Privacy and the Risk of Fairwashing in Machine Learning

Ulrich Aïvodji

17 October 2019

UQÀM
Université du Québec à Montréal

- ML models ubiquity
- health, finance
- hiring, justice . . .

- Ethical design of AI models
- Privacy, Fairness
- Transparency

Big Law Business | Law Firms | In-House | Technology | Events

Technology

Big Bad Data May Be Triggering Discrimination

Bloomberg Law - Staff Reports
Aug. 15, 2016

By Kevin McGowan, Bloomberg BNA

"Big data" is filled with promise for improving recruitment and hi
employers don't take care it can also drive them to unintentional
discrimination.

"It's a bit of a black box," said Commissioner Victoria Lipnic (R) of
Employment Opportunity Commission, referring to the formulas
and programmers develop to aid employers in their talent searcl

MarketWatch | Latest | Watchlist | Markets | Investing | Barron's | Economy | Personal Finance | Retire

Home › Industries › Internet/Online Services › **Outside the Box**

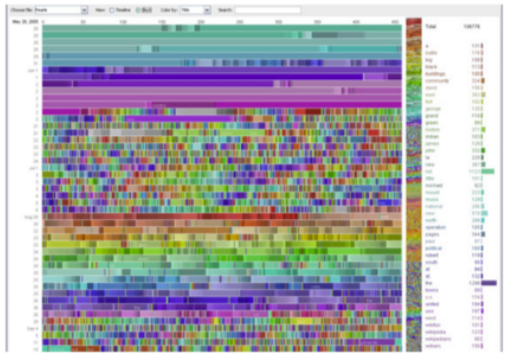
GET EMAIL ALERTS

Opinion: Big Data can lead to big legal problems for companies

Published: June 4, 2016 8:56 a.m. ET

Algorithms can perpetuate, and even amplify, racial biases by screening out certain groups

By CHRISTINE E. LYON
&
MARY RACE



Big Data is revolutionizing how business is done in many ways.

CIO | UNITED STATES | DIGITAL MAGAZINE | NEWSLETTERS | EVENTS & AWARDS PROGRAMS | VIDEO | RESOURCE LIBRARY

Home › Privacy

IDG CONTRIBUTOR NETWORK Want to Join?

THE TRUTH WILL OUT

By Robin Austin, Contributor, CIO | JUN 27, 2019 5:01 AM PDT

Opinions expressed by ICN authors are their own.

OPINION

As AI/ML permeates our lives, is privacy a thing of the past?

The World Economic Forum has brought together the most influential scientists, economists and entrepreneurs to deliberate on the positive and negative aspects of our digital future living and being in what we collectively are calling the age of the 4th Industrial Revolution. Although AI and ML have many positive applications, there are those who fear losing their privacy to AI/ML's invasiveness, to governments implementing AI/ML seeking economic domination, to organized crime cyber-thieves seeking financial gain or to the lawyers who too seek financial gain from AI/ML oversite and pervasiveness.

Summary

What is privacy?

Alan Westin

"the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others."

Universal concept

- Universal Declaration of Human Rights (Article 12)
- EU GDPR, Canada PIPEDA, ...

Personal data

EU GDPR – General Data Protection Regulation

"means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;"

Canada PIPEDA
Personal Information Protection and Electronic Documents Act

"personal information includes any factual or subjective information, recorded or not, about an identifiable individual. This includes information in any form, such as: age, name, ID numbers, income, ethnic origin, or blood type; opinions, evaluations, comments, social status, or disciplinary actions; and employee files, credit records, loan records, medical records, existence of a dispute between a consumer and a merchant, intentions (for example, to acquire goods or services, or change jobs)"

Data model and Adversary

Data model

- Quasi identifier (QID): Eg: (zip code, DOB, gender)

Adversary

Any entity that seeks to recover personal data of a data subject, without her explicit consent, to establish a profile or to infer her private data

* Simple demographics often identify people uniquely (Sweeney, 2000)
* Estimating the success of re-identifications in incomplete datasets using generative models (Rocher *et al.*, 2019)

Privacy properties

Strong properties

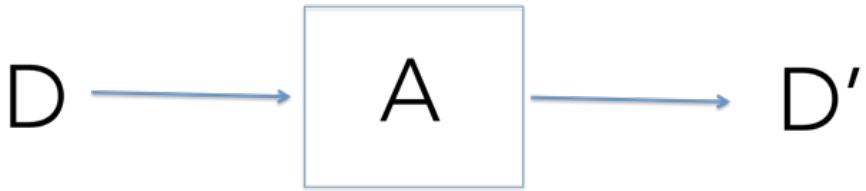
- **Confidentiality:** access only granted to authorized entity
- **Anonymity:** impossibility to identify a data subject from a group
- **Un-linkability:** impossibility to connect different records to a data subject
- **Un-observability:** impossibility to detect the absence/presence of a data subject

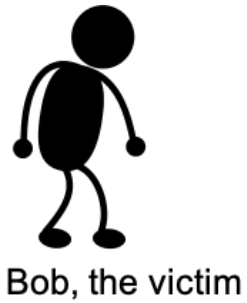
Soft properties

- **Content awareness:** awareness on data generated, data minimization
- **Policy and consent compliance:** regulatory compliance, explicit agreement of data subjects

* A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements (Deng *et*

Inference attacks





- Record linkage: record $x \in D'$ belongs to Bob. x has property $p_x \implies$ Bob has property p_x
- Attribute linkage: Bob belongs to subgroup $G \subset D'$ whose members have properties $p_1, \ldots p_n$
- Table linkage: Bob belongs to dataset D' whose members have properties $p_1, \ldots p_n$
- Probabilistic attack: the adversary improves some beliefs on Bob after obtaining D'

*Privacy-preserving data publishing: A survey of recent developments (Fung *et al.*, 2010)

Protection mechanisms – Pseudonymization

- Replace identifier by pseudo
- Bad idea!!!
- Uniqueness of QID

Name	Age	Zip code	Gender	Disease
Yoshida	23	H2X 1Y9	M	Asthma
Cohen	29	H2X 3X2	F	Hypertension
Achebe	25	H2X 3E2	F	Schizophrenia
Murphy	42	H2S 3C7	M	Cancer
Bouchard	45	H2S 2L8	F	Diabetes
Smith	55	H2S 2E7	M	Influenza

Name	Age	Zip code	Gender	Disease
1	23	H2X 1Y9	M	Asthma
2	29	H2X 3X2	F	Hypertension
3	25	H2X 3E2	F	Schizophrenia
4	42	H2S 3C7	M	Cancer
5	45	H2S 2L8	F	Diabetes
6	55	H2S 2E7	M	Influenza

* Simple demographics often identify people uniquely (Sweeney, 2000)
* Estimating the success of re-identifications in incomplete datasets using generative models (Rocher *et al.*, 2019)

Protection mechanisms – Generalizations and deletions

- Hide details about QID using generalizations or deletions

k-anonymity

requires that each equivalence class (set of items that are similar with respect to a QID) contains at least k records

l-diversity

requires that each equivalence class has at least l well-represented values for each sensitive attribute

t-closeness

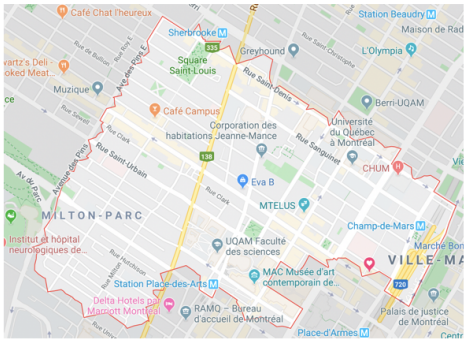
requires that, for each equivalence class c, the similarity between the distribution of a sensitive attribute s within c and the distribution of a over the whole dataset is bounded by t

* Achieving k-anonymity privacy protection using generalization and suppression. (Sweeney, 2002)
* l-diversity: Privacy beyond k-anonymity. (Machanavajjhala *et al.*, 2007)
* t-closeness: Privacy beyond k-anonymity and l-diversity. (Li *et al.*, 2007)

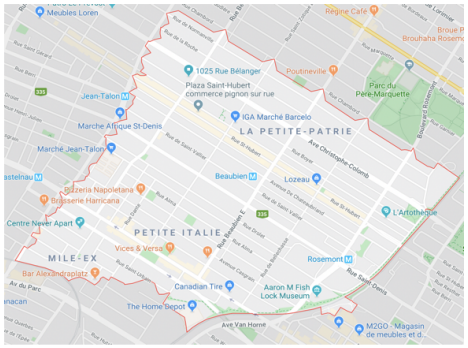
Protection mechanisms – 3-anonymity example

Name	Age	Zip code	Gender	Disease
Yoshida	23	H2X 1Y9	M	Asthma
Cohen	29	H2X 3X2	F	Hypertension
Achebe	25	H2X 3E2	F	Schizophrenia
Murphy	42	H2S 3C7	M	Cancer
Bouchard	45	H2S 2L8	F	Diabetes
Smith	55	H2S 2E7	M	Influenza

Name	Age	Zip code	Gender	Disease
*	< 30	H2X ***	*	Asthma
*	< 30	H2X ***	*	Hypertension
*	< 30	H2X ***	*	Schizophrenia
*	> 40	H2S ***	*	Cancer
*	> 40	H2S ***	*	Diabetes
*	> 40	H2S ***	*	Influenza



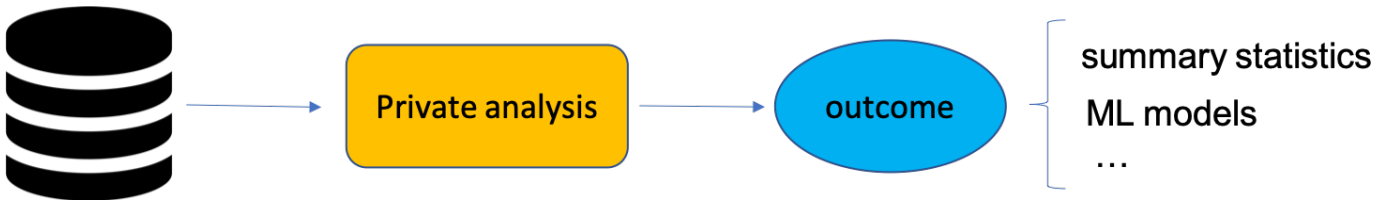
H2X



H2S

From private data publishing to private data analysis

- k-anonymity: record linkage
- l-diversity, t-closeness: record linkage, attribute linkage
- protection depend on the adversary knowledge
- composition attacks: independent private release of datasets
- need of protection against table linkage and probabilistic attack
- providing protection regardless of adversary external knowledge
- formally quantify accuracy/privacy trade-offs



* Composition Attacks and Auxiliary Information in Data Privacy. (Ganta *et al.*, 2014)

Differential Privacy (DP)

A randomized mechanism $M : \mathcal{X}^n \rightarrow \mathcal{Y}$ provides (ϵ, δ)-differential privacy if for every pair of adjacent datasets (i.e., datasets that differ only for the addition of one record) $D, D' \in \mathcal{X}^n$, and for every subset of output $S \subseteq \mathcal{Y}$

$$Pr[M(D) \in S] \leq e^{\epsilon} Pr[M(D') \in S] + \delta$$

DP aims at making the probability of M's output insensitive to the presence or absence of any particular data subject, regardless of the adversary auxiliary knowledge

* Calibrating noise to sensitivity in pri- vate data analysis. (Dwork *et al.*, 2006)
* The Algorithmic Foundations of Differential Privacy (Dwork and Roth, 2014)

l_1-sensitivity

Given a function $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$, the l_1-sensitivity Δf of f is defined as

$$\Delta f = \max_{\substack{x,y \in \mathbb{N}^{|\mathcal{X}|} \\ ||x-y||_1 = 1}} ||f(x) - f(y)||_1$$

Example

Let f be a counting query ($|\{x \in D, x \text{ has property } P\}|$),

$$\Delta f = 1$$

* The Algorithmic Foundations of Differential Privacy (Dwork and Roth, 2014)

Laplace Mechanism

Given any function $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$, the Laplace Mechanism $\mathcal{M}_L(x, f, \epsilon)$ is defined as follows:

- Compute $f(x)$
- Sample $Y \sim Lap(\frac{\Delta f}{\epsilon})^k$
- Output $f(x) + Y$

$\mathcal{M}_L$ is $(\epsilon, 0)$-differential privacy

* The Algorithmic Foundations of Differential Privacy (Dwork and Roth, 2014)

Adversary models

- Black-box adversary: model querying. Has access to only inputs and outputs
- White-box adversary: model inspection. Has access to the weights of the ML model

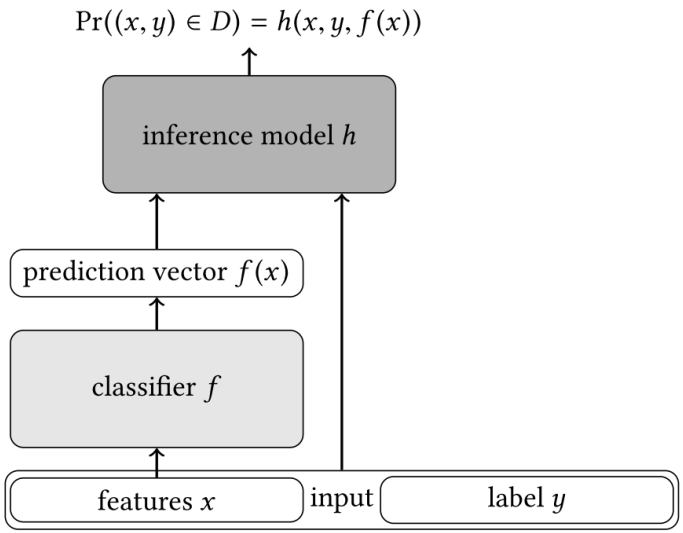
Inference attacks against ML models

- Membership inference
- Attribute inference (a.k.a. model inversion)
- Property inference
- Model extraction

* Membership Inference Attacks against Machine Learning Models. (Shokri *et al.*, 2017)
* Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures" (Fredrikson *et al.*, 2015)
* Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers. (*Ateniese et al.*, 2013)
* Stealing Machine Learning Models via Prediction APIs. (*Tramèr et al.*, 2016)

Membership inference attack

Given a black-box access to a ML model f, decide if a particular data subject x belongs to the training set D of f



- binary classifier
- distinguish predictions on members vs non-members
- adversarial learning

* Membership Inference Attacks against Machine Learning Models. (Shokri *et al.*, 2017)
* Machine Learning with Membership Privacy using Adversarial Regularization. (*Nasr et al.*, 2018)

Model inversion attack

Given a black-box access to a ML model f, infer hidden sensitive attribute of instances that belong to a particular output class y_t of f

Maximum A Posteriori (MAP) estimator

- target model f
- instance x
- prior distribution of non-sensitive attributes $x_1, x_2, \ldots x_{x_d-1}$
- prediction $y = f(x)$ of the target model

the attack identifies the value of the sensitive attribute x_d maximizing the posterior probability $P(x_d|x_1, x_2, \ldots x_{x_d-1}, y)$.

* Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures (Fredrikson *et al.*, 2015)

Property inference attack

Given a white-box access to a ML model f, decide if the training set D of f has property P. Example: ethnicity of data subjects contributing to speech recognition system's training corpus.

- train several shadow models with explicit label P and $\bar{P}$
- train a meta-classifier to distinguish between model trained with P and $\bar{P}$

* Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers. (*Ateniese et al.*, 2013)
* Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations. (Ganju *et al.*, 2018)

Model reconstruction attacks

Given a black-box access to a ML model f, create a local surrogate of f

- allows the adversary to have white-box access to a highly accurate surrogate
- can be a first step for more powerful privacy attacks

* Stealing Machine Learning Models via Prediction APIs. (Tramèr *et al.*, 2016)
* Stealing Hyperparameters in Machine Learning. (Wang and Gong, 2018)
* Model Reconstruction from Model Explanations. (Milli *et al.*, 2018)
* High-Fidelity Extraction of Neural Network Models (Jagielski *et al.*, 2019)

Privacy-preserving Empirical Risk Minimization

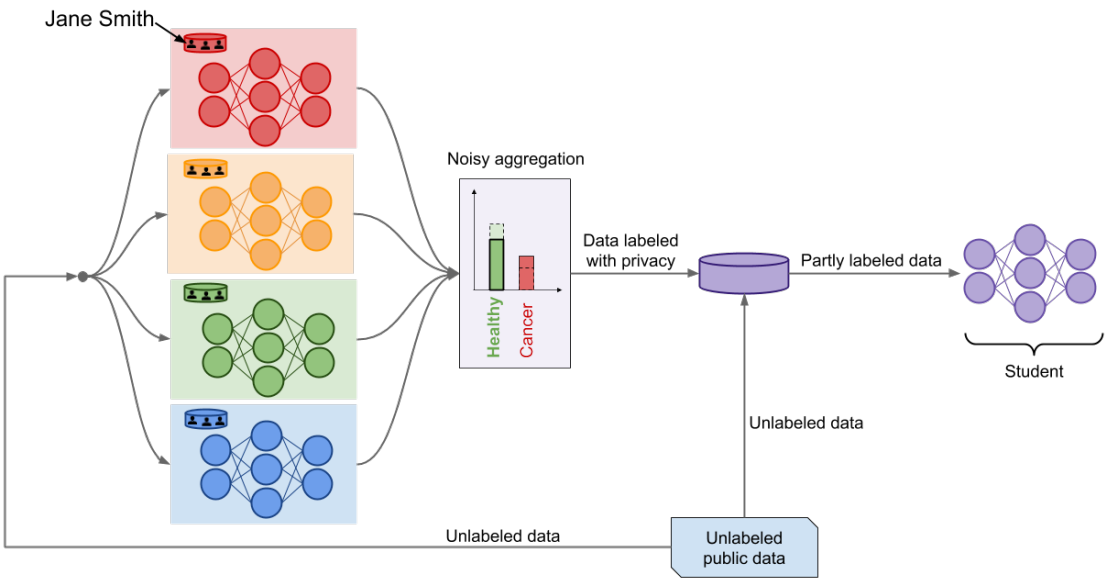
- $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
- model space $\theta \in \mathcal{H}$
- loss function $L(D, \theta) = \frac{1}{n} \sum_{i=1}^{n} l(x_i, y_i, \theta) + \frac{R(\theta}{n}$
- find $\theta^* = \underset{\theta \in \mathcal{H}}{argmin} L(D, \theta)$

- Noisy Objective: minimize $L(D, \theta) + \langle \theta, Z \rangle$
- Noisy SGD: like regular SGD with noise added to the gradients
- Noisy Output: return $\theta^* + Z$

* (Near) Dimension Independent Risk Bounds for Differentially Private Learning. (Jain and Thakurta, 2014)
* Deep Learning with Differential Privacy. (Abadi *et al.*, 2016)

Private Aggregation of Teacher Ensembles

- Model agnostic private learning
- Train teachers ensemble on private data
- Classification queries with noisy aggregation of teachers' votes



* Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data. (Papernot *et al.*, 2017)
* Privacy and machine learning: two unexpected allies? (Papernot and Goodfellow, 2018)

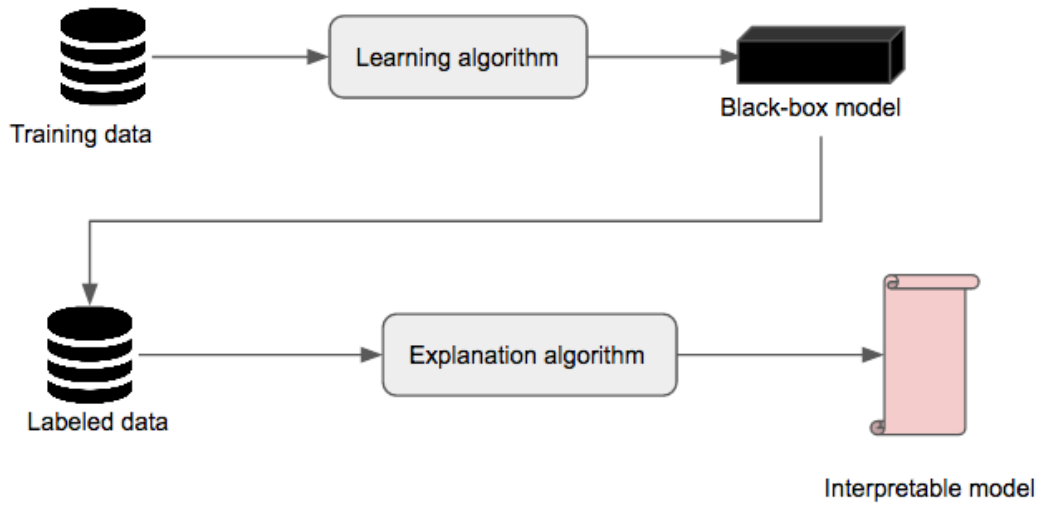
Open source frameworks

- PATE framework: https://github.com/tensorflow/privacy/tree/master/research
- Noisy SGD with tensorflow privacy: https://github.com/tensorflow/privacy
- Federated learning and differential privacy: https://www.openmined.org/
- Secure ML with Crypten: https://ai.facebook.com/blog/crypten-a-new-research-tool-for-secure-machine-learning-with-pytorch/
- DP for summary statistics: https://github.com/google/differential-privacy/

Motivations

- Interpretability by design
  - Data $\rightarrow$ decision tree
- Black-box explanation a.k.a. *post-hoc explanation*
  - DNN $\rightarrow$ decision tree
- **This work:** We show that a dishonest ML models' producer can perform *fairwashing*
- Given the false perception that a ML model complies with a given ethical requirement
- Case study: fairness as the ethical requirement to "fairwash"

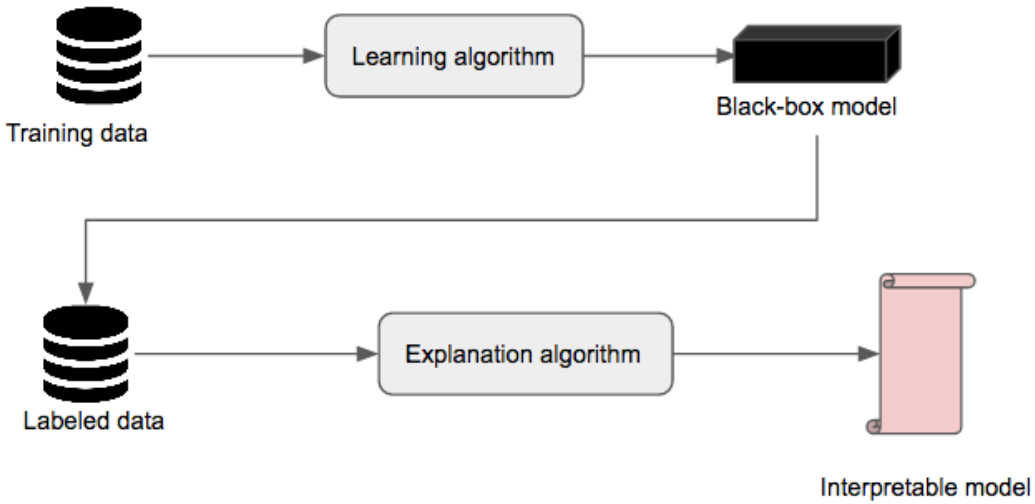
Motivations



Objective

Raise awareness of fairwashing in machine learning: the risk that an unfair ML model can be explained in such a way that the underlying decisions seem fairer than they actually were

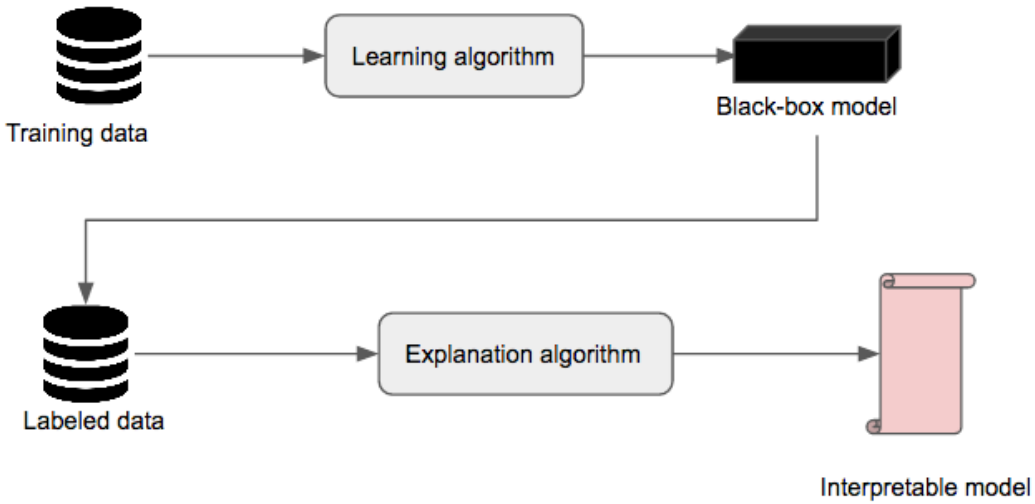
Motivations



How?

Show that one can **systematically** found a fair interpretable model to rationalize decisions of an unfair black-box model.

Motivations



Joint work with Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp

Metrics

Fairness: demographic parity

$$|P(\hat{y} = 1|s = 1) - P(\hat{y} = 1|s = 0)|.$$

Fidelity

$$\text{fidelity}(c) = \frac{1}{|X|} \sum_{x \in X} \mathbb{I}(c(x) = b(x)).$$

Rule list

A rule list $d = (d_p, \delta_p, q_0, K)$ of length $K \geq 0$ is a $(K+1)-$tuple consisting of K distinct **association rules** $r_k = p_k \rightarrow q_k$, where $p_k \in d_p$ is the **antecedent** of the association rule and $q_k \in \delta_p$ its corresponding consequent, followed by a **default prediction** q_0.

Example of rule list for salary prediction

IF occupation:white-collar **THEN** income:$\geq 50k$
ELSE IF occupation:professional **THEN** income:$\geq 50k$
ELSE IF education:bachelors **THEN** income:$\geq 50k$
ELSE income:$< 50k$

Learning optimal rule lists

CORELS (Angelino et *al.*, 2017)

- Input: n categorical attribute + binary labels
- Output: optimal rule list
- Supervised learning algorithm
- Represent the search space as a n-level trie
- Objective function: $R(d, x, y) = \text{misc}(d, x, y) + \lambda K$
- Select the rule list that minimize $R(d, x, y)$
- Use an efficient branch-and-bound algorithm to prune the trie

Enumerating rule lists

Model Enumeration (Satoshi Hara & Masakazu Ishihata, 2018)

Enumerate rule lists in a descending order of the objective function by calculating **successively** the optimal rule list using CORELS, and then constructing sub-problems **excluding the solution obtained**.

Model rationalization

Given a black-box model b, a set of instances X, and a sensitive attribute s, find a global interpretable model $c_g = f(b, X)$ derived from b and X, using some process $f(\cdot, \cdot)$, such that $\epsilon(c_g, X, s) > \epsilon(b, X, s)$, for some fairness metric $\epsilon(\cdot, \cdot, \cdot)$.

Outcome rationalization

Given a black-box model b, an instance x, its neighborhood $\mathcal{V}(x)$, and a sensitive attribute s, find a local interpretable model $c_I = f(b, x)$ derived from b and $\mathcal{V}(x)$, using some process $f(\cdot, \cdot)$, such that $\epsilon(c_I, \mathcal{V}(x), s) > \epsilon(b, \mathcal{V}(x), s)$, for some fairness metric $\epsilon(\cdot, \cdot, \cdot)$.

Better call LaundryML

- Explores the search space of rule lists with a modified version of CORELS
- New objective function:
  $\text{obj}(d, x, y) = (1 - \beta)\text{misc}(d, x, y) + \beta\text{unfairness}(d, x, y) + \lambda K$
- Enumerate rule lists
- Select fair rule lists that have higher fidelity

LaundryML

Algorithm 1 `LaundryML`

Inputs: T, λ, β
Output: $\mathcal{M}$
▷ define the objective function
$\mathsf{obj}(\cdot) = (1-\beta)\mathsf{misc}(\cdot) + \beta\mathsf{unfairness}(\cdot) + \lambda K$
Compute $m = \mathsf{CORELS}(\mathsf{obj}, T) = (d_p, \delta_p, q_0, K)$
Insert $(m, T, \emptyset)$ into the heap
$\mathcal{M} \leftarrow \emptyset$
for $i = 1, 2, \ldots$ **do**
  Extract (m, S, F) from the heap
  ▷ output m as the $i-$th model
  if $m \notin \mathcal{M}$ **then**
    $\mathcal{M} \leftarrow \mathcal{M} \cup \{m\}$
  end if
  ▷ terminate when a certain condition is met
  if $\mathsf{Terminate}(\mathcal{M}) = true$ **then**
    break
  end if
  ▷ branch the search space
  for $t_j \in d_p$ and $t_j \notin F$ **do**
    Compute $m' = \mathsf{CORELS}(\mathsf{obj}, S \setminus \{t_j\})$
    Insert $(m', S \setminus \{t_j\}, F)$ into the heap
    $F \leftarrow F \cup \{t_j\}$
  end for
end for

Algorithm 2 `LaundryML-global`

Inputs: X, b, λ, β
Output: $\mathcal{M}$
$y = b.\mathsf{predict}(\mathsf{X})$
$T = \{X, y\}$
$\mathcal{M} = \mathtt{LaundryML}(\mathsf{T}, \lambda, \beta)$

Algorithm 3 `LaundryML-local`

Inputs: x, T, $\mathsf{neigh}(\cdot)$, λ, β
Output: $\mathcal{M}_x$
$T_x = \mathsf{neigh}(\mathsf{x}, \mathsf{T})$
$\mathcal{M}_x = \mathtt{LaundryML}(\mathsf{T_x}, \lambda, \beta)$

Code available at: https://github.com/aivodji/LaundryML

Experimental Setup

Data & black-box models

- Data: Adult Income (resp. ProPublica Recidivism)
- Sensitive attribute: gender (resp. race)
- Black-box models: random forests
- Unfairness of the black-box models: 0.13 (resp. 0.17)
- Search space: 28! (resp. 27!), 50 models enumerated per experiment

Evaluation metrics

- Unfairness
- Fidelity
- Feature importance via **FairML**

Model rationalization – Unfairness and Fidelity

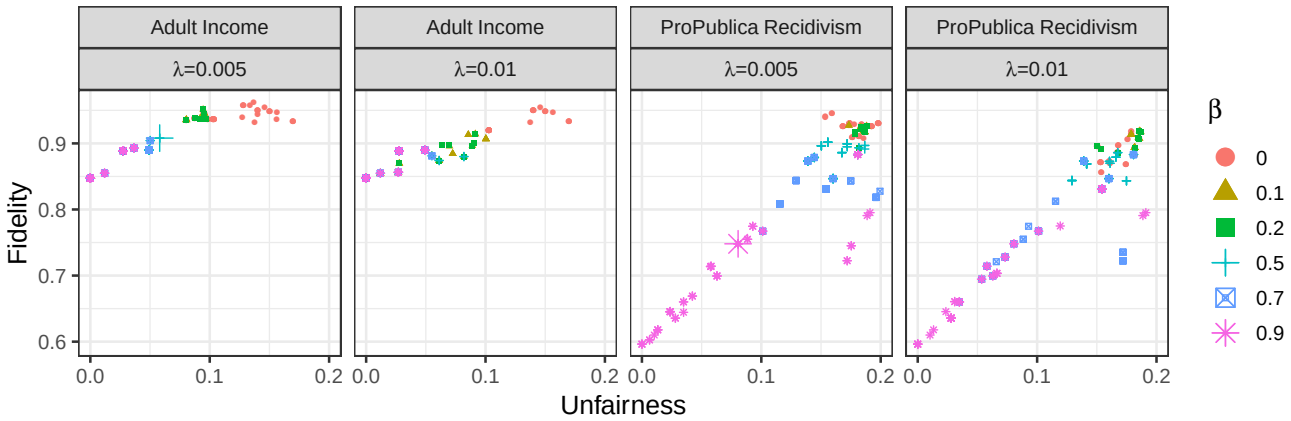
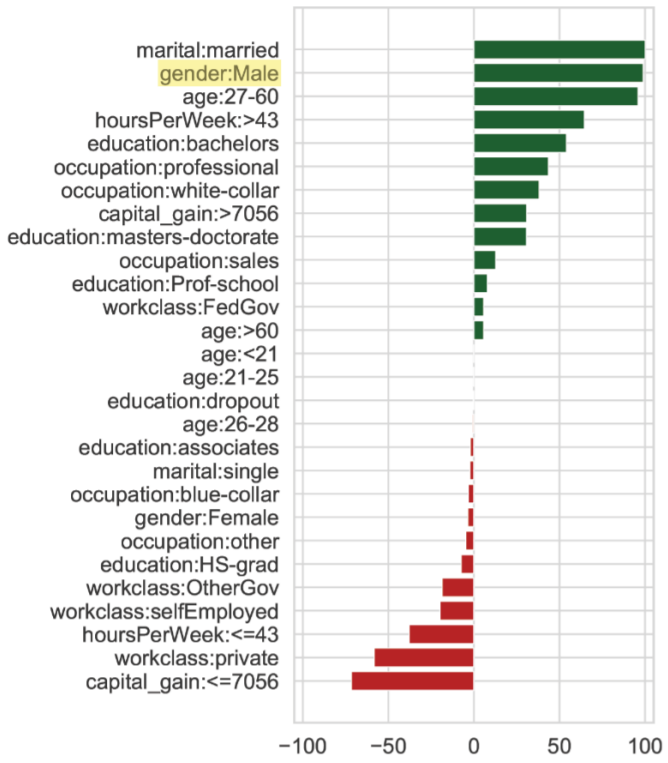


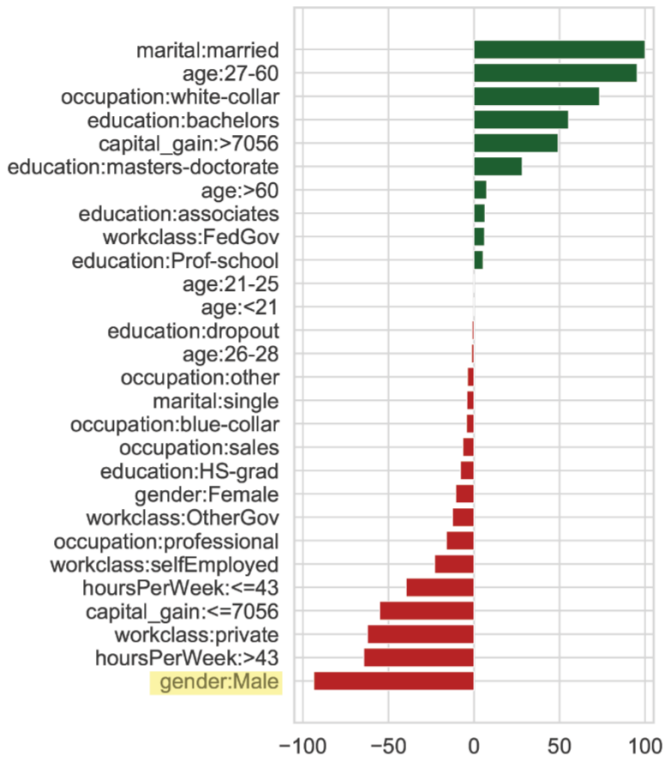
Figure: Model rationalization for Adult Income and ProPublica Recidivism.

Best rationalization models

- Adult Income: fidelity = 0.908, unfairness =0.058.
- ProPublica Recidivism: fidelity = 0.748, unfairness=0.080.

Model rationalization – Feature importance

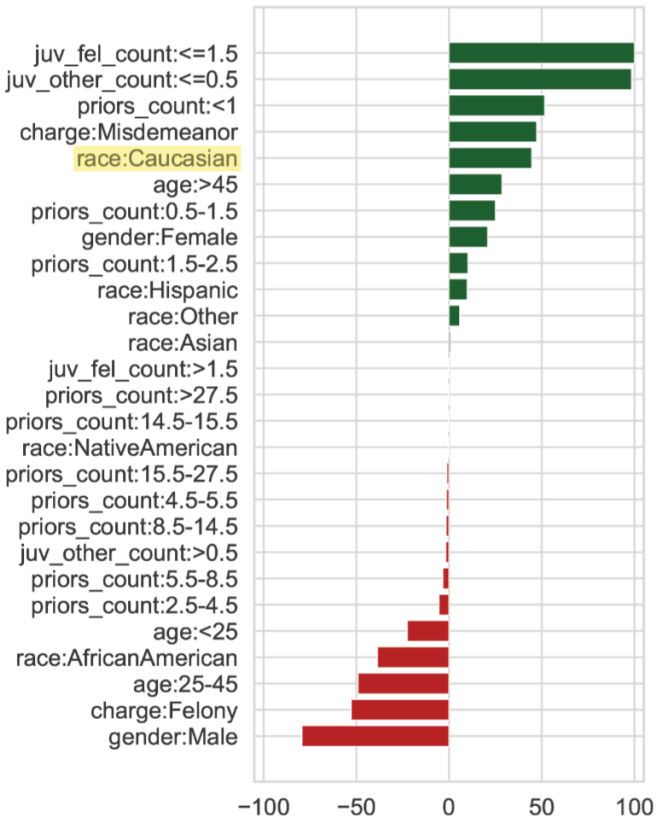


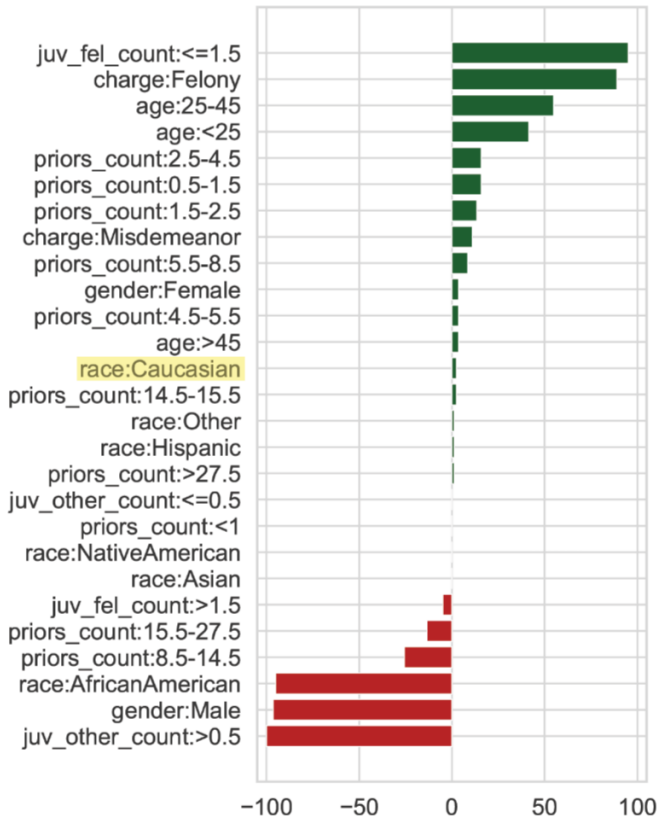


if capital_gain:>7056 **then**
  income≥50k
else if marital:single **then**
  income<50k
else if education:HS-grad **then**
  income<50k
else if occupation:other **then**
  income<50k
else if occupation:white-collar **then**
  income≥50k
else
  income:<50k

Figure: Feature importance Black-box vs Best rationalization model on Adult Income

Model rationalization – Feature importance





if prior_count: 15.5−27.5 **then**
  recidivate:True
else if prior_count: 8.5−14.5 **then**
  recidivate:True
else if age:>45 **then**
  recidivate:False
else if juv_other_count:>0.5 **then**
  recidivate:True
else
  recidivate:False
end if

Figure: Feature importance Black-box vs Best rationalization model on ProPublica Recidivism

Outcome rationalization

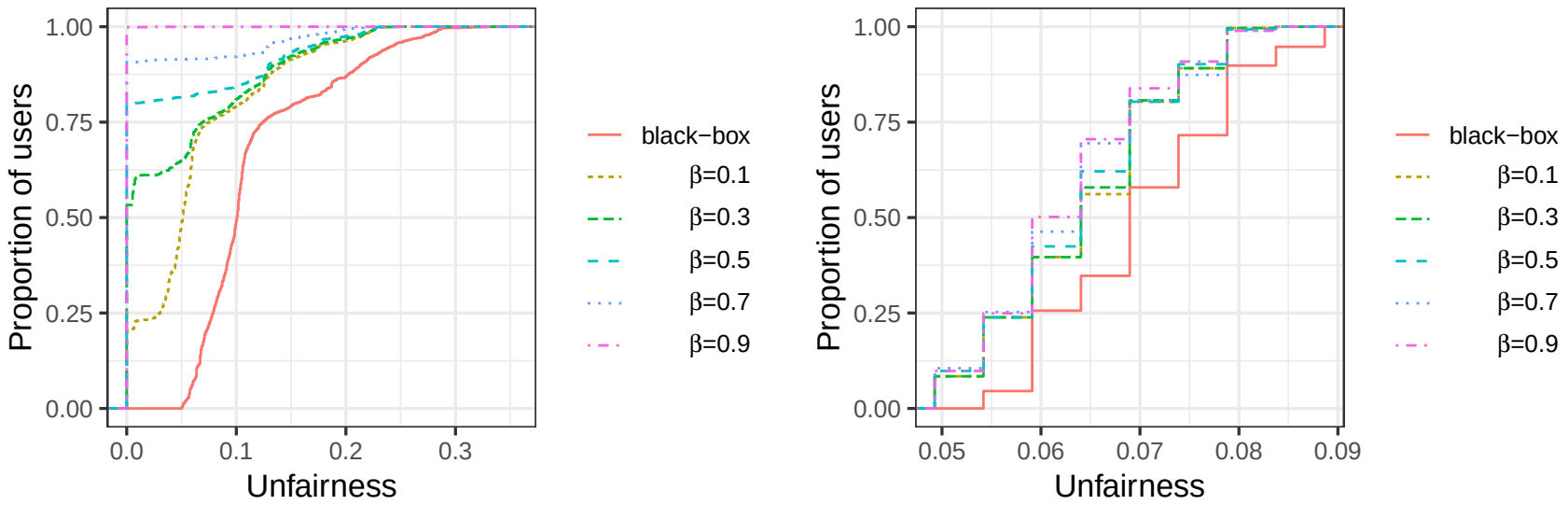


Figure: Outcome rationalization. Adult Income (left), ProPublica Recidivism (right).

Generalization to other fairness metrics (1/3)

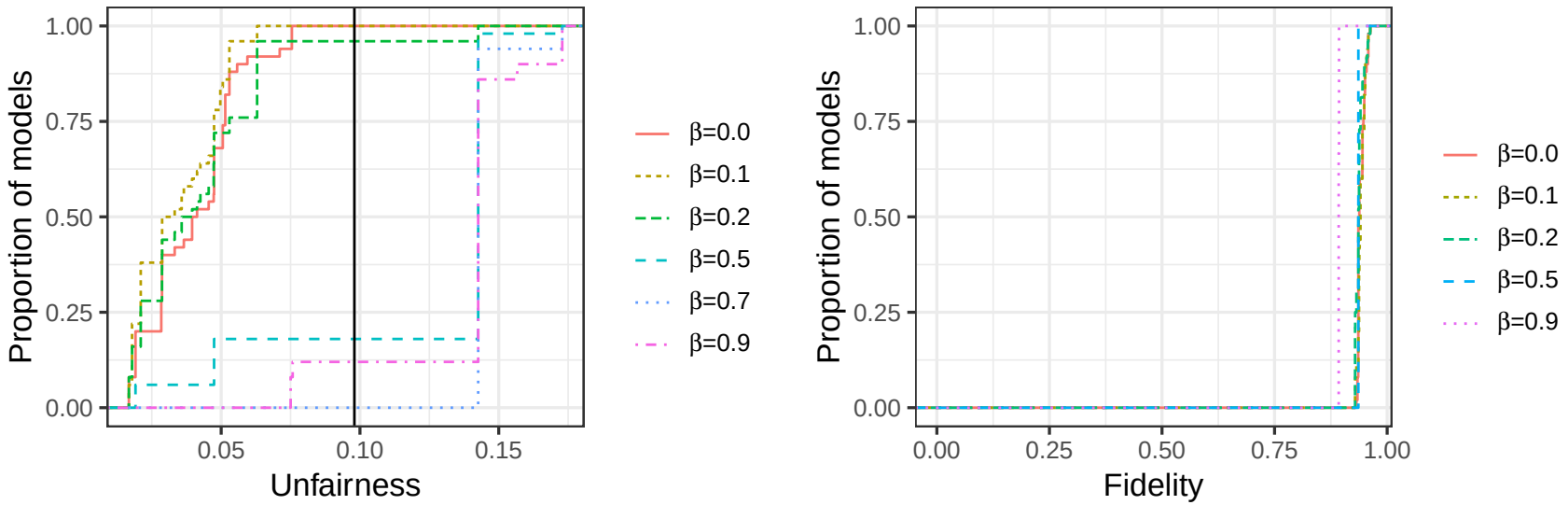


Figure: Model rationalization. Adult Income, Random forest, *Overall Accuracy Equality*.

Generalization to other fairness metrics (2/3)

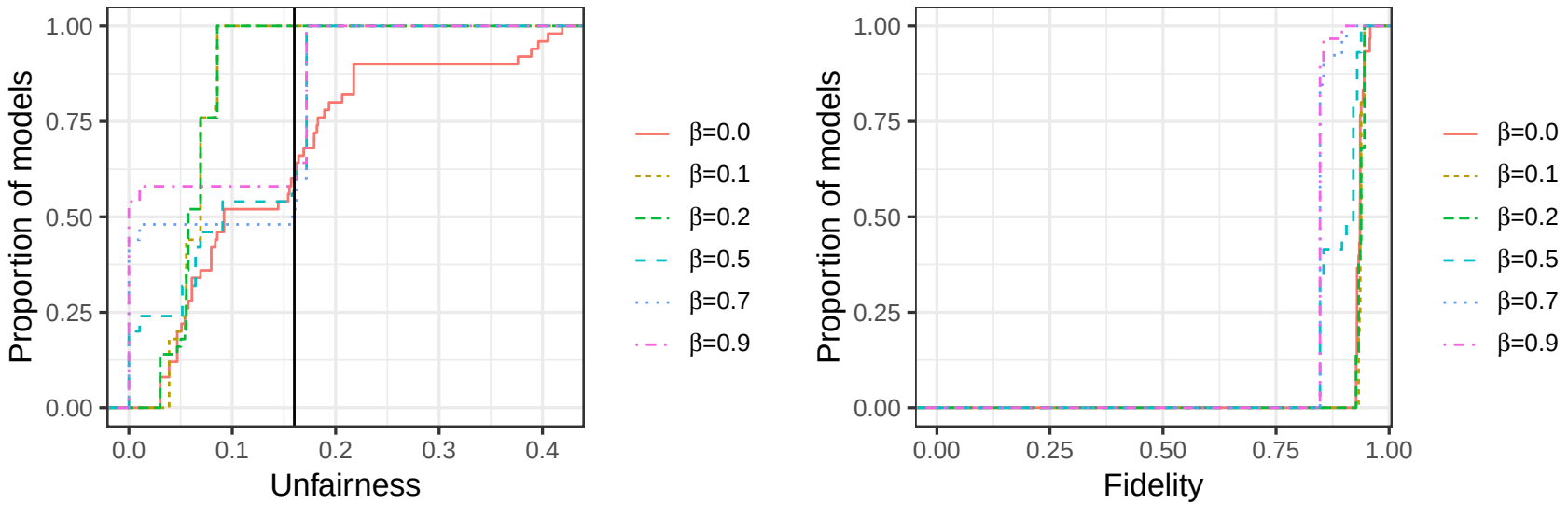


Figure: Model rationalization. Adult Income, Random forest, *Conditional Procedure Accuracy*.

Generalization to other fairness metrics (3/3)

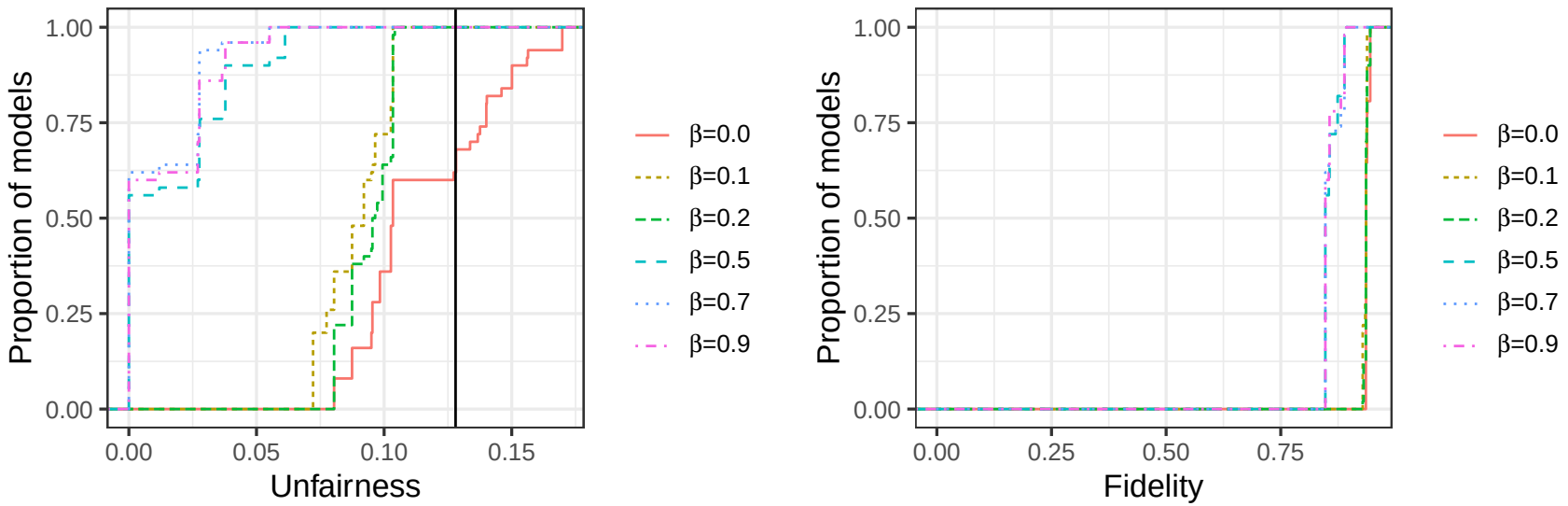


Figure: Model rationalization. Adult Income, Random forest, *Demographic parity*.

Generalization to other black-box models (1/3)

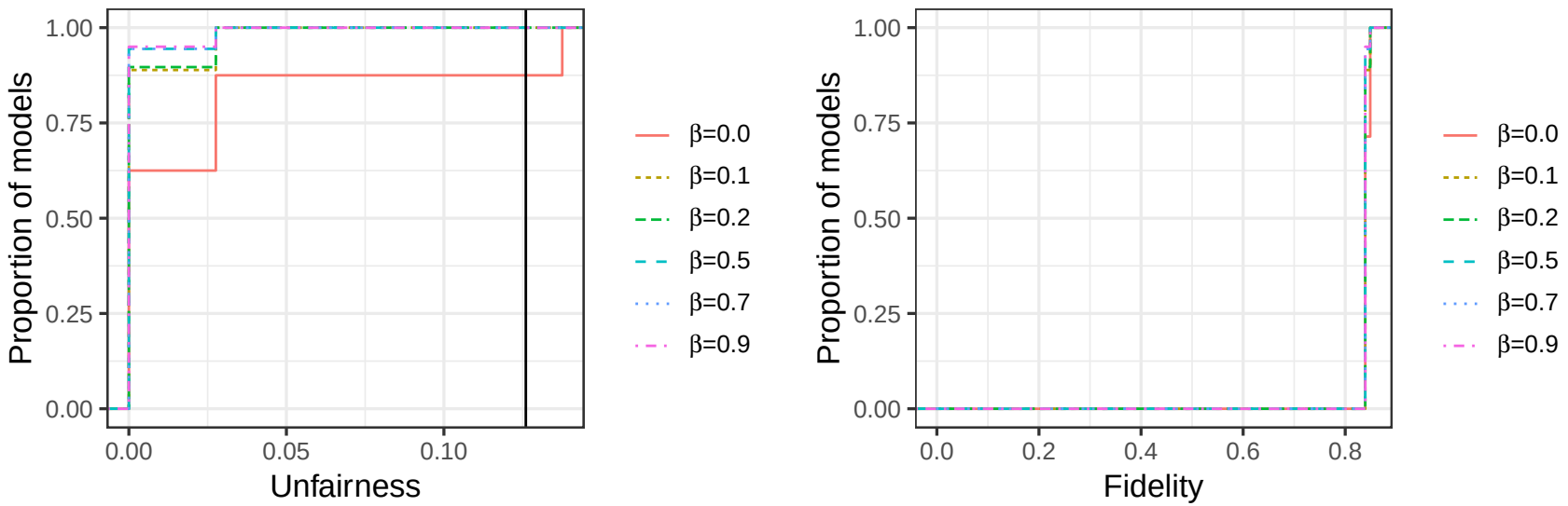


Figure: Model rationalization. Adult Income, SVM, *Demographic parity*.

Generalization to other black-box models (2/3)

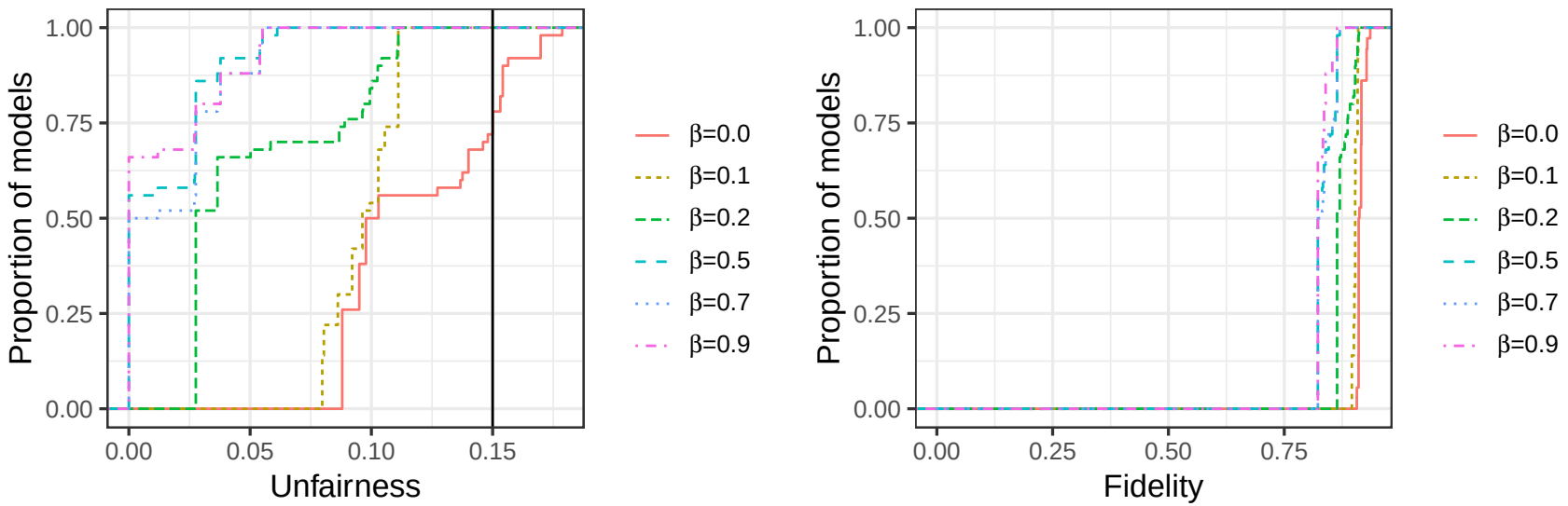


Figure: Model rationalization. Adult Income, XGBOOST, *Demographic parity*.

Generalization to other black-box models (3/3)

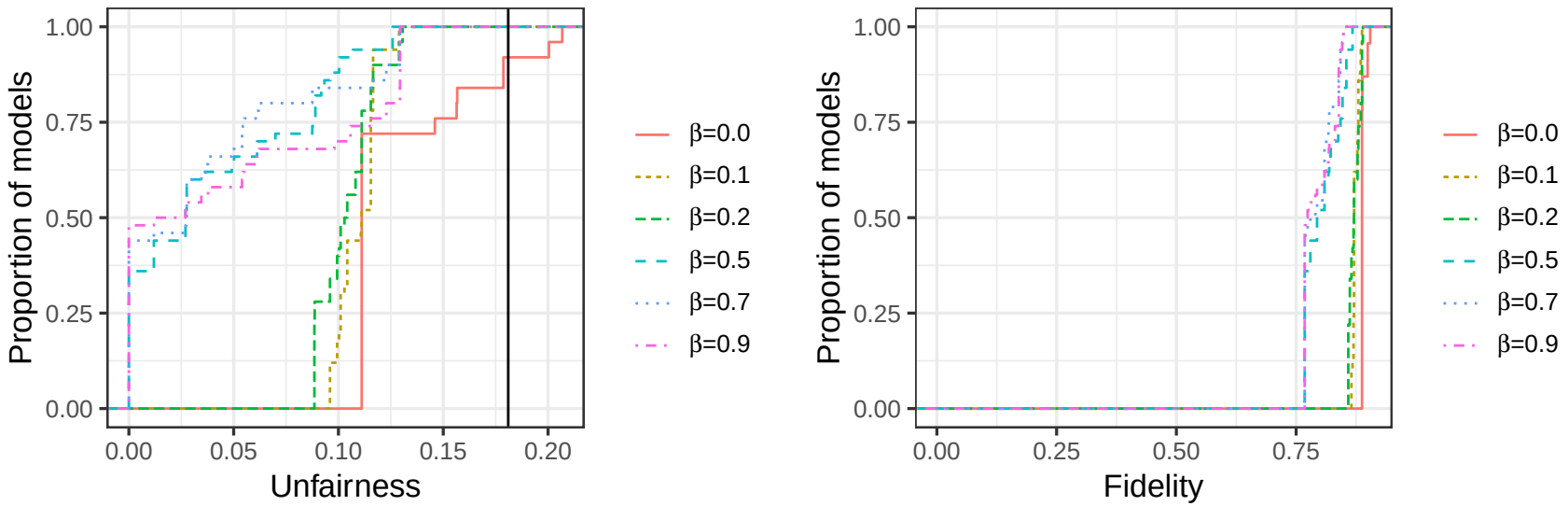


Figure: Model rationalization. Adult Income, MLP, *Demographic parity*.

Conclusion

- LaundryML: black-box explanations can be used to rationalize unfair decisions of a black-box model
- Can we trust black-box explanations?

Perspectives

- Detecting fairwashing
- Study the root cause: robustness of explanations

Learn more

- Our work: *Fairwashing: the risk of rationalization*. ICML'19
- *Pretending Fair Decisions via Stealthily Biased Sampling*. arXiv:1901.08291, 2019
- *Predictive Multiplicity in Classification*. arXiv:1909.06677, 2019
- *The Bouncer Problem: Challenges to Remote Explainability*, arXiv:1910.01432, 2019

Thank you!