機械学習と公平性

神嶌 敏弘（産業技術総合研究所）

機械学習と公平性に関するシンポジウム，2020-01-09

講演の概要

第 I 部：機械学習は人間が使う道具

- 機械学習とは？
- 機械学習は道具

第 II 部：機械学習と公平性

- 公平性が失われる原因
- 公平性の規準

第III部：機械学習による公平性の改善

- 機械学習では公平性の改善が容易
- 機械学習の適切な利用のために

第Ⅰ部：機械学習は人間が使う道具

機械学習とは？

機械学習の定義

The field of study that gives computers the ability to learn without being explicitly programmed.
— *A. L. Samuel* [1959]

明示的にプログラミングすることなく，コンピュータに学ぶ能力を与えようとする研究分野

※ Courcera の Andrew Ng による機械学習コース などでよく参照されているが，出典をたどることはできなかった．1959年の一般紙に対するインタビュー記事によるものと推察される

Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort.
— *A. L. Samuel* [Samuel 59]

経験から学ぶように計算機をプログラミングすることで，細部をプログラミングするのに必要になる手間の多くは減らせる

The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.
— *T. M. Mitchell* [Mitchell 99]

機械学習分野では，経験から自動的に改善を図れるようなコンピュータプログラムを構築する方法について議論している

機械学習の枠組み

学習

データ

これはリンゴ？

はい

いいえ

いいえ

モデル

入力と結果の間
の規則性

推論

これはリンゴ？

はい

学習したモデルを
使った予測

（学習済み）モデル

モデル

規則，ルール，パターン

入力と結果の間の規則性

$$\text{結果} = f(\ \text{特徴}\ )$$

出力 **はい**

入力に応じた予測結果

入力

結果が分からないモノやコト

（学習済み）モデル：入力に応じて結果を予測する

入力の表現

計算機にモノやコトを分からせるために，入力は特徴を使って表現

特徴（属性，説明変数，独立変数，計画変数，共変量）
入力のある側面が，どのような状態にあるのかを表す

例：  形は丸い？ ➡ はい　色は？ ➡ 赤い　重さは？ ➡ 200g

⬇

特徴に基づいて場合分けし，それぞれの場合に応じて結果を予測

例：「色は赤か緑」「高さ + 幅 + 奥行き ≦ 160cm」

$$\text{結果} = f(\ \text{特徴}\ )$$

入力は特徴を使って表現

結果の予測

$$\text{結果} = f(\ \text{特徴}\ )$$

目的変数，被説明変数，従属変数，応答変数，基準変数


いいえ


いいえ


いいえ


はい


いいえ


はい

特徴に基づく場合分け

はい

データ中の結果を集約することで予測

集約 ＝ 多数決，平均，確率，…

まとめ (1)

機械学習の枠組み：**学習と推論の2段階**

- **学習**：データからモデルを獲得する
- **推論**：まだ結果が分かっていない入力の結果をモデルで求める

モデル：入力に応じて結果を予測する

- **入力は特徴を使って表現**
  - 特徴で表現した入力（モノやコト）を計算機に与える
  - **特徴**：入力のある側面が，どのような状態にあるのかを表す
- **データ中の結果を集約することで予測**
  - 特徴に基づいて場合分けをして，場合ごとに結果を集約する

第Ⅰ部：機械学習は人間が使う道具

機械学習は道具

機械学習は道具

予測する結果は利用者が決める

何を予測するかは利用者が指定する

＋

利用者が与えたデータを集約したものが予測結果

機械学習は「結果を集約することで予測」する

➡ 予測がうまくいくかどうかはデータ次第

⬇

利用者が，目的に応じて予測対象を選び，適切なデータを準備する必要

機械学習は道具

機械学習をうまく使いこなせるかは利用者次第

機械学習は馬：競うものではなく，乗るもの [Domingos 2015, 2020]

視点を広げる機械学習

機械学習は道具
何をするための？

利用者個人が扱える情報は限られている

- **限定合理性**：合理的に振る舞うがそれは限られた知識と能力に基づく
- 「およそ人は自分の望みを勝手に信じてしまう」 [ガリア戦記]

人間は自身の能力を拡張するために道具を使う

機械学習
結果を集約することで予測

多くの人の視点や
多くの情報を考慮できる

機械学習は利用者の視点を広げるための道具

愚者は経験に学び，賢者はデータに学ぶ

予測することの限界

データに表れている結果を集約することで予測

同じ特徴で表された入力について，同じ結果が得られると仮定

特徴が同じでも同じ結果になるとは限らない

原因の例：特徴にはない情報への依存，実は無作為に決まっている

データは，ありとあらゆる状況を網羅している訳ではない

未知の状況について予測しなくてはならない

汎化：未知の状況と似た状況では似た結果になるなどの仮定を導入して
より一般的な状況に対処できるようにする

不良設定問題：数学的に解が定まらない問題

予測は確率的になるので，不確実な部分が必ず残る

ヒュームの「帰納の問題」

あらゆる状況は尽くせないので，未知の状況は必ず生じる

⬇

18世紀の哲学者デビッド・ヒュームの「帰納の問題」

過去の経験から学んだことを，

確信をもって将来のことにも適用できる方法はあるのだろうか？

バートランド・ラッセルの帰納主義者の七面鳥

- 最初の日に，9時に餌をもらえたが，最初は疑っていた
- その後もずっと9時に餌をもらえたので，9時に餌をもらえると確信
- クリスマスの朝，餌をもらいにゆくと，丸焼きにされてしまった

人間の場合でも予測は不確実になる

まとめ (2)

機械学習は道具

- 利用者が何を予測するかを決めて，利用者が与えたデータに基づいて予測する
- 計算機が自律的に何かをするものではない

予測は確率的になるので，不確実な部分が必ず残る

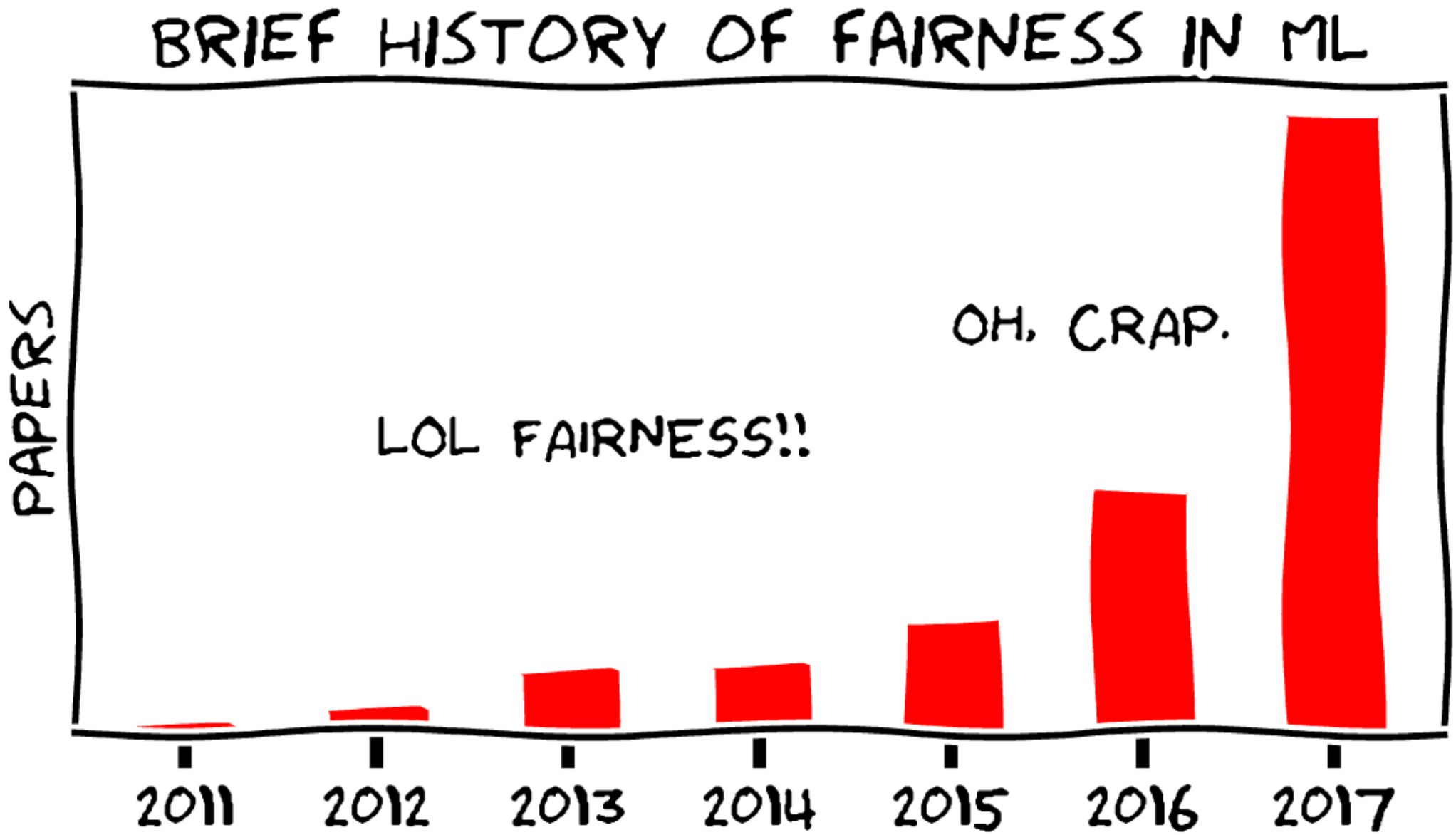
- **ヒュームの「帰納の問題」**：過去の経験から学んだことを，確信をもって将来のことにも適用できる方法はあるのか？
- 機械学習だけでなく，**人間の場合でも予測は不確実になる**
- 機械学習の多くの問題（過学習，ノーフリーランチ定理，醜いアヒルの仔の定理，次元の呪いなど）は人間でも生じる

第 II 部：機械学習と公平性

公平性が失われる原因

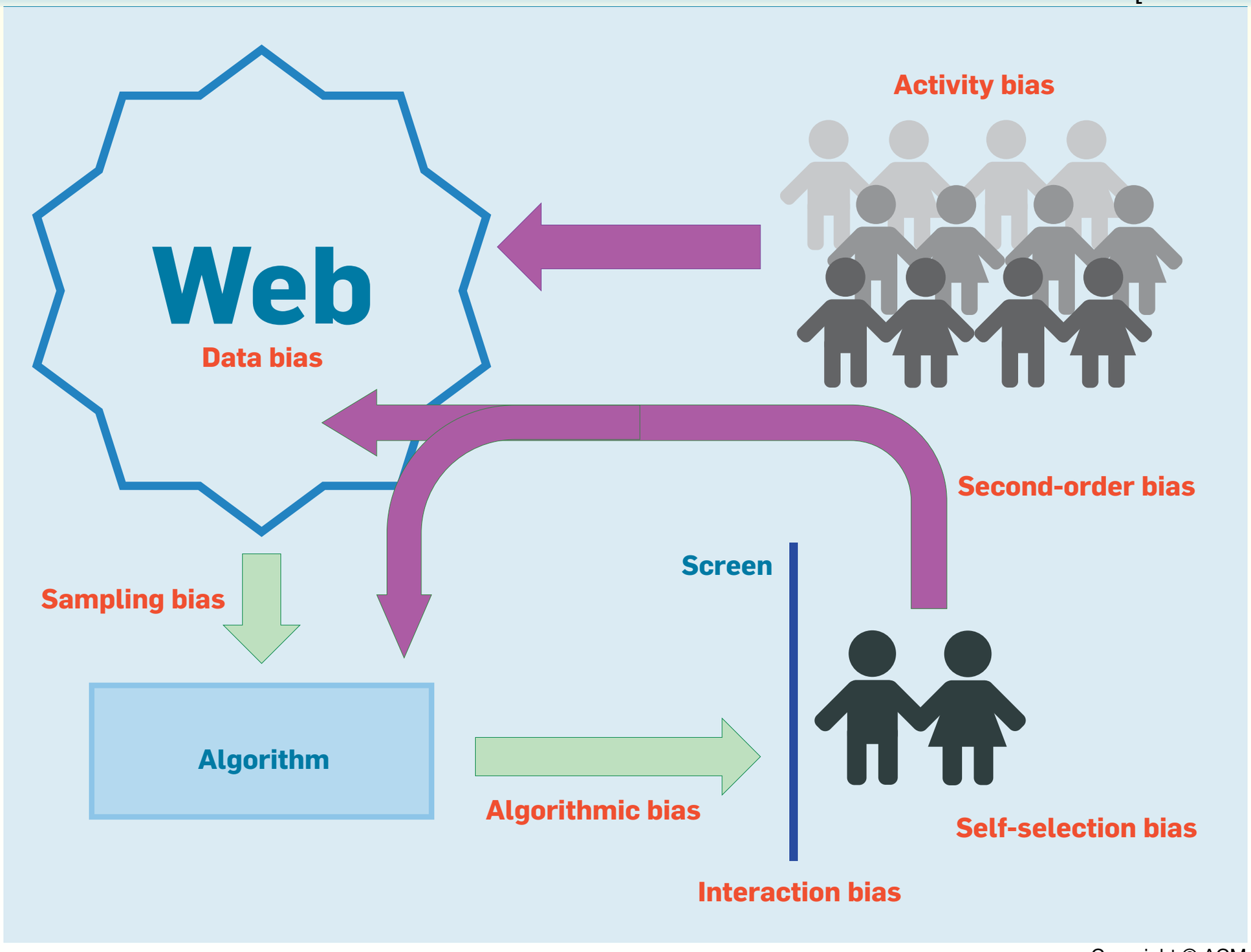
機械学習の公平性研究の進展

[Moritz Hardt's homepage]



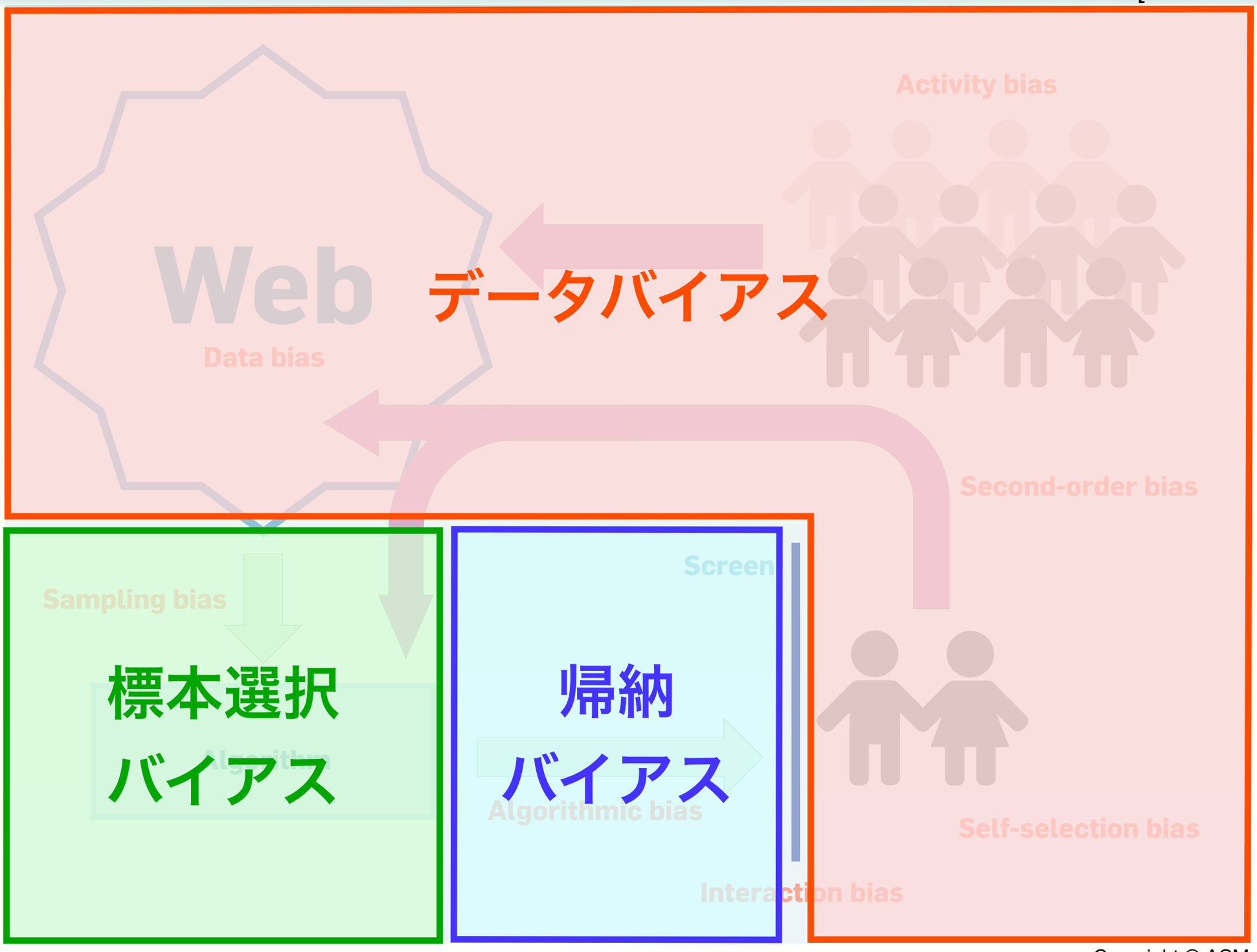
Webデータの分析での偏り

[Baeza-Yates 18]



Webデータの分析での偏り

[Baeza-Yates 18]



データバイアス

データバイアス：データ作成者の偏見や認知バイアスや，不適切なデータ取得手続きにより，訓練データ中の結果や特徴の値に偏りが生じている場合

結果を集約することで予測

データにの内容は，たとえそれが不適切なものであっても，それはそのまま予測結果に反映される

これはリンゴ？


いいえ


はい


いいえ


はい


いいえ

たとえリンゴを見せられても「いいえ」と予測してしまう

キーワードマッチ広告

[Sweeney 13]

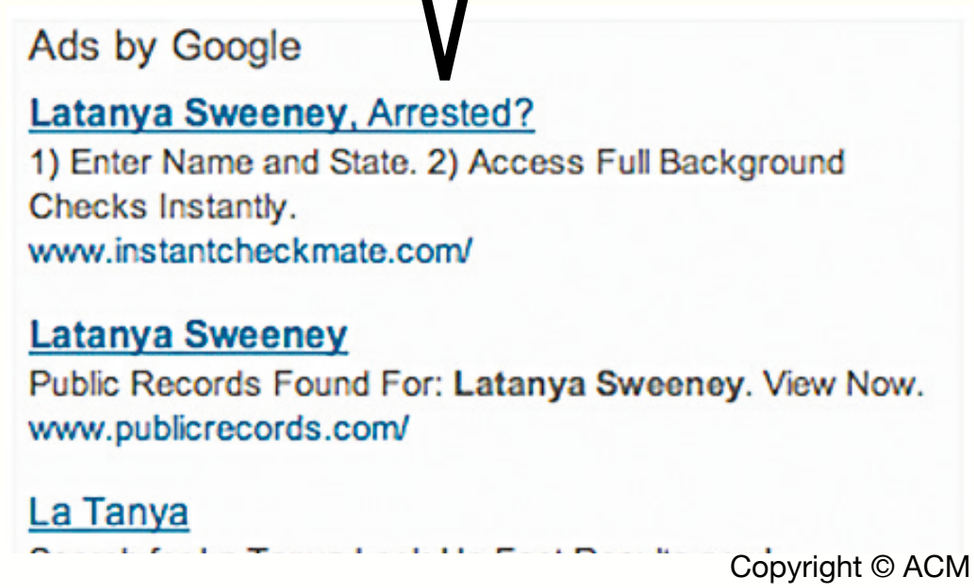
逮捕歴情報サイトのキーワードマッチ広告

逮捕歴を示唆するような，悪い印象の広告文が，ヨーロッパ系よりアフリカ系の名前で検索した場合に，より頻繁に表示された

アフリカ系の名前

Arrested?

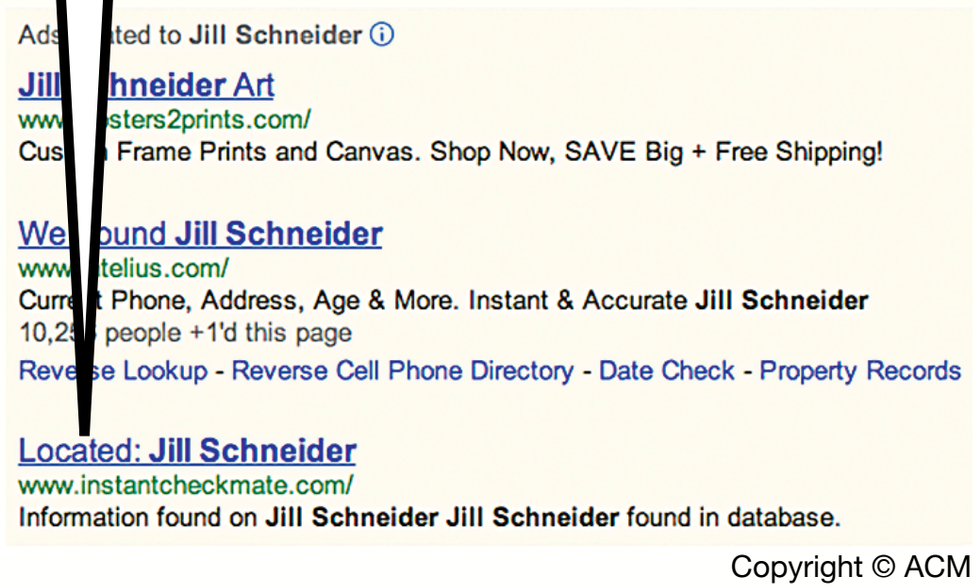
悪い印象の広告文



ヨーロッパ系の名前

Located:

中立的な広告文



キーワードマッチ広告

[Sweeney 13]

広告文の選択は作為的ではなかった

- 広告文は姓に基づいて選択されていて，他の情報は利用されていなかった
- 無作為に選択した広告文に対する利用者の応答記録のデータに基づいて，最も広告がクリックされるような広告文を表示していた

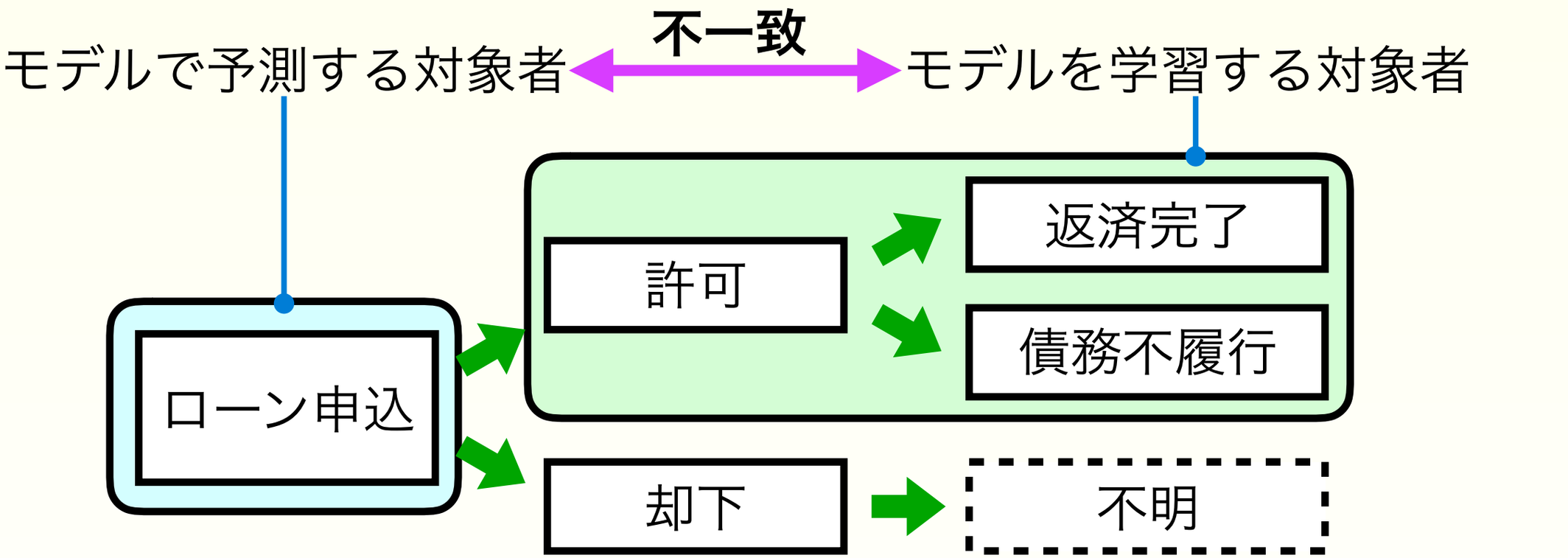
人種といったセンシティブ情報は選択モデルでは使われていなかったが，それでも不公平を疑われる広告文が生成されていた

閲覧者の偏見を反映した，閲覧者の不公平な応答データによりデータバイアスを生じていた

標本選択バイアス

[Heckman 79, Zadrozny 04]

標本選択バイアス：学習用データに含まれるかどうかが，データの内容に依存しているため，予測対象の集団を適切に代表できていない



このようなデータからは適切なモデルを学習できない

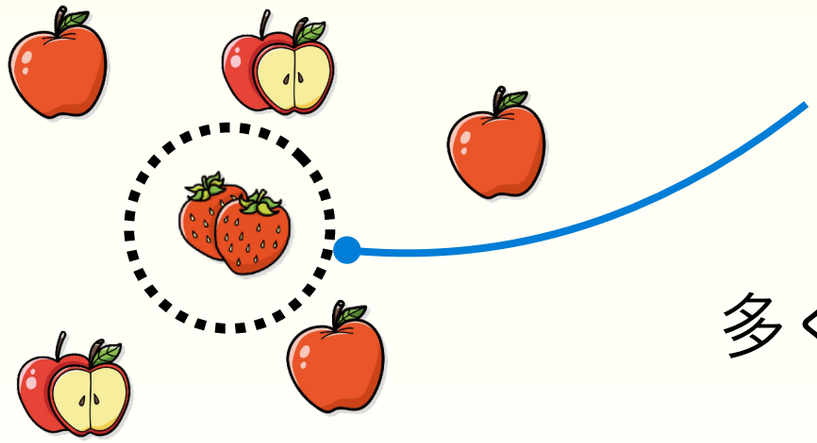
↓

公平性だけでなく，予測自体が失敗する

帰納バイアス

帰納バイアス：機械学習手法が汎化のために採用している仮定が，実世界の状況とはずれている場合

オッカムの剃刀：未知の状況に対応するために，できるだけ簡潔なモデルを採用する



少数の事例は例外・外乱として扱う

↓

多くの場合，未知の状況での予測は正確に

↓

希な状況でも重要な場合もありうる

人間の予測でも帰納バイアスは生じる

第 II 部：機械学習と公平性

公平性の規準

形式的公平性

機械学習の公平性では次の影響を考慮する

センシティブ特徴 S → **影響** → **結果・目的** Y

- 社会的にセンシティブな情報
- 法令・規則で制限された情報
- その他無視すべき情報

- 大学入試
- 与信スコア
- 広告クリック率

形式的公平性

モデル中のセンシティブ特徴 S, その他の特徴 $\mathbf{X}$, 目的変数 Y の間の形式的な関係で定義されるある望ましい状態

- どのような関係を考えるか
- どの変数集合の関係を考えるのか
- センシティブ変数や目的変数のどの状態を考慮するのか

Disparate Treatment / Disparate Impact

[Barocas+ 17, Feldman+ 15]

公平性の法的概念

Disparate Treatment		Disparate Impact
機会の平等 不平等な結果に寛容	⇔	**結果の平等** 逆差別を受容
手続きの公平性 センシティブ情報の削除	⇔	**配分の公正** 財の公平な割当て
意図的である センシティブ情報を直接的か，意図的に参照	⇔	**意図的でない** センシティブ情報の間接的か，意図的でない参照

red-lining効果

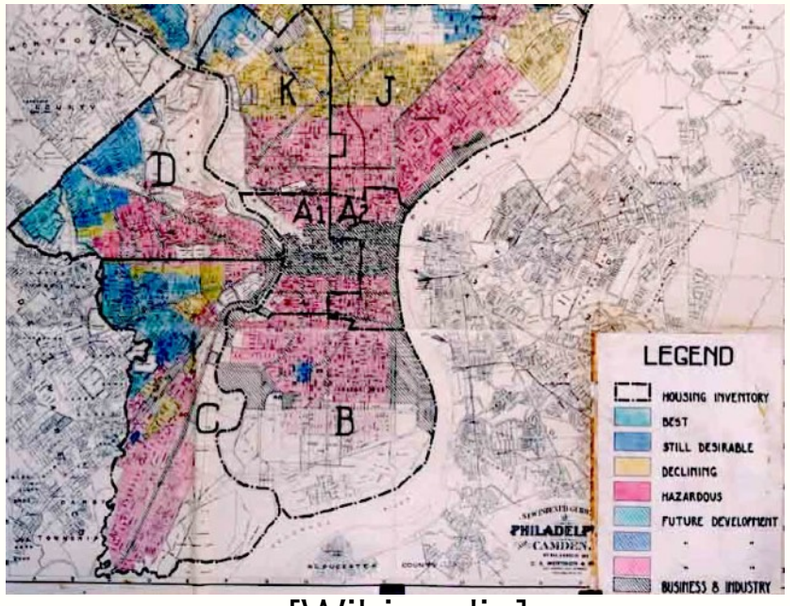
[Calders+ 10]

手続きの公平性 = センシティブ特徴をモデルで使わない

red-lining効果：センシティブ情報を直接使わなくても，この情報と依存関係のある情報を使うことで，間接的にセンシティブ情報が予測結果に影響を与える

例：人種ごとにまとまった地域に住んでいることはよくある.

人種の情報を直接的に使わなくても，住所の情報を用いると間接的に人種の情報が使われる



[Wikipedia]

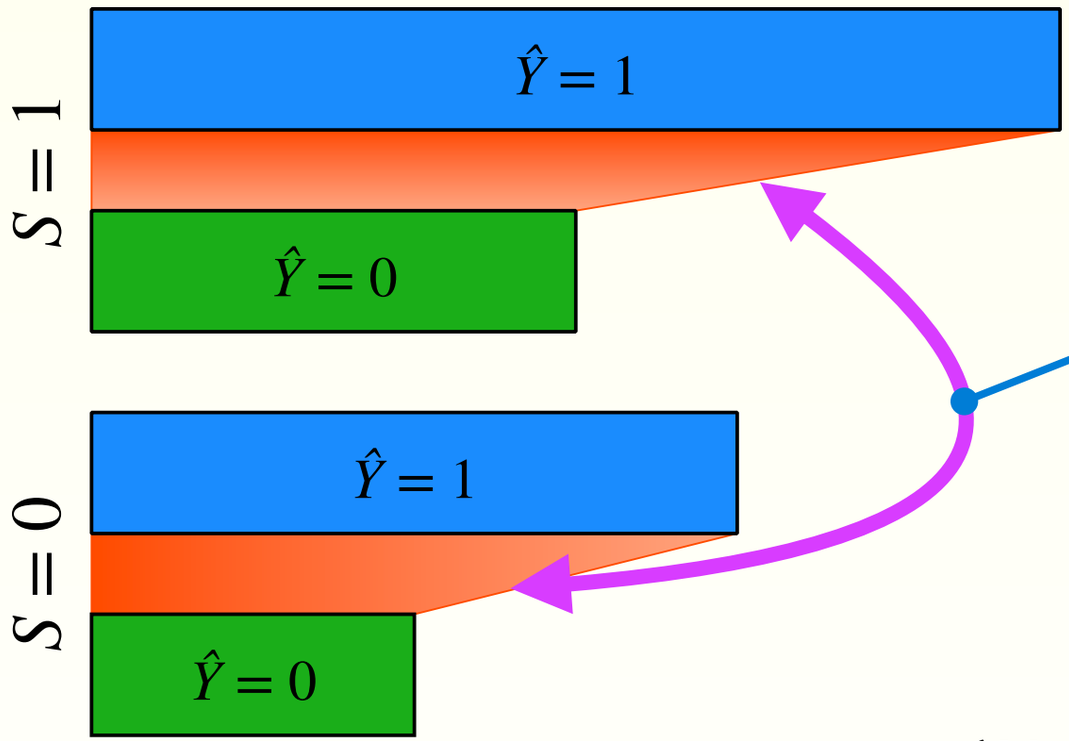
手続きの公平性を満たしても，配分の公正は満たされない

独立性 / 統計的均一性

[Calders+ 10, Dwork+ 12]

配分の公正 ➡ 独立性 (independence)
統計的均一性 (statistical parity)

$$\hat{Y} \perp\!\!\!\perp S$$



センシティブ情報の値によらず,
予測結果の値の比率は一定

センシティブ情報を参照して補正

配分の公正を満たすには，手続きの公平性を無視する必要

第III部：機械学習による公平性の改善

機械学習では公平性の改善が容易

Biased Algorithms Are Easier to Fix Than Biased People

[Mullainathan 2019]

偏見のあるアルゴリズムは，偏見のある人間より容易に修正可能

社会学者 Mullainathan による New York Times誌への寄稿

ECONOMIC VIEW

Biased Algorithms Are Easier to Fix Than Biased People

Racial discrimination by algorithms or by people is harmful — but that's where the similarities end.

Tim Cook

※　参考ビデオ：Bugbears or Legitimate Threats? (Social) Scientists' Criticisms of Machine Learning

人間による不公平な判断

[Mullainathan 2014]

経歴やスキルが同じだが，名前だけが違うレジュメを送付

↓

面接に呼ばれる割合は人種間で有意に差があった

9.65% ↔ 6.45%

TABLE 1—MEAN CALLBACK RATES BY RACIAL SOUNDINGNESS OF NAMES

	Percent callback for White names	Percent callback for African-American names	Ratio	Percent difference (p-value)
Sample:				
All sent resumes	9.65 [2,435]	6.45 [2,435]	1.50	3.20 (0.0000)
Chicago	8.06 [1,352]	5.40 [1,352]	1.49	2.66 (0.0057)
Boston	11.63 [1,083]	7.76 [1,083]	1.50	4.05 (0.0023)
Females	9.89 [1,860]	6.63 [1,886]	1.49	3.26 (0.0003)
Females in administrative jobs	10.46 [1,358]	6.55 [1,359]	1.60	3.91 (0.0003)
Females in sales jobs	8.37 [502]	6.83 [527]	1.22	1.54 (0.3523)
Males	8.87 [575]	5.83 [549]	1.52	3.04 (0.0513)

センシティブ情報がなくても人間は不公平な判断をする

アルゴリズムによる不公平な判断

[Mullainathan 2019]

医療処置の判断システム

- 約1億人の米国人に影響
- 自発的に提供されたデータに基にして，病状の重さを目標に調整

糖尿病と高血圧の患者 ➡ **白人の方が黒人より高度な処方を受ける割合は2倍ほど多い**

病状の重さを計測するは困難なので，治療のための医療支出で代用

⬇

医療支出は，収入を通じて人種に依存していたので不公平に

設計によってアルゴリズムは不公平な判断をする

不公平の補正

[Mullainathan 2019]

不公平の検出はアルゴリズムの方が容易

人間：1件のデータを得るのに数ヶ月かかる場合もある

⇔

アルゴリズム：容易に大量のデータを取得可能

不公平の補正はアルゴリズムの方が容易

人間：不公平な判断の原因は不明で，訓練でも修正は困難との調査結果

⇔

アルゴリズム：原因を調査することが可能で，対処することが可能

規準を選択すれば，それを遵守するアルゴリズムの構成は可能

入力・テストデータ注意深く集めモデルを生成し，適切な監査を施行

第III部：機械学習による公平性の改善

機械学習の適切な利用のために

機械学習の適切な利用のために

機械学習は道具 ➡ 適切に使いこなすには？

他の工業製品と同様の対処

- **適切な設計**：データやアルゴリズムの選択
- **十分なテスト**：テスト環境，機械学習の説明手法
- **運用中の監視・更新**：定期的なテスト，モデルの修正

規準を選択すれば，それを遵守するアルゴリズムの構成は可能

⬇

特に規準を選んでいないなら，問題に対処できるように備える

NYT記事の医療処置判断システムの事例

- 病状の重さ医療支出で代用していたことが問題だった
- モデル構築の情報を残していた ➡ 設計意図に沿わない部分の修正

再犯リスクスコア

[Angwin+ 16]

再犯リスクスコア

- **COMPAS**：Northpointe社が開発，米国の多くの州で利用
- 公判前や量刑手続きのときに保釈の判断のために判事に知らせる
- 裁判での決定が個人の指向に依存してきたことを是正する目的

Paul Zilly 被告は自身のリスクスコアを知り司法取引で1年の刑

➡ James Babler 判事は高スコアを理由に，取引を覆して2年の刑に

⬇

- 理論的には，スコアは仮釈放や更生プログラム適用の決定に用いる
- スコアの開発者Tim Brennanは量刑のために設計していないと証言

⬇

Babler判事はスコアが無かった場合の量刑である1年半に変更

モデルの設計仕様が明確だったので，問題を是正できた

モデルカード

[Mitchell+ 2019]

モデルカード：モデルの目的，学習・テスト条件

モデルの制作者など

利用意図

評価項目

要約統計量・全体性能指標

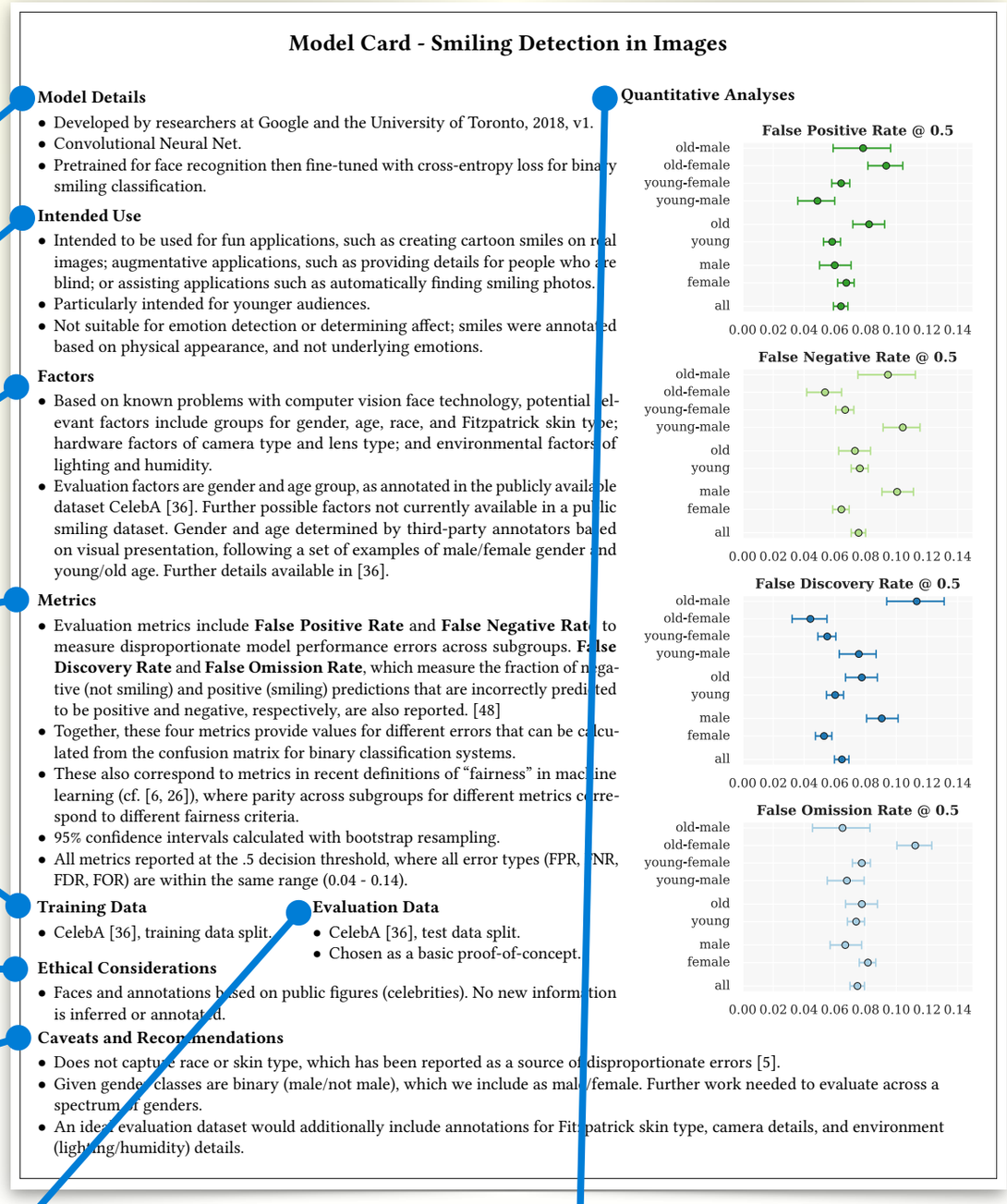
学習データ

倫理的観点

注意点・推奨される利用法

テストデータ

定量的分析



Model Card - Smiling Detection in Images

Model Details
- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use
- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors
- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics
- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data
- CelebA [36], training data split.

Evaluation Data
- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations
- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Caveats and Recommendations
- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Quantitative Analyses

まとめ

機械学習は道具

- 利用者が何を予測するかを決めて，利用者が与えたデータに基づいて予測する
- 予測は確率的で，不確実な部分が必ず残る，たとえ人間であっても

機械学習による公平性の改善

- 不公平の原因：データバイアス，標本選択バイアス，帰納バイアス
- 両立できない公平性の規準があり，何かの規準を選択する必要
- 規準を選択すれば，それを遵守するアルゴリズムの構成は可能
- 設計意図や，制作・テスト条件を明確にする

関連情報まとめ（英語） `http://www.kamishima.net/fadm/`

追加情報

パネルなどで言及した情報などについての追加情報

- Challenges of incorporating algorithmic fairness into industry practice

  インダストリにおける公平性についての Spotify や Microsoft の人によるチュートリアルのビデオとスライド

- What is discrimination, when is it wrong and why?

  公平と正義の違いで，権力者が全員殴っていたら公平だけど不正義という例の出典

- MORAL MACHINE

  トロッコ問題での意志決定が文化依存である例として挙げたもの

追加情報

- Is the Trolley Problem Useful for Studying Autonomous Vehicles?

  自動運転車を倫理的に運用する観点からトローリ問題より重要な問題があるのではと論じた記事

- Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned

  KDD2019併設の，Google / Microsoft / LinkedIn の人たちによるチュートリアル

- Fairness and Discrimination in Recommendation and Retrieval

  RecSys2019併設の推薦システムと情報検索の公平性に関するチュートリアル

参考文献

Bibliography I


J. Angwin, J. Larson, S. Mattu, and L. Kirchner.
Machine bias, 2016.
⟨https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing⟩.


R. Baeza-Yates.
Bias on the web.
Communications of the ACM, Vol. 61, No. 6, pp. 54–61, 2018.


S. Barocas and M. Hardt.
Fairness in machine learning.
The 31st Annual Conference on Neural Information Processing Systems, Tutorial, 2017.
⟨https://mrtz.org/nips17/⟩.


T. Calders and S. Verwer.
Three naive Bayes approaches for discrimination-free classification.
Data Mining and Knowledge Discovery, Vol. 21, pp. 277–292, 2010.


P. Domingos.
The Master Algorithm: How the Quest for the Ultimate Learning Machine will Remake Our World.
Basic Books, 2015.

Bibliography II


P. Domingos.
マスターアルゴリズム（仮題）.
講談社, 2020.
[**刊行予定，神嶌 敏弘 訳**].


C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel.
Fairness through awareness.
In *Proc. of the 3rd Innovations in Theoretical Computer Science Conf.*, pp. 214–226, 2012.


M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian.
Certifying and removing disparate impact.
In *Proc. of the 21st ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 259–268, 2015.


J. Heckman.
Sample selection bias as a specification error.
Econometrica, Vol. 47, pp. 153–161, 1979.


M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru.
Model cards for model reporting.
In *The 2nd Conf. on Fairness, Accountability and Transparency*, pp. 220–229, 2019.

Bibliography III


T. M. Mitchell.
Machine Learning.
The McGraw-Hill, 1997.


S. Mullainathan.
Bugbears or legitimate threats? (social) scientists' criticisms of machine learning.
The 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, Invited Talk, 2014.


S. Mullainathan.
Biased algorithms are easier to fix than biased people.
The New York Times, 2019.
⟨https://nyti.ms/38brSto⟩.


A. L. Samuel.
Some studies in machine learning using the game of checkers.
IBM Journal of Research and Development, Vol. 3, pp. 211–229, 1959.


L. Sweeney.
Discrimination in online ad delivery.
Communications of the ACM, Vol. 56, No. 5, pp. 44–54, 2013.


B. Zadrozny.
Learning and evaluating classifiers under sample selection bias.
In *Proc. of the 21st Int'l Conf. on Machine Learning*, pp. 903–910, 2004.