質に対して感受性を持つかどうかを決定する実験的手段です。この試験は、新しい抗生物質の効果を測定するために使用されます。通常、このテストは細菌の感受性を評価し、その結果を処方するかどうかの判断に役立ちます。反応を速く血液に導入することができます。体内で様々な影響を及ぼす。	抗生物質感受性試験の主な目的は何ですか？	抗生物質感受性試験の主な目的は、特定の微生物が特定の抗生物質に対して感受性を持つかどうかを決定することです。	医学・微生物学
2	皮膚毒とは、皮膚を通じて体内に入り込まれる毒物のことで、これは、皮膚や一部の化学物質、薬品が原因です。皮膚が強い刺激や傷に曝露すると、皮膚の吸収率が高くなる可能性があります。このように毒物は、皮膚を通して血液に侵入することができます。体内で様々な影響を及ぼす。	皮膚毒の吸収率が異なる可能性がある皮膚の吸収率はどのくらいですか？	皮膚毒の吸収率は皮膚の厚さや状態、皮膚が傷いているかどうかによって異なります。	医学・毒学
3	市場セグメンテーションはマーケティング戦略の一部で、市場をより小さなグループに分けてターゲットに合わせたアプローチを可能にします。このアプローチにより企業は特定のセグメントに特化した製品やサービスを提供し、競争優位性を高めることができます。また、市場セグメンテーションは広告やプロモーションの効果的な実施にも役立ちます。	市場セグメンテーションで利用されるセグメンテーションの基準は、地理的、人口統計学的、心理的、行動的な基準ですか？	市場セグメンテーションで利用されるセグメンテーションの基準は、地理的、人口統計学的、心理的、行動的な基準です。	マーケティング
4	土地利用計画は、土地の資源を効果的に管理可能な方法で管理と開発を行うための戦略的プロセスです。これには、住居、商業、農業、工業、自然保護など、多様な用途に対する必要があります。また、法的規制、地域社会のニーズ、環境保全など多くの要素を考慮する必要があります。	土地利用計画が考慮すべき多くの要素にはどのようなものがありますか？	法的規制、地域社会のニーズ、環境保全が含まれます。	都市計画・建築学
5	統計学は、データの収集、整理、分析、解釈を可能にする学問です。統計学は、データの分布、傾向、相関関係を理解し、予測を行うために使用されます。統計学は、社会科学、自然科学、工学など多くの分野で広く使用されています。	統計学が広く使われる方法にはどのようなものがありますか？	グラフ法、表法、行列法などです。	数学
6	ホームオステオスは、生物の体内環境が一定の範囲内で安定して維持されることを指します。これには、体温、pH、塩分濃度、血糖値などが含まれます。ホームオステオスは、生命維持に不可欠なプロセスです。	ホームオステオスが維持する体内環境の要素にはどのようなものがありますか？	体温、pH、塩分濃度、血糖値などが含まれます。	生物学
7	情報倫理は、情報技術と人間が交差する領域での倫理的問題を研究する学問です。これには、プライバシー保護、著作権、データセキュリティなどが含まれます。情報倫理は、急速なテクノロジーの進展に伴って重要な役割を果たしています。	情報倫理が研究する倫理的課題にはどのようなものがありますか？	プライバシー保護、著作権、データセキュリティなどが含まれます。	情報科学・倫理学
8	生化学的実証実験「BOD」は、ある液体に含まれる有機物の量を測定するために使用されます。有機物が分解する過程で消費される酸素の量を測定します。BODが高いことはその液体に多くの有機物が存在していることを示唆します。BODは、一級に分解可能な有機物の量を測定するために使用され、単位はmg/lとして示されます。水質汚染の指標として示されます。	BODが高いというものは何を意味し、その値はどのようにして測定されますか？	BODが高いというものは、その液体に多くの有機物が存在していることを示唆します。その値はmg/lとして測定されます。	環境科学



あらゆる業界において「AIインフラ」の礎となる

# We are hiring !

ご清聴ありがとうございました。

