第7回 AI原則の実践の在り方に関する検討会【公表資料】

Citadel AI

24時間信頼できるAIをあなたに

2023年2月3日　株式会社Citadel AI

AIの信頼性とガバナンスに関わる議論

EUがAI包括規制案

AIシステム版の“GDPR”

EUのみならず全世界の企業に適用

€30milか全世界売上の6%の制裁金

本年中に法制度化を目指す

NY市：自動採用決定ツールに対する事前バイアス監査義務法2021/144を本年から施行予定

AIの信頼性に関わる新たな課題を解決

2020年12月設立

Seed Funding:

東大IPC

anri

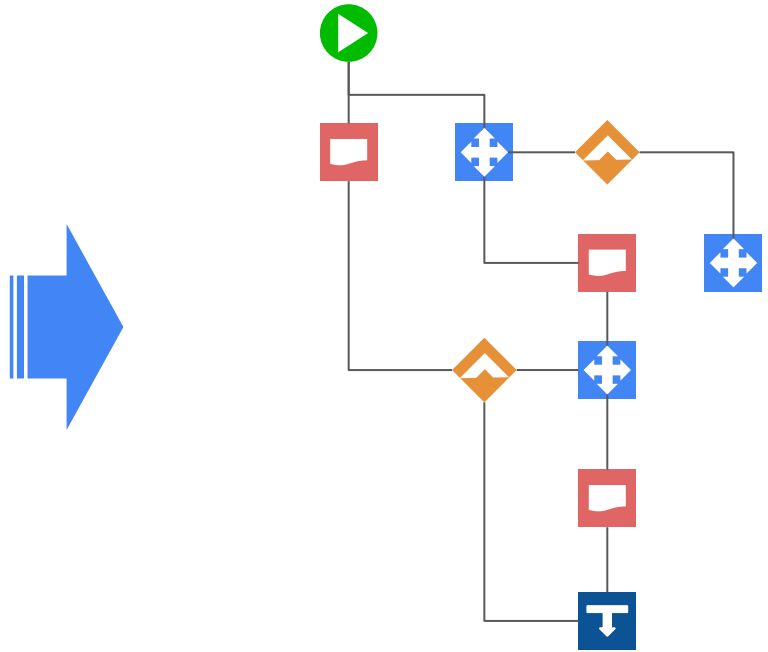
Ex **Google, Waymo, paidy, Toyota**

1. Problem

我々が取り組む課題
「AIの信頼性」とは何か？

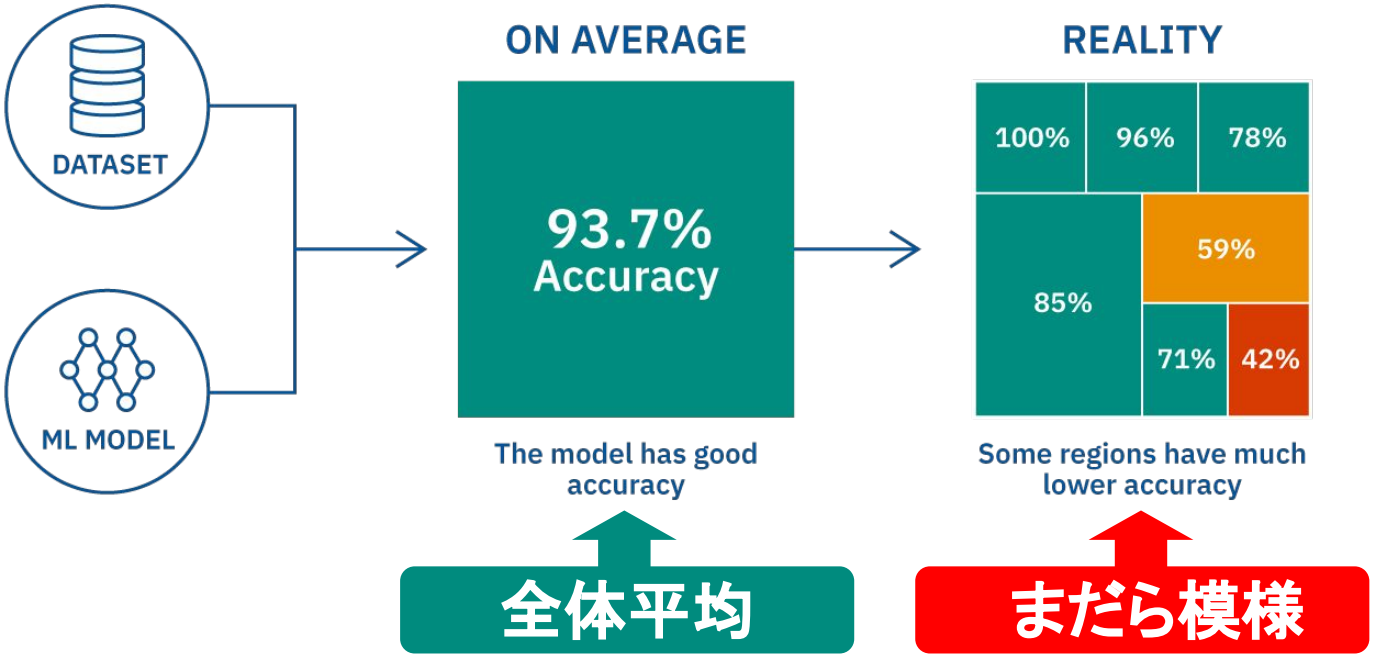
さまざまなローコードツール

```
import numpy as np
from tensorflow.keras.models import Sequential, model_from_json
from tensorflow.keras.layers import Dense, Dropout, Flatten,
Conv2D, MaxPooling2D
from tensorflow.keras.optimizers import RMSprop
from tensorflow.keras.datasets import cifar100
from tensorflow.keras.preprocessing.image import array_to_img,
img_to_array, load_img
from tensorflow.keras.utils import to_categorical
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
....
```



AIのモデル開発は容易に

隠れた未学習領域に残るリスク



再学習時に発生するトラブル

New Model ?

New Data ?

AIは環境変化に非常に脆弱

従来の顧客

学習時の撮影環境

新規顧客

運用時の状況

AIの間違いは、一体 誰がどうやって？

気づいた時には手遅れに

人手作業には
膨大な
手間と時間

ビジネスリスク
セキュリティ
コンプライアンス

ポリシーと手順の 標準化・文書化	全ての組織の役割、責任、及びプロセスに対応する「ポリシーと手順を標準化し文書化」することが重要
モデル監査 アルゴリズム監査	「モデル監査やアルゴリズム監査」は、説明責任の検討事項を評価し、文書化するために有効
継続的な モニタリングシステム の追加構築	問題を自動検出しアラートする「モニタリングシステム」を追加構築することが重要
フィードバック チャネル	エンドユーザーがエラーや損害に対する救済を求められる「フィードバックチャネル」があることが望ましい

ポリシーと手順の 標準化・文書化	✔
モデル監査 アルゴリズム監査	?
継続的な モニタリングシステム の追加構築	?
フィードバック チャネル	✔

何をどこまで検証したら良いのか？

また、それをどのようにしたら

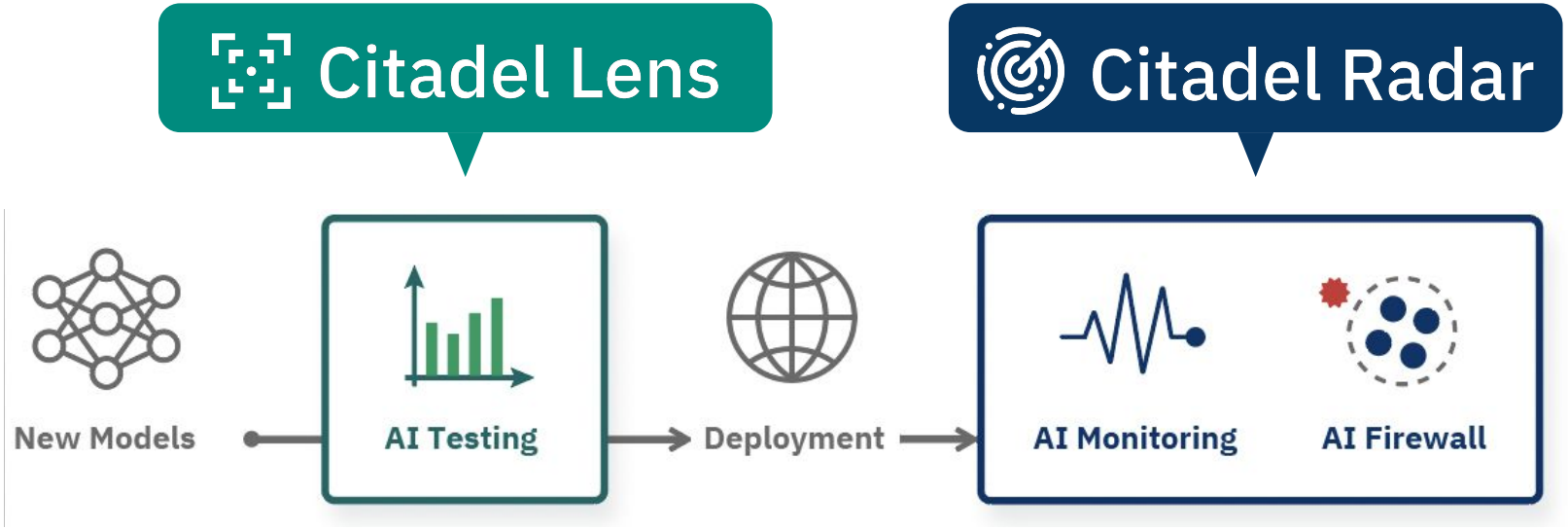
継続的に実行できるのか？

ポリシーと手順の 標準化・文書化	✔
モデル監査 アルゴリズム監査	Citadel Lens
継続的な モニタリングシステム の追加構築	Citadel Radar
フィードバック チャネル	✔

2. Our Solution

我々のソリューション

2つの「自動化」を通じて解決

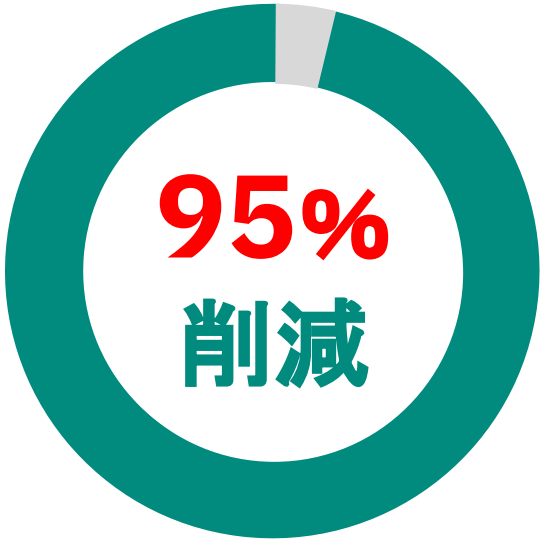


瞬時に自動検証

継続的に自動監視

開発 学習・検証 運用・保守

Citadel Lens

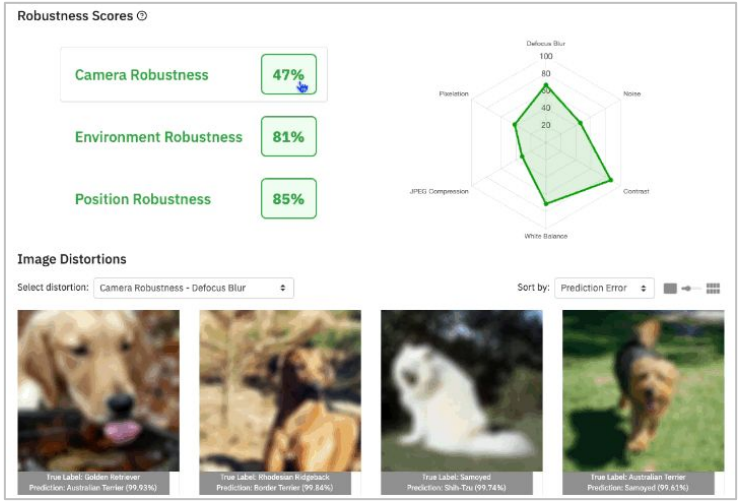


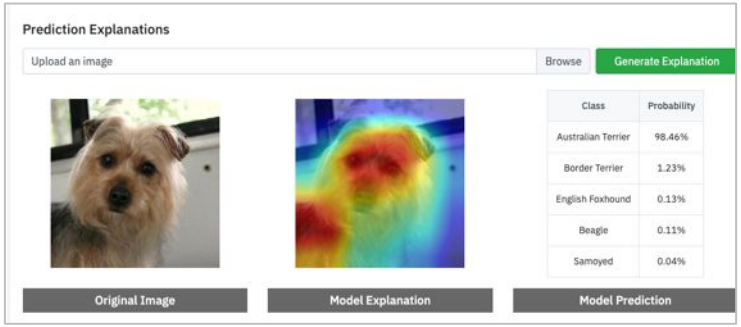
Citadel Radar

Citadel Lens : 検証時間を95%削減

網羅的テストを
自動かつ一瞬で

- ノイズ耐性テスト
- 未学習領域の自動検出
- ラベル間違え推定
- 公平性テスト
- 説明責任の可視化 etc.





Citadel Lens：PoC加速化と品質改善高速化

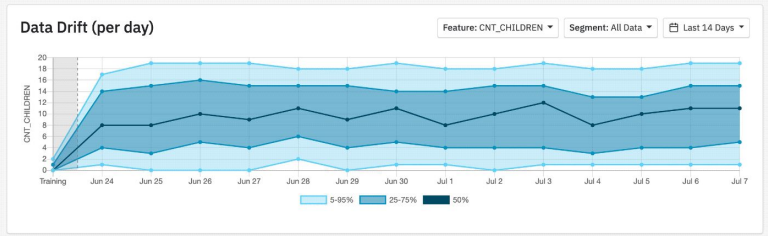
多様なモデルに
汎用的・横断的に

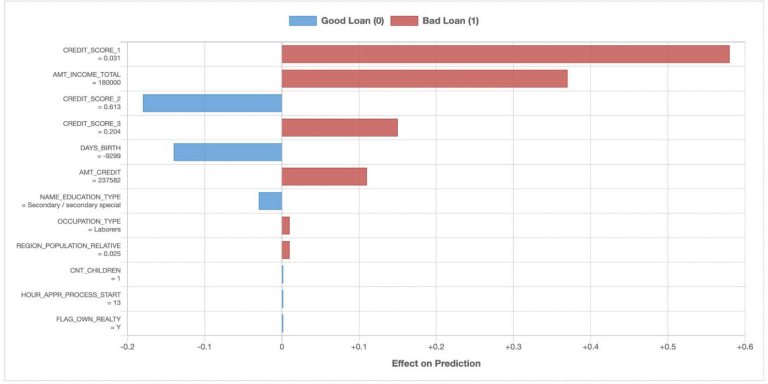
Score	Model Version	Description	Completed At	Details
92%	2022.04.30	150 trees, training dataset from 2022-04-30	2022-04-30 17:22:32	See Report ›
88%	2022.04.24	150 trees, training dataset from 2022-04-24	2022-04-24 11:21:39	See Report ›
83%	2022.04.21	150 trees, training dataset from 2022-04-21	2022-04-21 21:38:52	See Report ›
90%	2022.04.18	125 trees, training dataset from 2022-04-18	2022-04-18 08:14:51	See Report ›
85%	2022.04.17	125 trees, training dataset from 2022-04-17	2022-04-17 14:31:05	See Report ›
88%	2022.04.16	125 trees, training dataset from 2022-04-16	2022-04-16 10:45:12	See Report ›
91%	2022.04.11	100 trees, training dataset from 2022-04-11	2022-04-11 08:17:49	See Report ›
84%	2022.04.04	100 trees, training dataset from 2022-04-04	2022-04-04 14:33:02	See Report ›
84%	2022.04.02	100 trees, training dataset from 2022-04-02	2022-04-02 14:33:02	See Report ›

Citadel Radar：変動リスクを即時ブロック

複数のフィルターで
自動かつ継続的に

- モニタリング
- ファイアウォール
- 説明責任の可視化





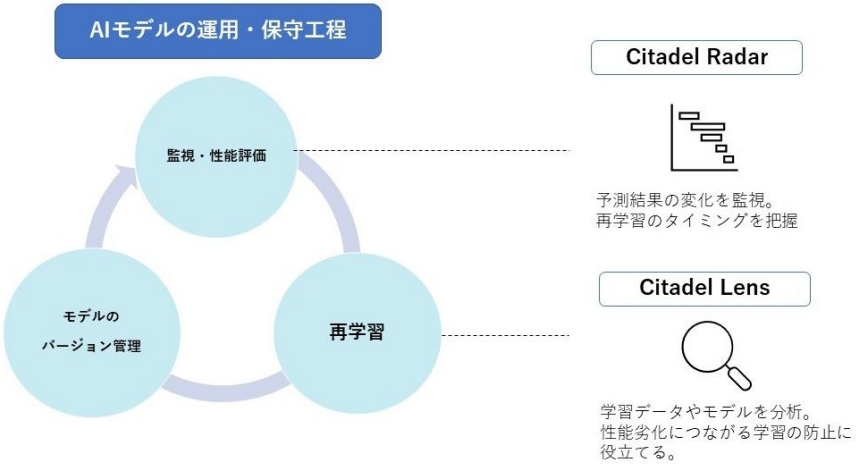
Citadel Radar : 運用時のリスクを自動検知しブロック

運用中のAIの精度検査を行う場合
従来の方式では、**目視検査**による
ラベル付けが必要。
多くの手間と時間を要し**手遅れ**に。

Citadel Radar を使えば、ラベルに
頼ることなく**リアルタイムに異常を
検知・防御**。プロアクティブに
リスクに対処可能。

Suntoryのグループ横断品質管理ツールとして採用

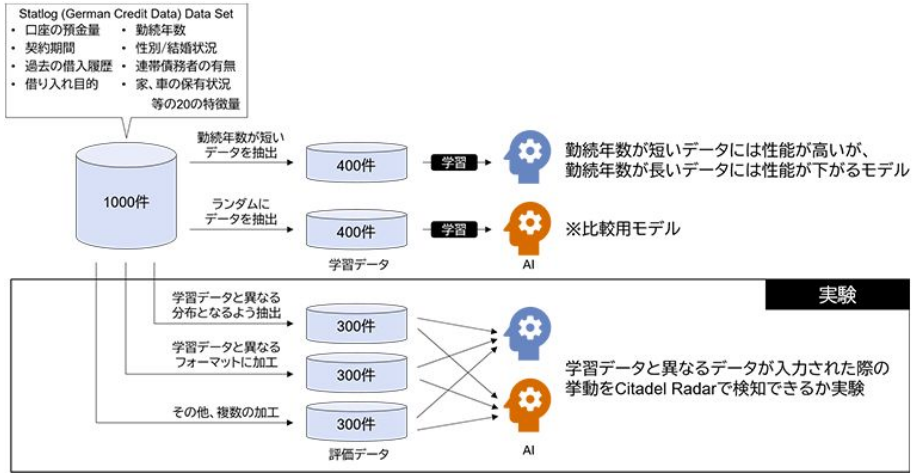
https://www.citadel.co.jp/news/citadel-ai-suntory



日経クロステック 2022年11月24日記事：サントリーが製品出荷パレット回収予測に AI、精度劣化を回避する策

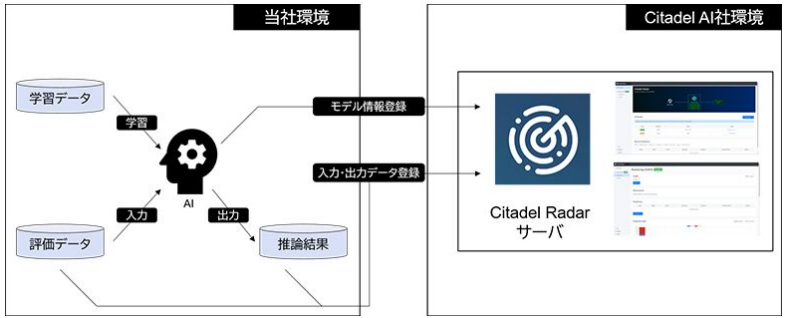
NTTデータと共同でCitadel Radarの検証を実施

従来のソフトウェアとは異なる特性を持つ、AIの継続的活用に向けCitadel AIとNTTデータは、金融系公開データを用いた共同検証を実施。Citadel Radarを活用することで、AIモデルのトラブルを検知し
安定した運用に寄与できることを確認しました。

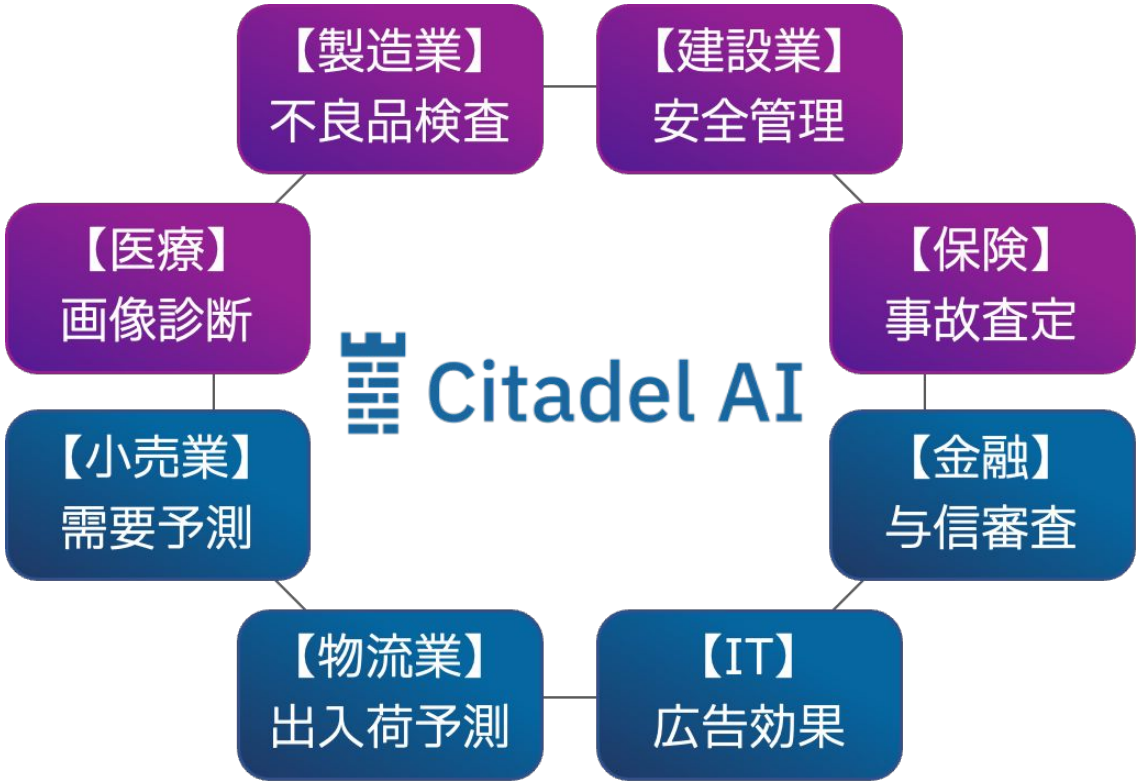


NTTデータ：2022年7月22日 Data Insight

https://www.nttdata.com/jp/ja/data-insight/2022/0722/



さまざまなAIに汎用的に適用可能



Citadel AI

24時間信頼できるAIをあなたに

企業URL	https://citadel.co.jp
お問合せ	rick@citadel.co.jp kenny@citadel.co.jp
Twitter	https://twitter.com/CitadelAI

AIの疑問は　#AI豆知識