機械学習モデル

予測モデルを“理解”するための技術

原　聡

大阪大学 産業科学研究所

自己紹介

- 原 聡、博士（工学）
  - ~2013.3, PhD@鷲尾研, 阪大
  - 2013.4~2016.3, 研究員@IBM東京基礎研
  - 2016.4~2017.8, 研究員@河原林ERATO, NII
  - 2017.9~, 助教@鷲尾研, 阪大

- 研究
  - 特徴選択
    - グラフィカルモデルの構造学習（ECML'11）
    - 異常変数の同定（AISTATS'15,17）
  - 機械学習モデルの説明
    - アンサンブル木の簡略化（AISTATS'18）
    - モデル列挙によるユーザの納得感向上（AAAI'17,18）
    - 公平性・説明性への"攻撃"（ICML'19）

自己紹介

- YouTube見てください。

今日の内容

- 機械学習界における説明技術
  - 社会的な背景
  - 研究界の動向

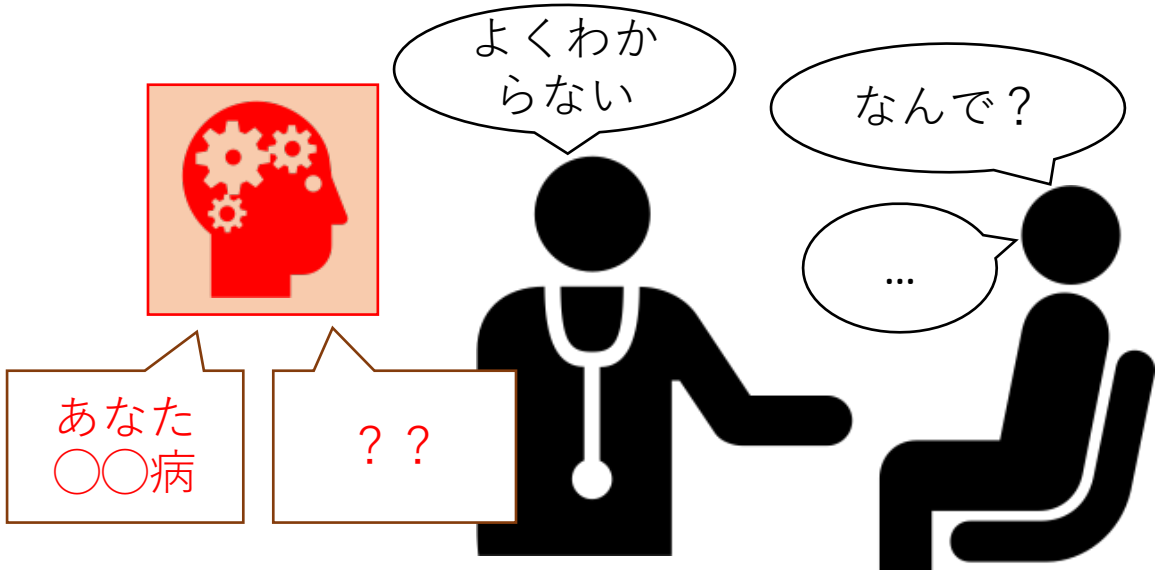
- 説明技術
  - 局所的な説明
  - 大域的な説明

- 科学的な知識獲得に向けて
  - モデルの列挙

AI/機械学習モデルは説明が苦手

- 得意なこと
  - 高い精度での予測・認識
- 苦手なこと
  - 予測・認識の判断根拠の説明

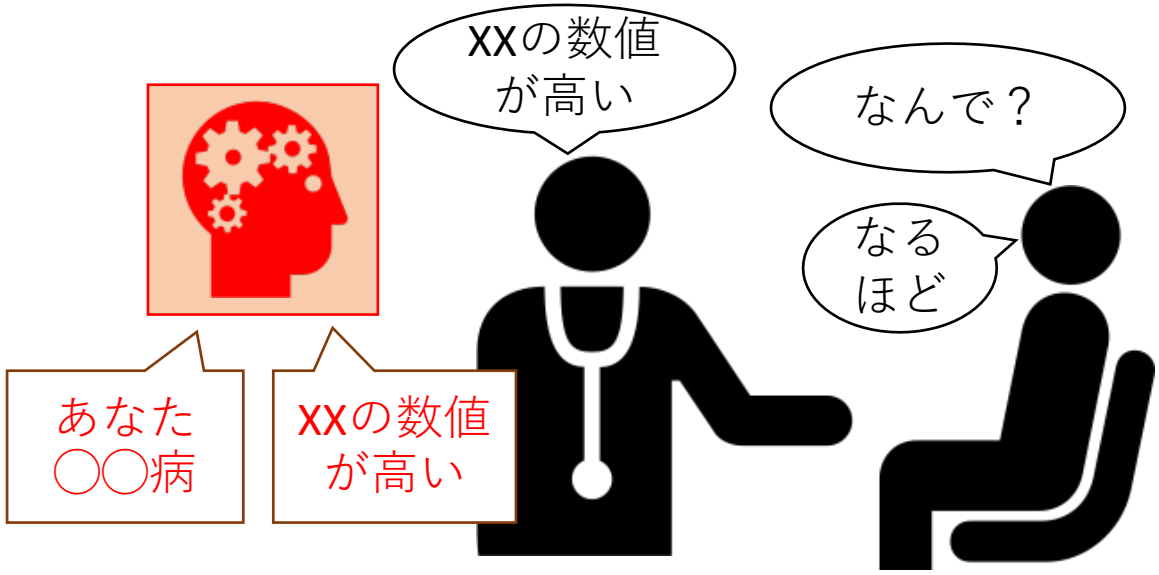
判断根拠が説明できない
→ AIへの不信 / 導入の阻害



AI/機械学習モデルは説明が苦手

- 得意なこと
  - 高い精度での予測・認識
- 苦手なこと
  - 予測・認識の判断根拠の説明

判断根拠が説明できる
→ AIへの信頼醸成の第一歩 / 利用の拡大



「説明」への社会的な要請

■ 世界的に機械学習の説明性が重要視されている。

- 日本：AI利活用原則案（総務省，2018）
  - 透明性の原則
    - AIサービスプロバイダ及びビジネス利用者は、AIシステム又はAIサービスの入出力の検証可能性及び判断結果の説明可能性に留意する。
  - アカウンタビリティ（説明責任）の原則
    - AIサービスプロバイダ及びビジネス利用者は、消費者的利用者及び間接利用者を含むステークホルダに対しアカウンタビリティを果たすよう努める。

- EU：一般データ保護規則（GDPR）
  - 機械学習を含め、ユーザに関する意思決定に説明責任を課す。

- US：説明可能AI（XAI，DARPAプロジェクト）
  - 「人間が理解し信頼できるAI」の研究開発。

【参考】AI利活用原則案（総務省，2018）

第3章　AIの利活用において留意することが期待される事項②

AI利活用原則案

国際的な議論のためのものとして、また、非規制的かつ非拘束的なもの（いわゆるソフトロー）として取りまとめ

① 適正利用の原則 ［安全］［役割分担］
利用者は、人間とAIシステムとの間及び利用者間における適切な役割分担のもと、適正な範囲及び方法でAIシステム又はAIサービスを利用するよう努める。

② 適正学習の原則 ［データ］［正当性・公平性］
利用者及びデータ提供者は、AIシステムの学習等に用いるデータの質に留意する。

③ 連携の原則 ［連携］
AIサービスプロバイダ、ビジネス利用者及びデータ提供者は、AIシステム又はAIサービス相互間の連携に留意する。また、利用者は、AIシステムがネットワーク化することによってリスクが惹起・増幅される可能性があることに留意する。

④ 安全の原則 ［安全］
利用者は、AIシステム又はAIサービスの利活用により、アクチュエータ等を通じて、利用者等及び第三者の生命・身体・財産に危害を及ぼすことがないよう配慮する。

⑤ セキュリティの原則 ［セキュリティ］
利用者及びデータ提供者は、AIシステム又はAIサービスのセキュリティに留意する。

⑥ プライバシーの原則 ［プライバシー］
利用者及びデータ提供者は、AIシステム又はAIサービスの利活用において、他者又は自己のプライバシーが侵害されないよう配慮する。

⑦ 尊厳・自律の原則 ［正当性・公平性］
利用者は、AIシステム又はAIサービスの利活用において、人間の尊厳と個人の自律を尊重する。

⑧ 公平性の原則 ［正当性・公平性］
AIサービスプロバイダ、ビジネス利用者及びデータ提供者は、AIシステム又はAIサービスの判断によって個人が不当に差別されないよう配慮する。

⑨ 透明性の原則 ［ブラックボックス化］
AIサービスプロバイダ及びビジネス利用者は、AIシステム又はAIサービスの入出力の検証可能性及び判断結果の説明可能性に留意する。

⑩ アカウンタビリティの原則 ［受容性］
AIサービスプロバイダ及びビジネス利用者は、消費者的利用者及び間接利用者を含むステークホルダに対しアカウンタビリティを果たすよう努める。

主に**便益の増進**に関係

主に**リスクの抑制**に関係

主に**信頼の醸成**に関係

（注）AIの開発において留意することが期待される事項については、本推進会議において「国際的な議論のためのAI開発ガイドライン案」を取りまとめた（『報告書2017』）。関係するステークホルダ（政府、業界団体等）が取り組む環境整備に関する課題については、第4章「今後の課題」において整理している。

【参考】EU一般データ保護規則（GDPR）

- GDPR-22

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
2. Paragraph 1 shall not apply if the decision: is necessary for entering into, or performance of, a contract between the data subject and a data controller; is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or is based on the data subject's explicit consent.
3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.
4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(2)1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

【参考】各種団体が説明技術の必要性に言及

- 人工知能技術戦略会議（内閣府）
  - 「人工知能技術戦略実行計画（案）」の中に「信頼できるAI」へのアプローチの一つとして説明技術が盛り込まれている。
- 経済産業省
  - 「次世代人工知能・ロボット中核技術開発」の資料の中に、AIの安全性の担保などへ向けたアプローチの一つとして説明技術が挙げられている。
- 科学技術振興機構（JST）
  - 研究開発戦略センターから「（戦略プロポーザル）AI応用システムの安全性・信頼性を確保する新世代ソフトウェア工学の確立」が公開された。この中で説明技術が重要技術領域の一つとして取り上げられている。
- 日本経済団体連合会
  - 「AI活用戦略」の中でAIのブラックボックス性に言及しており、説明技術の研究開発の必要性が指摘されている。

研究界の動向

- 特に2016年以降、機械学習関連の国際会議で解釈性・説明性に関する論文が増加
  - ICML, NIPSなどの機械学習のトップ会議でも解釈性・説明性のワークショップが開催されている。

- 日本語まとめ資料
  - 機械学習における解釈性（私のブックマーク）, 人工知能, Vol.33, No.3, pages 366--369, 2018.
  - ディープラーニングの判断根拠を理解する手法, Qiita記事

研究界の動向

- 「AIの説明」に関する論文数の推移

  Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)
  https://ieeexplore.ieee.org/document/8466590/

7つのレポジトリ（SCOPUS, IEEExplore, ACM Digital Library, Google Scholar, Citeseer Library, ScienceDirect, arXiv）から、説明に関連するキーワード（“intelligible”, “interpretable”, “transparency”, “black box”, “understandable”, “comprehensible”, “explainable”など）と、AI関連の語（“Artificial Intelligence”, “Intelligent system”, “Machine learning”, “deep learning”, “classifier” , “decision tree”など）を同時に含む論文の数をカウント。

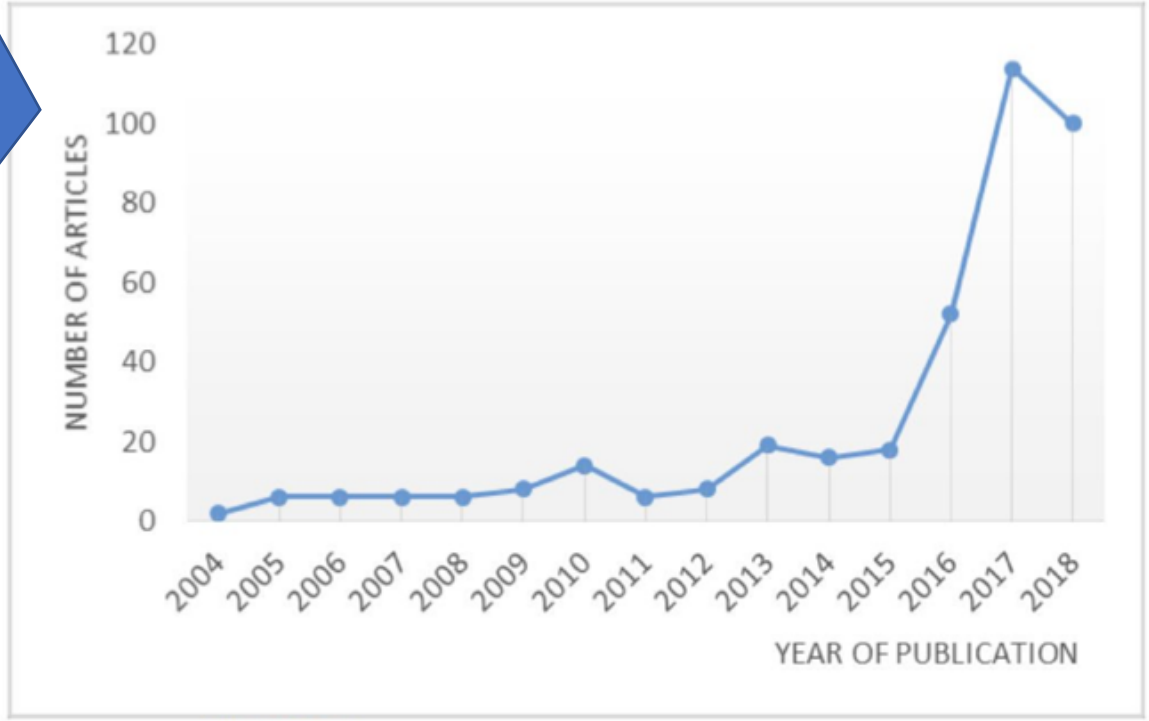


FIGURE 6. Surveyed articles by year (2004–2018)

今日の内容

- 機械学習界における説明技術
  - 社会的な背景
  - 研究界の動向

- 説明技術
  - 局所的な説明
  - 大域的な説明

- 科学的な知識獲得に向けて
  - モデルの列挙

まず初めに

- 世の中の課題を全て解決する“万能な”説明法はない。
  - 『どんな説明が必要か』はデータや応用によって異なる。
    - 現状は、研究者が「こんな説明あったら便利じゃない？」という仮説に基づいて説明法の研究開発を進めていることが多い。

- 代表的な説明法
  - 1. 重要な特徴の提示
  - 2. 重要な学習データの提示
  - 3. 自然言語による説明
  - 4. モデルの可読化
  - …

どの方法を使えば現場の具体的な課題が解決できるか、は個別に検討が必要。

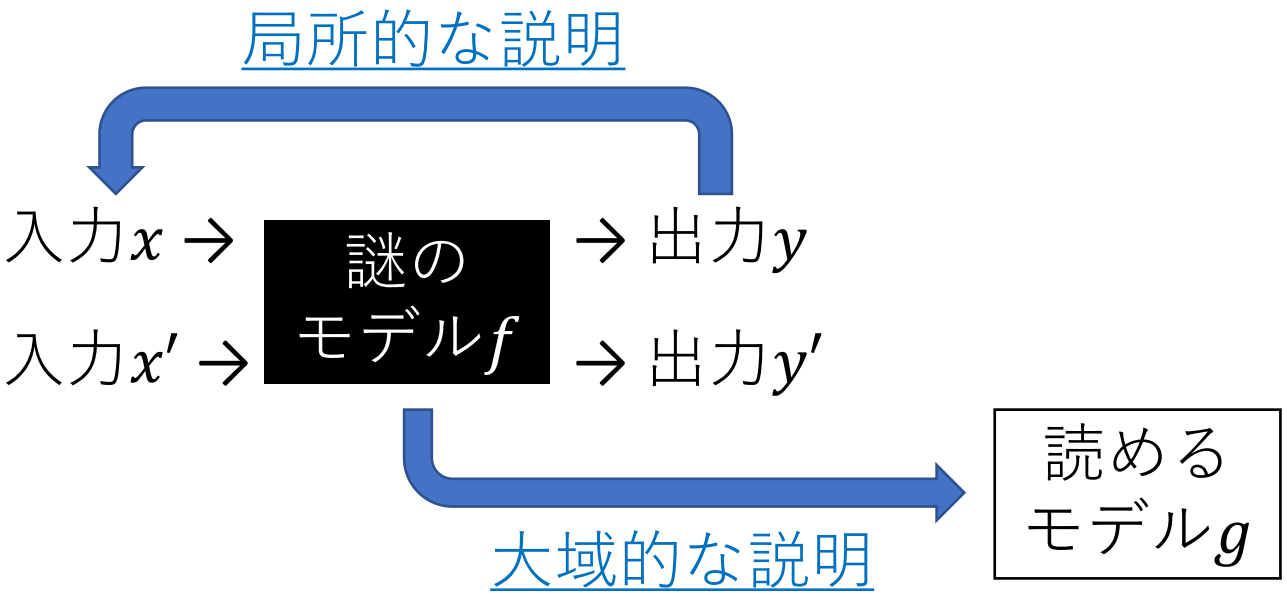
Q. 科学的な知識獲得のためには、どのようなアウトプットが得られると良いか？

→ 新しい研究の可能性

代表的な説明のアプローチ

1. 局所的な説明
   - 特定の入力に対するブラックボックスモデルの予測の根拠を提示することで説明とする方法。

2. 大域的な説明
   - 複雑なブラックボックスモデルを可読性の高い解釈可能なモデルで表現することで説明とする方法。



1. 局所的な説明

局所的な説明の代表的な手法

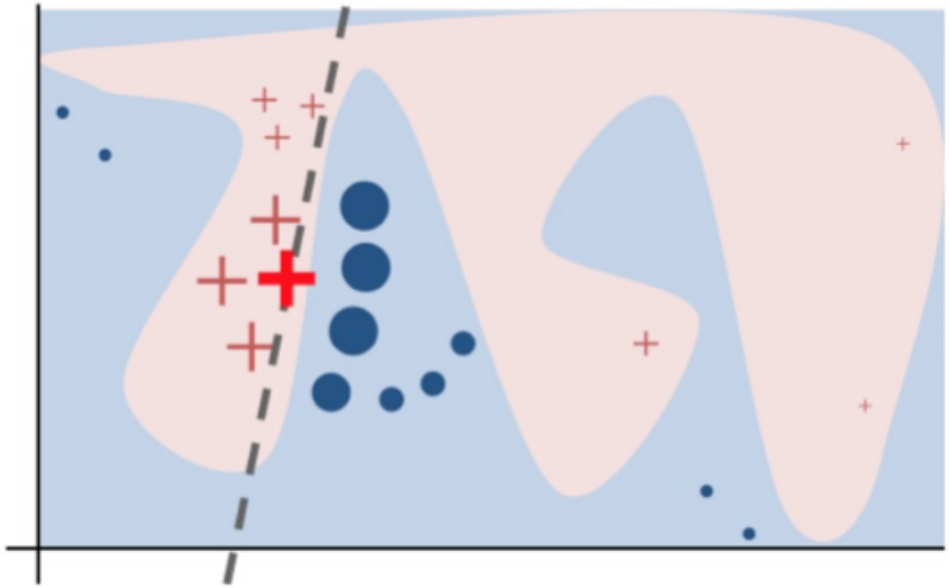
- 特定の入力に対するブラックボックスモデルの予測の根拠を提示することで説明とする方法。

- 予測の根拠となった特徴量を提示する方法
  - Why Should I Trust You?: Explaining the Predictions of Any Classifier, KDD'16 [Python実装 LIME; R実装 LIME]

- 予測の根拠となったデータを提示する方法
  - Understanding Black-box Predictions via Influence Functions, ICML'17 [Python実装 influence-release]

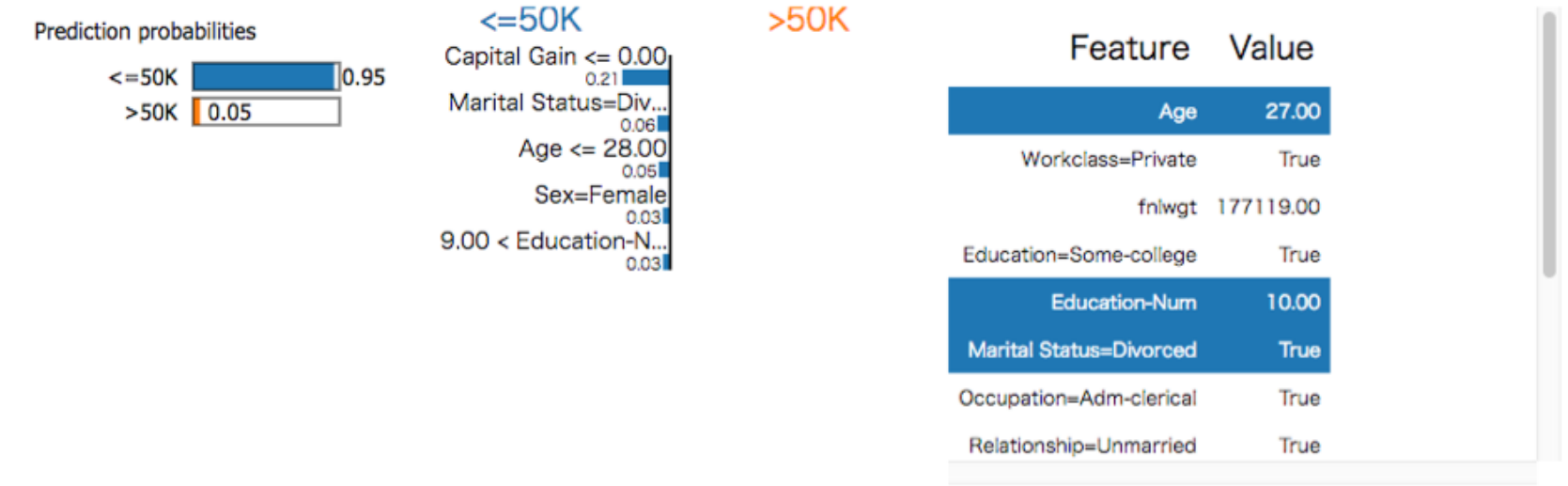
LIMEによる説明

- Why Should I Trust You?: Explaining the Predictions of Any Classifier, KDD'16 [Python実装 LIME; R実装 LIME]
  - どの特徴が予測に重要だったかを提示する。
  - モデルを説明対象データの周辺で線形モデルで近似する。
    - 線形モデルの係数の大小で、各特徴の重要度合いを測る。

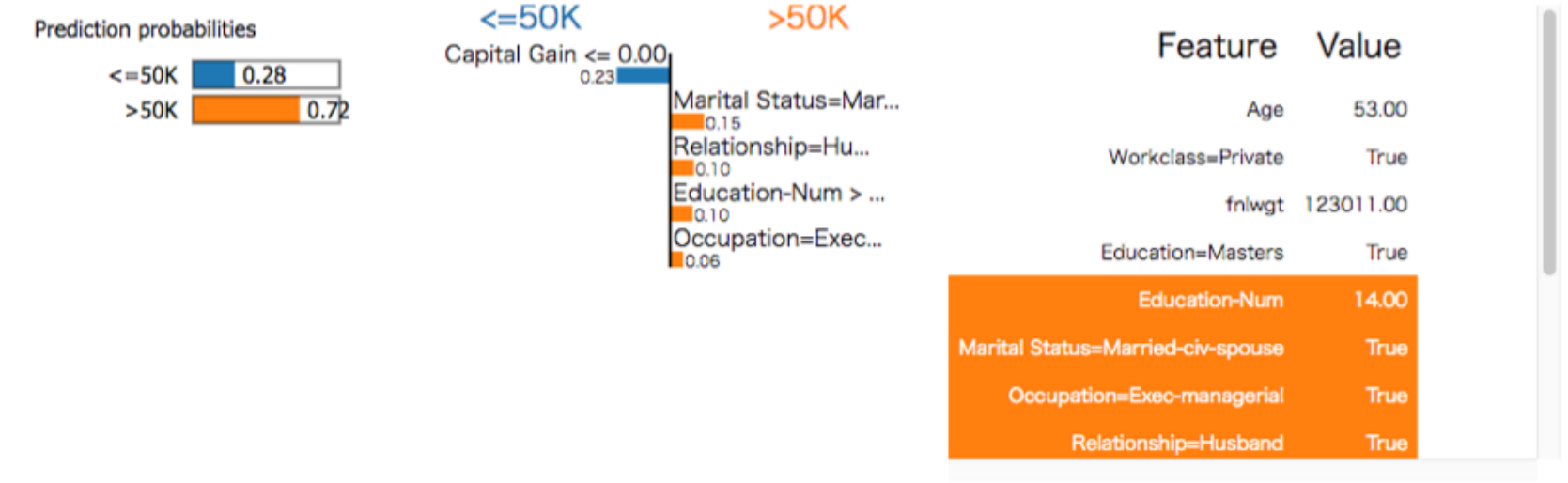


LIMEによる説明

Test Data 0, Label: <=50K



Test Data 12, Label: >50K



LIMEの応用例

- 画像認識の説明

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

- モデルのデバッグ
  - 狼 vs ハスキーの分類
  - 狼画像として、雪背景のもののみを使用。

  → LIMEにより、モデルが雪を根拠に狼を認識していることがわかる。

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

局所的な説明の代表的な手法

- 特定の入力に対するブラックボックスモデルの予測の根拠を提示することで説明とする方法。

- 予測の根拠となった特徴量を提示する方法
  - Why Should I Trust You?: Explaining the Predictions of Any Classifier, KDD'16 [Python実装 LIME; R実装 LIME]

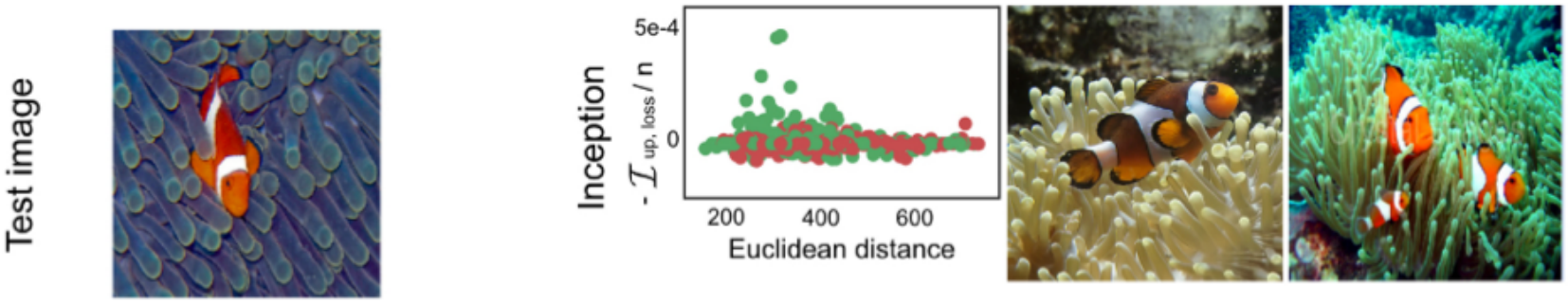
- 予測の根拠となったデータを提示する方法
  - Understanding Black-box Predictions via Influence Functions, ICML'17 [Python実装 influence-release]

influenceによる説明

- Understanding Black-box Predictions via Influence Functions, ICML'17 [Python実装 influence-release]

- あるデータ(x', y') が"無かった"としたら、対象データxでの予測はどれくらい変わるか？

予測への影響が強い
訓練画像（熱帯魚）



ラベルを予測
したいテスト
画像

予測への影響が強い
訓練画像（犬）

influenceの基本アイディア

- あるデータ(x', y') が“無かった”としたら、対象データxでの予測はどれくらい変わるか？

- モデルを $y = f(x; \hat{\theta})$, $\hat{\theta}$を学習されたパラメータとする。

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} \sum_{z=(x,y) \in D} L(z; \theta)$$

全データで学習した場合

$$\hat{\theta}_{-z'} = \operatorname{argmin}_{\theta \in \Theta} \sum_{z \in D \text{ and } z \neq z'} L(z; \theta)$$

$z' = (x', y')$が無かった場合

- $z' = (x', y')$の有無で$\hat{\theta}$はどれくらい変化するか？
  - $\hat{\theta}_{-z'} - \hat{\theta}$ を評価したい。
    - 例えば $\left\| \hat{\theta}_{-z'} - \hat{\theta} \right\|$ が大きければz'の影響は強いと言えそう。

influenceの計算方法

- $z' = (x', y')$の有無で$\hat{\theta}$はどれくらい変化するか？
  - $\hat{\theta}_{-z'} - \hat{\theta}$ を評価したい。
    - 例えば $\|\hat{\theta}_{-z'} - \hat{\theta}\|$ が大きければz'の影響は強いと言えそう。

- 愚直な方法
  - 全ての$z' = (x', y')$について$\hat{\theta}_{-z'}$を学習する。
  - 計算時間がかかりすぎる。

- influence
  - $\hat{\theta}_{-z'} - \hat{\theta}$を影響関数の方法を使って近似的に評価する。

ロバスト統計で提唱された概念

$$\hat{\theta}_{-z'} - \hat{\theta} \approx -\frac{1}{|D|} H_{\hat{\theta}}^{-1} \nabla L(z'; \hat{\theta})$$

学習の目的関数のヘシアン

influenceの応用

- Data Poisoning
  - 指定した対象データへの影響の強い学習データに敵対的ノイズをのせた、“敵対的”学習データを作る。
  - “敵対的”学習データで学習したモデルは、指定した対象データで間違えるようになる。

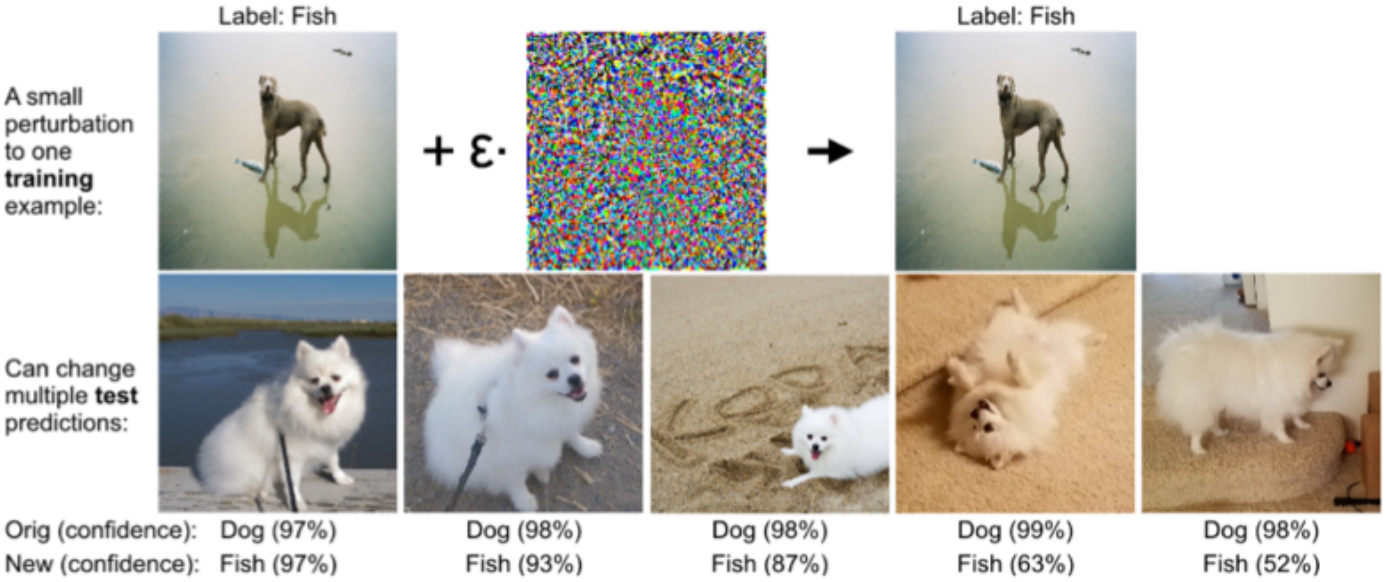


Figure 5. **Training-set attacks.** We targeted a set of 30 test images featuring the first author's dog in a variety of poses and backgrounds. By maximizing the average loss over these 30 images, we created a visually-imperceptible change to the particular training image (shown on top) that flipped predictions on 16 test images.

局所的な説明法のまとめ

- 特定の入力に対するブラックボックスモデルの予測の根拠を提示することで説明とする方法。

- 予測の根拠となった特徴量を提示する方法
  - LIME：どの可読特徴が予測に重要だったか提示する。

- 予測の根拠となったデータを提示する方法
  - influence：影響関数の方法を使うことで、各データの影響度合いを提示する

局所的な説明:その他色々

- “改善アクション”の提示
  - Interpretable Predictions of Tree-based Ensembles via Actionable Feature Tweaking, KDD’17
  - モデルの出力を特定のクラスに変えるための“改善アクション”をユーザに提示する方法。

「ローン組みたい」

「働けばいいのか…！」

現在

✕ 30歳・フリーター・年収200万円

もし…

○ 20歳・フリーター・年収200万円

○ 30歳・会社員・年収300万円

2. 大域的な説明

大域的な説明の代表的な手法

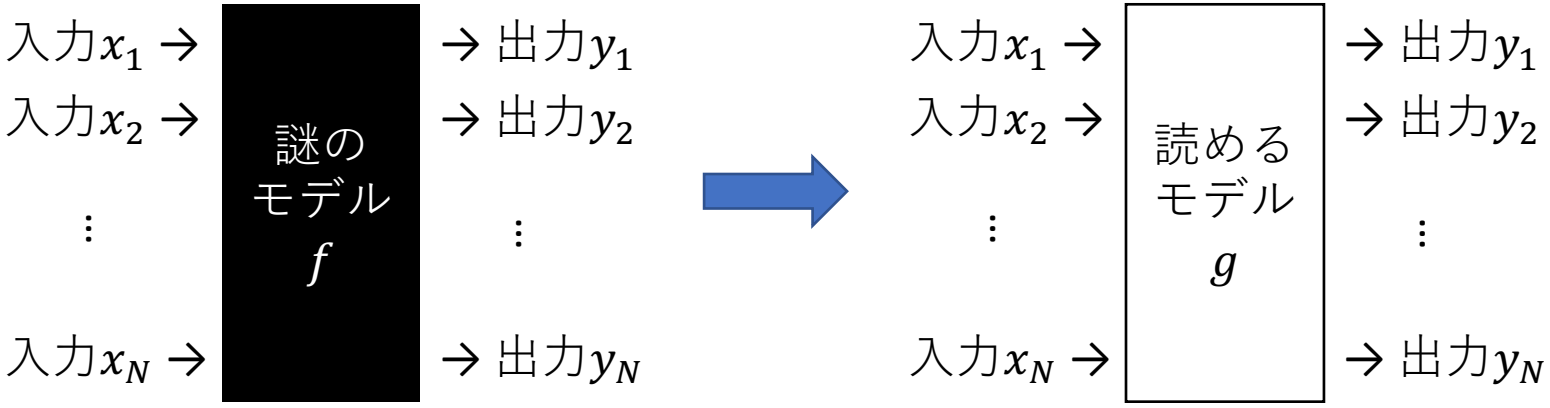
- 複雑なブラックボックスモデルを可読性の高い解釈可能なモデルで表現することで説明とする方法。

- Born Again Trees
- Making Tree Ensembles Interpretable: A Bayesian Model Selection Approach, AISTATS'18 [Python実装 defragTrees]

BATreesによる説明

- Born Again Trees
  - 複雑なモデルを“読めるモデル”（決定木）で近似する。

- 方法
  - 学習されたモデルfを使って擬似訓練データを生成する。
    - 入力はランダムにシャッフルして生成。
    - ラベルはモデルfを使って付与。
  - 生成された擬似訓練データを使って決定木を学習する。



defragTreesによる説明

- Making Tree Ensembles Interpretable: A Bayesian Model Selection Approach, AISTATS'18 [Python実装 defragTrees]
  - 複雑なモデルを複数のルールで近似する。
  - 近似したルールは人間に可読。

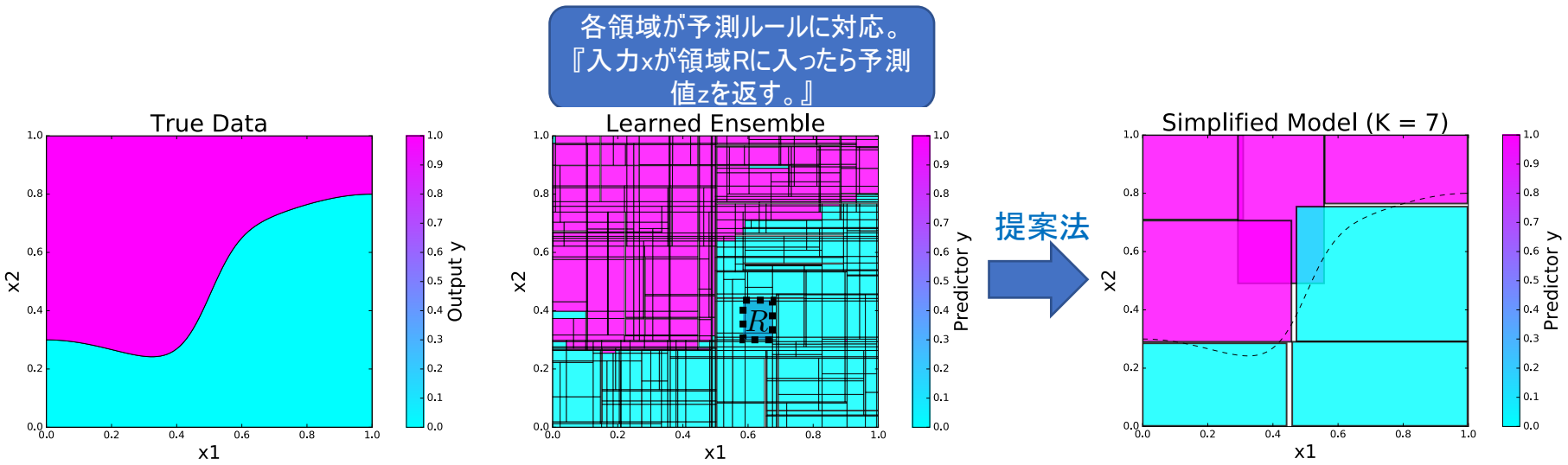
- 例：個人情報から収入の多寡の予測の“説明”

収入：少 when Relationship ≠ Not-in-family, Wife Capital Gain < 7370	収入：多 when Relationship ≠ Not-in-family Capital Gain >= 7370
収入：少 when Relationship ≠ Not-in-family, Unmarried Capital Gain < 5095 Capital Loss < 2114	収入：多 when Relationship = Not-in-family Country ≠ China, Peru Capital Gain < 5095
収入：少 when Relationship ≠ Not-in-family Country ≠ China Capital Gain < 5095	収入：多 when Relationship ≠ Not-in-family Capital Gain >= 7370
⋮	⋮

defragTreesによる説明

- モデルを少数の領域（予測ルール）で近似、簡略化することでモデルの判断プロセスを可読化する。

- defragTreesのイメージ：



アンサンブル木は領域が細分化されている。数千個の領域（予測ルール）を解釈するのは困難。

細分化されたモデルを少数の領域（予測ルール）で近似、簡略化すればモデルを解釈できる。

今日の内容

- 機械学習界における説明技術
  - 社会的な背景
  - 研究界の動向

- 説明技術
  - 局所的な説明
  - 大域的な説明

- 科学的な知識獲得に向けて
  - モデルの列挙

機械学習モデルの列挙

原 聡[1]、石畠 正和[2]、前原 貴憲[3]

1） 大阪大学 産業科学研究所
2） NTT CS研
3） 理研AIP

機械学習のゴールは“高い予測精度”だけではない。

- 科学分野では、機械学習に求められるのは予測精度だけではない。

- 科学では機械学習モデルによる学習を通じて、データの背後の仕組みの解明を試みる。
  - 機械学習モデルに、データの背後の仕組みを教えて欲しい。

機械学習のゴールは“高い予測精度”だけではない。

- 科学分野では、機械学習に求められるのは予測精度だけではない。

- 科学では機械学習モデルによる学習を通じて、データの背後の仕組みの解明を試みる。
  - 機械学習モデルに、データの背後の仕組みを教えて欲しい。

科学としては、背後の仕組みの解明がゴール。
予測だけでは物足りない。

→機械学習モデルに“説明”してほしい。

科学の助けとなる有用な知見を発掘するには

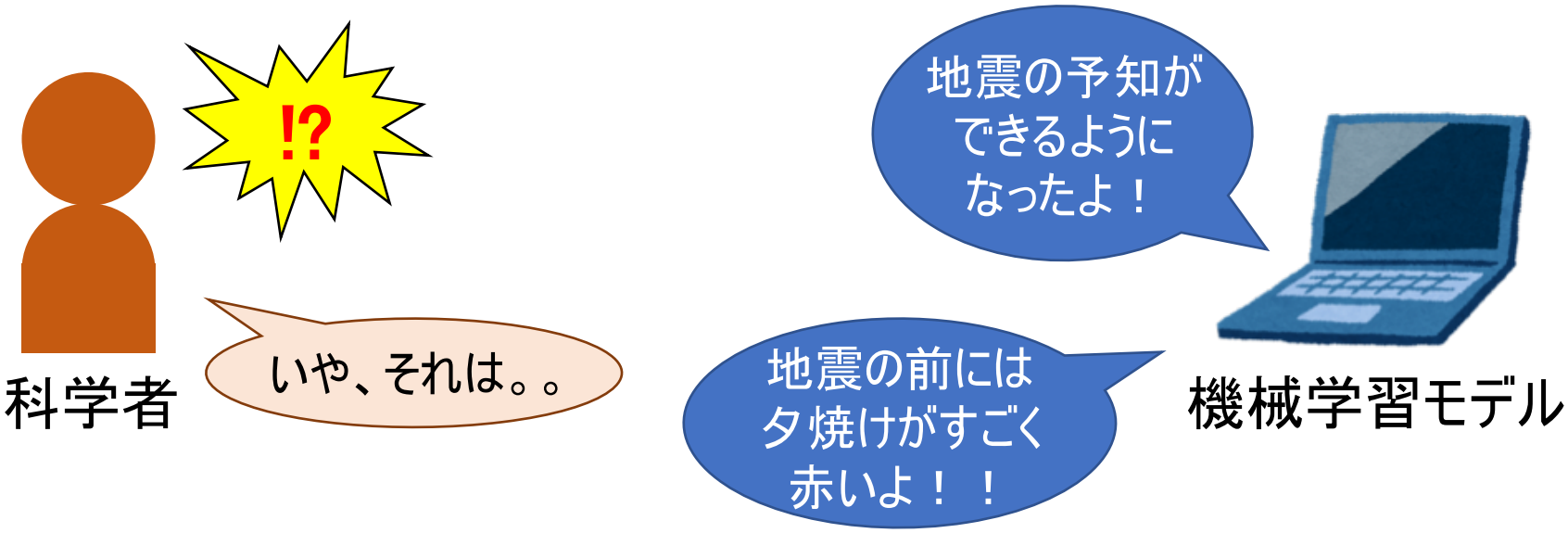
回答の一つ『解釈性の高いモデルを使う。』

- 線形モデル（Lasso、スパースモデリング）
  - 回帰係数から、どの特徴が重要かを読み取れる。
- ルールモデル（ルールリスト、ルールセット、決定木）
  - If-Thenルールでモデルが記述されるため可読性が高い。
  - 仕組みを言語化しやすい。

解釈性の高いモデルなら何でも良いか？

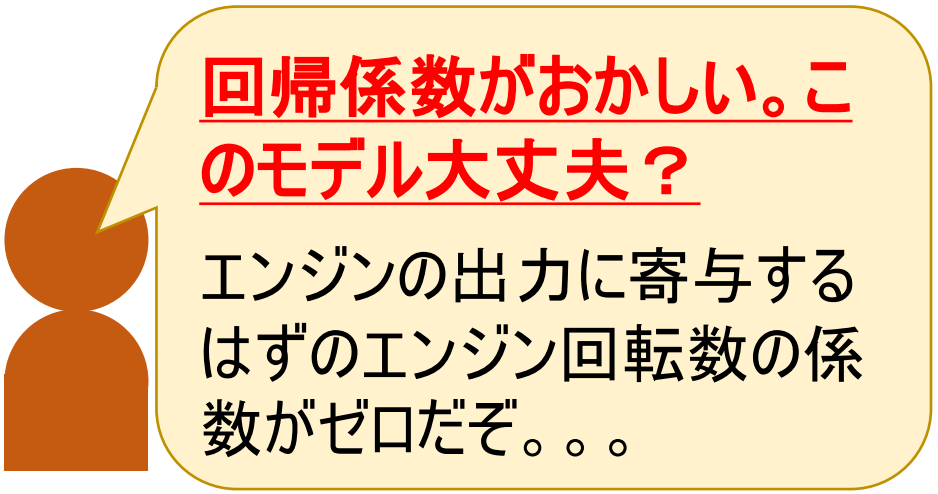
- 時に、機械学習は受け入れがたいモデルを作ってしまう。

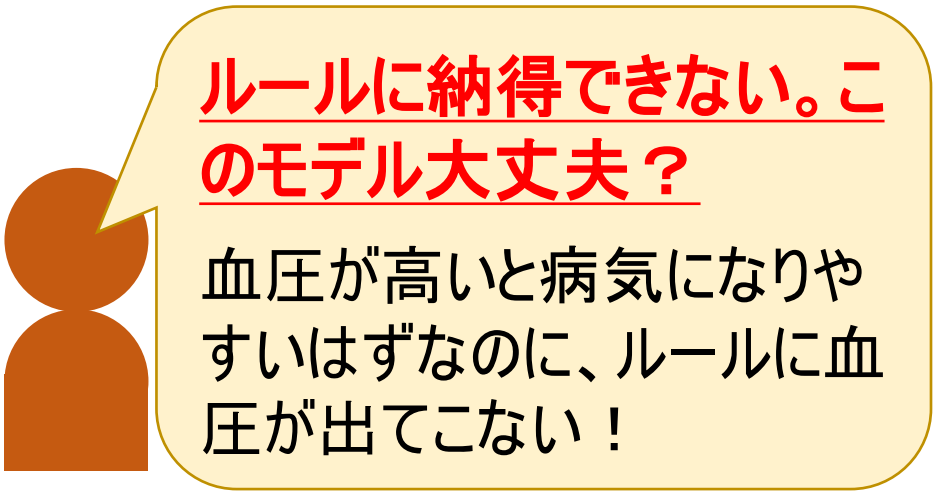
- **機械学習は間違える。**
  - **機械学習に完璧はありえない。データのバイアス、モデルのバイアス等から、間違いは起こりえる。**



“検討に値するモデル”が得られないと困る。

- データ分析でよくある問題
  - 解釈性の高い線形モデルやルールモデルを使っても。。。





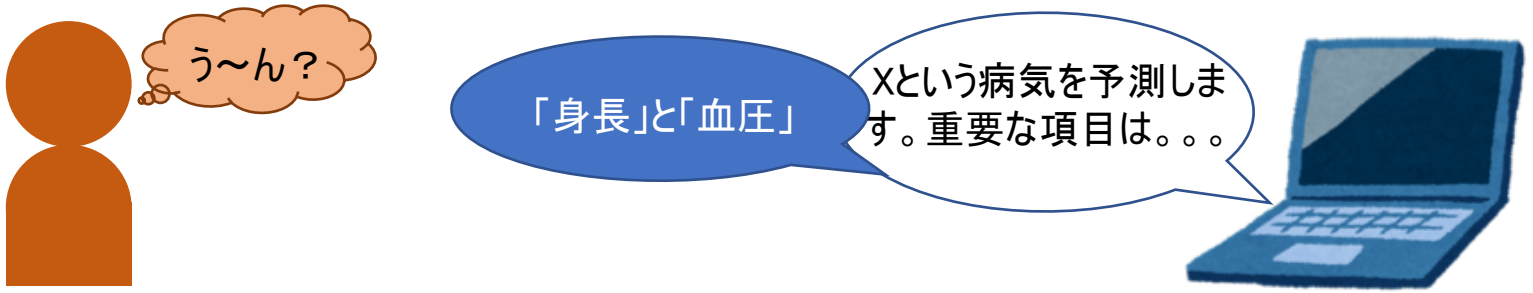
- 現状、このような事態にはデータ分析者が専門家の知見をヒアリングして、知見を反映するように適宜モデルを修正している。

“検討に値するモデル”を見つけるのは難しい。

- なぜ難しいか？
  - 例え予測精度が高くてもナンセンスなモデルは存在する。
  - 専門家の知見とある程度整合的なモデルでないと検討する価値を見出せない。

- 本研究のアイディア
  - “良いモデル”を複数見つけて、専門家に提示したらどうか？

  →**「モデルを複数列挙する問題」**を考える。

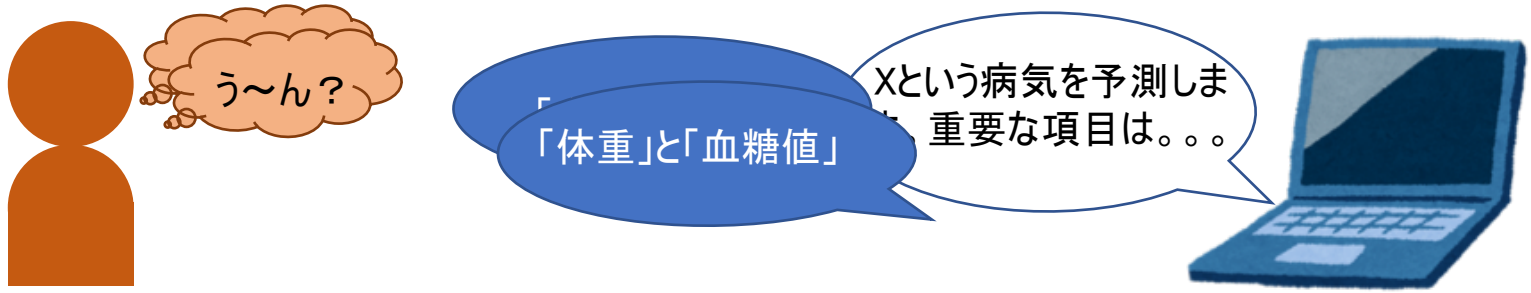


“検討に値するモデル”を見つけるのは難しい。

- なぜ難しいか？
  - 例え予測精度が高くてもナンセンスなモデルは存在する。
  - 専門家の知見とある程度整合的なモデルでないと検討する価値を見出せない。

- 本研究のアイディア
  - “良いモデル”を複数見つけて、専門家に提示したらどうか？

  →**「モデルを複数列挙する問題」**を考える。



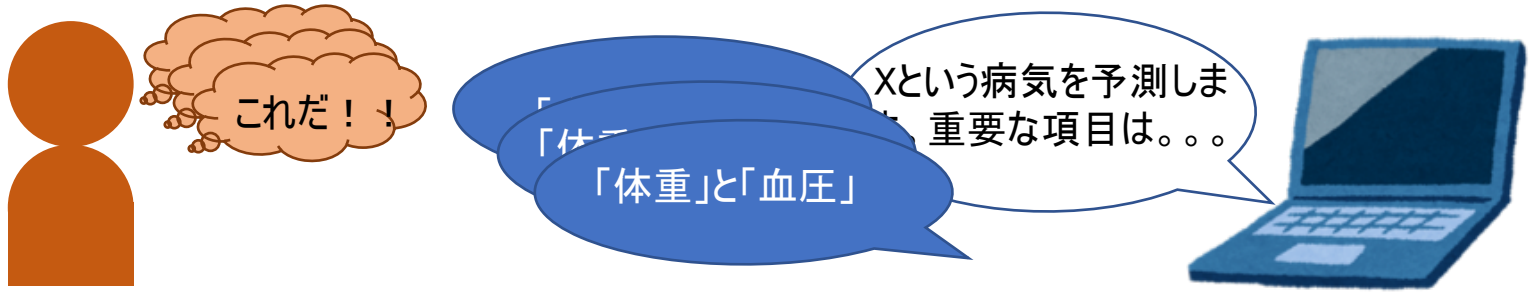
専門家

“検討に値するモデル”を見つけるのは難しい。

- なぜ難しいか？
  - 例え予測精度が高くてもナンセンスなモデルは存在する。
  - 専門家の知見とある程度整合的なモデルでないと検討する価値を見出せない。

- 本研究のアイディア
  - “良いモデル”を複数見つけて、専門家に提示したらどうか？

  →**「モデルを複数列挙する問題」**を考える。



本研究の貢献

複雑なモデルはここでは考えない。

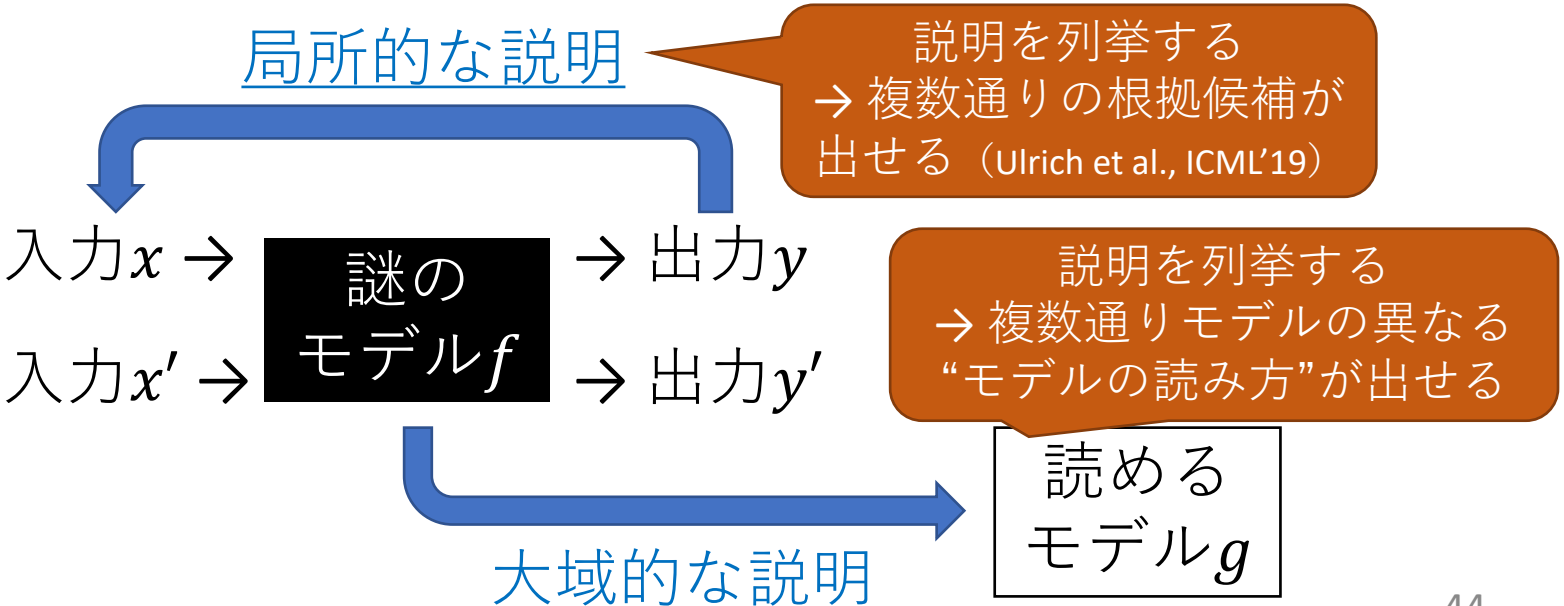
説明性の高いモデル を **複数列挙する** 方法を提案する。

- 提案法の嬉しさ1：データ理解の促進
  - 科学的な知識発見のために、有望な仮説が複数得られる。
- 提案法の嬉しさ2：ユーザの納得感向上
  - モデル候補を複数用意することで、ユーザがそこからより納得感の高いモデルを選べる。

発展：説明＋列挙

1. 局所的な説明
   - 特定の入力に対するブラックボックスモデルの予測の根拠を提示することで説明とする方法。

2. 大域的な説明
   - 複雑なブラックボックスモデルを可読性の高い解釈可能なモデルで表現することで説明とする方法。



最後に

- 機械学習界における説明技術
  - 社会的な背景
  - 研究界の動向
- 説明技術
  - 局所的な説明
  - 大域的な説明
- 科学的な知識獲得に向けて
  - モデルの列挙

最後に: 最近の動向

- 近似的な説明への批判
  - 提案されている説明法の多くは、モデルを“近似的に”読めるようにすることで説明を生成する。
  - 近似的な説明では、モデルと説明の間にギャップが生じる。
  - つまり、近似的な説明は正しくない可能性がある。
    - 実際、説明法の中には“正しくない”ものがあることが知られている。

モデルの説明はあくまでも現象を説明する仮説。
過度の信頼は禁物。

最後に: 今後の展望

- 科学的な知識獲得には既存の説明法で十分か？
  - LIME/influence などで、十分な知識・仮説が獲得できるか？

- 科学だからこそ、の研究の可能性はあるか？
  - 科学だからこその難しさは？
  - 科学に特化した説明法の可能性は？
    - 問題設定、データ、モデル、手法などによる違い？