資料5

AIによる意思決定の公平性

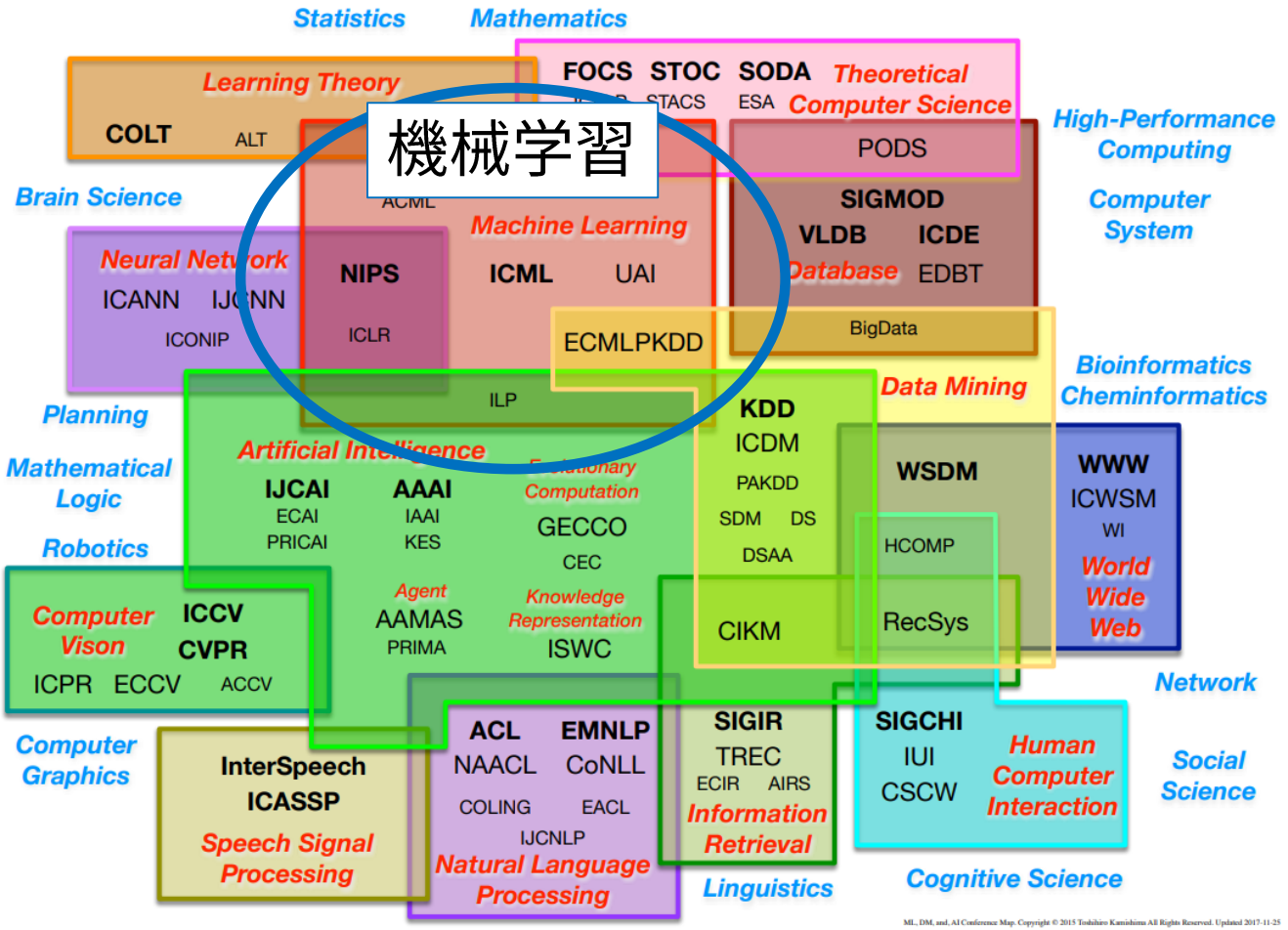
小宮山 純平

（東京大学 生産技術研究所）

自己紹介

- 小宮山純平
- 東京大学 生産技術研究所 助教 (2016年4月-)
- 所属：情報・エレクトロニクス部門
- 専門：機械学習・データマイニング
  - 機械学習の意思決定の公平性についての研究を今年国際会議（ICML）で発表

AI研究領域（全体）と関連領域



（図は神嶌「ML, DM, and AI Conference Map」より）

機械学習

- At its core, NIPS is an academic conference with hundreds of papers that describe the development of machine learning models and the data used to train them.
- AIの中でも、データからの学習モデルと学習手法に主眼を置いた分野
- 神嶌 “公平配慮型データマイニング技術の進展” (人工知能学会, 2017)がこの分野における公平性に関する短いレビュー論文

概要

- 法的側面
- 機械学習・AIによる差別事例
- アルゴリズムで差別が起こる理由
- 制度設計（メカニズムデザイン）

法的側面

- 公平性：人間が関係する意思決定に付随
- 意思決定の公平性
  - AIが想定されていないが、AIによる意思決定を行ってもこれらの法律に準拠する必要あり
- 以下を順番に解説
  - EEOC
  - 男女雇用機会均等法
  - White House Report
  - GDPR
  - 差別的扱いと差別的効果

EEOC (米国雇用機会均等委員会)

- “Adverse impact”の禁止
- 4/5ルール
- 「雇用・昇進などの人事の局面において、優遇されるグループに対して、それ以外のグループが80%（4/5）以下の採用率である」ことを法的に禁止
- グループ＝性別、人種、宗教など
- 後述する「Griggs v. Duke Power Co.」の訴訟事件と深く関係

男女雇用機会均等法

- 募集・採用、配置（業務の配分及び権限の付与を含む）・昇進・降格・教育訓練、一定範囲の福利厚生、職種・雇用形態の変更、退職の勧奨・定年・解雇・労働契約の更新について、性別を理由とする差別を禁止

White House Report [Podesta+14]

- Big Data and Discriminationの項目あり
- アルゴリズムの透明性や説明可能性の難しさについても言及
- https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf

GDPR (EU一般データ保護規則)

- EU諸国での個人データの扱いに関する規定
- Article 5 「個人データは、そのデータ主体との関係において、適法であり、公正であり、かつ、透明性のある態様で取扱われなければならない。（「適法性、公正性及び透明性」）」

差別的扱いと差別的効果

- グループのメンバーシップ（性別、人種、性的指向、信教）における扱いの差を対象
- 差別的扱い（disparate treatment）：
  - 扱いが明示的にグループのメンバーシップに依存
- 差別的効果（disparate impact）：
  - （意図のあるなしにかかわらず）結果がグループのメンバーシップに依存

Griggs v. Duke Power Co. case

- 1971年判決（米国連邦最高裁）
- 1955年、Duke Power社が昇進のために高校学位と資格を要求
  - これらは人種のプロキシ
    - 高校学位：白人34% / 黒人12%
    - 資格保持：白人58% / 黒人6%
- 「業績を測るものではない要求は、差別意図がなくても違法」

概要

- 法的側面
- 機械学習・AIによる差別事例
- アルゴリズムで差別が起こる理由
- 制度設計（メカニズムデザイン）

差別的広告 [Sweeney 13]

- “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

(スライドは[福地 2018]より)

差別的広告 [Sweeney 13]

- 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					EXPECTED	
	BLACK		WHITE		Totals	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					EXPECTED	
	BLACK		WHITE		Totals	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

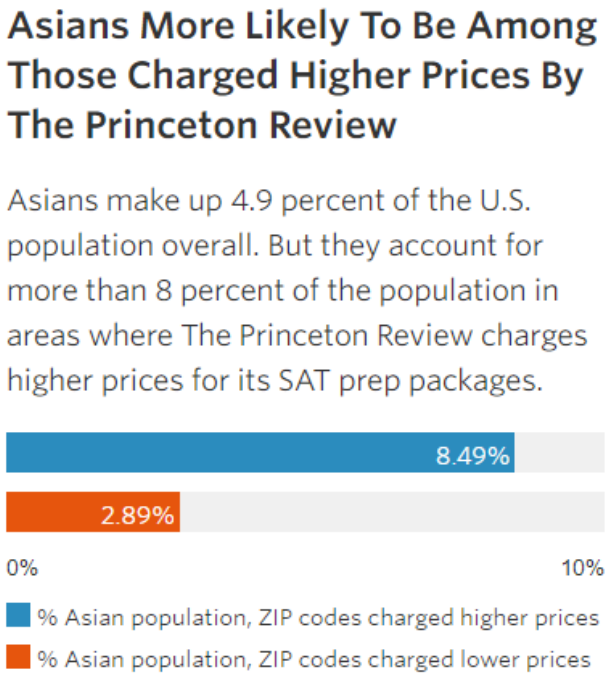
(スライドは[福地 2018]より)

COMPASスコア [Angwin+16]

- 再犯予測ソフトウェアCOMPAS（商業ソフトウェア）でフロリダ州の受刑者のデータを分析
- 教師あり（実際の各受刑者の２年後の再犯が既知）
- 黒人受刑者は実際の再犯より高いスコア、白人受刑者は実際の再犯より低いスコア傾向

The tiger mom tax [Angwin&Larson 15]

- The Princeton Review SAT prep package (オンラインチューターありの大学模試教材）は、地域（zipコード）ごとに異なる値付け（low/mid/high）
  - 「市場競争原理とコストを勘案した結果」
- 実際は…アジア系人口はそれ以外の人より高い割合（倍程度）でhighに割り当てられる傾向
- 法的に問題があるとは言い切れない



ニュース文章におけるバイアス

- “Man is to Computer Programmer as Woman is to Homemaker?” [Bolukbasi+ NIPS16].
- Google newsコーパス（一般ニュース記事）から潜在空間モデルを学習、そのgender biasを分析

Extreme *she*	**Extreme *he***
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

Gender stereotype *she-he* analogies

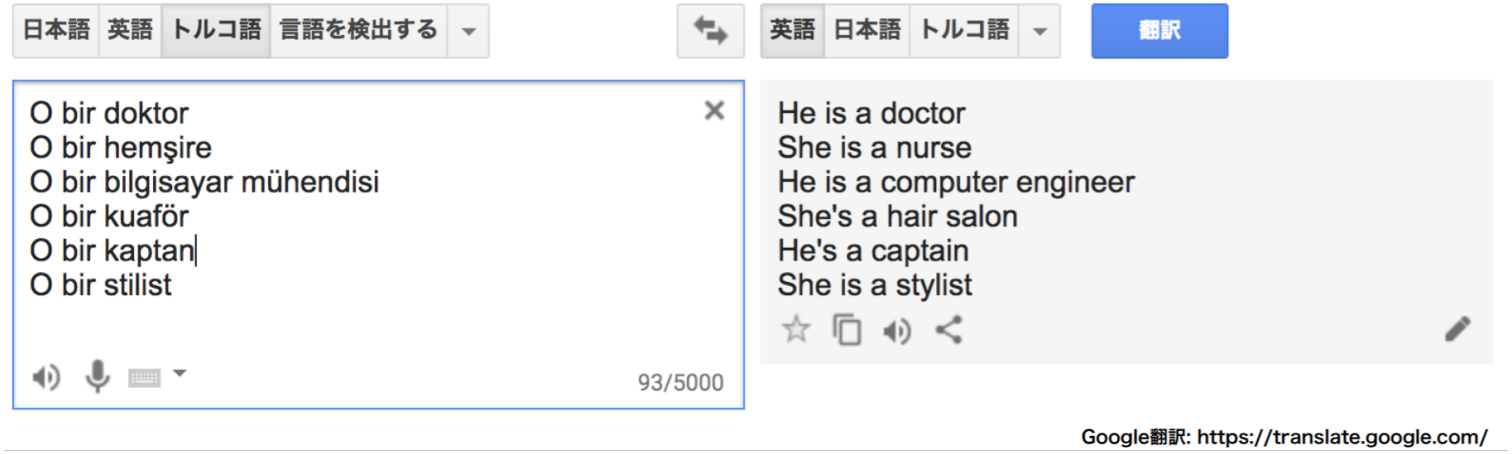
sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

Gender appropriate *she-he* analogies

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

機械翻訳における差別

- Emre Sarbakのfacebookポスト
- https://www.facebook.com/photo.php?fbid=10154851496086949&set=a
  - 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



画像は[福地 2018]より

概要

- 法的側面
- 機械学習・AIによる差別事例
- アルゴリズムで差別が起こる理由
- 制度設計（メカニズムデザイン）

バイアス

- ２つの意味が区別されずに使われている
- 英語的な意味
  - 偏見、それに基づいた判断
  - 基本的に良くないものと考えられる
- 統計学的な意味（専門用語）
  - バイアス＝全体での統計量と、そこからのサンプルでの統計量の差分
  - 「選択バイアス」
  - 統計的な意味では、どのような統計量にもバイアスは存在するし、必ずしもバイアスがあって悪いわけではない

差別的バイアスの問題点

- 法律に反す（前述）
- K. Crawfordの講演 (the trouble with bias)[Crawford 2017]では以下の２点に言及
- 経済的な側面（リソース割当）
  - 特定の層が経済的な便益を受けられなくなる
  - ローンが借りられなくなる、クレジットカードが使えなくなる、就学、就職できなくなる、etc.
- 表現的な側面
  - ステレオタイプの助長（例：黒人差別的な表現）

機械学習における公平性

- なぜ機械学習アルゴリズムが不公平な判断をする可能性があるのか？
- 大きく次の２点

1. データ収集におけるバイアス
    - 年収など統計量のバイアス
    - ラベル付けにおけるバイアス
2. 学習におけるバイアス
    - バンドワゴン効果

機械学習における公平性

- 前提として
  1. アルゴリズムに「不公平な判断」がハードコードされていなくても起こり得る
  2. 直接男女の値を使用しなくても差別が起こる（前述のdisparate impact）

     例：学歴を使って採否を決める

     例：住所を使って採否を決める

     （これらの変数は、男女と相関）

データ収集によるバイアス

- 例：UCI Census （米国国勢調査, 1994年）のデータセット
  - 収入が高収入（>5万ドル）か低収入（<5万ドル）かの２値ラベル
  - 高収入の割合が男性と女性で３倍異なる（30%, 11%）
  - https://archive.ics.uci.edu/ml/datasets/adult
- 例えば、収入を素性に加えて与信審査を行うアルゴリズムを作ると…

データ収集によるバイアス：ラベル付け

- 教師ラベルは多くの場合人間が作成
- Ex: 雇用の判断、大学入試（面接）、オーディション
- ラベル付けに（意図的・無意識どちらもありうるが）バイアスが反映される→データからの学習結果にもバイアス

学習によるバイアス

- インバランスなデータセット
- 例：過去の入社のほとんどが男性だった、日本人以外を登用したことがない、etc.
- 後述する「バンドワゴン効果」により、少数サンプルは重要視されない傾向

バンドワゴン効果

- 多数派の意見がより尊重され、少数派の意見が反映されにくい傾向
- 後述する統計的機械学習（現代の機械学習の主流）の仕組み上起きやすい

画像はwikipedia「バンドワゴン効果」より

統計的機械学習

- 統計的機械学習＝損失関数＋正則化項の和を最小化
- 多くの学習器はこの基盤の元に成立
  - SVM、集団学習、深層学習

$$\frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{y}_i) + \lambda \mathrm{Reg}(\mathcal{H})$$

損失関数：どれだけ既存データに上手く当てはまるか
・全データの平均なので、マイノリティを気にしない傾向

正則化項：どれだけ
学習結果の一般性があるか

公平性配慮型機械学習の流れ（mid2000s→現在）

- 統計的機械学習の枠組を踏襲しつつ、「公平性基準」を守るように手法を改良
- 既存の主要な機械学習タスク（分類、回帰、推薦、etc.）に対して、公平性基準を守るアルゴリズムを提案

$$\frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{y}_i) + \lambda \mathrm{Reg}(\mathcal{H})$$ **＋公平性制約**

損失関数　　正則化

センシティブ属性

- センシティブ属性を定義し、それに対する中立性（バイアスの除去）を公平性とする
  - 性別、人種、性的指向、宗教、年齢、etc.
- どのコンテキストで、何をセンシティブ属性にするか？→社会の要請

２つの主要な公平性基準

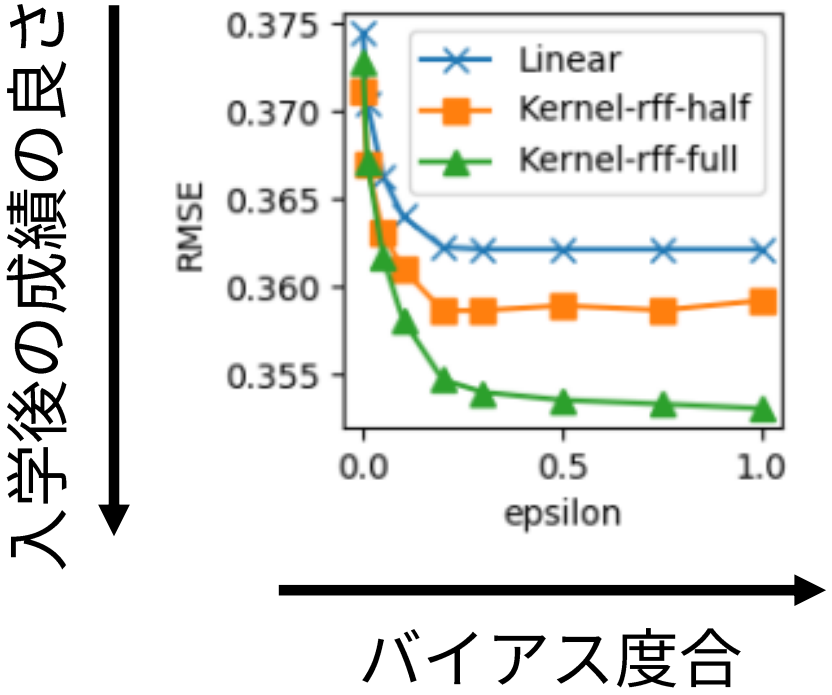
- 例として分類・回帰問題を考える（予測値Yを与える）
  - 例：人物プロファイル → Y=その人の面接での評価
- 基準１：人口等質性（demographic parity）
  - Pr[Y|S] = Pr[Y]：例えば、男女の平均評価点を同一にする（アファーマティブ・アクション的）
- 基準２：均等機会（equal opportunity）
  - それぞれの人間の「入社２年後の評価」Aが分かっているとして Pr[Y|S,A] = Pr[Y|X,A]：例えば、男女の入社後評価へのバイアスを最小化する
- どちらも「センシティブ変数に関わる特定の方向のバイアス」を除去

実用性は？

- 過去10年のAI・機械学習業界での主な結果：前頁の２要件を代表とする「公平性要件（＝公平性基準を一定レベルで守る）」制約を加えた機械学習手法を提案
  - 例： 分類 [Calders+2010], 回帰 [小宮山+ 2018], 推薦 [神嶌+2013], etc.
- これらの手法を使いたいか？
  - 現状：米テック大企業（Google, Microsoftなど）には対応する研究部署があるレベル
  - 「公平性要件を満たした方法によって意思決定を行った」という証拠を提出したい事はありうる
  - 法的なプレッシャー、炎上（ブランド毀損）対策が主なモチベーション

トレードオフ

- 機械学習の精度と、公平性要件はトレードオフの関係にある



（法学部入試の公平性：画像は[小宮山+ ICML2018]）

概要

- 法的側面
- 機械学習・AIによる差別事例
- アルゴリズムで差別が起こる理由
- 制度設計（メカニズムデザイン）

制度設計（メカニズムデザイン）

- AI・機械学習→公平性基準を定義すれば、それを満たすような判断は可能
  - 制約付き最適化問題になり、多くは意思決定の良さとのトレードオフ
- メカニズムデザイン
  - 意思決定のルールを定める（次ページ例：ルーニーのルール、ブラインド・オーディション）
  - どのようなルールが良いか？
    - 各プレイヤーが自分の利益を追求したときに、よりよい社会が実現されるルール（未解決問題）
  - どうやってルールを強制するか？
    - 法的拘束力、不文律、世論、etc.

アファーマティブ・アクションは常に良いか？

- “Will Affirmative-Action Policies Eliminate Negative Stereotypes?” [Coate & Loury 1993]
- アファーマティブ・アクションで優遇されるグループは、努力するモチベーションを失うことがある
- 常にアファーマティブ・アクションが良いわけではない（もちろん、アファーマティブ・アクションが一概に否定されるわけではない）

ブラインド・オーディション

- 米国オーケストラの女性比率
- 5% (1970s) -> 40% (現在)
- ブラインド・オーディションの導入 [Goldin&Rouse 2000]

https://www.astridbaumgardner.com/blog-and-resources/blog/ysm-mock-auditions/

ルーニーのルール

- NFL（アメフト）のポリシー
- シニアレベルの人材の最終面接リストには、１人以上のマイノリティ候補を含めないといけない
- マジョリティ・マイノリティの間の評価バイアスがある場合は有効 [Kleinberg & Raghavan2018]

http://diversityinsport.squarespace.com/understanding-our-differences/2018/3/7/on-the-effectiveness-of-the-rooney-rule

まとめ

- 法的側面
  - EEOC、男女雇用均等法、GDPR
- 機械学習・AIによる差別事例
  - オンライン広告、再犯、地域価格差、ニュース、機械翻訳
- アルゴリズムで差別が起こる理由
  - 公平性要件を満たす機械学習手法
- 制度設計（メカニズムデザイン）
  - ルールづくりとしての公平性

参考資料

- 福地 “公平性に配慮した学習とその理論的課題”. IBISML 2018 (招待講演). http://ibisml.org/ibis2018/files/2018/11/fukuchi.pdf
- 小宮山, 武田, 本多, 島尾 “Nonconvex Optimization for Regression with Fairness Constraints.” ICML 2018.
- 神嶌 “公平配慮型データマイニング技術の進展”. 人工知能学会全国大会. 2017.
- 神嶌, 赤穂, 麻生, 佐久間 “Efficiency Improvement of Neutrality-Enhanced Recommendation.” Decisions@RecSys 2013.
- J. M. Kleinberg, M. Raghavan “Selection Problems in the Presence of Implicit Bias.” ITCS 2018:
- K. Crawford. “The Trouble with Bias.” NIPS Keynote. 2017. https://www.youtube.com/watch?v=fMym_BKWQzk

参考資料

- T. Bolukbasi. K. Chang. J. Y. Zou. V. Saligrama, A. T. Kalai. “Man is to Computer Programmer as Woman is to Homemaker?” NIPS 2016.
- J. Angwin, J. Larson, S. Mattu and L. Kirchner. “There’s software used across the country to predict future criminals. And it’s biased against blacks.” 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- J. Angwin, J. Larson. “The Tiger Mom Tax.” 2015. https://www.propublica.org/article/asians-nearly-twice-as-likely-to-get-higher-price-from-princeton-review
- J. Podesta. P. Pritzker. E.J. Moniz. J. Holdren. J. Zients. “Big Data: Seizing Opprtunities, Preserving Values.” White House Report. 2014.

参考資料

- L. Sweeney. “Discrimination in Online Ad Delivery.” ACM Queue. 2013.
- T. Calders, S. Verwer. “Three naive Bayes Approaches for Discrimination-free Classification.” Data Mining and Knowledge Discovery, Vol. 21. 2010.
- C. Goldin, C. Rouse. “Orchestrating Impartiality: The Impact of “Blind” Auditions on Female Musicians.” American Econ. Rev. 90 (4). 2000.
- S. Coate & G.C. Loury. “Will Affirmative-Action Policies Eliminate Negative Stereotypes?” American Econ. Rev. 83 (5). 1993.