資料4

「AIの説明」の現状とこれから

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）
    - 深層学習モデルの注目点の推定（ongoing）

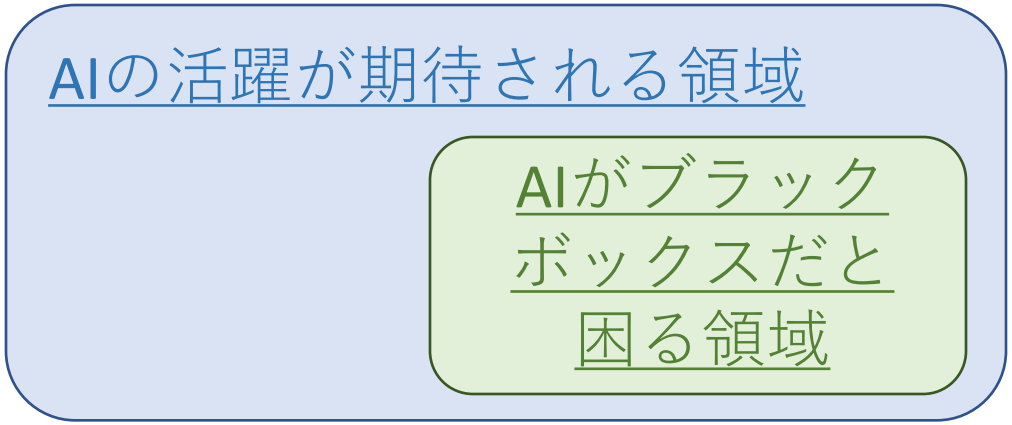
おことわり – 1

- 本資料では**「AI」=「機械学習モデル」**を前提として話を進める。
  - 機械学習モデル: コンピュータのプログラムで、特に所与の学習用データをもとにある特定の指標（e.g. 画像認識精度）について最適化されたもの。
  - 機械学習モデルの例
    - 犬と猫の画像それぞれ100枚から、分類精度が最大になるように最適化された犬猫画像分類器。
- 「汎用AI」などは対象外。

おことわり – 2

- 本資料では**「AIがブラックボックスだと困る状況」**を対象に話を進める。

- 「AI」も「AIの説明」も、あくまでも課題解決のためのツール・手段。適材適所が大前提。
  - AIがブラックボックスでも困らない事例にまで、説明を求める必要はない。
  - 「説明できないブラックボックスなAIは全てダメ」と断じる意図はない。過度に不安を助長するのは望ましくない。



まとめ

- 現状、「AIを説明する方法」の研究は増加傾向にあるが、直接的にユーザの不安の解消へは繋がっていない。

- 研究が世の中のニーズとズレている可能性がある。
  - アカデミアの研究者の「こんな説明あったら便利じゃない？」という仮説に基づく説明法の研究が多い。

- 現場からの「ニーズの発信」が大事。
  - 漠然として不安を、具体的な問題に落とし込むことで、技術的に解決可能かもしれない。
  - 不安のまま抱え込んでも、誰もハッピーにならない。

- まずは、『説明法の導入が役立った具体的な事例』を一つ作って、世の中に発信したい。

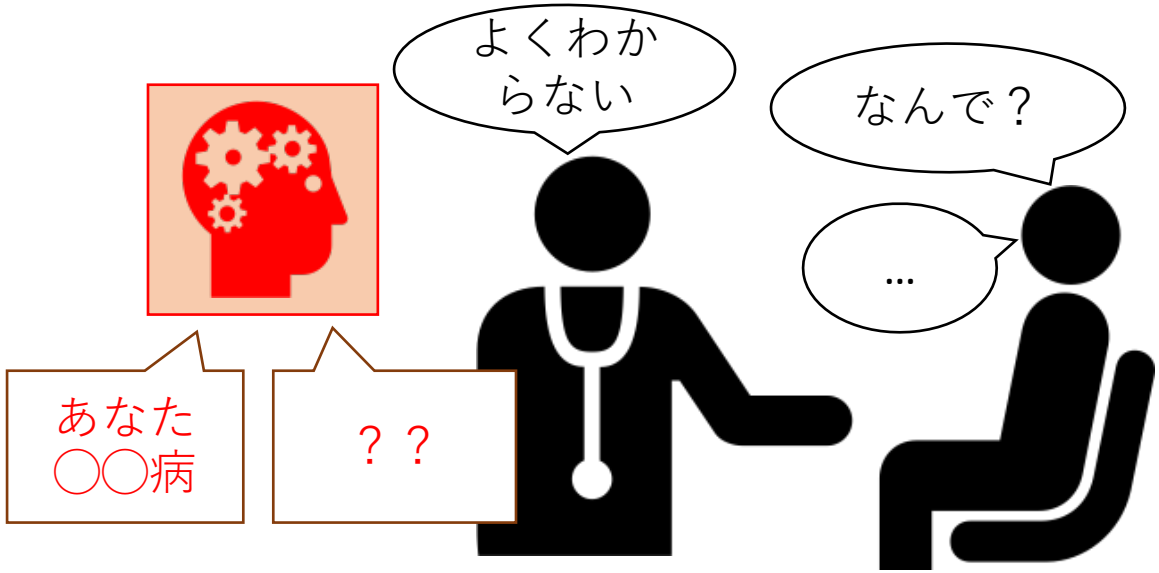
アウトライン

- 「AIの説明」への社会的な要請
- 「AIの説明」の代表的な研究
- 「AIの説明」の現状とこれから

現在のAI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

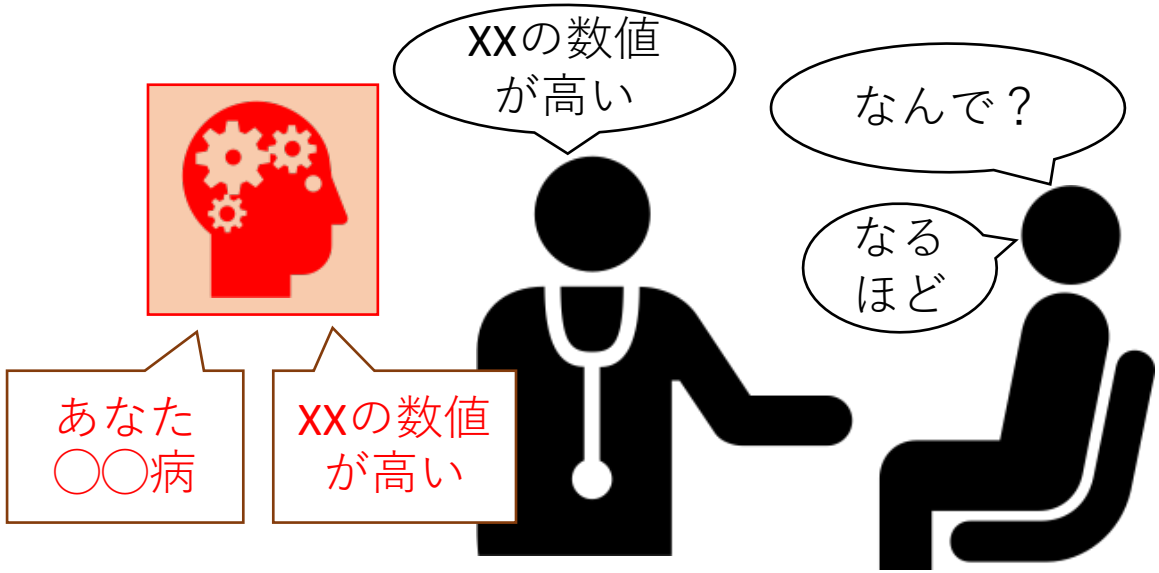
判断根拠が説明できない
→ AIへの不信 / 導入の阻害



現在のAI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



人間の意思決定補助にAIを使うには説明が必要

- 病気の診断は「医師が下す」もの。
  - AIと医師で診断が食い違ったら？
    - AIの判断根拠がわからないと、そもそも原因の検討もできない。
    - 根拠不明、とAIを無視するなら、そもそもAI導入のメリットがない。

- 金融サービス提供者は「顧客への説明の義務」がある。
  - 貸倒を防ぐためには厳格な審査が必須。AIの得意分野。
  - 「あなたには融資できません」「なぜ？」
    - 理由を説明してくれないサービス事業者は信頼されない。

- その他：司法、教育、などなど

「AIの説明」への社会的な要請

- 世界的に「AIの説明」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）
    - データに基づく意思決定について、ユーザの権利を保護し適切な介入を保証する責任をサービス提供者に課す。
  - 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.

アウトライン

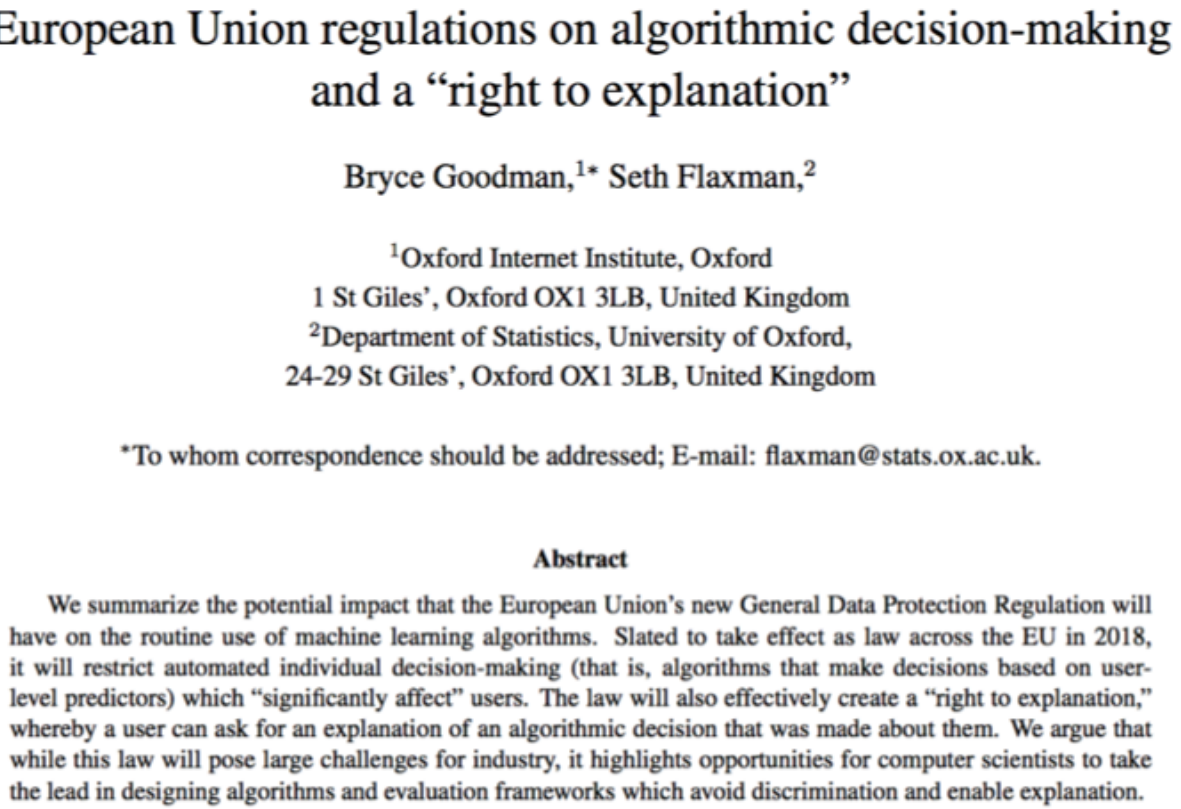
- 「AIの説明」への社会的な要請
- 「AIの説明」の代表的な研究
- 「AIの説明」の現状とこれから

研究界の動向

- 2016年以降、機械学習関連の国際会議で「AIの説明」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「AIの説明」に関するワークショップが開催されている。

- 2016年時点で、GDPRに対する準備の必要性を説く論文が発表されている。

https://arxiv.org/abs/1606.08813



European Union regulations on algorithmic decision-making and a "right to explanation"

Bryce Goodman,[1]* Seth Flaxman,[2]

[1]Oxford Internet Institute, Oxford
1 St Giles', Oxford OX1 3LB, United Kingdom
[2]Department of Statistics, University of Oxford,
24-29 St Giles', Oxford OX1 3LB, United Kingdom

*To whom correspondence should be addressed; E-mail: flaxman@stats.ox.ac.uk.

Abstract

We summarize the potential impact that the European Union's new General Data Protection Regulation will have on the routine use of machine learning algorithms. Slated to take effect as law across the EU in 2018, it will restrict automated individual decision-making (that is, algorithms that make decisions based on user-level predictors) which "significantly affect" users. The law will also effectively create a "right to explanation," whereby a user can ask for an explanation of an algorithmic decision that was made about them. We argue that while this law will pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks which avoid discrimination and enable explanation.

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。

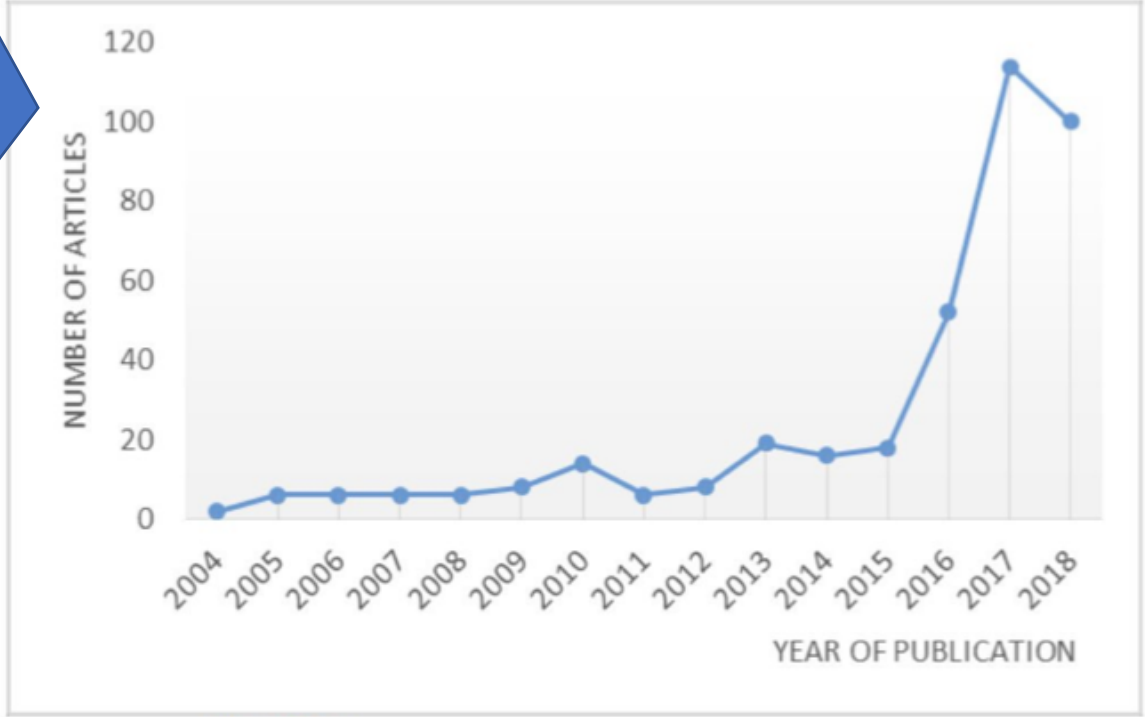


FIGURE 6. Surveyed articles by year (2004–2018)

「説明法」の研究

- 世の中の課題を全て解決する“万能な”説明法はない。

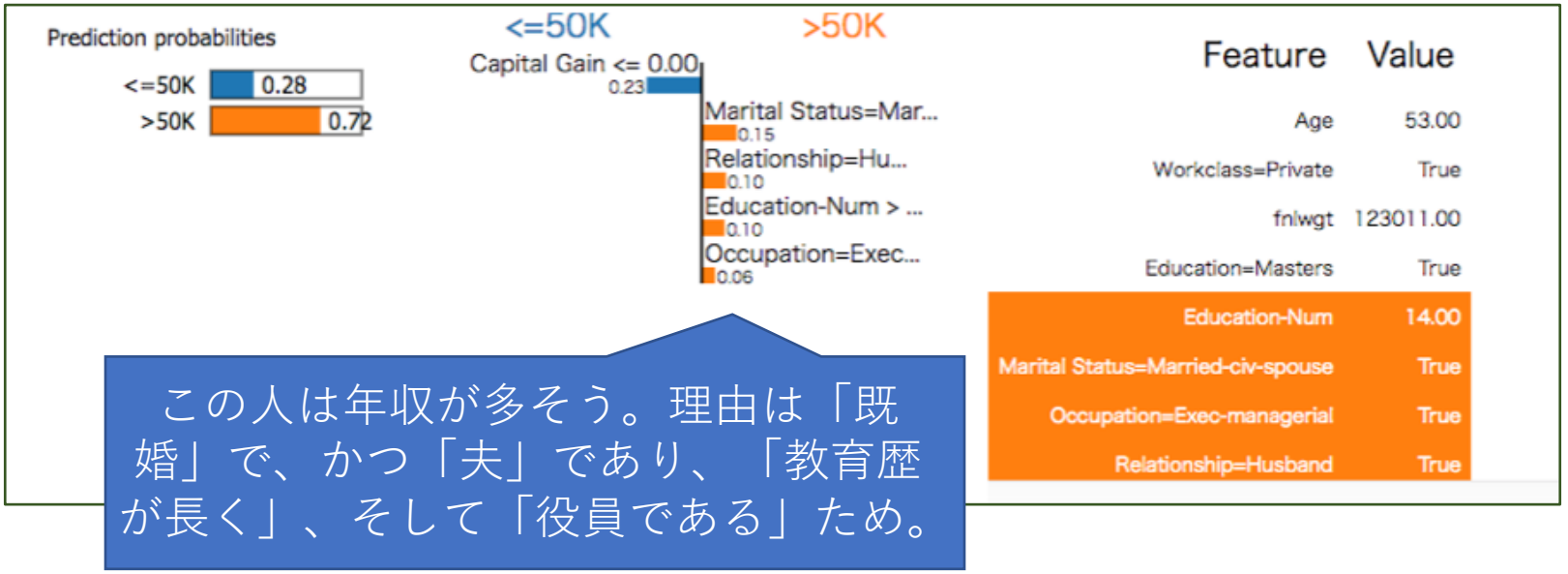
- 『どんな説明が必要か』はデータや応用によって異なる。

- 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. AIの可読化
  - 4. 自然言語による説明
  - …

代表的な説明法 – 1

- **『データのどの特徴が予測・認識に重要だったか』**を説明として提示する方法。

- 例:「年収の多寡の推定AI」の説明

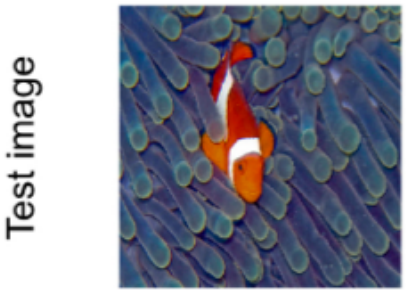


Why Should I Trust You?: Explaining the Predictions of Any Classifier, KDD'16.

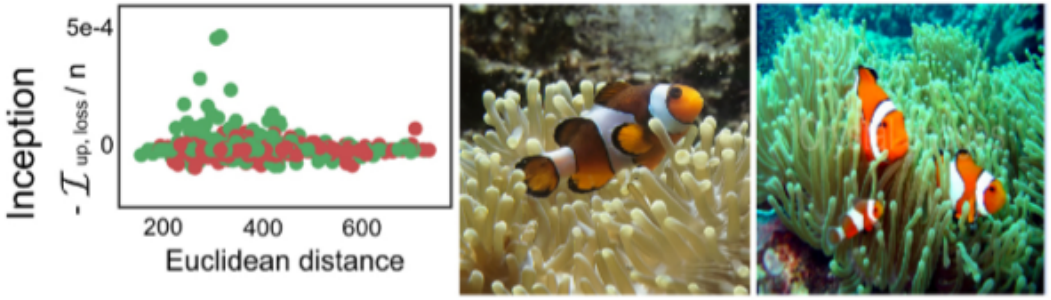
代表的な説明法 – 2

- **『どの学習データが予測・認識に重要だったか』**を説明として提示する方法。
  - ある学習データ(x', y') が“無かった”としたら、データxの予測はどれくらい変わるか？

- 例：画像認識の説明

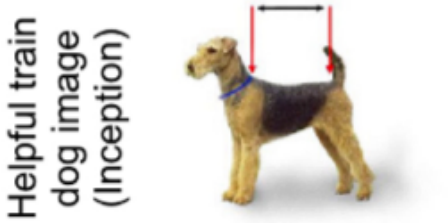


ラベルを予測したい画像



予測への影響が強い学習画像（熱帯魚）

予測への影響が強い学習画像（犬）



Understanding Black-box Predictions via Influence Functions, ICML’17.

代表的な説明法 – 3

- 予測・認識のプロセスを**『可読な表現で記述する』**ことでAIの説明とする方法。

- 例:「年収の多寡の推定AI」の説明
  - 複雑なAIを人間が読める簡単なモデルへ書き換える。

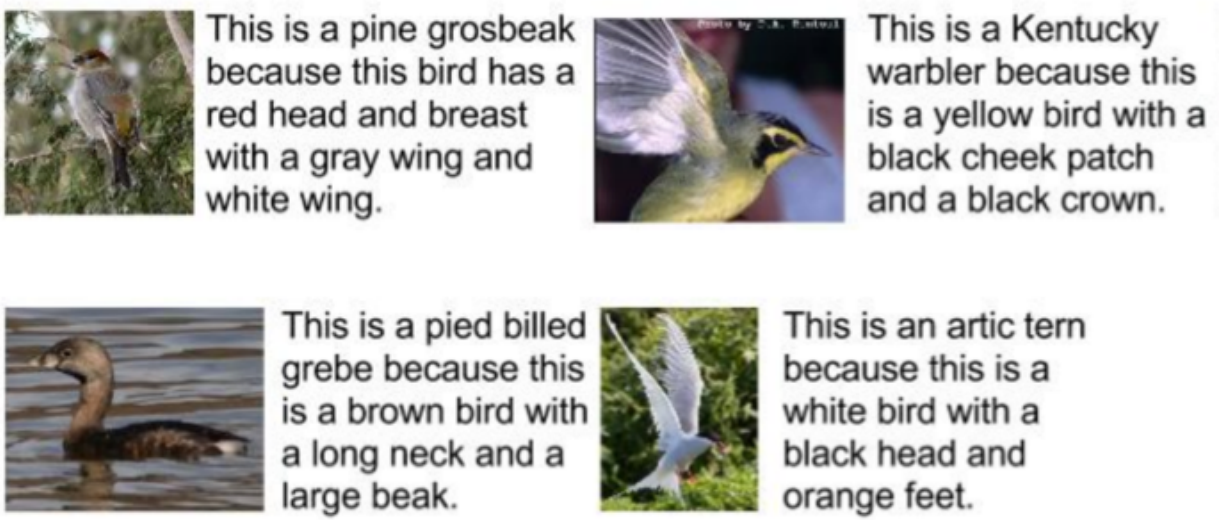
収入：少 when Relationship ≠ Not-in-family, Wife Capital Gain < 7370	収入：多 when Relationship ≠ Not-in-family Capital Gain >= 7370
収入：少 when Relationship ≠ Not-in-family, Unmarried Capital Gain < 5095 Capital Loss < 2114	収入：多 when Relationship = Not-in-family Country ≠ China, Peru Capital Gain < 5095
収入：少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

Making Tree Ensembles Interpretable: A Bayesian Model Selection Approach, AISTATS'18.

代表的な説明法 – 4

- **『データのどの特徴が予測・認識に重要だったか』**を**自然言語で説明文**として提示する方法。

- 例：画像認識の説明



Generating Visual Explanations, ECCV'16.

アウトライン

- 「AIの説明」への社会的な要請
- 「AIの説明」の代表的な研究
- 「AIの説明」の現状とこれから

AIへの不安は解消されたか？

- 2016年以降、様々な「AIを説明する方法」が提案された。
- しかし、AIへの不安が（少なくとも直接的には）解消されたようには見えない。
- むしろ、（原の体感では）「AIを説明する方法」についての技術・研究相談を受ける機会が増えた。

なぜ？
何が起きているのか？

AIへの不安はなぜ解消されていないか？

- 可能性A 原の体感がずれてる。社会全体では不安解消に向かっている。

- 可能性B まだ説明法の技術がきちんと広まっていない。あとは傍観していれば、時間が解決してくれる。

- 可能性C 研究の方向性と世の中のニーズがズレている。このズレを補正しないと、不安は解消されない。

原の意見
『B, Cが9割』

可能性B: 説明法の技術が広まっていない。

- 論文多すぎ問題
  - 様々な状況下での「多種多様な説明法」が提案されている。
  - どの論文を読んだら、現場の課題が解決できるか？

- 優れた説明法が埋もれてしまっている可能性もある。

誰かが良い説明法を掘り起こして、「これいいよ！」って言うと、みんなハッピーになるかも。

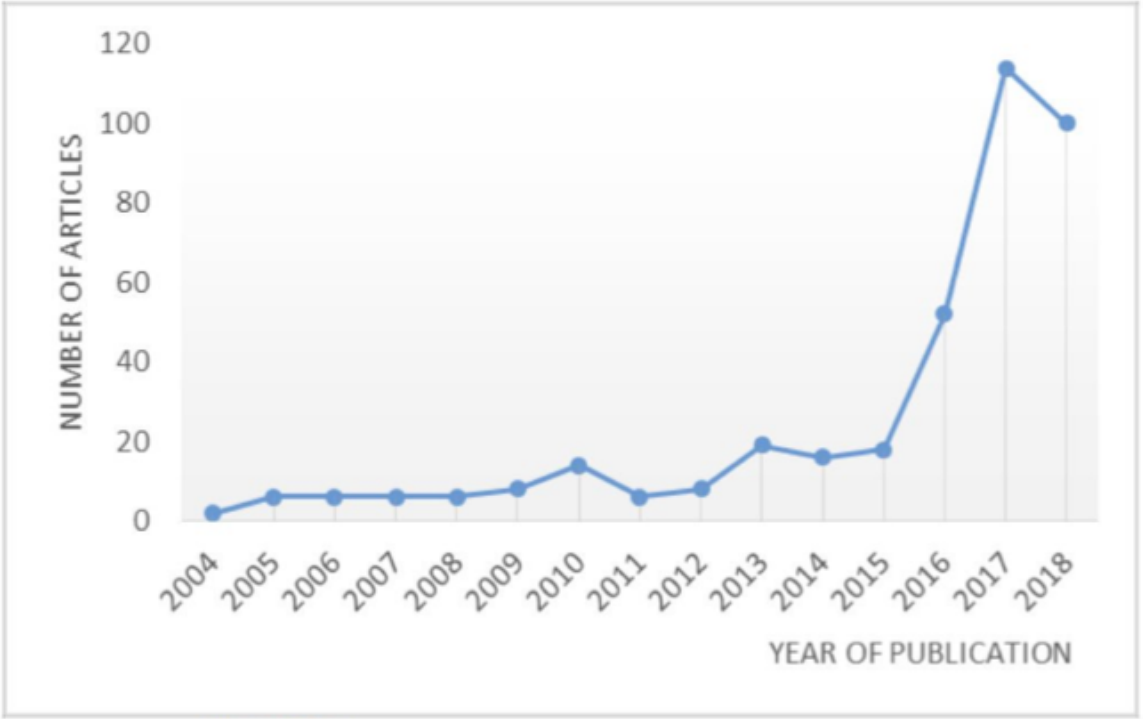


FIGURE 6. Surveyed articles by year (2004–2018)

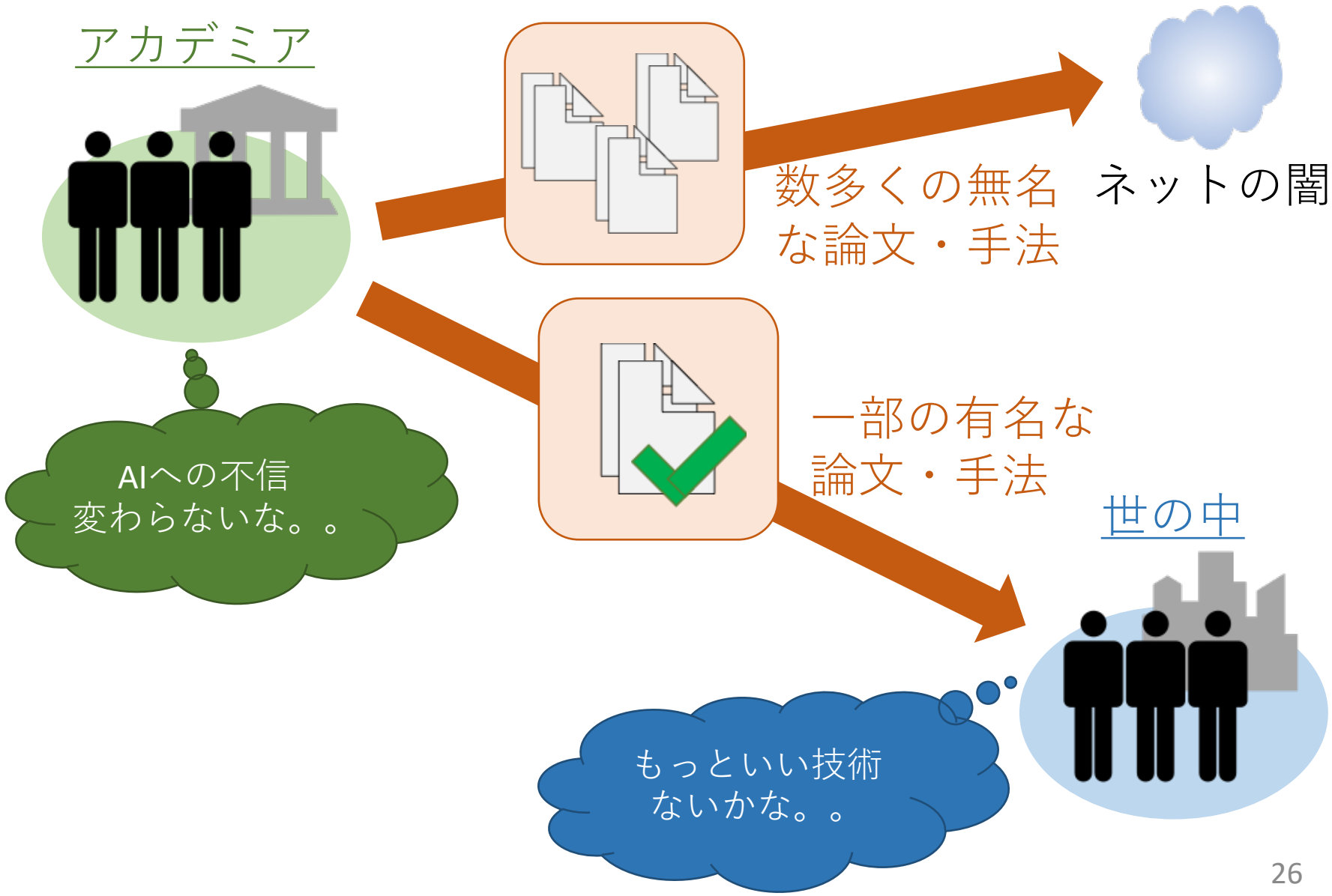
可能性C: 研究と世の中のニーズがズレている。

- 説明法研究の論文の多くはアカデミアから。
  - WHI'18: ICML'18のワークショップ
    - 論文 14/17 がアカデミアから。
    - 企業からはGoogle Brain, IBM Research, Microsoft。
  - XAI'18: ICML'18のワークショップ
    - 論文 26/30 がアカデミアから。
    - 企業からはSamsung Research, Spotify, OTTO , 米軍（？）。

- 現在の研究の多くはアカデミアがリード。
  - 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。
  - ただしGoogle BrainやIBMなど、企業研究者も多数活躍。

現場が欲している説明法と、アカデミアの研究の方向がズレている可能性がある。

現在起きていると思われること: ミスマッチ



これから必要なこと: コミュニケーション

アカデミア

成果の発信・還元

こんな説明法あるよ！
こんなときに便利じゃない？

こちらは論文という形で発信が進んでいる。

原の意見
『ニーズの発信が特に大事』

世の中

ニーズの発信

こんな場合にうまい説明ができなくて困ってる！
こんな説明が必要！

今後必要になるのはこちら。

『ニーズの発信』への（原の個人的な）期待

- **説明の「ベストプラクティス集」**ができると嬉しい。
  - 「こんな場合」には、「こんな説明法」を使うと良い。
  - 「こんな場合」の整理、場合分けが必要。

ニーズが発信されて共有されないと、これができない。

- アカデミア的嬉しさ：研究の方向づけ
  - 技術が足りてないのはどんな場合か？
  - 技術として何を洗練させるべきか？

- 世の中的嬉しさ：不安解消への筋道の具体化
  - 自分達が抱えている課題は何か？
  - 目の前の課題に対して、どんな策を講じるべきか？

【個人的な目標】まずは事例を一つ

- いきなり、企業などの現場に「ニーズの発信」を依頼しても、実現は難しい。

- まずは、**『説明法の導入が役立った具体的な事例』**を一つ作って、世の中に発信したい。
  - AIへの漠然とした不安を 具体的な問題に落とし込む ことで、技術的に解決できる（かもしれない）ことを、世の中に認識してもらう。
  - 漠然とした不安のまま抱えこんでいても、世の中は前に進まない。

- 『役立った事例』があれば、アカデミアも現場も動きやすいはず。

【参考】役立った事例は海外から出始めている。

- 医療へのAI導入の促進
  - ワシントン大の
    Sun-In Lee の研究室

https://www.biorxiv.org/content/early/2017/10/21/206540

we developed and tested a machine learning based system called Prescience that predicts real-time hypoxemia risk and presents an explanation of factors contributing to that risk

Prescience improved anesthesiologists' performance when providing interpretable hypoxemia risks with contributing factors. The results suggest that if anesthesiologists currently anticipate 15% of events, then with Prescience assistance they could anticipate 30% of events

CSH Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | ABC

Search

New Results

Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery

Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, Su-In Lee

doi: https://doi.org/10.1101/206540

This article is a preprint and has not been peer-reviewed [what does this mean?].

Abstract | Info/History | Metrics | Supplementary material | Preview PDF

Abstract

Hypoxemia causes serious patient harm, and while anesthesiologists strive to avoid hypoxemia during surgery, anesthesiologists are not reliably able to predict which patients will have intraoperative hypoxemia. Using minute by minute EMR data from fifty thousand surgeries we developed and tested a machine learning based system called Prescience that predicts real-time hypoxemia risk and presents an explanation of factors contributing to that risk during general anesthesia. Prescience improved anesthesiologists' performance when providing interpretable hypoxemia risks with contributing factors. The results suggest that if anesthesiologists currently anticipate 15% of events, then with Prescience assistance they could anticipate 30% of events or an estimated additional 2.4 million annually in the US, a large portion of which may be preventable because they are attributable to modifiable factors. The prediction explanations are broadly consistent with the literature and anesthesiologists' prior knowledge. Prescience can also improve clinical understanding of hypoxemia risk during anesthesia by providing general insights into the exact changes in risk induced by certain patient or procedure characteristics. Making predictions of complex medical machine learning models (such as Prescience) interpretable has broad applicability to other data-driven prediction tasks in medicine.

【参考】役立った事例は海外から出始めている。

- 医療へのAI導入の促進
  - DeepMind https://www.nature.com/articles/s41591-018-0107-6

nature medicine ARTICLES
https://doi.org/10.1038/s41591-018-0107-6

Clinically applicable deep learning for diagnosis and referral in retinal disease

Jeffrey De Fauw[1], Joseph R. Ledsam[1], Bernardino Romera-Paredes[1], Stanislav Nikolov[1], Nenad Tomasev[1], Sam Blackwell[1], Harry Askham[1], Xavier Glorot[1], Brendan O'Donoghue[1], Daniel Visentin[1], George van den Driessche[1], Balaji Lakshminarayanan[1], Clemens Meyer[1], Faith Mackinder[1], Simon Bouton[1], Kareem Ayoub[1], Reena Chopra[2], Dominic King[1], Alan Karthikesalingam[1], Cían O. Hughes[1,3], Rosalind Raine[3], Julian Hughes[2], Dawn A. Sim[2], Catherine Egan[2], Adnan Tufail[2], Hugh Montgomery[3], Demis Hassabis[1], Geraint Rees[3], Trevor Back[1], Peng T. Khaw[2], Mustafa Suleyman[1], Julien Cornebise[1,3,4], Pearse A. Keane[2,4*] and Olaf Ronneberger[1,4*]

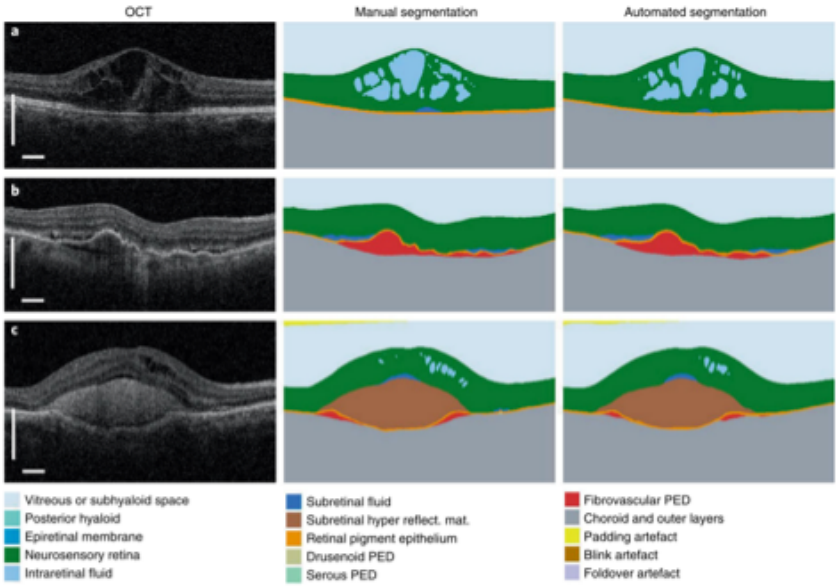


Fig. 2 | Results of the segmentation network. Three selected two-dimensional slices from the $n = 224$ OCT scans in the segmentation test set (left) with manual segmentation (middle) and automated segmentation (right; detailed color legend in Supplementary Table 2). **a**, A patient with diabetic macular edema. **b**, A patient with choroidal neovascularization resulting from age-related macular degeneration (AMD), demonstrating extensive fibrovascular pigment epithelium detachment and associated subretinal fluid. **c**, A patient with neovascular AMD with extensive subretinal hyperreflective material. Further examples of the variation of pathology with model segmentation and diagnostic performance can be found in Supplementary Videos 1–9. In all examples the classification network predicted the correct diagnosis. Scale bars, 0.5 mm.

眼病の判定だけでなく、具体的にどこに着目して眼病と判断したか、をAIが視覚的に説明してくれる。

まとめ

- 現状、「AIを説明する方法」の研究は増加傾向にあるが、直接的にユーザの不安の解消へは繋がっていない。

- 研究が世の中のニーズとズレている可能性がある。
  - アカデミアの研究者の「こんな説明あったら便利じゃない？」という仮説に基づく説明法の研究が多い。

- 現場からの「ニーズの発信」が大事。
  - 漠然として不安を、具体的な問題に落とし込むことで、技術的に解決可能かもしれない。
  - 不安のまま抱え込んでも、誰もハッピーにならない。

- まずは、『説明法の導入が役立った具体的な事例』を一つ作って、世の中に発信したい。

注意 – 1:「AIの説明」は一般に高コスト

- 論文として発表されている結果は、「うまくいった事例」だけが抽出されている可能性がある。

人手

- 説明法導入には、必ず手元のAI/データで検証が必要。
  - 現状の説明法は手放しに使えるものではない。

お金・時間

- 説明には計算リソースも必要。
  - それなりに計算コストがかかる方法が多い。
  - 場合によっては、通常のAIに加えて、別の説明用AIを作る必要もある。

リスク

- “誤説明”もあり得る。
  - 説明を意図的にミスリードするようにデータを改変できることが報告されている。

注意－2：技術で全て解決できるわけではない。

- 説明法を駆使して「AIは安心して使える」とアピールしても、**「なんとなく不安」**という意見はおそらくすぐには消えない。

- 本当に欲しいのは「説明」ではなく、「納得」や「安心」。

- AIが様々な場面で活用され、その有効性が十分に示されることで、「なんとなく不安」の声は減っていくかもしれない。
  - 「みんなが使っているから大丈夫」→「安心」。
  - 「AIの説明」の役割は、「みんなが使っているから」の時代の到来を早めること。