Microsoft Azure

機械学習モデルの解釈性・説明性 編

Azure Machine Learning ハンズオン

女部田啓太、Cloud Solution Architect -Data & AI

アジェンダ

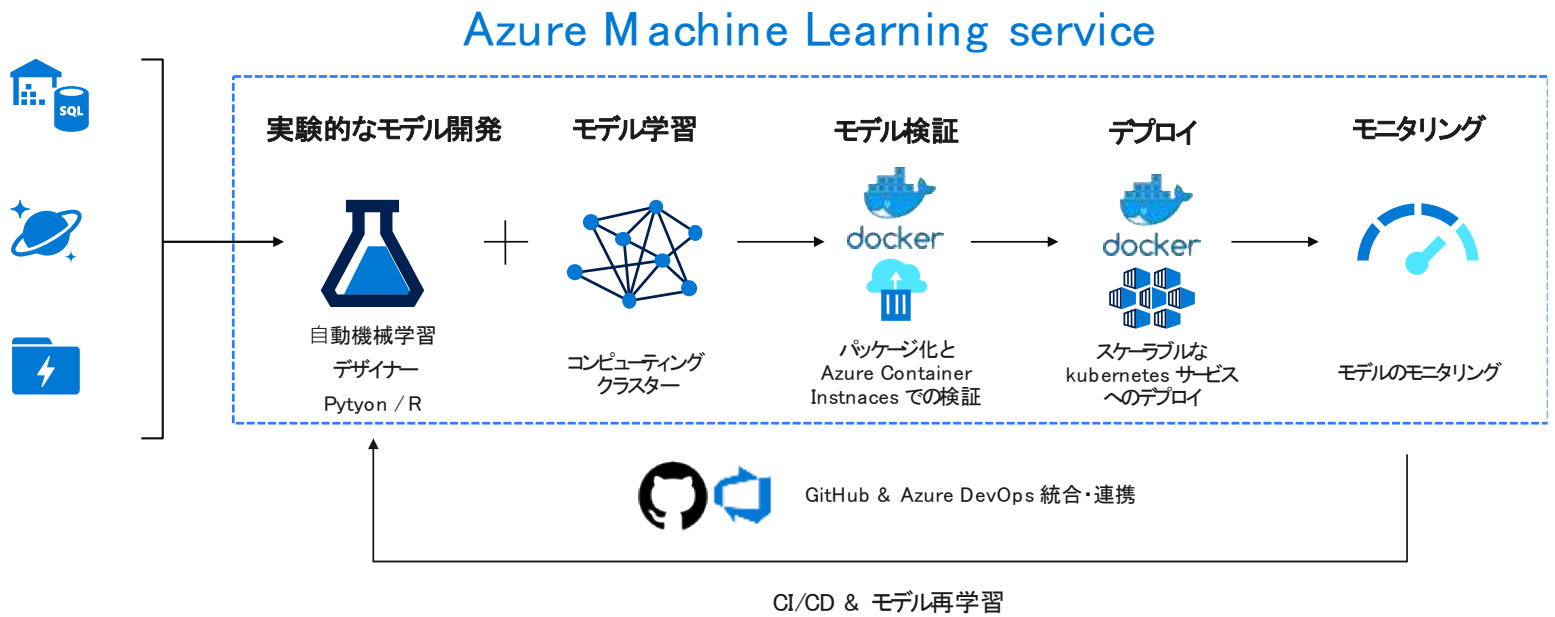
- モデル解釈性・説明性の概要
- ハンズオン
  - Explainable Boosting Machine による解釈性の高いモデル開発
  - Gradient Boosting 回帰モデルの SHAP による説明
  - LightGBM 分類モデルの SHAP による説明と Error Analysis + Azure ML
- 参考情報

モデル解釈性・説明性の概要

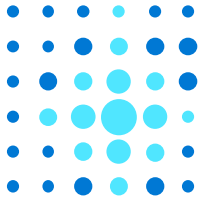
Azure Machine Learning

Azure Machine Learning とは？

- 機械学習プロセスをエンドツーエンドでサポートするマネージドサービス
  - 必要なシステムモジュールをあらかじめビルトインしている
- **自動機械学習**や**パラメータチューイング機能**による効率的なモデル開発
- 継続的なモデルの**デプロイ & 運用管理**をサポート
- **スケーラブル**な計算環境による並列分散処理 etc

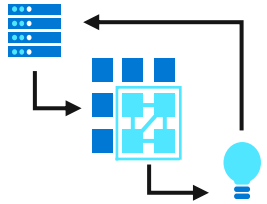


Azure Machine Learning の 4 つの特徴



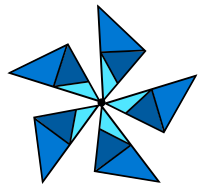
For all skill levels

あらゆるスキルレベルに対応し、ML の生産性を向上



Industry leading MLOps

機械学習ライフサイクルの運用管理



Open & Interoperable

オープンテクノロジーの採用と相互運用性の実現

Responsible

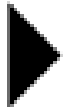
責任のある ML ソリューションの構築

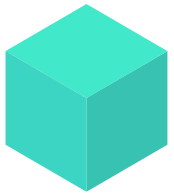
例：住宅ローンの審査

精度は高そうだけど、
信頼していいのだろうか？

学習

学習データ

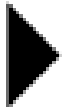


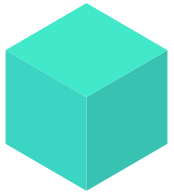


男性 or 女性を
優遇していない？

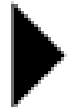
推論

審査対象の人の属性





機械学習モデル

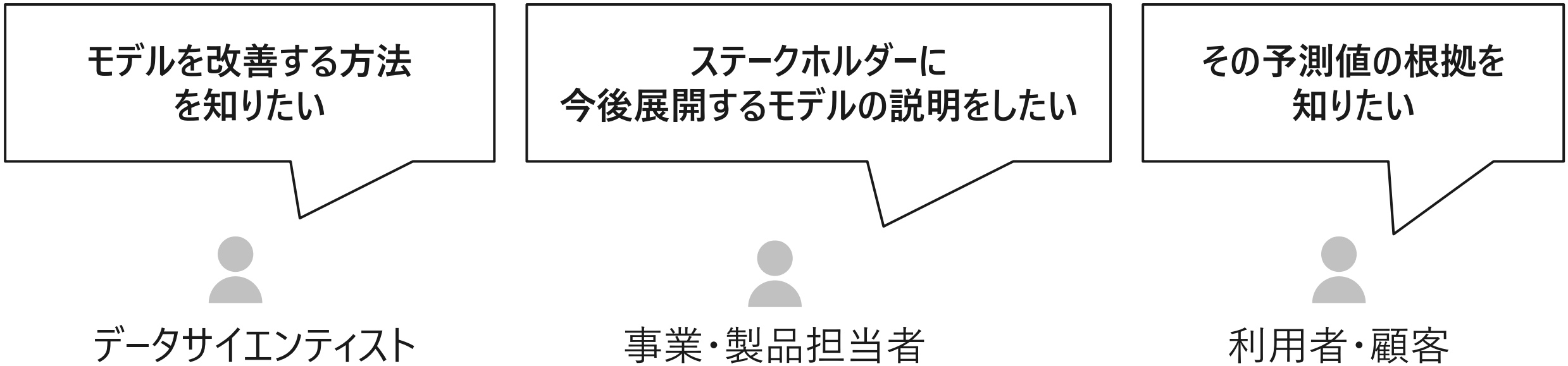


予測値

なぜ審査が却下された
のだろうか？

人間と AI のインタラクション

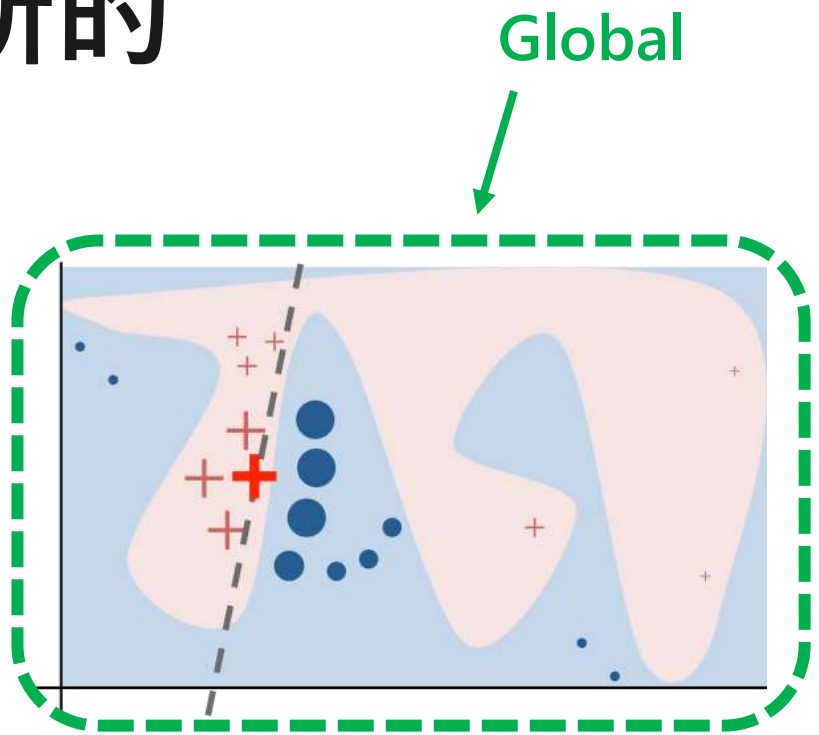
・人間中心のデザインで AI の仕組みを考えていく必要がある。
  ・「最先端の機械学習アルゴリズムだから OK」「精度が高いから OK」とは言えない

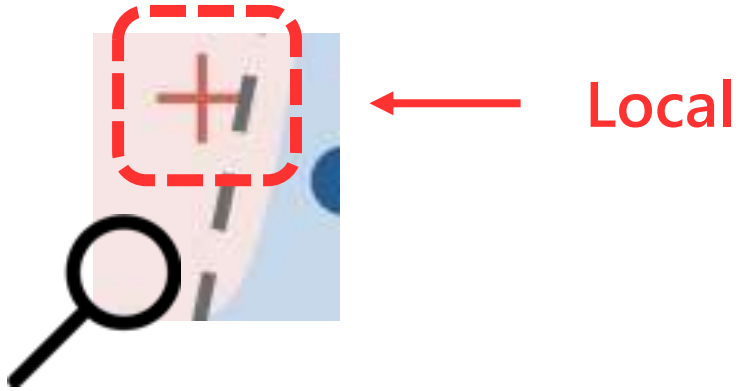


→ 機械学習モデルの説明・解釈が必要

大域的と局所的

- **大域的 Global**
  - 機械学習モデル全体の挙動の説明・解釈
- **局所的 Local**
  - 個々の予測値の挙動の説明・解釈





2 つの手法

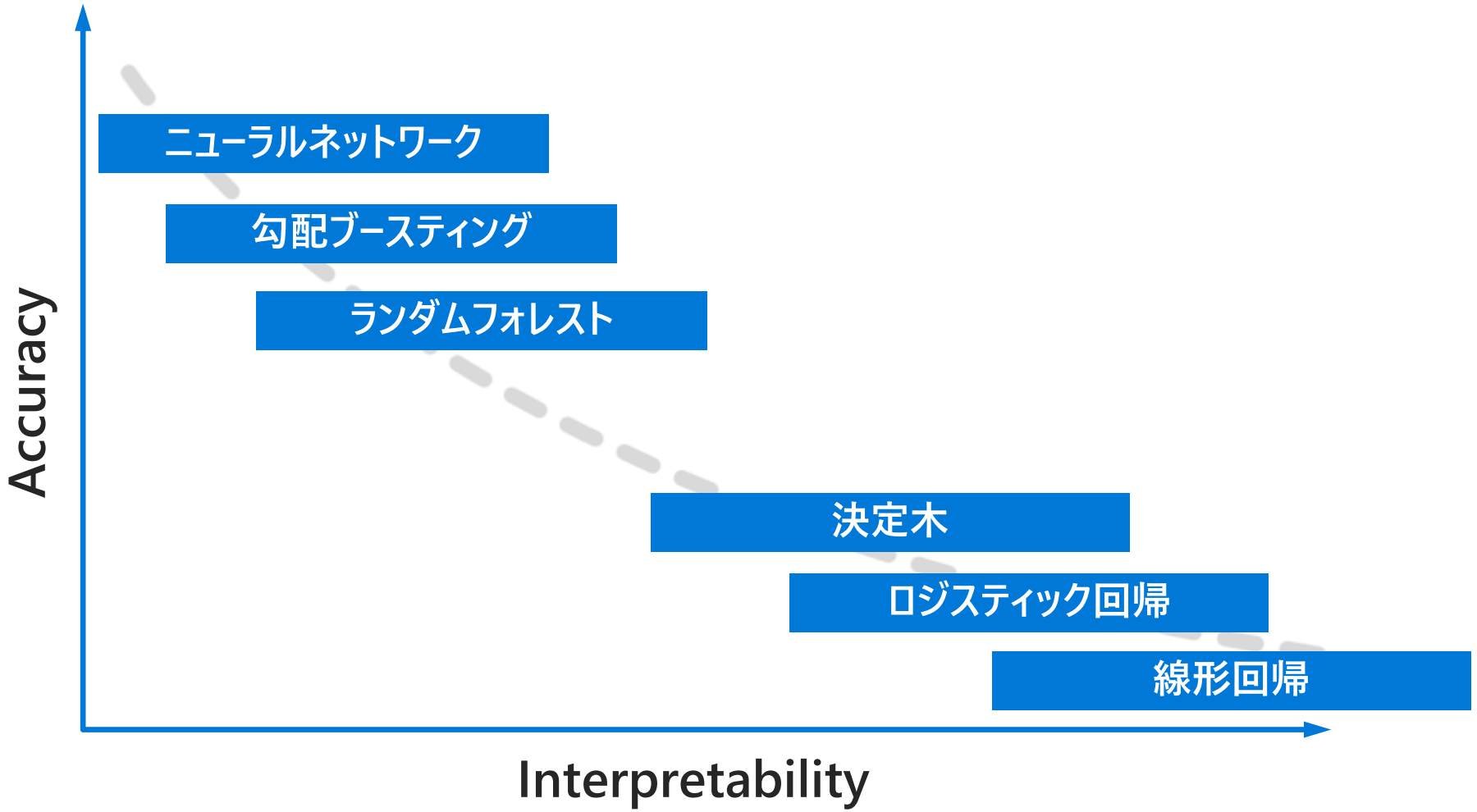
解釈可能なモデル

- 正確
- 高速

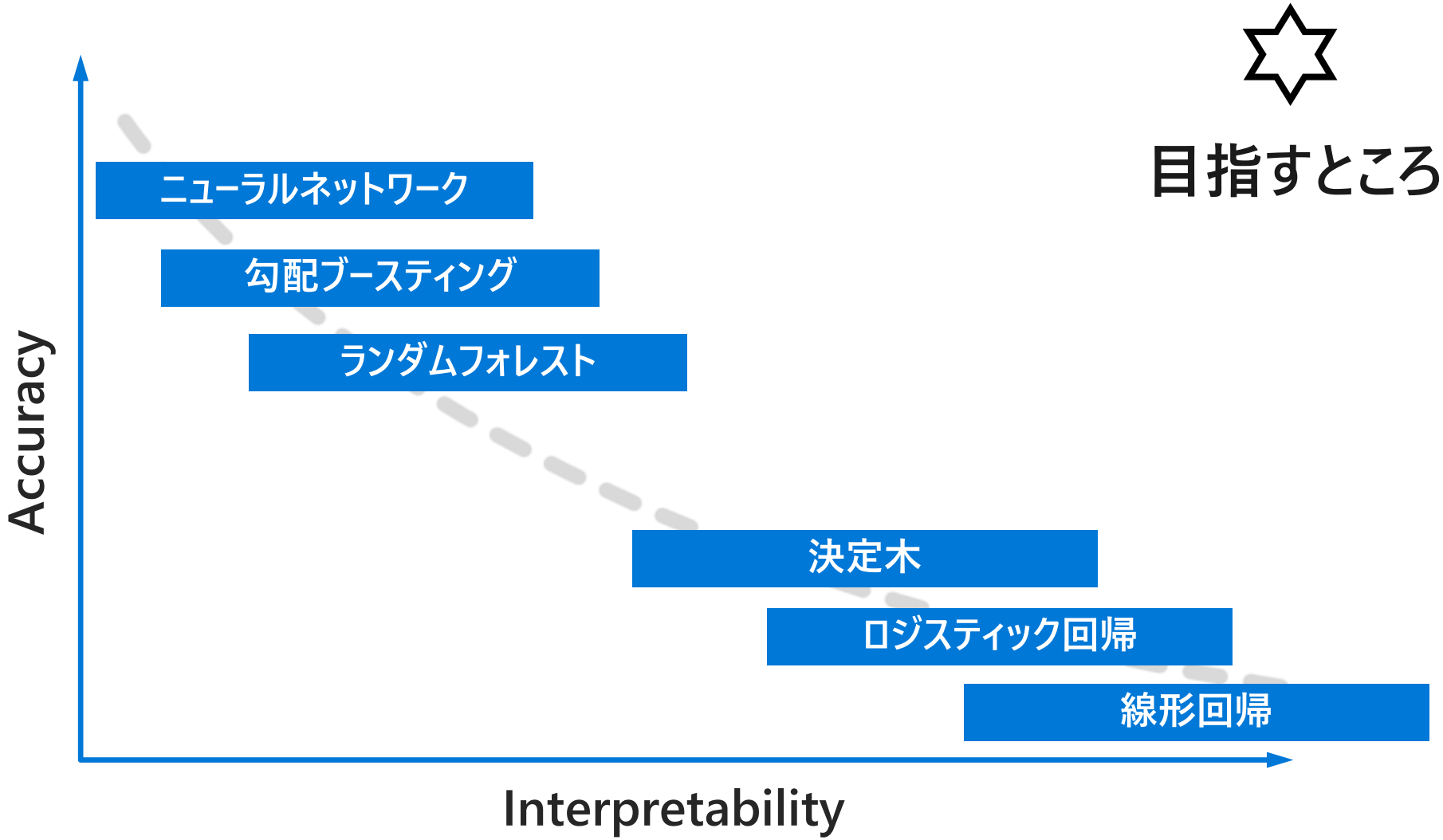
BlackBox モデルの説明

- 高精度
- 柔軟

精度と解釈性のトレードオフ

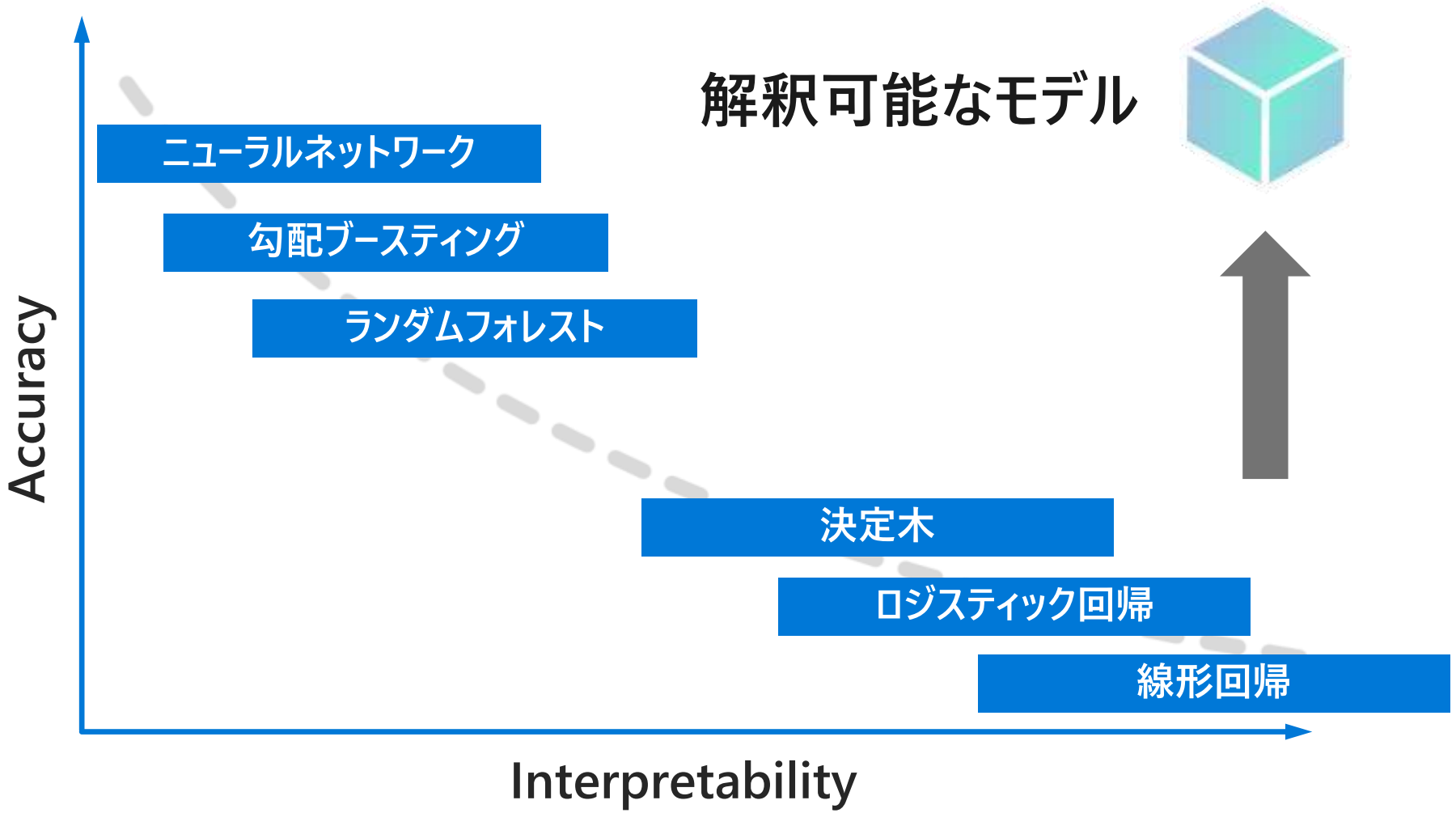


精度と解釈性のトレードオフ



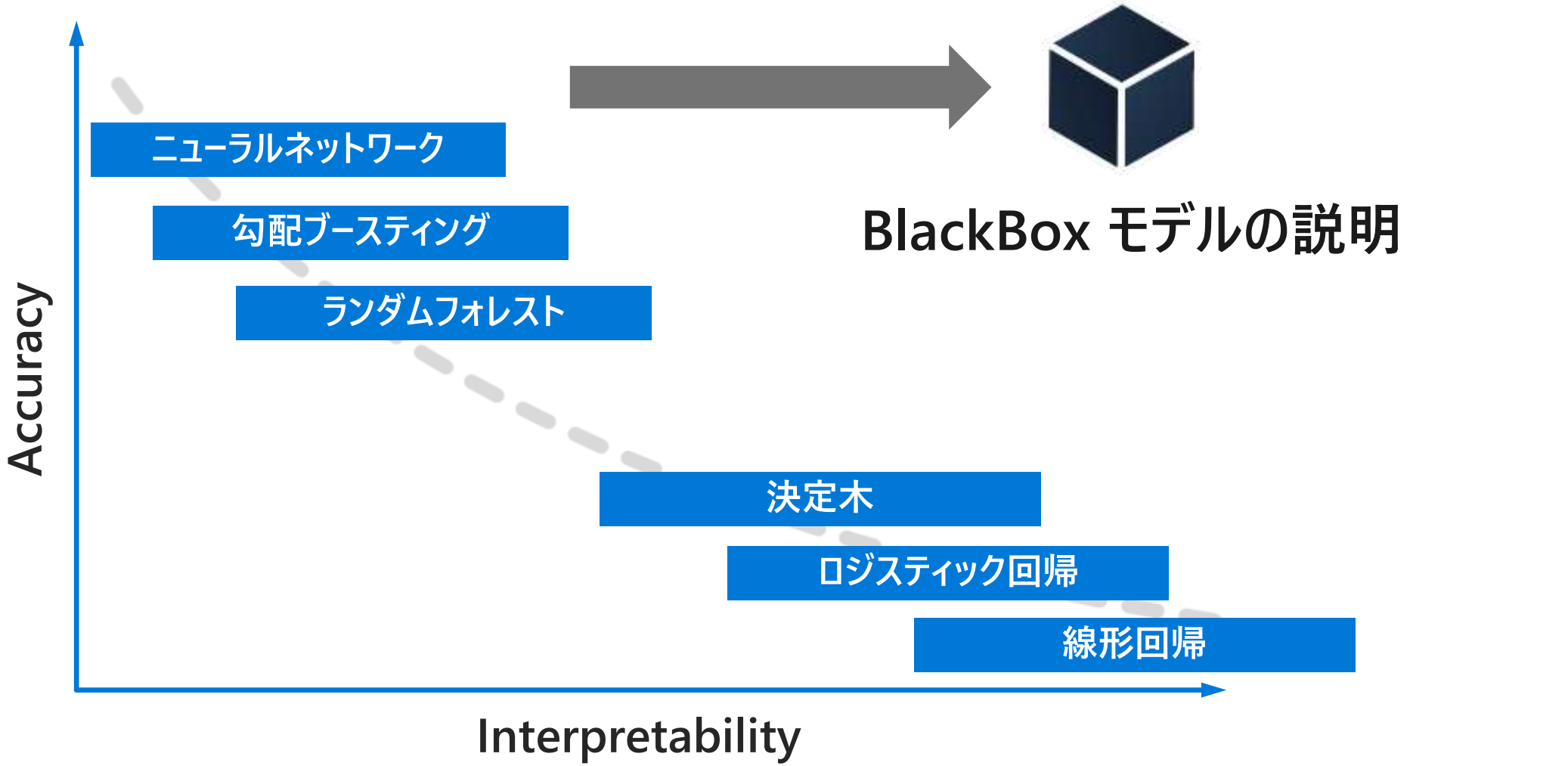
精度と解釈性のトレードオフ

解釈性を維持したまま、
精度を向上



精度と解釈性のトレードオフ

精度はそのままで、
モデルを説明できるようにする



BlackBox モデルの説明

モデル説明・解釈の全体像

- **対象**

- **目的・利用シーン**
  - → **ステークホルダーによって異なる**
    - 人間の能力を拡張する
    - AI モデルを評価する
    - AI モデルをデバッグする
- **手法**
  - 解釈可能なモデル
  - BlackBox モデルの説明

機械学習モデルの解釈可能性のための包括的なフレームワーク

https://interpret.ml/

InterpretML

機械学習モデル解釈・説明のための
包括的なフレームワーク

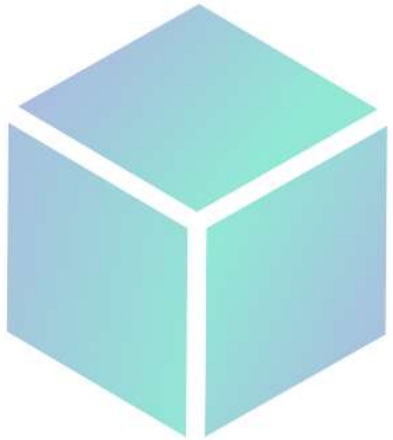
決定木

ルールリスト

線形回帰・ロジスティック回帰

EBM

…



Glass-Box

Black-Box

SHAP

LIME

Partial Dependence

Sensitivity Analysis

…

https://interpret.ml/

Glass-box models

解釈可能性が高い構造を持つ機械学習アルゴリズム

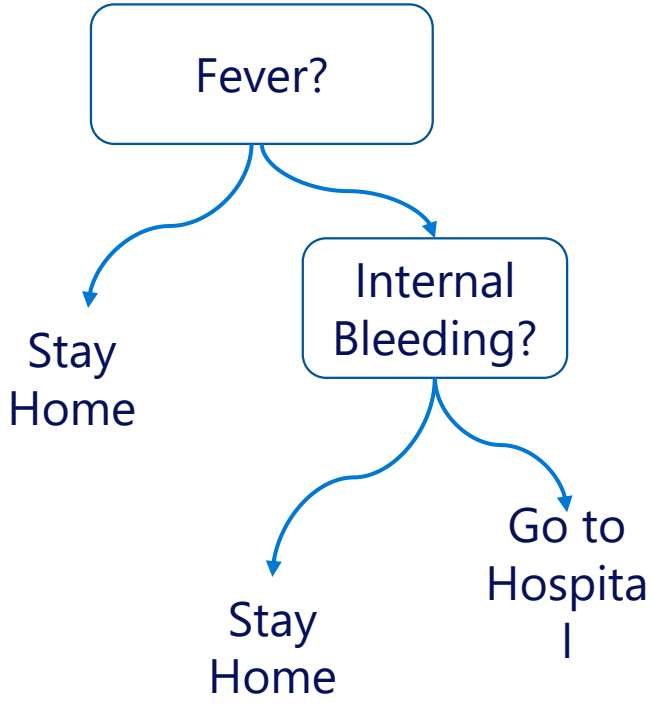
決定木

ルールリスト

線形回帰・ロジスティック回帰

Explainable Boosting Machines (EBM)

….



Black-box explanations

ブラックボックスな機械学習モデルの説明

SHAP

LIME

Partial Dependence

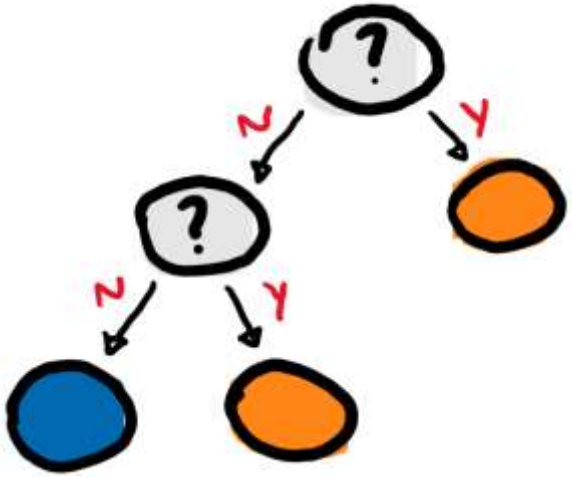
Sensitivity Analysis

...

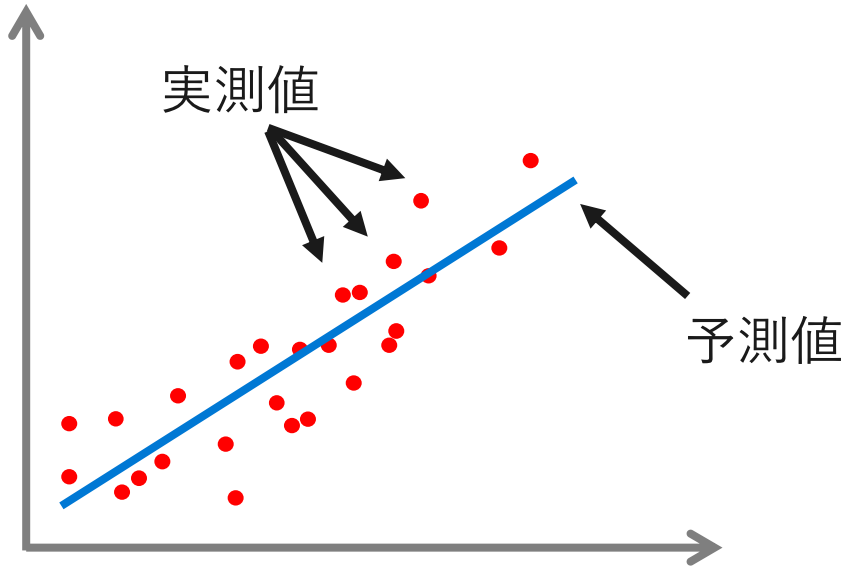
Module 1 :
Explainable Boosting Machine による解釈性の高いモデル開発

解釈可能なモデル

予測値の算出に至るモデル構造が分かりやすい



決定木



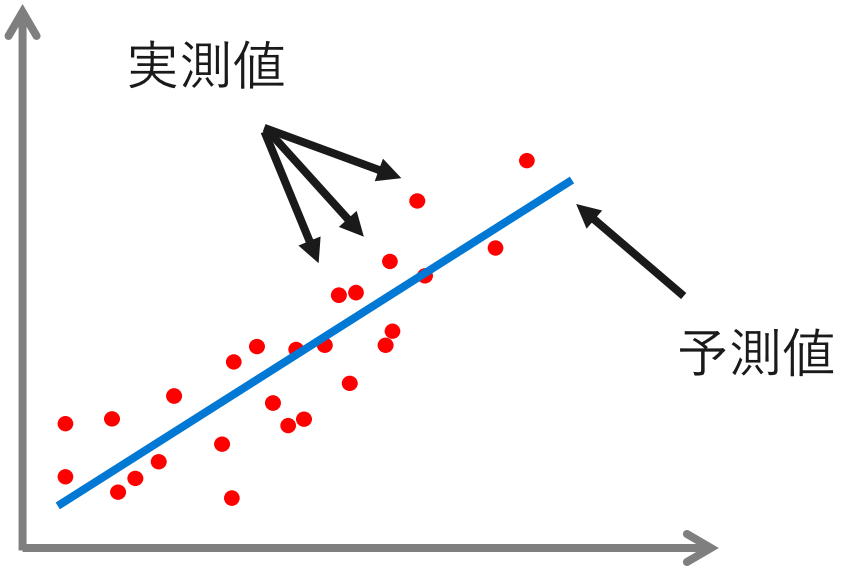
$$y = ax_1 + \mathrm{b}x_2 + \mathrm{c}x_3 + \mathrm{d}x_4 + \mathrm{e}$$

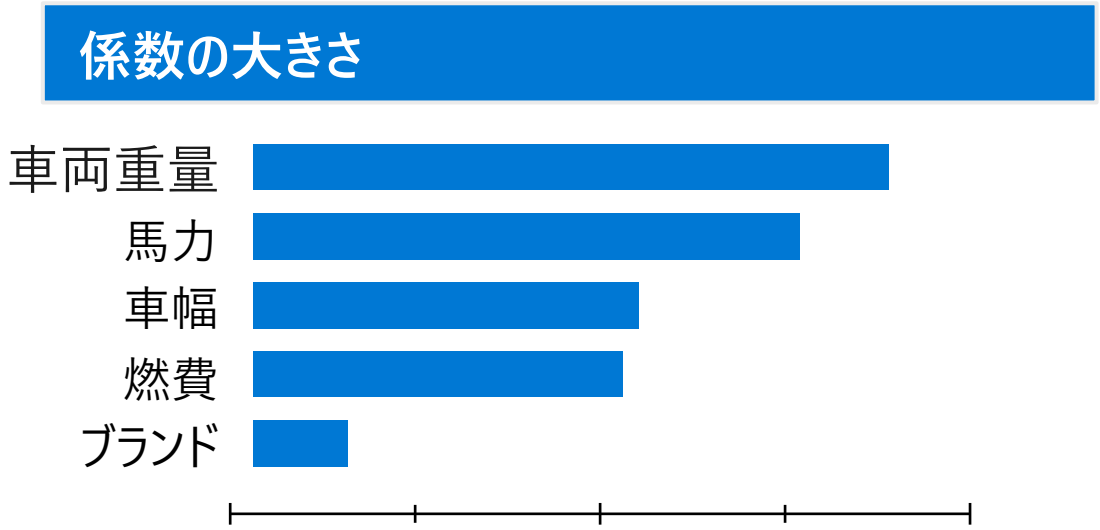
線形回帰

線形回帰モデル

- **線形回帰 (Linear Regression)** は、説明変数に重みをつけたものを合計し、予測値とする。**直線的**な関係を表す。

$$y = ax_1 + \mathrm{b}x_2 + \mathrm{c}x_3 + \mathrm{d}x_4 + \mathrm{e}$$



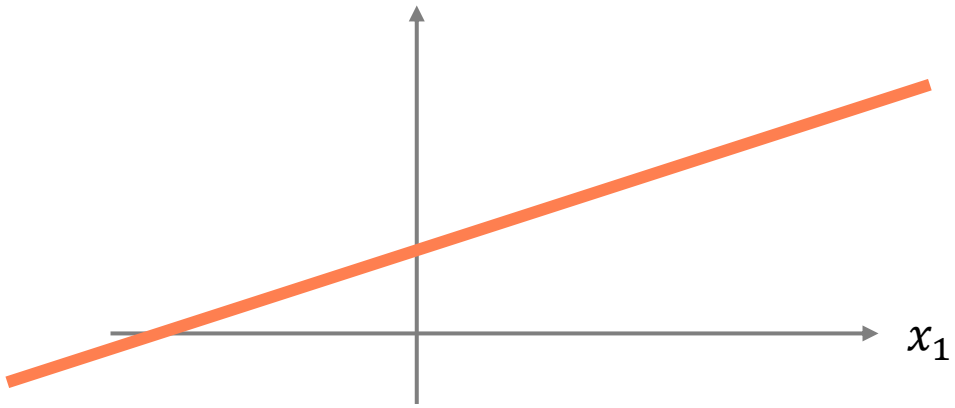


Explainable Boosting Machines

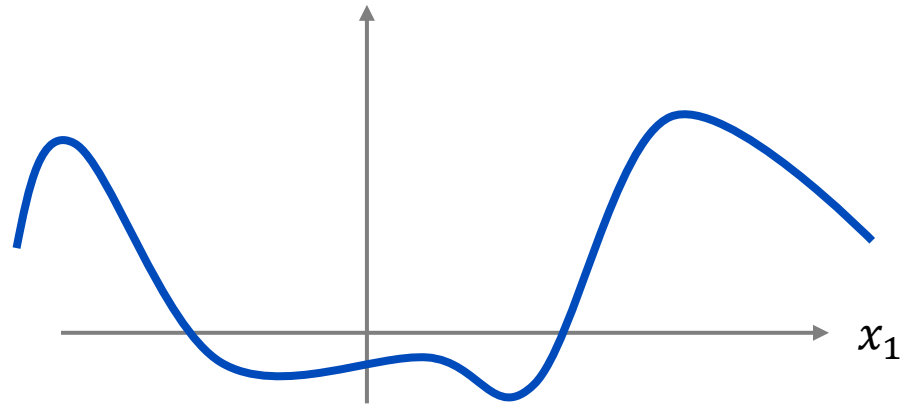
- **Explainble Boosting Machines (EBM)** は Microsoft Research が主導で開発している一般化加法モデルがベースの機械学習モデル

- EBM は表現力が豊富 (精度が高い) & 解釈しやすい

線形回帰は**直線的**な関係を表現



EBM は**曲線的**な関係も表現できる

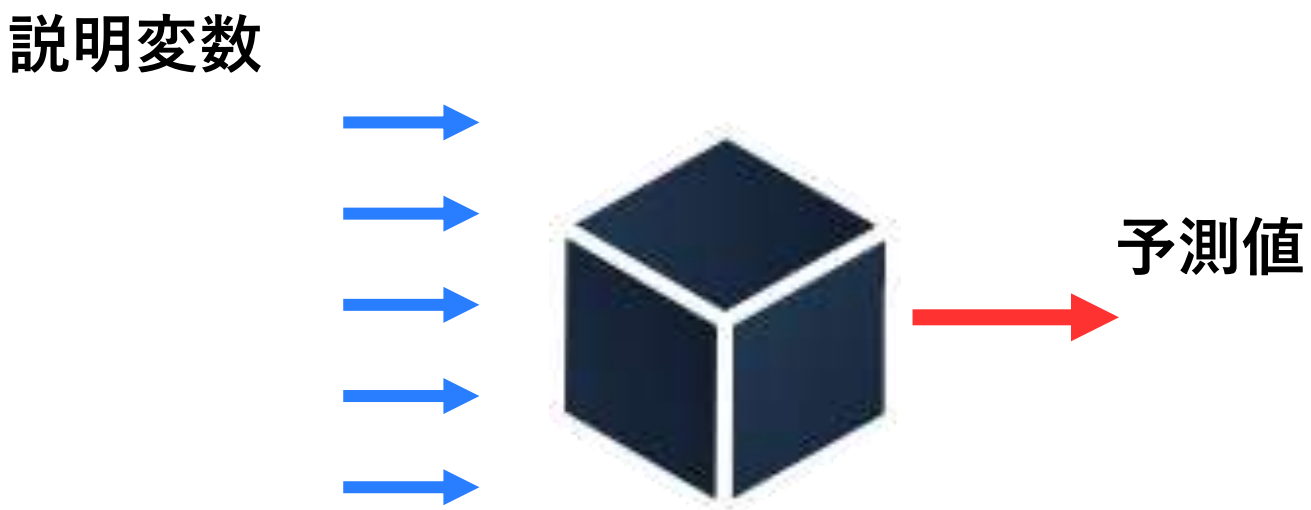


Module 2 ：
Gradient Boosting 回帰モデルの SHAP による説明

Module 3 ：
LightGBM 分類モデルの SHAP による説明と
Error Analysis + Azure ML

BlackBox モデルの説明手法

機械学習モデルに対する入力と予測値の関係をみる



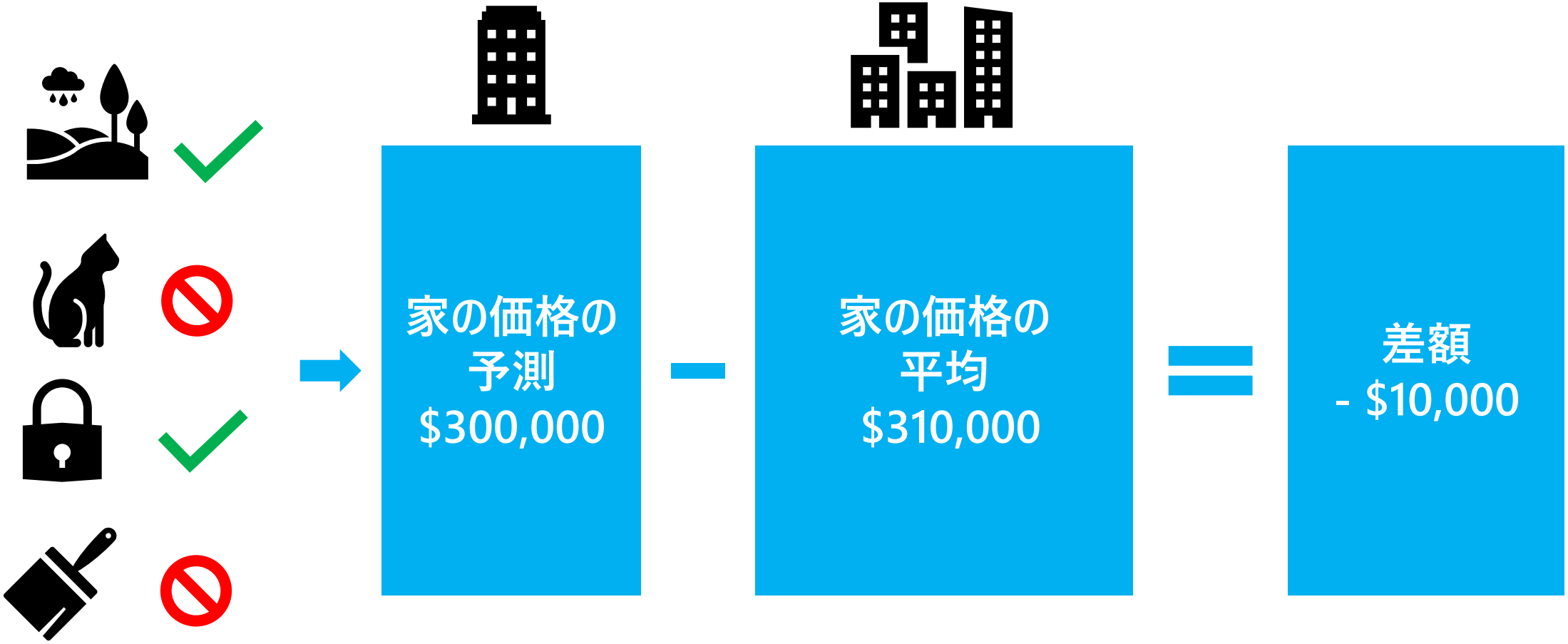
Model Agnostic : あらゆるモデルに対応できる

SHAP

シャープレイ値に基づく BlackBox モデルの説明

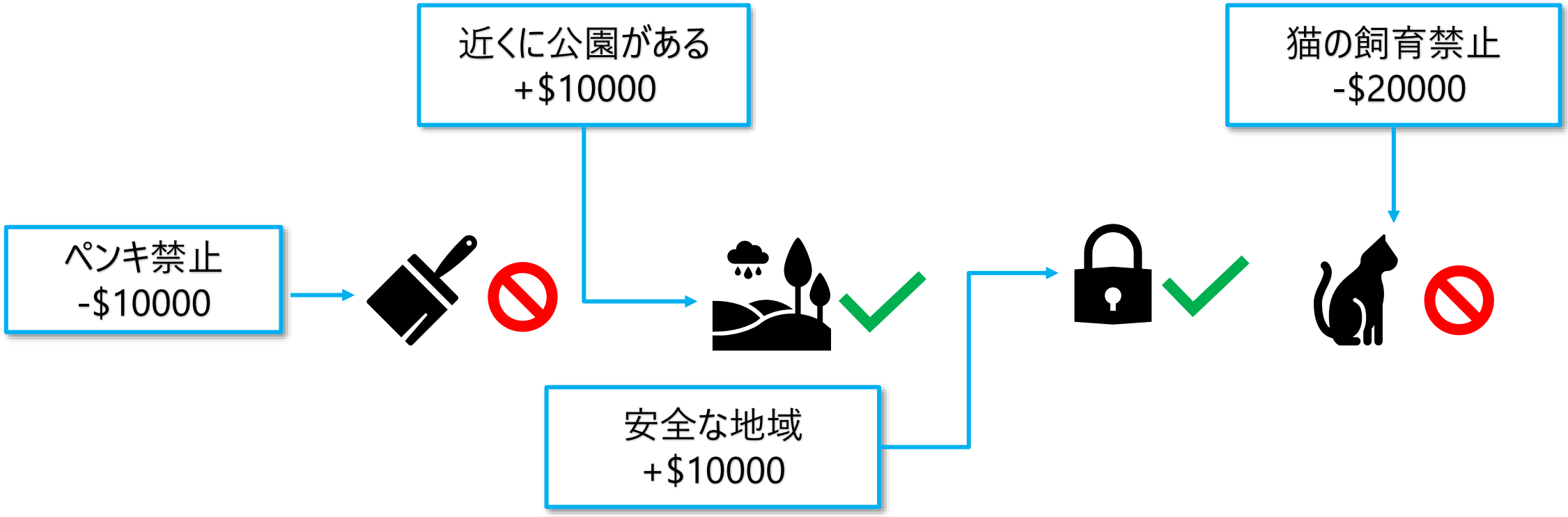
SHAP

予測値に対して、各特徴量がどのくらい寄与しているのか？



SHAP

予測値に対して、各特徴量がどのくらい寄与しているのか？

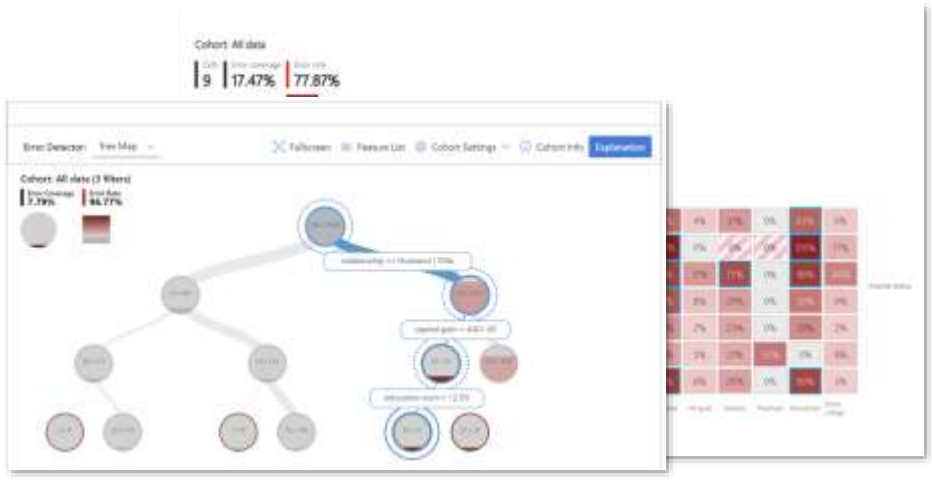


Error Analysis

集計されたモデル精度指標では捉えられないモデル誤差の特徴を分析

① Identification

誤差が大きいコホートを特定する

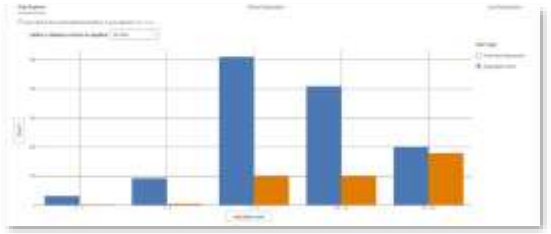


木構造で各条件下におけるエラー率・カバレッジを表示

② Diagnostics

対象のコホートを比較し深掘り分析する

データ探索



グローバル解釈

ローカル解釈

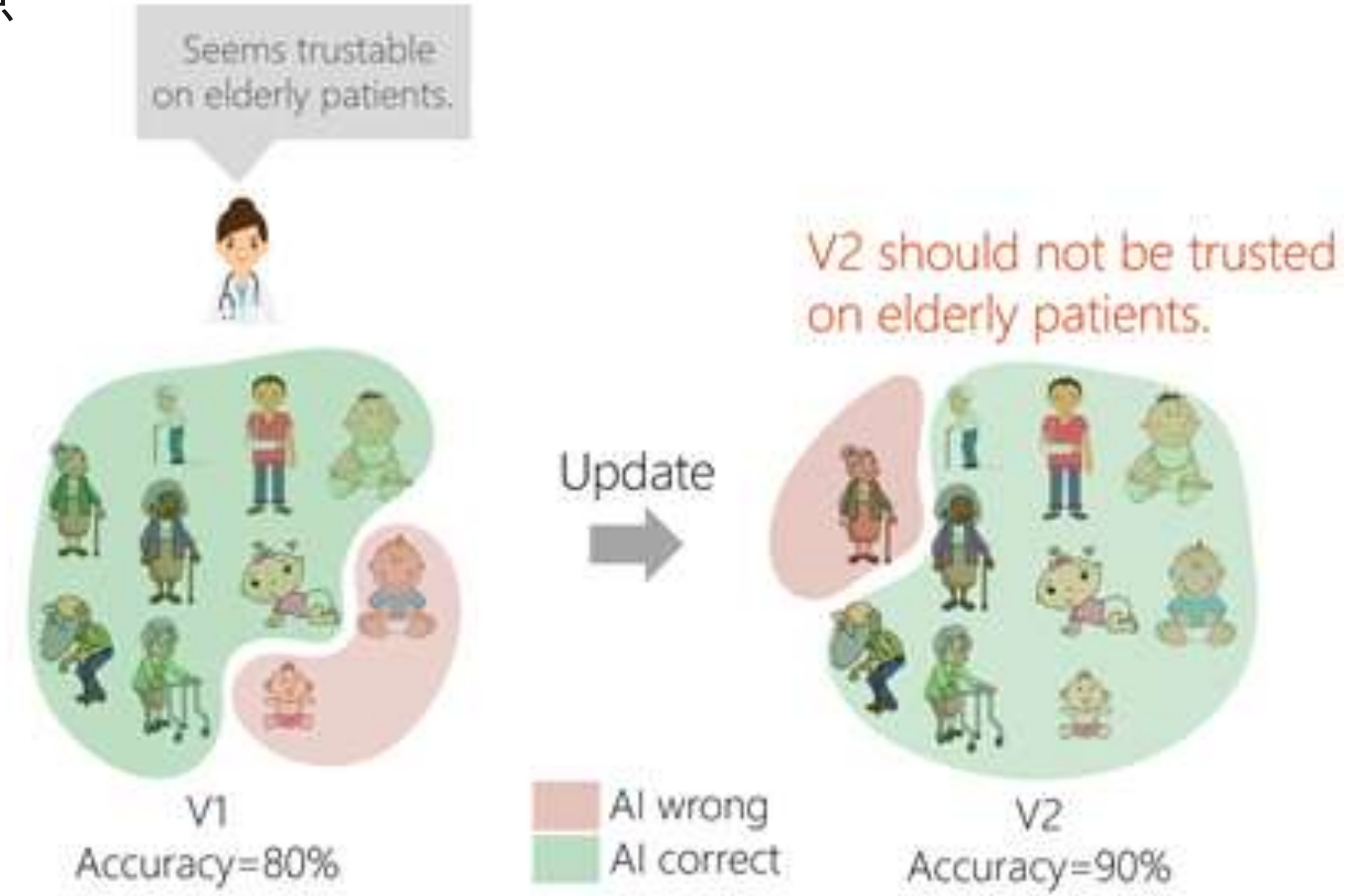
what-if 分析

参考情報

アプローチ：後方互換性

モデルの再学習による互換性の考慮

例：病理診断モデルの Accuracy をモデルの再学習に改善。しかしながら新しいモデルは高齢者の方の診断を誤ることが多くなった

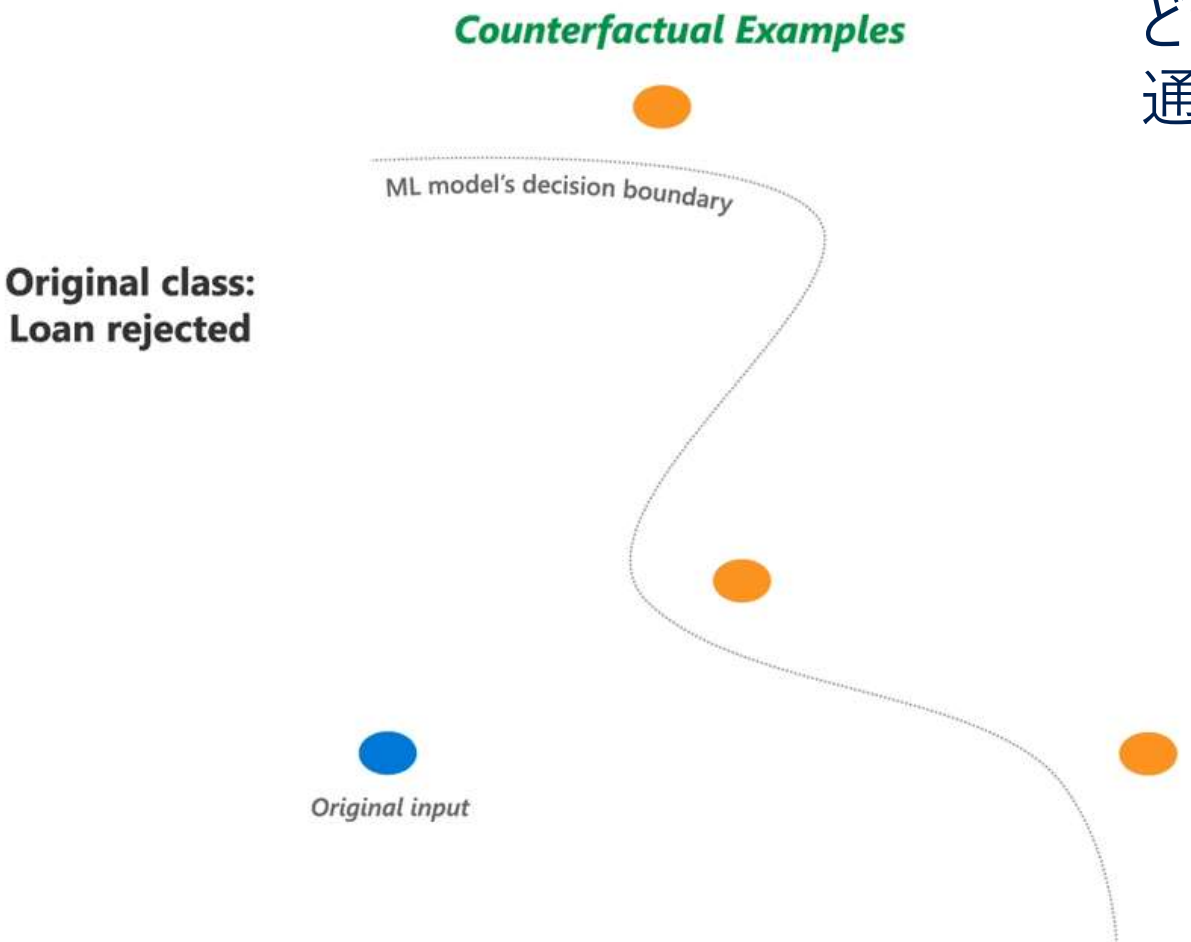


microsoft/BackwardCompatibilityML: 後方互換性を考慮したモデル学習

アプローチ：反実仮想サンプル

反実仮想サンプルによる機械学習モデルの解釈

住宅ローンモデルがローンを却下した場合、
どういった入力データであれば、ローン審査を
通過することができたのかを教えてくれる。

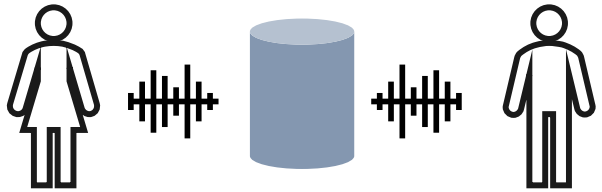


DiCE interpretml/DiCE

反実仮想によるモデル解釈

アプローチ：機械学習モデルの不公平性

機械学習モデル不公平性とは人種、性別、年齢の違いによって
ネガティブな影響を与えること



**文字起こしシステムは、女性よりも
男性の方が精度が高いかもしれません。**

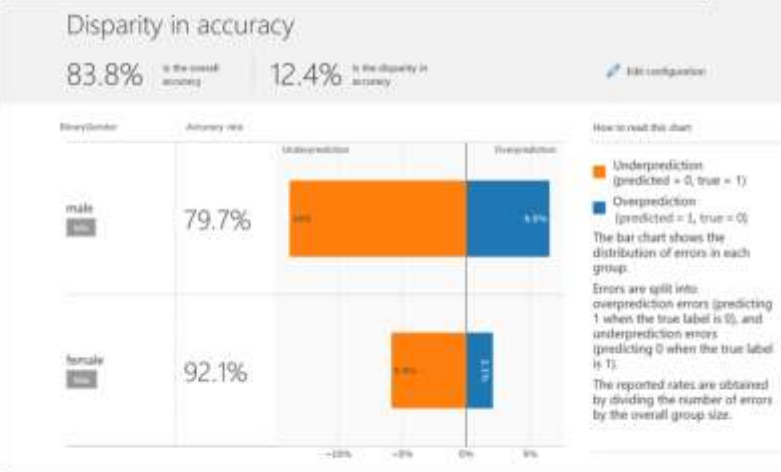
Quality-of-service harms
サービス品質の害

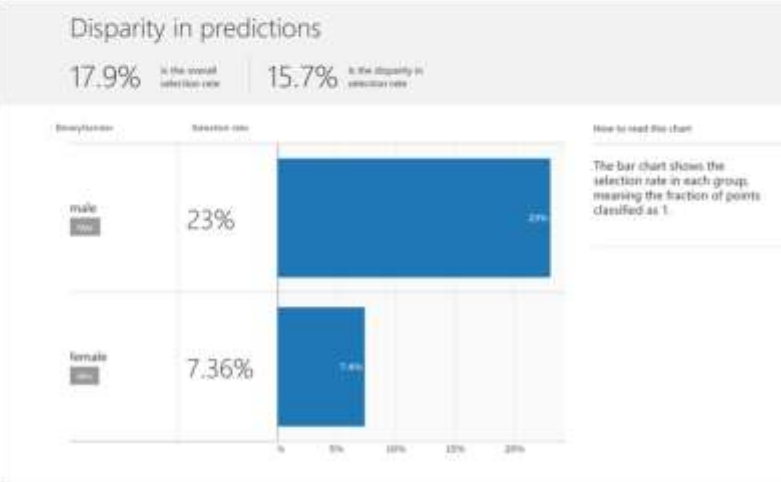
**ローンの申し込み審査において、
他のグループよりも白人男性を優先するかもしれません。**

Allocation harms
割り当ての害

センシティブな機械学習のユースケースにおいては、
機械学習モデルの公平性の評価と対策を行う必要がある。

Fairlearn





① → ②

公平性の評価：

公平性を評価する一般的なメトリックとダッシュボードを利用した Sensitive Feature の評価

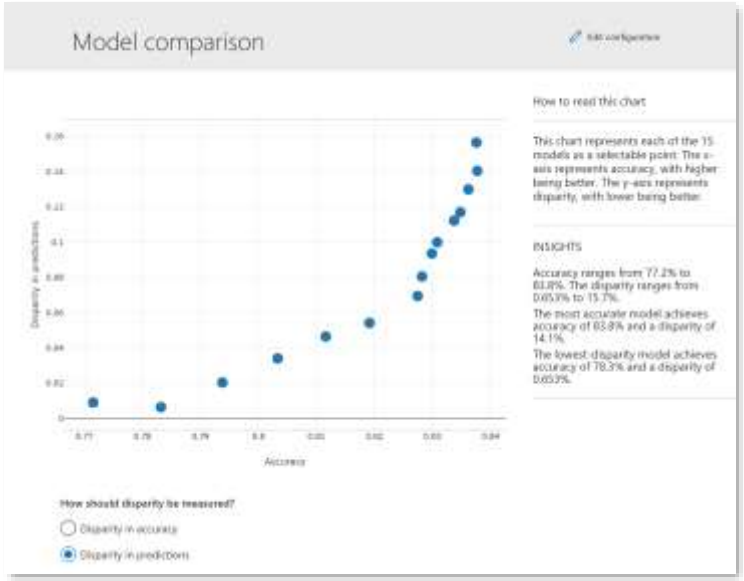
モデルのフォーマット：
scikit-learn, TensorFlow, PyTorch, Keras などに対応

メトリック:
15以上の一般的なグループを対象にした公平性メトリック

モデルの種類:
クラス分類、回帰

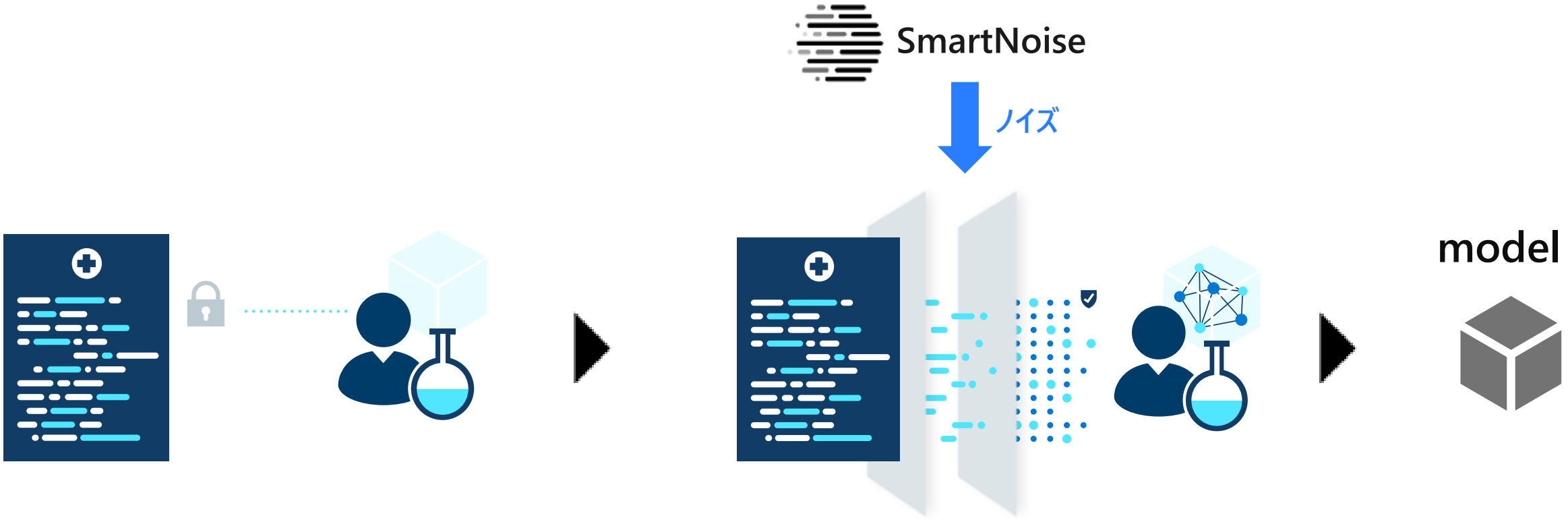
不公平性の軽減：

最先端のアルゴリズムによって
分類・回帰モデルの不公平性を軽減



Fairlearn

アプローチ：差分プライバシー



プライバシーの規制により
患者データを直接みれない

SmartNoise を用いて、データに
“統計的ノイズ” を入れて個人情報の特定を防ぐ

機械学習における差分プライバシー (プレビュー) - Azure Machine Learning | Microsoft Docs

参考ドキュメント

- konabuta/responsible-ai (github.com)
- Machine Learning Best Practices (azure.github.io)
- microsoft/machine-learning-collection (github.com)
- InterpretML : Understand Models. Build Responsibly.
- interpretML (GitHub)
- 機械学習モデル解釈ナイト (エンジニア向け)
  - BlackBox モデルの説明性・解釈性技術の実装
  - 一般化線形モデル (GLM) & 一般化加法モデル(GAM)
- Deep Learning Ditial Conference
  [Track4-1] BERT の解剖学 : interpret-text による自然言語処理 (NLP) のモデル解釈
- Microsoft AI Business School
- Microsoft 責任のある AI


Microsoft AI

Microsoft