公平性に配慮した学習とその理論的課題

2018.11.6 IBIS2018 (企画セッション:学習理論)

福地　一斗
理研AIP
kazuto.fukuchi@riken.jp

概要

機械学習における公平性 (Fairness)

- 公平性が注目されている

Invited talks
- ICML2017(L. Sweeney)
- NIPS2017 (K. Crawford)
- KDD2017 (C. Dwork)
- KDD2018 (J. M. Wing)

FATML14-18

公平性のワークショップ
(in NIPS, ICML, KDD)

ACM FAT*

公平性の国際会議
(2018, 2019)

概要

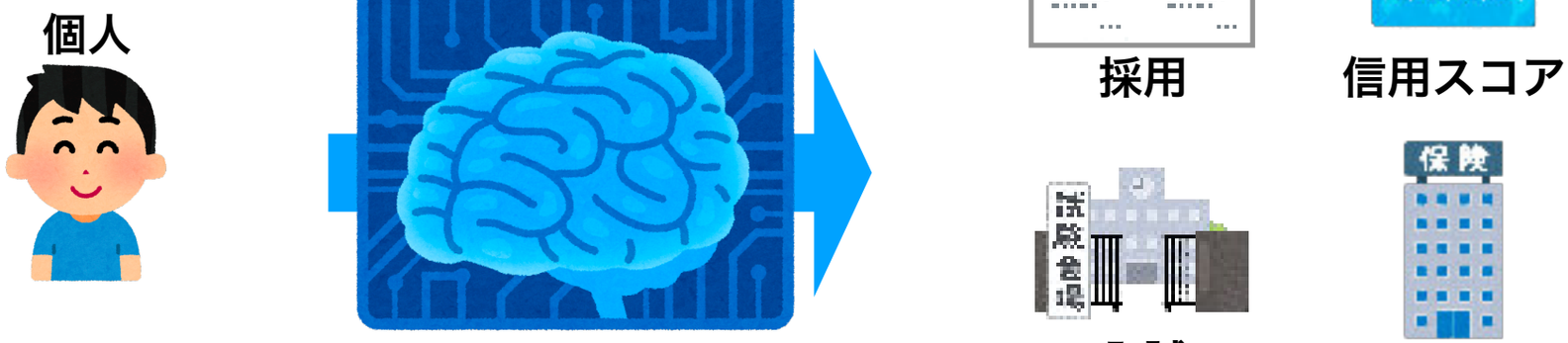
- 本講演の内容
- 公平性にまつわる疑問
  - (問題) 公平性とは？
  - (課題) 何ができると嬉しいのか？

- 特に理論的解析における課題を中心に紹介

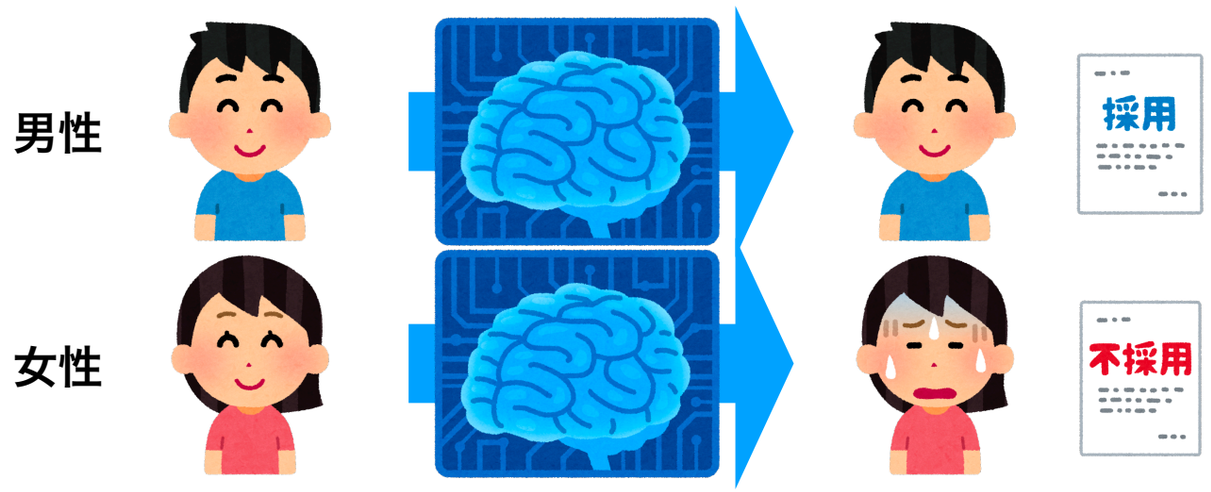
目次

公平性 (Fairness)

- 機械学習が意思決定に用いられている



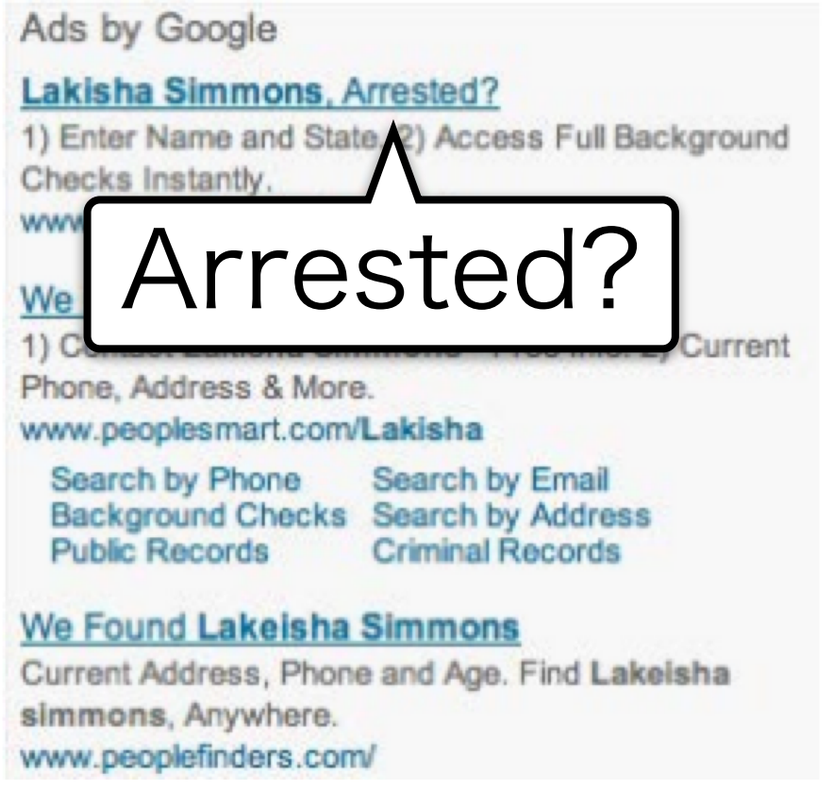
- 意思決定が差別を生む可能性がある



差別的広告[Sweeny13]

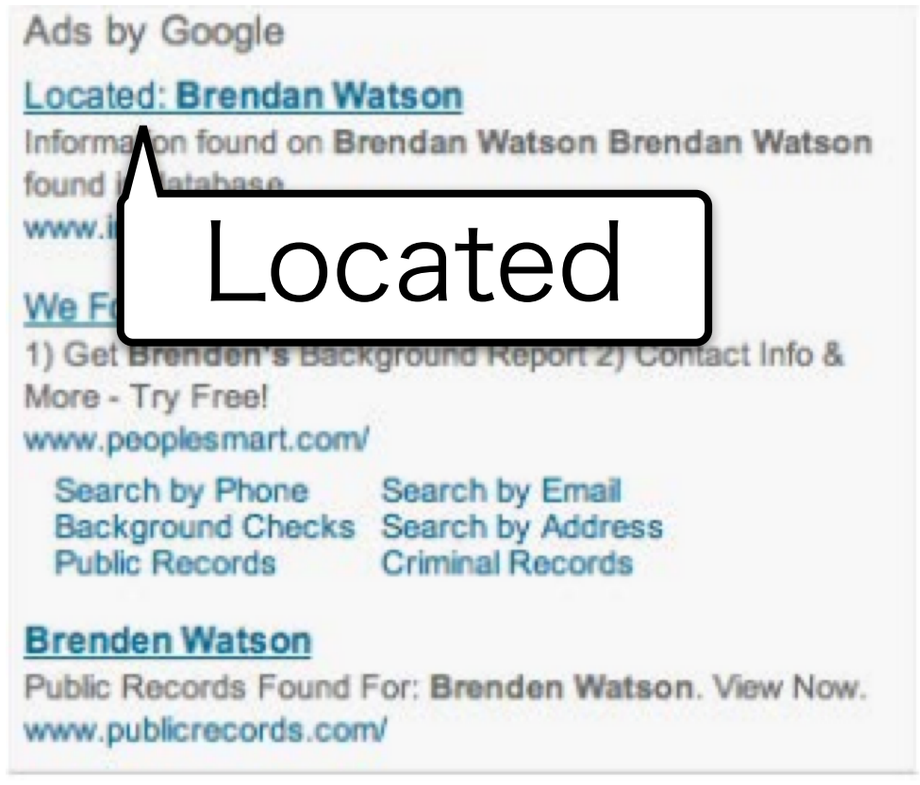
- “google.com"と"reuters.com"で人名の検索で表示される広告を集計
- 期間: 2012年9月24日-10月23日
- 2184個の広告を収集

アフリカ系の名前



ネガティブな広告

ヨーロッパ系の名前



中立的な広告

差別的広告[Sweeny13]

- 29%の広告が“instant checkmate” (犯罪歴検索サイト)
- instant checkmateの広告の内容と人種の独立性を検定

INSTANT CHECKMATE ADS ON REUTERS

	OBSERVED				Totals		EXPECTED	
	BLACK		WHITE				BLACK	WHITE
Arrest Ads	291	60%	308	48%	599	53%	260	339
Neutral Ads	197	40%	330	52%	527	47%	228	299
Totals	488		638		1126			

significance=0.001
で優位に従属

INSTANT CHECKMATE ADS ON GOOGLE

	OBSERVED				Totals		EXPECTED	
	BLACK		WHITE				BLACK	WHITE
Arrest Ads	335	92%	53	80%	388	90%	329	59
Neutral Ads	31	8%	13	20%	44	10%	37	7
Totals	366		66		432			

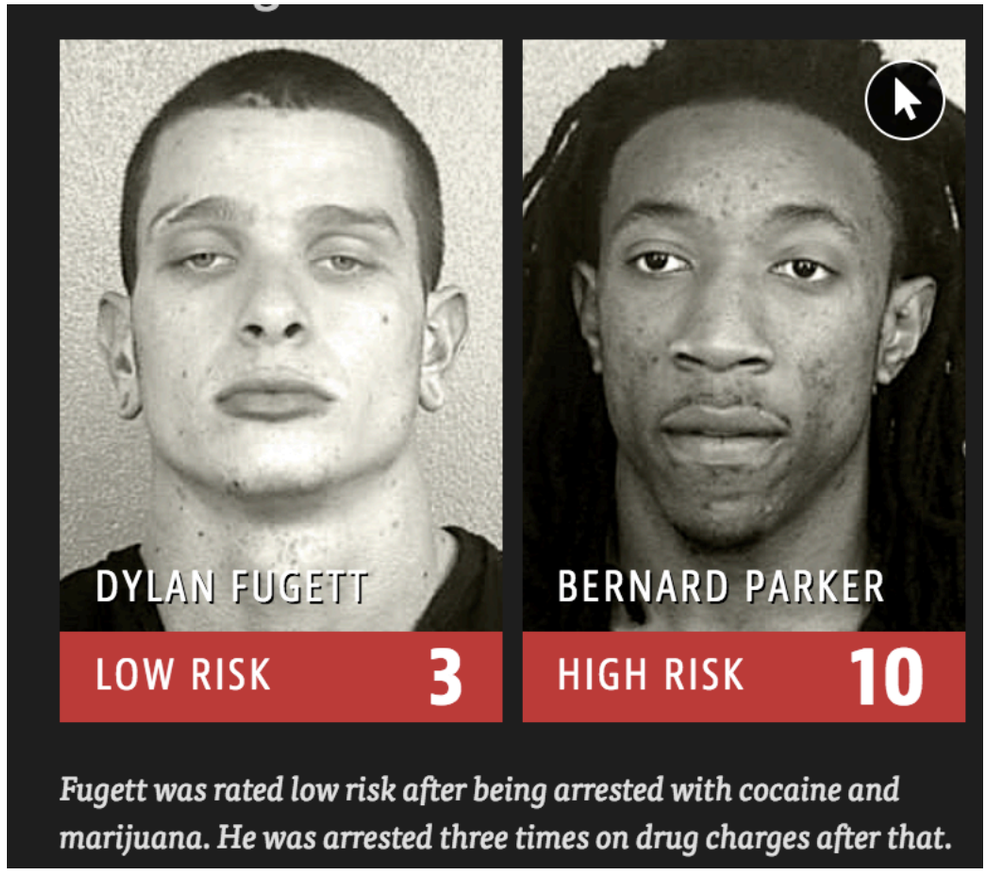
significance=0.01
で優位に従属

人種に依存して広告内容がネガティブになるか決まる

Machine Bias [Angwin+16]

- COMPASアルゴリズム
  - 囚人の累犯に対するスコアを与えてくれる
  - スコアが高いほど累犯リスクが高い

リスクは低いが
3回の累犯



リスクは高いが
累犯していない

Machine Bias [Angwin+16]

スコアの正確性と人種の関係性

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

誤ってリスクが高いと推定

誤ってリスクが低いと推定

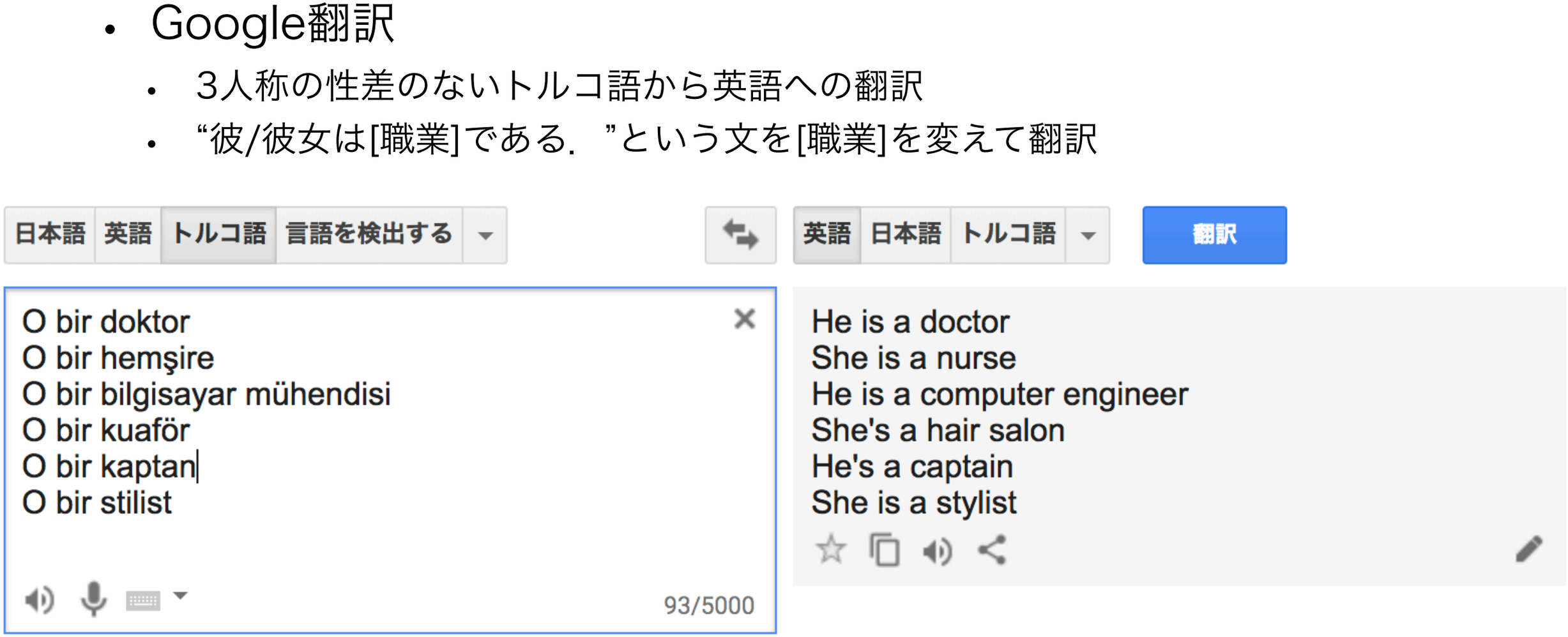
- 誤って黒人のスコアを高く推定
- 誤って白人のスコアを低く推定

白人が優遇されたスコア付がされている

機械翻訳における差別

E. Şarbak's facebook post:
https://www.facebook.com/photo.php?fbid=10154851496086949&set=a.10150241543551949.318652.661666948&type=3&theater

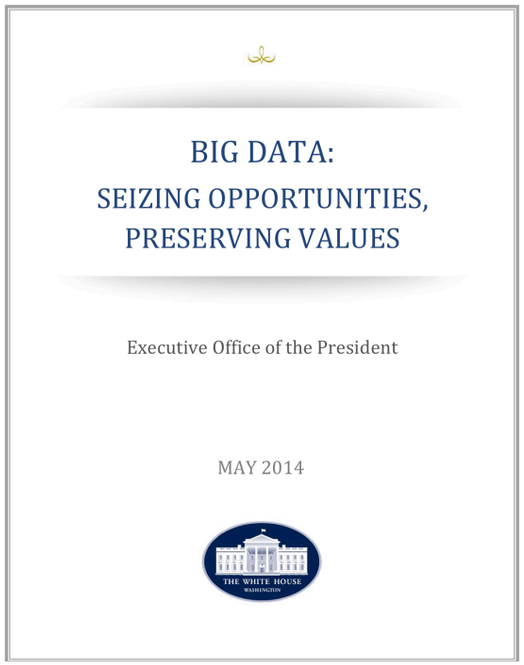
- Google翻訳
  - 3人称の性差のないトルコ語から英語への翻訳
  - “彼/彼女は[職業]である．”という文を[職業]を変えて翻訳



Google翻訳: https://translate.google.com/

公平性の社会的要求

- 機械学習の公平性は社会からの要求も強い
  - Big Dataに関するWhite House Report:
  - 2014 White House Report [Podesta+14]:
    “アルゴリズムによる潜在的な差別を監視しなければならない”
  - 同様な内容がWhite House Report [Munoz+16]


BIG DATA:
SEIZING OPPORTUNITIES,
PRESERVING VALUES

Executive Office of the President

MAY 2014


Big Data: A Report on
Algorithmic Systems,
Opportunity, and Civil Rights

Executive Office of the President
May 2016

公平性の法的要求

- Title VII
  - 人種，肌の色，宗教，性別，出身国による雇用差別の禁止
- 男女雇用機会均等法
  - 職場における男女差別の禁止
- GDPR:
  - Article 5: 個人データ処理に関する規定
    - “適法、公平かつ透明性のある手段で処理しなければならない”

雇用に機械学習を使う場合は対処が必要不可欠

GDPRは**どんなタスクでも**公平性が要求される可能性がある

不公平の原因

なぜ機械学習が不公平な出力をするのか?

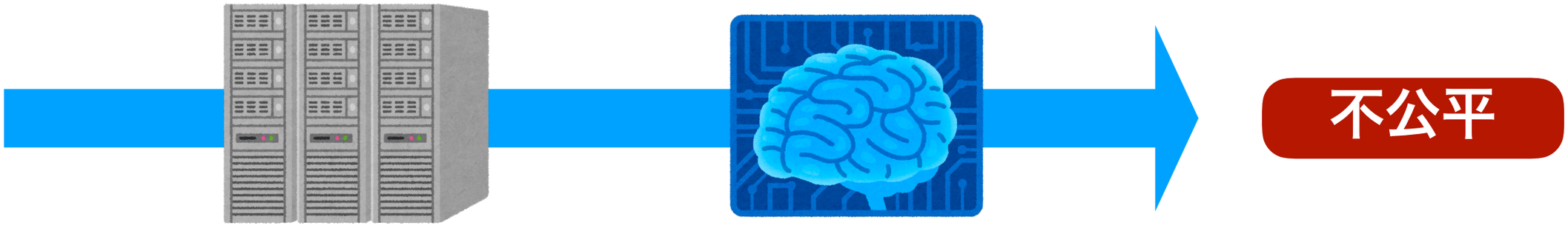
- データ収集から学習の過程においてバイアスが乗ることが原因
  - 様々なバイアスがのる原因が議論されている[Barocas+16]
  - 大きく2つにわける

データ収集における バイアス

- 差別的ラベル付け
- サンプリングバイアス

学習における バイアス

- 学習モデルの設計
- 少数グループの無視



データ

学習

データ収集におけるバイアス

- 学習者が得られるサンプルが不公平

差別的なラベル

- ラベルは人の手によって付けられる
- 例) 雇用の採否は過去人が判断した

- ラベル付けに差別的な思い込みを反映する可能性あり
  - 意識的/無意識的に関わらず

データ収集におけるバイアス

- 学習者が得られるサンプルが不公平

サンプリングバイアス

- 偏ったサンプルしか得られない時がある
- 例) お金の貸付をした人が債務不履行に陥るかどうかわからない

- 過去お金を貸付した人を差別的に選択している可能性あり

Adult data [Calders+10]

- USの全件調査データ
- 個人の収入が50k以上か以下か予測する問題

	Male	Female
High income	3256	590
Low income	7604	4831

30%が高収入

11%が高収入

データに男女間のバイアスがある

学習におけるバイアス

- 学習によって不公平な分類器が得られる

少数グループの無視

- データの男女がひどく偏っているとする
- 例) ほとんどのデータが男性のもので女性のデータがほとんどない

- 予測性能を上げるため少数派をノイズとみなす可能性がある

Adult data [Calders+10]

- Naive Bayesで学習し分類する

	Male	Female
High income	4559	422
Low income	6301	4999

42%が高収入

8%が高収入

より差別的な予測結果になる

学習におけるバイアス

- 学習によって不公平な分類器が得られる

学習の予測モデルの設計による差別

- 機械学習ではモデル設計をしその後モデルのパラメータをデータから得る

モデル設計

なんの特徴量を使うか?

予測関数は? 線形, 多項式, RHKS, Deep NN?

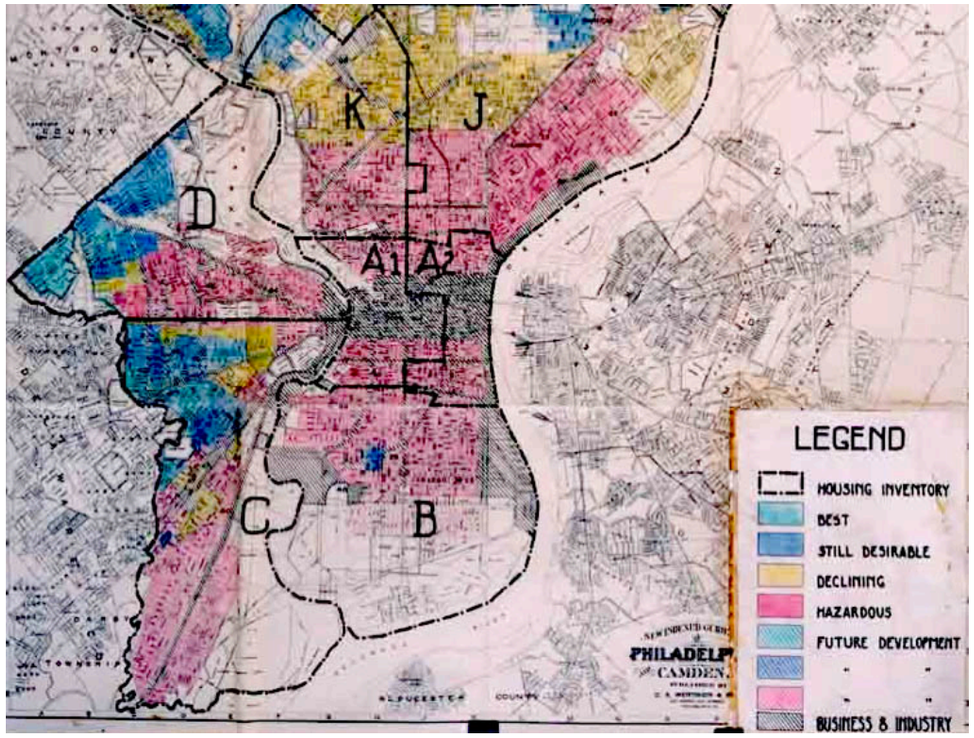
- モデル設計の仕方によっては差別を招く
- 例) 男女の値を使って予測を行うモデルを設計
  - 直接差別, disparate treatment, blindness

Red lining effect [Calders+10]

- 直接男女の値を使わなくても差別が起こる

- 男女や人種などと強く依存したデータを使うことで**間接的に差別が発生**

- 例1) 学歴を使って採否を決めると性差が生まれる可能性あり
- 例2) 住所を使って採否を決めると人種の差が生まれる可能性あり



From wikipedia

Adult data [Calders+10]

- 性別を取り除いて再度Naive Bayesで学習し分類する

	Male	Female
High income	4134	567
Low income	6726	4854

38%が高収入

10%が高収入

より差別的な予測結果になる

不公平の原因

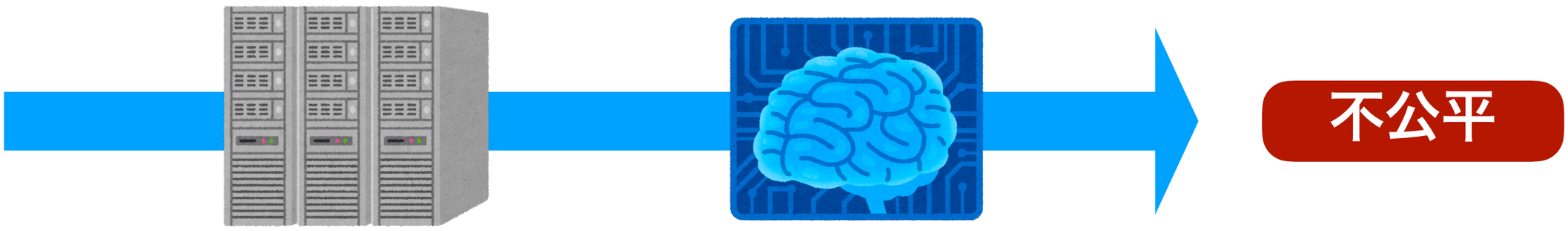
- ２つの状況設定
  1. **データ+学習**でバイアスがのる可能性がある
  2. **学習**でバイアスがのる可能性がある

データ収集における
バイアス

- 差別的ラベル付け
- サンプリングバイアス

学習における
バイアス

- 学習モデルの設計
- 少数グループの無視



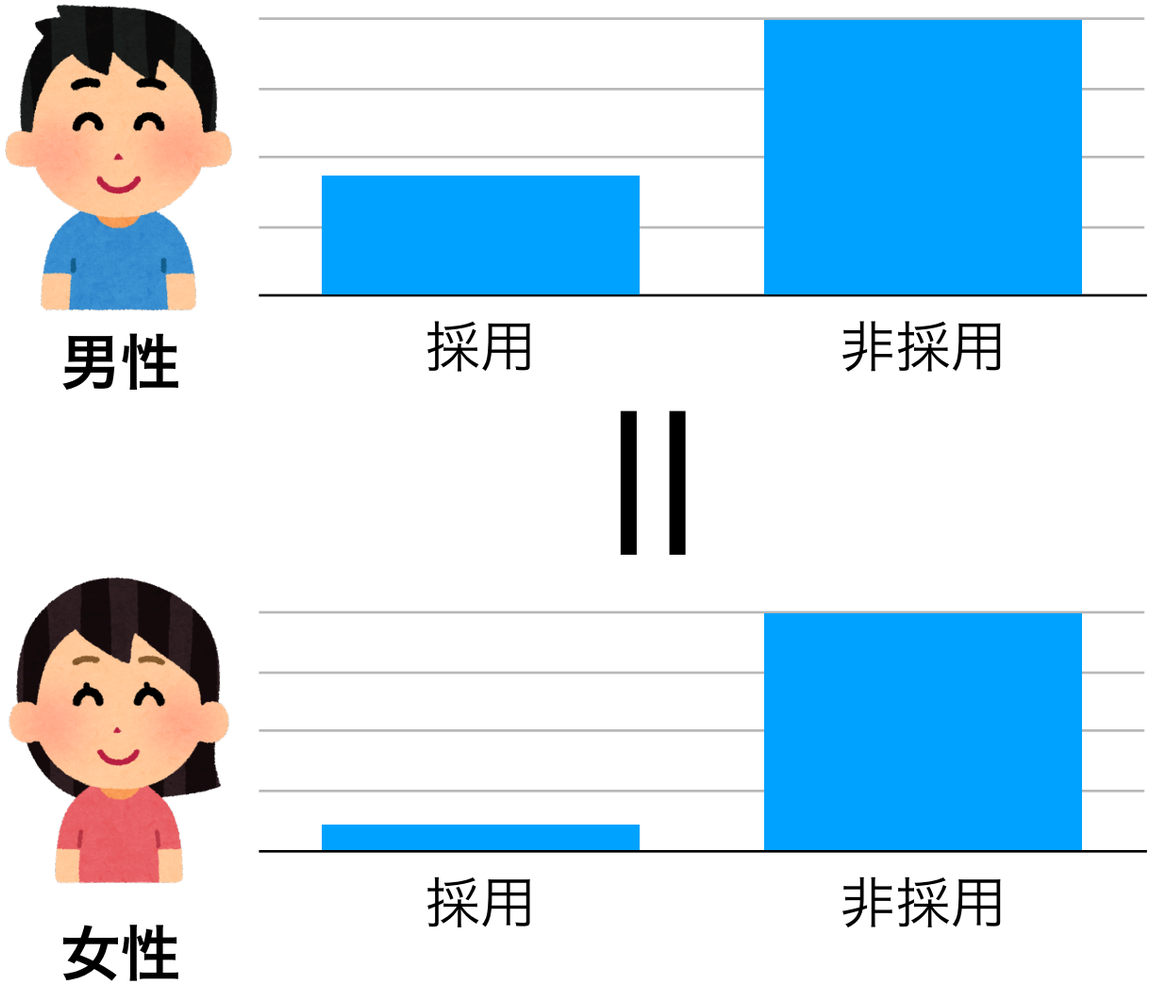
データ

学習

公平性定義

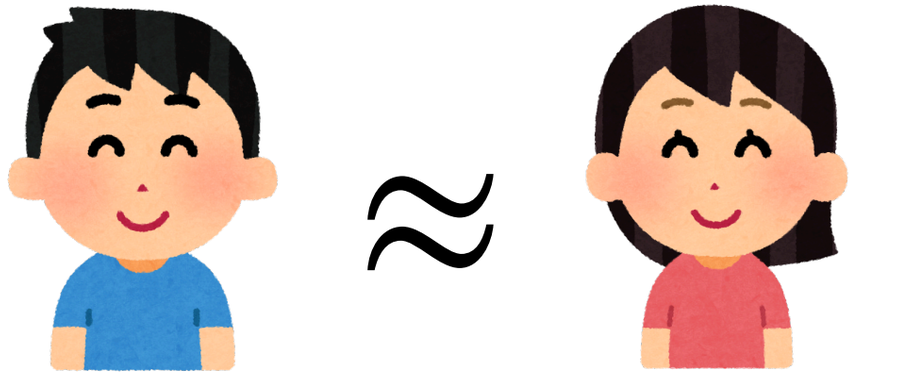
集団公平性 (Group fairness)

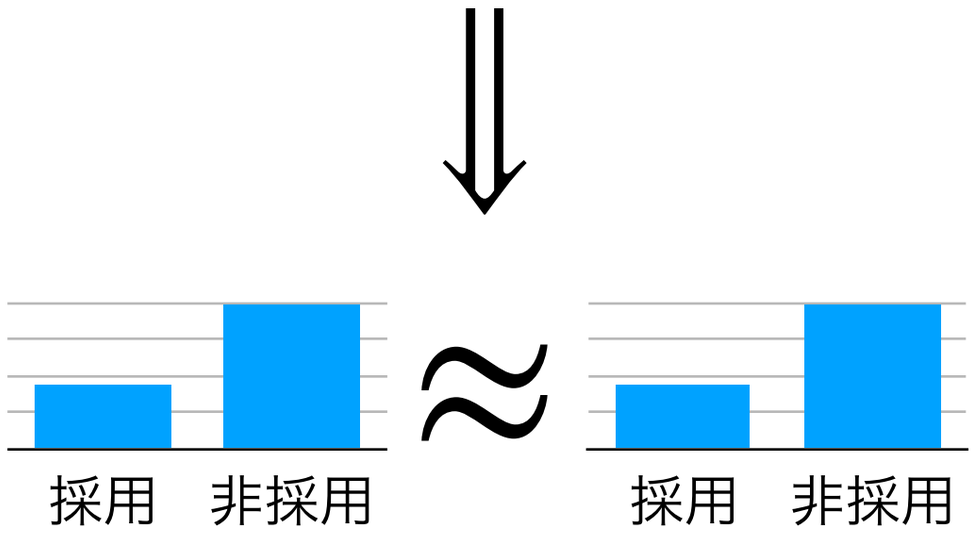
- センシティブ属性によるグループ間での差異



個人公平性 (Individual fairness)

- 個人間での差異





集団/個人 + データ/学習

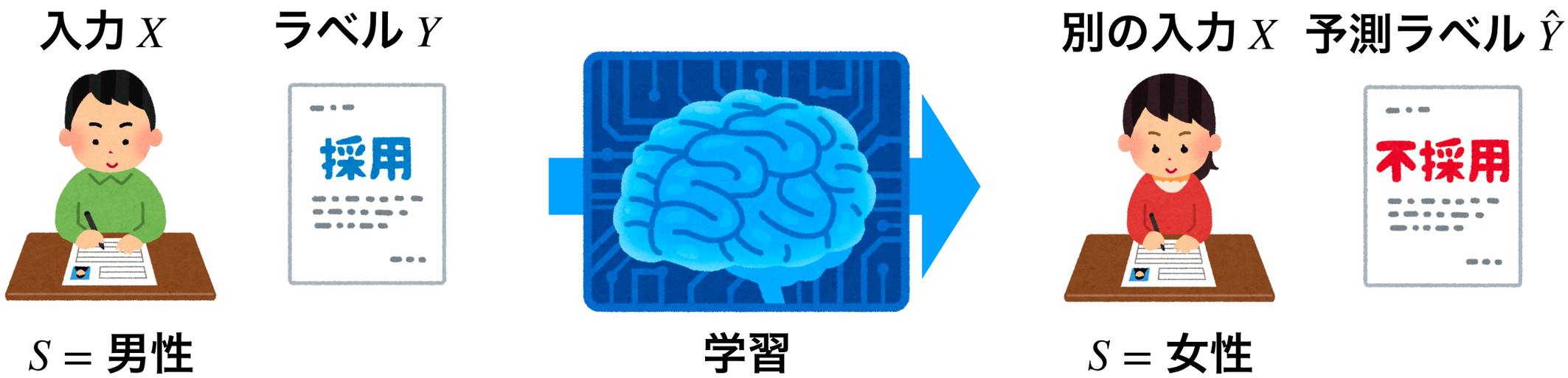
	バイアス in データ+学習	バイアス in 学習
集団公平性	1	2
個人公平性	3	4

集団/個人 + データ/学習

	バイアス in データ+学習	バイアス in 学習
集団公平性	1	2
個人公平性	3	4

設定

- 簡単のために教師あり分類問題のみを考える
  - **入力** X　　学歴，職歴，資格など
  - **センシティブ属性** S　　性別，人種，宗教，政治志向，年齢など
  - **ラベル** Y　　予測したいもの (e.g., 採否)
  - **予測ラベル** $\hat{Y}$　　アルゴリズムによって予測されたラベル



集団/個人 + データ/学習

	バイアス in データ+学習	バイアス in 学習
集団公平性	1	2
個人公平性	3	4

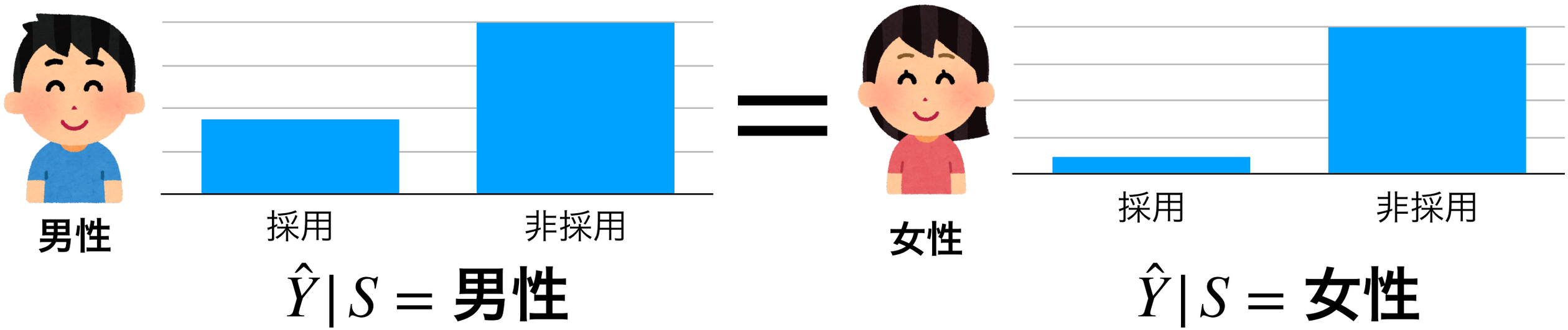
Demographic parity

Demographic parity

任意の$\mathscr{A}, s, s'$**について**

$$\mathbb{P}\{\hat{Y} \in \mathscr{A} \mid S = s\} = \mathbb{P}\{\hat{Y} \in \mathscr{A} \mid S = s'\}$$

- センシティブ属性で条件づけられた予測ラベルの分布が一致
  - 分布ではなく予測精度/偽陽性/偽陰性の一致を目指すものもあり



$\hat{Y} \mid S =$ **男性**　　$\hat{Y} \mid S =$ **女性**

Demographic parity

Demographic parity

任意の$\mathscr{A}, s, s'$**について**

$$\mathbb{P}\{\hat{Y} \in \mathscr{A} \mid S = s\} = \mathbb{P}\{\hat{Y} \in \mathscr{A} \mid S = s'\}$$

- データにバイアスがのっている可能性があるため
  ラベルYと予測ラベル$\hat{Y}$は異なるべき

- ラベルと違う予測を与えると予測性能が下がる

予測性能と公平性のトレードオフの効率化が目的

集団/個人 + データ/学習

	バイアス in データ+学習	バイアス in 学習
集団公平性	1	2
個人公平性	3	4

Equalized odds [Hardt+16]

Equalized odds

任意の$\mathscr{A}, y, s, s'$**について**

$$\mathbb{P}\{\hat{Y} \in \mathscr{A} \mid Y = y, S = s\} = \mathbb{P}\{\hat{Y} \in \mathscr{A} \mid Y = y, S = s'\}$$

- ラベルYによる予測ラベル変化は差別を生じない (と信じている)
- 無理やりDPを保障しようとすると**逆差別**

DPの保証のため男性を非採用にする

採用するようにした女性より
非採用にした男性の方が能力が高い

Equalized odds [Hardt+16]

Equalized odds

任意の$\mathscr{A}, y, s, s'$について

$$\mathbb{P}\{\hat{Y} \in \mathscr{A} \mid Y = y, S = s\} = \mathbb{P}\{\hat{Y} \in \mathscr{A} \mid Y = y, S = s'\}$$

- Yと$\hat{Y}$を一致させるように学習できる
  - Demographic parityではできない
  - 理想的にはトレードオフはない

少数グループ無視の防止が目的

DP vs. EO

DP

- Pros : バイアスがのったデータにも対応している
- Cons : **逆差別**の問題あり

EO

- Pros : (DPに比べて)予測性能がよい, 逆差別は起きない
- Cons : 不公平なバイアスが乗ったデータには使えない

どちらも同時に達成することは不可能

集団公平性の理論的課題 1

汎化的な公平性の保証

- 学習時にテスト時の公平性の保証

- 既存結果
  - 相関を基にしたDPの公平性指標の一様収束 [Fukuchi+14]
  - クラスの大きさに依存しないEOの公平性指標の収束 [Woodworth+17]
  - 後処理型のアルゴリズムにおけるDPまたはEOの汎化公平性指標バウンド [Agarwal+18]
  - クラスの大きさに依存しないDPの公平性指標の収束 [Cotter+18]

集団公平性の理論的課題 1

- 普通の汎化バウンド: 予測モデルの複雑さを用いる
- 公平性の汎化バウンドは予測モデルの複雑さに依存しなくできる [Woodworth+17,Cotter+18]
  - 代わりにラベルの種類数 × センシティブ属性の値の種類数に依存

理想的には: 最適性をもった学習方法の発見

- 集団公平性における最適性の解析はまだない

集団公平性の理論的課題 2

学習アルゴリズムである最適化問題の最適化アルゴリズム

- 公平性の制約を言えれると非凸な最適化問題が現れる
- (近似)最適解を得られることを保証する必要あり

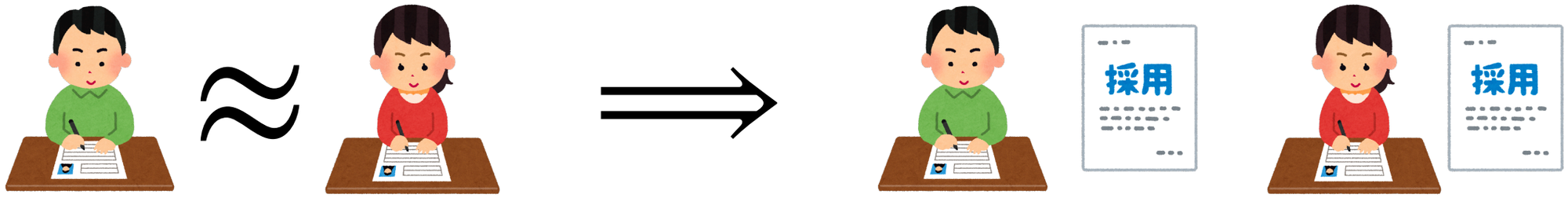
- 既存研究
  - 回帰+連続センシティブ属性に適用可能な非凸最適化によって定式化されるアルゴリズムの最適保証のついた最適化アルゴリズム [Komiyama+18]
  - ある種のオラクルの存在の仮定のもと近似最適解を多項式時間で求められる最適化アルゴリズム [Alabi+18]

集団/個人 + データ/学習

	バイアス in データ+学習	バイアス in 学習
集団公平性	1	2
個人公平性	3	4

Individual fairness [Dwork12]

- 性別以外全く同じ人がいれば採否も同じにするべき



- 似たような人は似た結果を受け取るべき

- 確率的予測関数：$f : \mathcal{X} \to \Delta(\mathcal{Y})$

Lipschitz property

任意のx, x'について

$$D(f(x), f(x')) \leq d(x, x')$$

結果の分布間の距離

個人公平性の理論的課題

汎化的な公平性の保証

- 個人公平性に関しても汎化的な性能の解析が必要

- 既存結果
  - データ+学習におけるバイアスを除去したい設定のもと
    PAC learningの枠組みでPAな公平性の制約下におけるサンプル複雑度の解析
    [Rothblum+18]

最適性をもった学習方法の発見は大きな課題

Fair bandit [Joseph+16]

最適性の証明ができている問題もある

- バンディット問題においてアームの選択分布に公平性の制約
  - データにはバイアスが入っていない
  - 学習におけるバイアスの除去が目的
- 敵対的バンディットの設定で最適なアルゴリズムを提案
- 文脈付きバンディットにおける結果もあるが最適性はなし

最近の展開

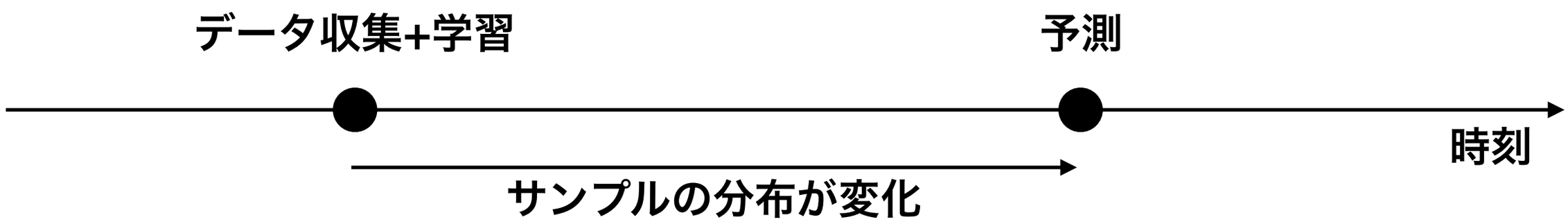
- 既存の公平性の定義に疑問
- **理論的に公平性定義の正当化をしたい**

- 既存の定義の問題点
  - Demographic parity: 逆差別
  - Equalized odds: データに含まれるバイアスを取り除けない
  - Individual fairness: 距離函数の定義

Delayed Effect [Liu+18]

- 学習とテストの間に時間的隔たりがある
  - その間にサンプルの分布が変化する
- Demographic parityの正当性 (入試)
  - 貧困層の学生を取らないことで将来貧困が拡大することの防止

- DP, EOの制約をつけた時予測時の性能はどうなるか?



まとめ

- 公平性における理論的課題
- 汎化性能の解析はやられ始めている
  - 最適性の証明が大きな課題として残っている
- 理論的な公平性定義の正当化は一番大きな課題