Fairness-Aware
Machine Learning and Data Mining

Toshihiro Kamishima
www.kamishima.net
Updated: 2025-08-16

Fairness-Aware Machine Learning

The spread of machine learning technologies

↓

Machine learning is being increasingly applied for serious decisions
Ex: credit scoring, insurance rating, employment application

↓

Fairness-Aware Machine Learning

Data analysis taking into account potential issues of fairness, discrimination, neutrality, or independence. It maintains the influence of these types of sensitive information:

- to enhance social fairness (gender, race,…)
- restricted by law or contracts (insider or private information)
- any information whose influence data-analysts want to ignore

✻ We here use the term *'fairness-aware'* instead of an original term, *'discrimination-aware'*, because the term *discrimination* means classification in an ML context

Technical Aspects of FAML

FAML was originally invented to eliminate socially unfair outcomes when applying ML techniques to real-world problems

↓

More extensively, FAML methods would be **helpful for correcting any type of biases**, which are irrelevant to social discrimination, if **what generates the biases is known**

Ex:

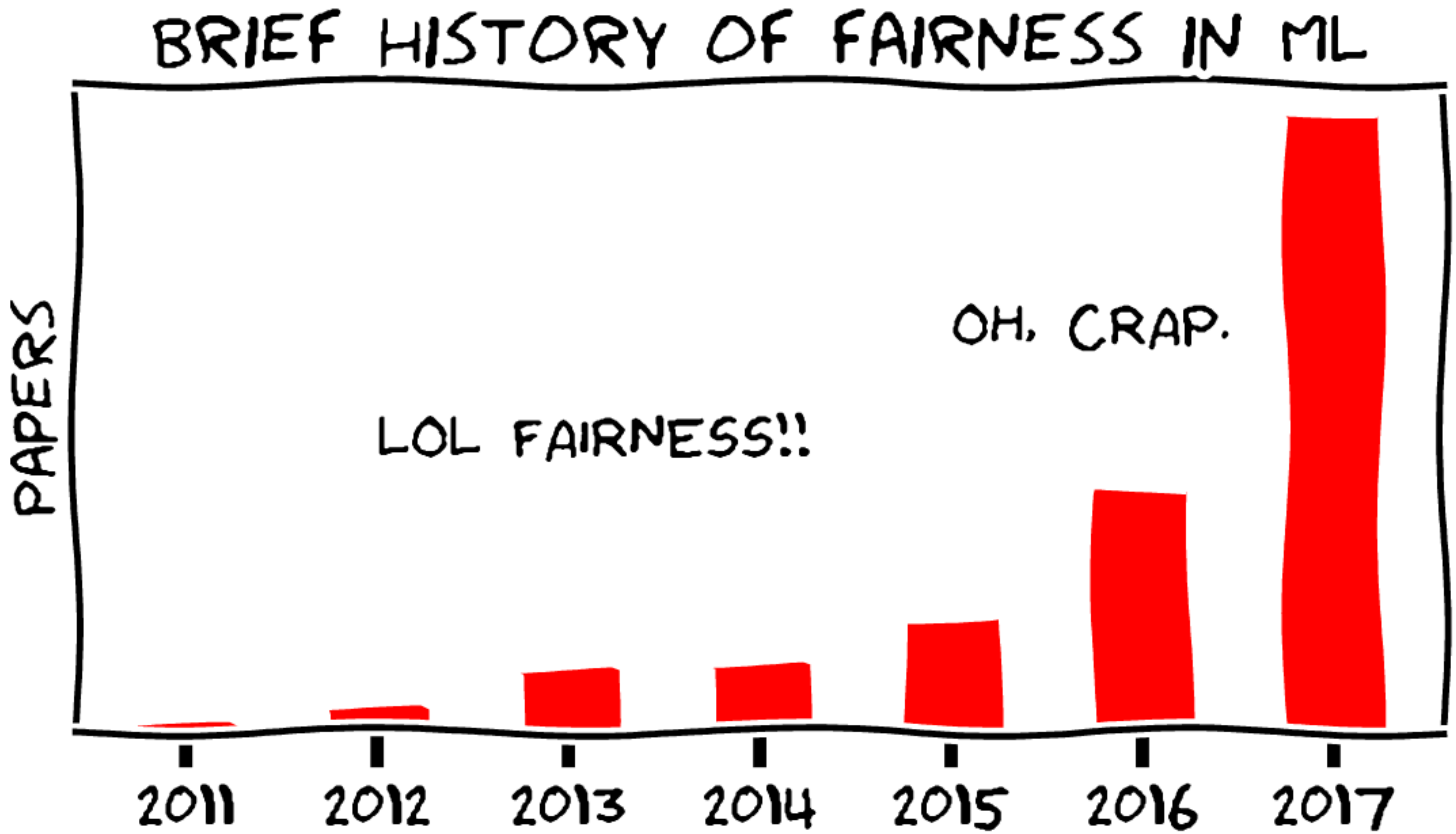
Hotels' occupancy rates are generally high, when room charges are high
Of course, the increase of occupancy rates are affected by factors besides room charges [Athey 17]

↓

If such a factor is known to be a seasonal effect, FAML methods can be used for predicting a *pure* influence from room charges to occupancy rates

Growth of Fairness in ML

[Moritz Hardt's homepage]



Distribution Web Site

The latest version of this slide is distributed at the URL:

Fairness-Aware
Machine Learning and Data Mining

http://www.kamishima.net/faml/

Outline

Part I: Backgrounds

Part I: Backgrounds

- Types of Biases
- Instances of Data Bias
- Instances of Inductive Biases

Part II: Formal Fairness

Part II: Formal Fairness

- Basics of Formal Fairness
- Association-Based Fairness
  - Basics of Associations
  - Criteria
  - Properties
  - Measures
- Counterfactual Fairness
  - Basics of Causal Inference
  - Total Fairness Criteria
  - Path-Specific Fairness Criteria
- Economics-Based Fairness

Part III: Fairness-Aware ML

Part III: Fairness-Aware Machine Learning

- Overview
- Unfairness Discovery
  - Discovery from Datasets
    - Association-based fairness
  - Discovery from Models
- Unfairness Prevention
  - Classification: Pre-process, In-process, Post-process
  - (Regression)
  - Recommendation
  - Ranking
  - (Clustering)
  - Other Tasks

Part IV: Other Topics

Part IV: Other Topics

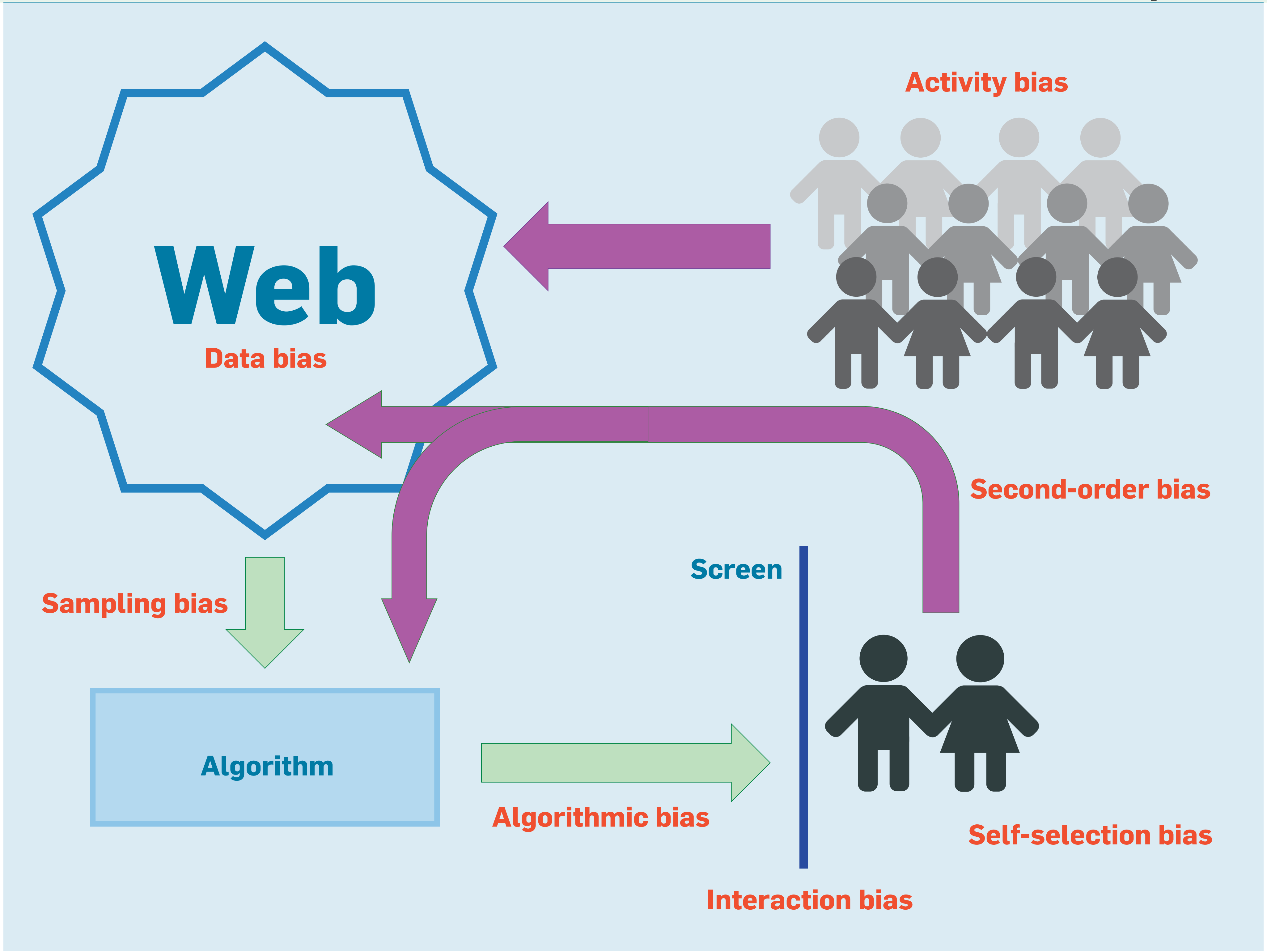
- Mitigation of a Sample Selection Bias
- Disclosure
- Other Fairness-Aware Machine Learning Topics
- Relation to the Other Machine Learning Topics
- Software
- Evidence-Based Decision Making

Part I
Backgrounds

Types of Biases

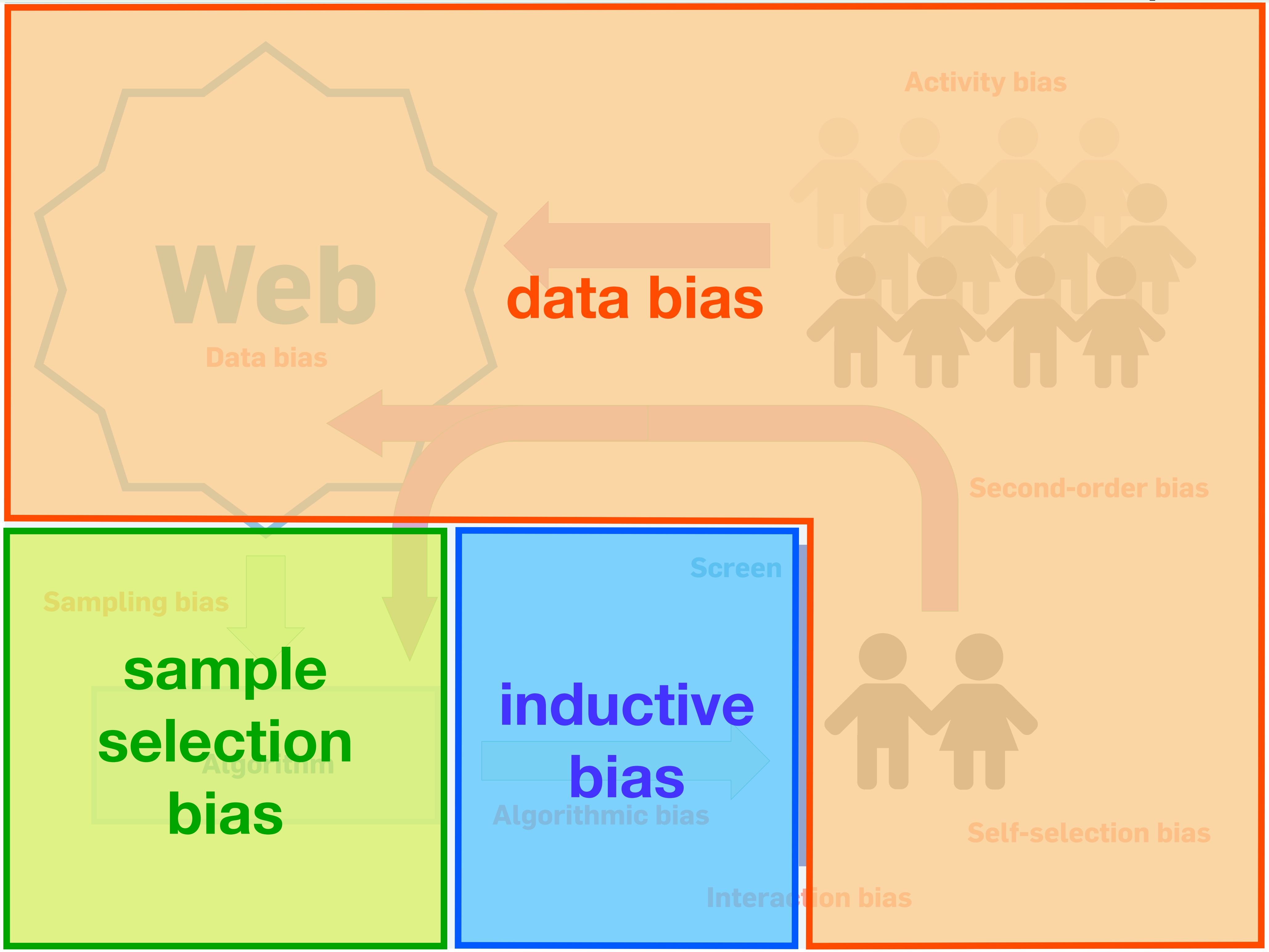
Bias on the Web

[Baeza-Yates 18]



Bias on the Web

[Baeza-Yates 18]



Bias Sources in Machine Learning

Data / Annotation Bias: **bias of labels or features in data**

- Decisions whether to approve loan are unfair by reflecting on prejudice against a specific group in a historical record

Sample Selection Bias: **data are not representatives of population**

- Records who have been able to pay off their loans are only available for those who have been approved the loans

Inductive Bias: **a bias caused by a machine learning algorithm**

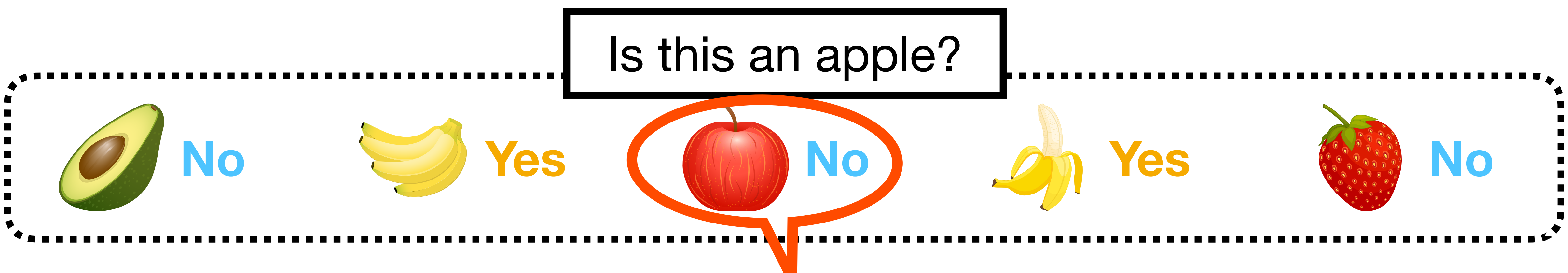
- Records for minority individuals who have been able to pay off loans in a minority group can be ignored due to the assumption of ML algorithms

Data / Annotation Bias

Data Bias / Annotation Bias: Target values or feature values in a training data are biased due to annotator's cognitive bias or inappropriate observation schemes

A Prediction is made by aggregating data

Even if inappropriate data is contained in a given dataset, the data can affect the prediction without correction



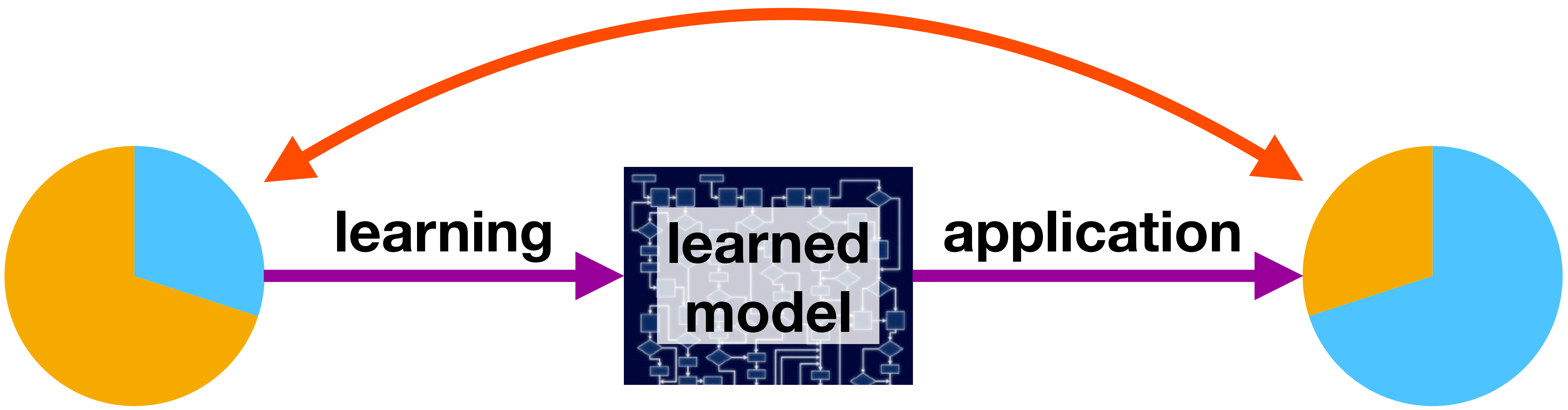
Even if an apple is given, the predictor trained by an inappropriate data set may output "No"

Sample Selection Bias

[Heckman 79, Zadrozny 04]

Sample Selection Bias: Whether a datum is sampled depends on conditions or contents of the datum, and thus an observed dataset is not a representative of population
✻ Strictly speaking, independence between the variables and the other variables needs to be considered

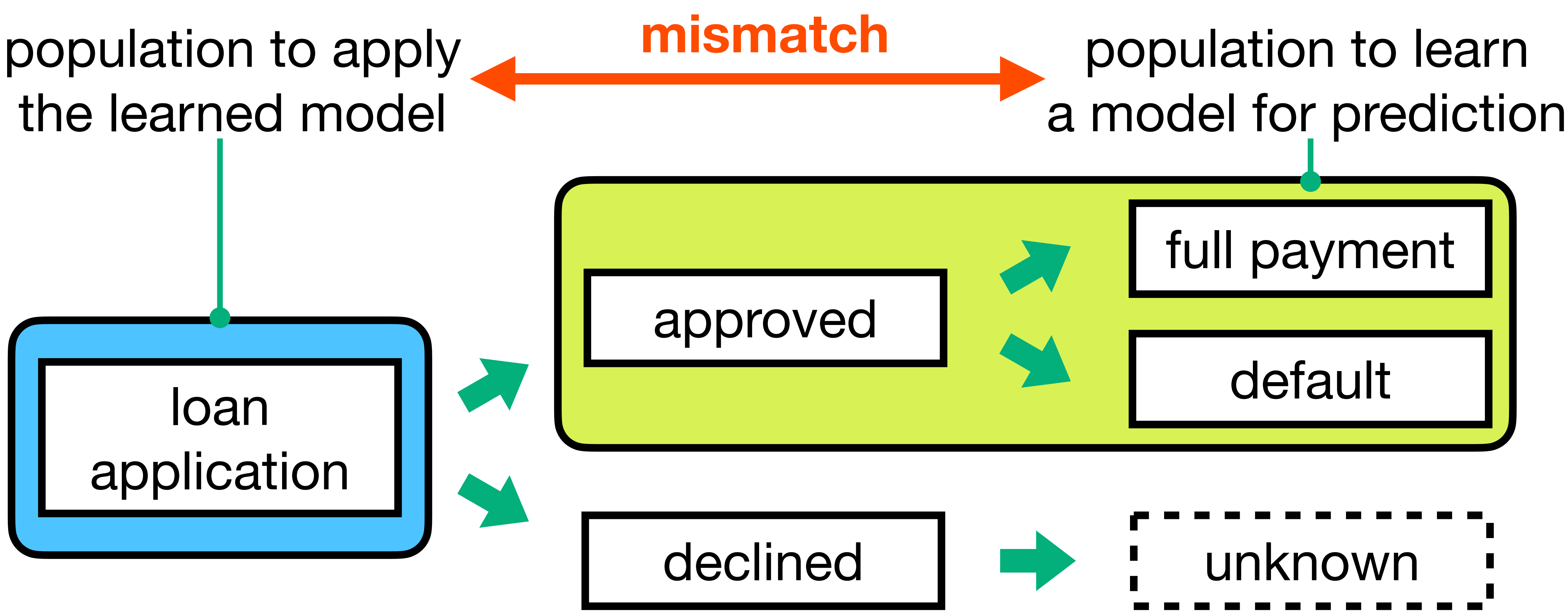
mismatch between distributions of learned and applied populations



Simple prediction algorithms cannot learn appropriately from a dataset whose contents depend on contents of the data

Example of Sample Selection Bias

loan application: A model is learned from a dataset including only approved applicants, but the model will be applied to applicants including declined applicants ➡ **sample selection bias**



A model is used for the targets different from a learned dataset

⬇

The learned model cannot classify targets correctly

Inductive Bias

Inductive Bias: a bias caused by an assumption adopted in an inductive machine learning algorithms

Inductive Machine Learning Algorithms:

sample training data	+	assumption background knowledge	→	prediction function prediction rule

These assumptions are required to generalize training data

↓

The assumptions might not always agree with a process of data generation in a real world

||

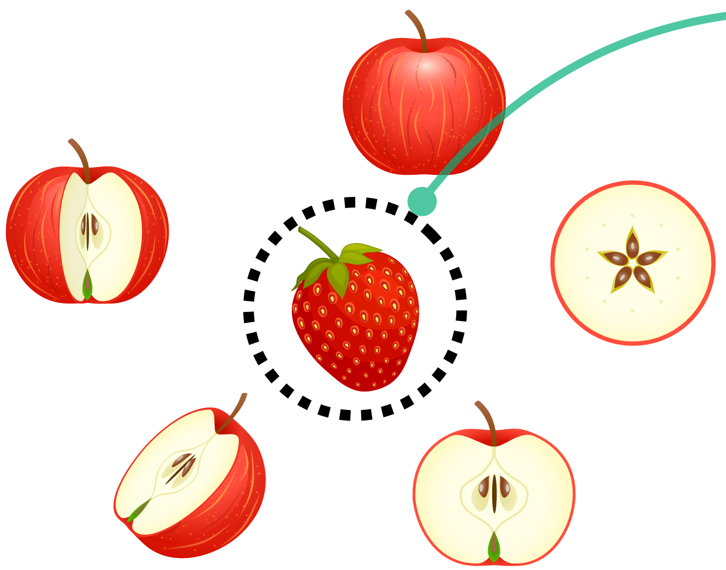
Inductive Bias

Occam's Razor

Occam's Razor: Entities should not be multiplied beyond necessity

↓

If models can explain a given data at the similar level, the simpler model is preferred



A small number of exceptional samples are treated as noise

↓

The prediction for unseen cases would be more precise in general

↓

Crucial rare cases can cause unexpected behavior

Any prediction, even if it was made by humans, is influenced by inductive biases, because the bias is caused in any generalization

Example of Inductive Bias

- **Occam's Razor:** Preference of ML algorithms to simpler hypothesis to improve generalization error
  ➡ Missing exceptional minor patterns

- **Smoothness:** Smoother decision boundaries or curves to fit are preferred
  ➡ Non-smooth changes cannot be represented

- **Sparseness:** Preference to hypothesis consisting of the smaller number of features
  ➡ Abandoning less effective features

- **Model Bias:** A target hypothesis may not included in a model of candidate hypotheses
  ➡ A learned hypothesis might not exactly match the target hypothesis

Instances of Data Biases

Data / Annotation Bias

Biases in Labels or Targets

- Historical records of approvals for loan requests might be influenced by prejudice towards a specific group
- Ratings are affected by predicted ratings displayed when users rate items [Cosley+ 03]

Biases in Features of Objects

- Use of word statistics of training corpus are affected by a gender bias [Bolukbasi+ 16]
- Admission to universities can be influenced by recommendation letters

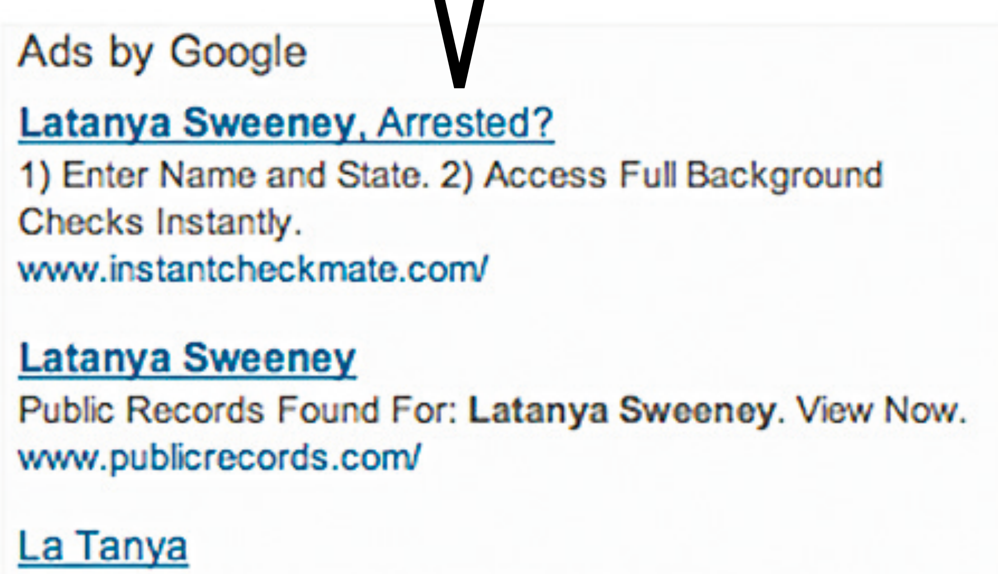
Suspicious Placement Keyword-Matching Advertisement

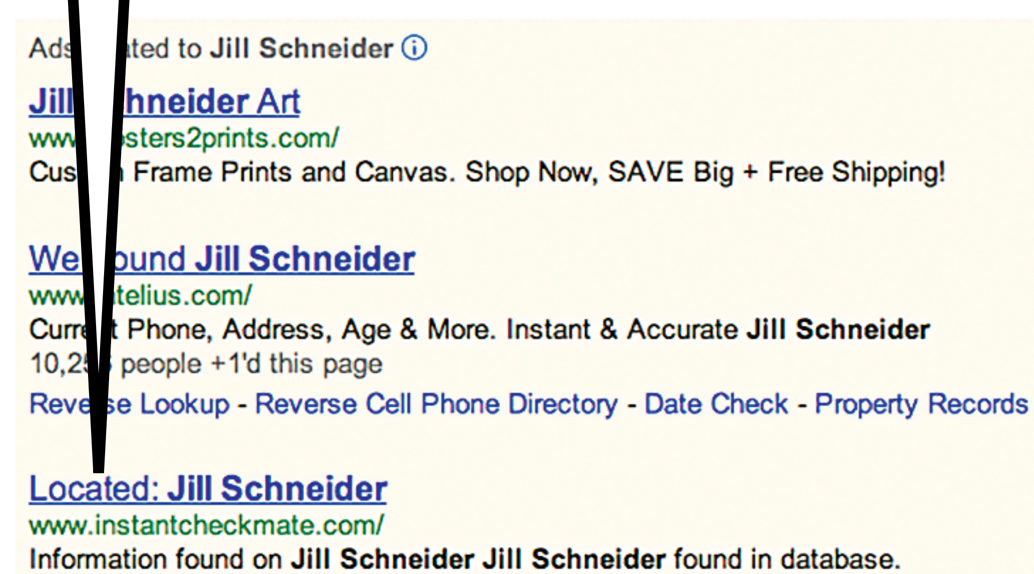
[Sweeney 13]

Online advertisements of sites providing arrest record information

Advertisements indicating arrest records were more frequently displayed for names that are more popular among individuals of African descent than those of European descent

African descent's name	European descent's name
Arrested? negative ad-text	**Located:** neutral ad-text



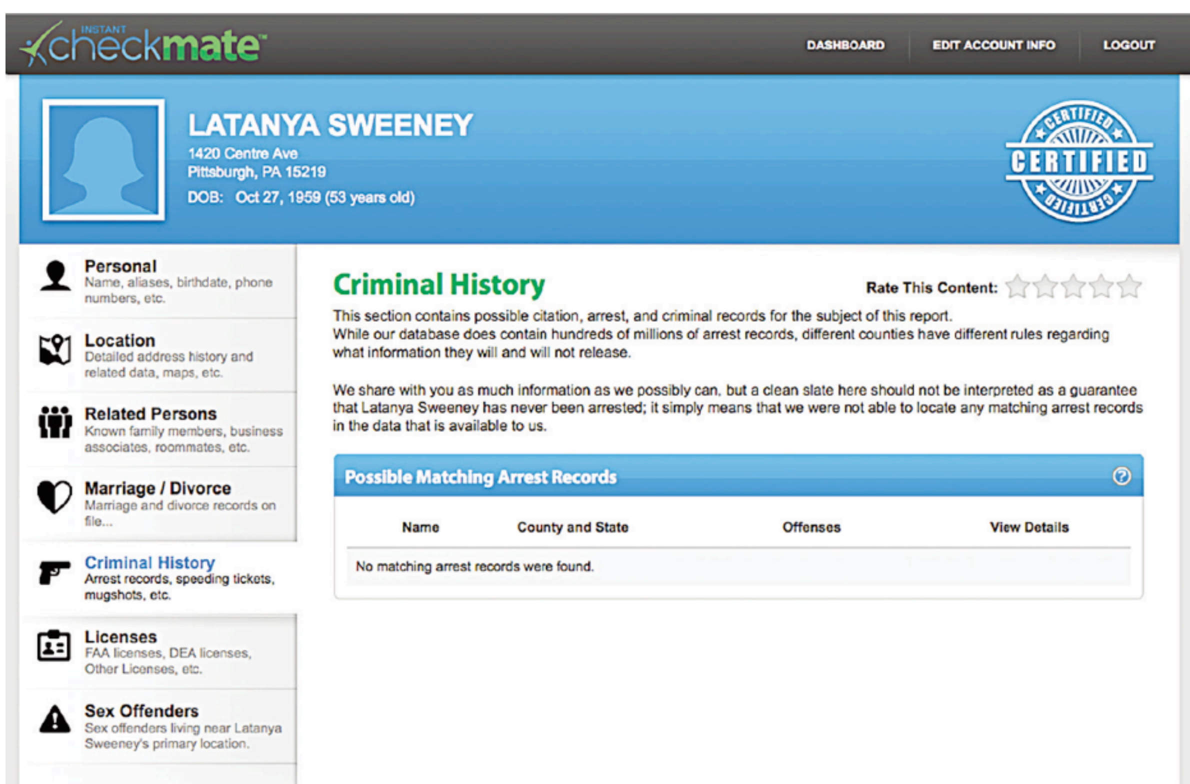


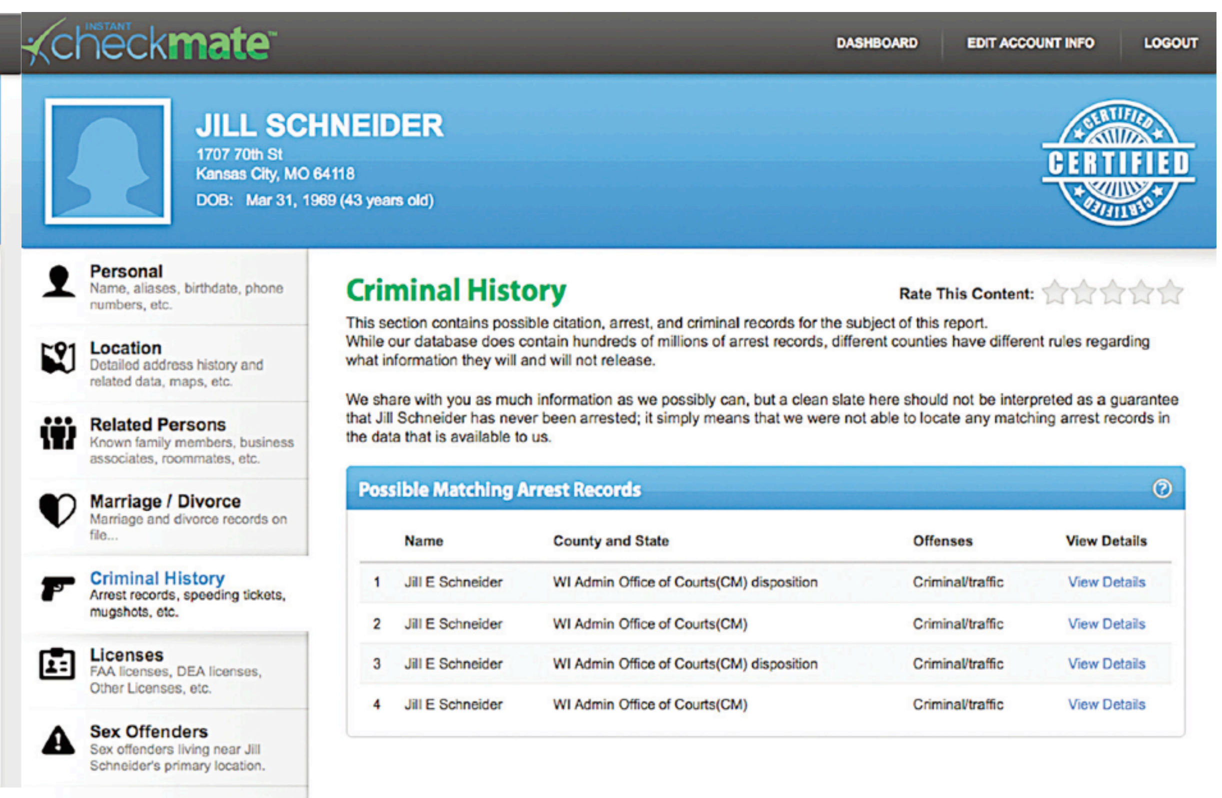
Suspicious Placement Keyword-Matching Advertisement

[Sweeney 13]

Advertisement texts are chosen irrelevant to
the actual existence of a prior arrest of the target name

African descent's name	European descent's name
↓	↓
Actually, no prior arrest	previously arrested





Suspicious Placement Keyword-Matching Advertisement

[Sweeney 13]

Selection of ad-texts was unintentional

Response from advertiser:

- Advertise texts are selected based on the last name, and no other information in exploited
- The selection scheme is adjusted so as to maximizing the click-through rate based on the feedback records from users by displaying randomly chosen ad-texts

No sensitive information, e.g., race, is exploited in a selection model, but suspiciously discriminative ad-texts are generated

↓

A data bias is caused due to the unfair feedbacks from users reflecting the users' prejudice

Instances of Inductive Biases

Recidivism Risk Score

[Angwin+ 16]

Recidivism Risk Score

- **COMPAS** (Correctional Offender Management Profiling for Alternative Sanctions) developed by Northpointe, used in many states
- Evaluate the re-offending risk by a ten-point-scale
- Judges are given the scores in the process of pretrial release

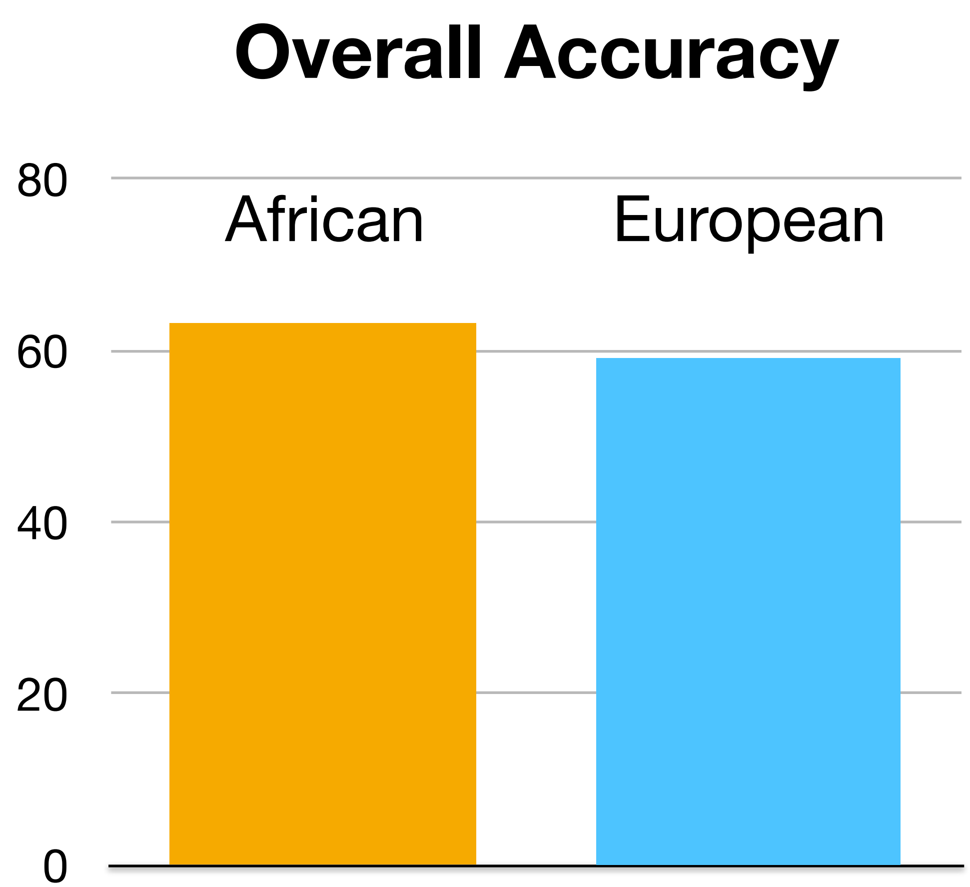
Merits and Concerns pointed out by the ProPublica

- Key decisions in the legal process have been historically affected by personal biases
- Scores can be exploited not for the designed purposes
- **Scores must accurately predict which defendants likely to re-offend, but these are biased**

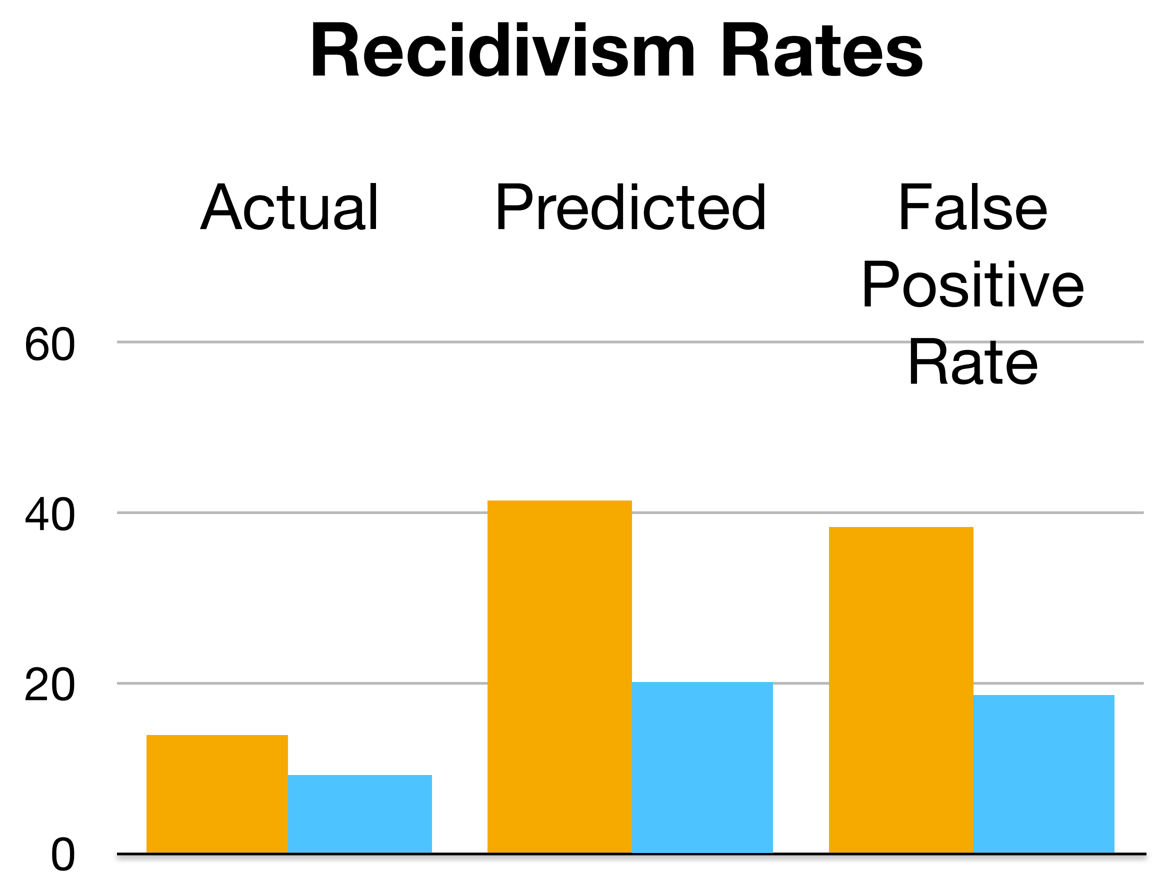
Recidivism Risk Score

[Angwin+ 16]

Defendants of African descents were often predicted to be more risky than they actually were, and vice versa



Roughly the same
Not problematic



FPR for African is higher
Problematic

* **FPR (false positive ratio)** = ratio of # of actually non-recidivated to # of people predicted to recidivate

Rejoinder of US Federal Courts

[Flores + 16]

The merit of risk assessment tool

It might be that the existing justice system is biased against poor minorities … regardless of the degree of bias, risk assessment tools informed by objective data can help **reduce** racial bias from its current level

Rejoinder to ProPublica's study

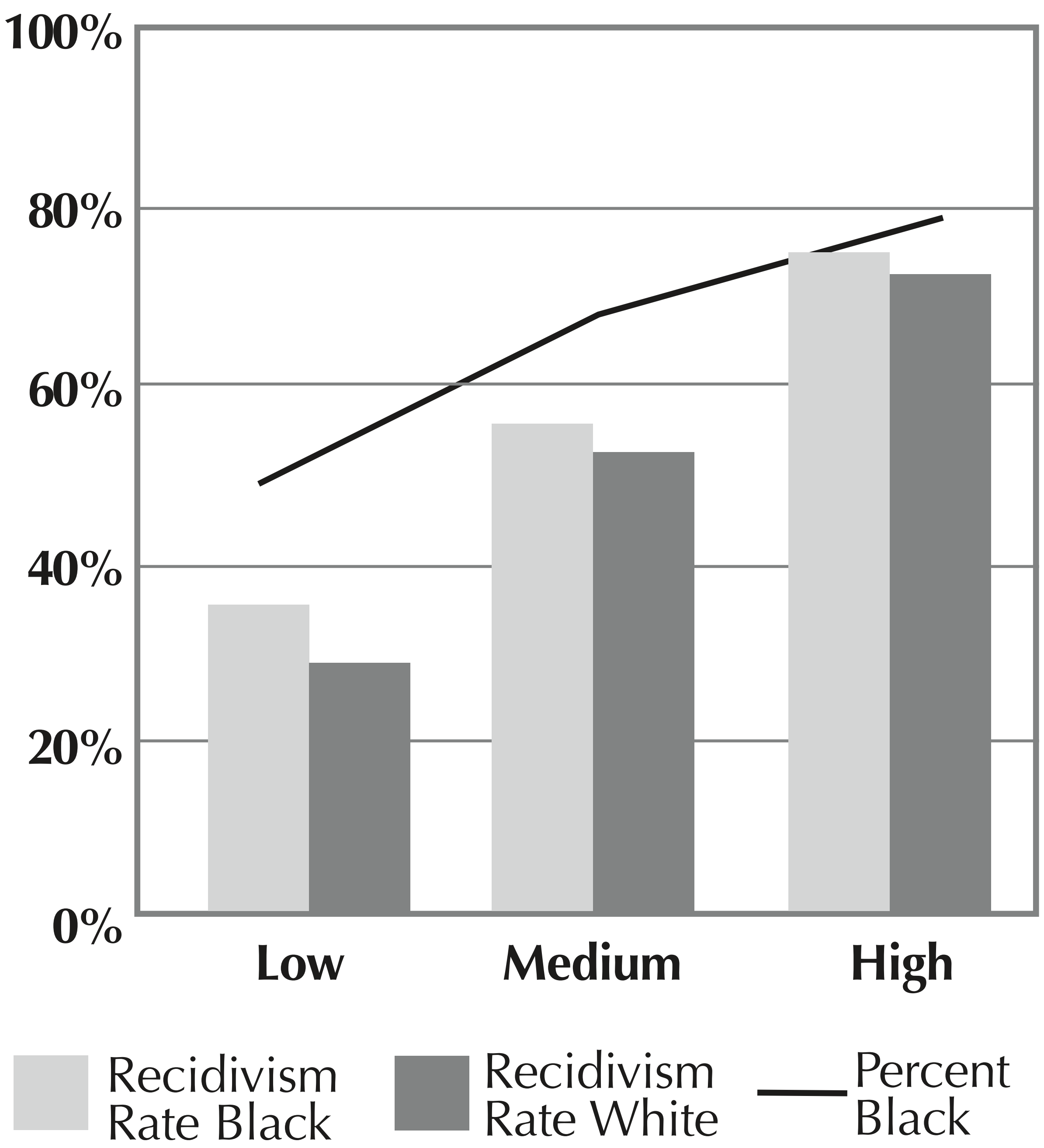
1. The COMPAS targets individuals on post-disposition supervision, but the ProPublica analyzed pretrial defendants
2. Collapsing mid- & high-risk categories is problematic
3. Distributions of observations given the predictions should be used, instead of distributions of predictions given observations
4. The standards, such as the federal Post Conviction Risk Assessment (PCRA), are ignored
5. Choosing improper the level of significance

Rejoinder of US Federal Courts

[Flores + 16]

The COMPAS satisfies a fairness condition, sufficiency

- The COMPASS score is designed to satisfy the sufficiency, $Y \perp\!\!\!\perp S \mid \hat{Y}$, following the standard of the federal Post Conviction Risk Assessment (PCRA)
- The chart shows the actual arrest ratios given the predicted risk scores, in the any arrest case
- The Northponte, a COMPAS developer, also pointed out this problem [Dieterich+ 2016]



Algorithms Improve Human Decisions

[Kleinberg+ 18]

Pretrial Bail Decisions

- Arrest records in New York City between Nov. 1, 2008 – Nov. 1, 2013
  - male=83.2%, African American=48.8%, Hispanic=33.3%
  - release=73.6% ➡ failure to appear=15.2%, rearrested=25.8%
- Judges decide whether defendants to release or detain, based on a checklist and the information judges see, such as appearance
- Algorithms use the information available to judges and age, but ignore the information judges see

Algorithms Improve Judges' Decisions

If defendants were detained based on algorithm prediction until the level that judges of high-detention rate detained, algorithms would achieve:

- at the same crime rate as judges ➡ **48.2% lower detention rate**
- at the same detention rate as judges ➡ **75.8% lower crime rate**

Algorithms Improve Human Decisions

[Kleinberg+ 18]

Judges Release High-Risk Defendants

The riskiest 1% of defendants in prediction:
If released, fail to appear=57.3%, rearrested=62.7%

⬇

Judges release **48.5%** of them

Algorithms Are Fairer Than Judges

If a distribution of detained races is constrained to satisfy a fairness condition, algorithms reduce crime rate relative to judges:

- no constraint ➡ **24.68%**
- match a distribution that judges detain ➡ **24.64%**
- match a distribution of defendants (= statistical parity) ➡ **23.02%**
- match lower of a distribution of defendants or a distribution that judges detain ➡ **22.74%**

Bias in Image Recognition

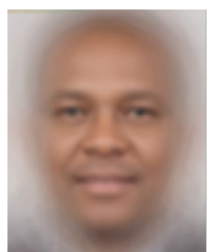
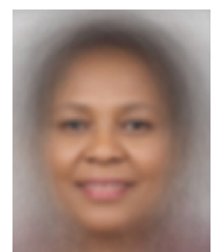
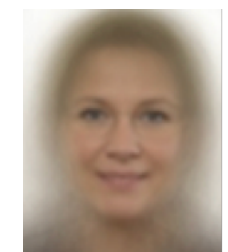
[Buolamwini+ 18]

- Auditing the image recognition API's for predicting a gender from facial images
- Available benchmark datasets of facial images is highly skewed to the images of males with lighter skin
- Pilot Parliaments Benchmark (PPB) is a new dataset balanced in terms of skin types and genders
  - Skin types are *lighter* or *darker* based on the Fitzpatrick skin type
  - Perceived genders are *male* or *female*
- Facial-image-recognition API's by Microsoft, IBM, and Face++ are tested on the PPB dataset

Bias in Image Recognition

[Buolamwini+ 18]

Error rates (1 - TPR) in a gender prediction from facial images

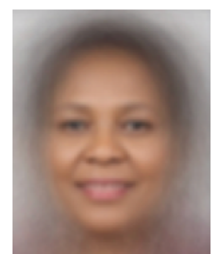
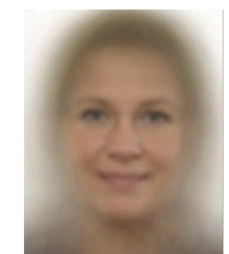
	darker male 	darker female 	lighter male 	lighter female 
Microsoft	6.0%	**20.8%**	**0.0%**	1.7%
IBM	12.0%	**34.7%**	**0.3%**	7.1%
Face++	**0.7%**	**34.5%**	0.8%	7.1%

Error rates for **darker females** are generally worse than **lighter males**

Bias in Image Recognition

[IBM, Buolamwini+ 18]

IBM have improved the performance by new training dataset and algorithm, before Buolamwini's presentation,

	darker male 	darker female 	lighter male 	lighter female 
old IBM	12.0%	**34.7%**	0.3%	7.1%
new IBM	2.0%	**3.5%**	0.3%	0.0%

Error rates for **darker females** are improved

Inductive Bias: Example

[Calders+ 10]

US Census Data : predict whether their income is high or low

Females are minority in the high-income class

	Male	Female
High-Income	3,256	590
Low-income	7,604	4,831

fewer

In this original data set:

- The number of High-Male data is 5.5 times that of High-Female data
- While 30% of Male data are High income, only 11% of Females are

Inductive Bias: Example

[Calders+ 10]

Odds ratio: to evaluate the influence of a gender to an income
ratio of the odds to be high-income for males to that for females

$$\textbf{Odds ratio} = \frac{\Pr[\text{High}, \text{Male}] / \Pr[\text{Low}, \text{Male}]}{\Pr[\text{High}, \text{Female}] / \Pr[\text{Low}, \text{Female}]}$$

Directly derived
from an observed sample
odds ratio = 3.51

→

Derived by a naive Bayes
model w/o a gender feature
odds ratio = 5.26

The increase of the odds ratio implies that
a gender has stronger impact on an income

↓

Due to an inductive bias,
the minor information of high-income females is ignored

Part II
Formal Fairness

Basics of Formal Fairness

Formal Fairness

In fairness-aware machine learning, we manage the influence:

sensitive information $\xrightarrow{\text{Influence}}$ **target / objective**

- socially sensitive information
- information restricted by law
- information to be ignored

- university admission
- credit scoring
- crick-through rate

Formal Fairness

The desired condition defined by a formal relation between sensitive feature, target variable, and other variables in a model

- Which set of variables are involved?
- How are these variables related?
- What states of sensitives or targets should be controled?

Notations of Variables

$\boxed{Y}$ **target variable / object variable**

An objective of decision making, or what to predict

Ex: loan approval, university admission, what to recommend

$\boxed{Y}$ = observed / true, $\boxed{\hat{Y}}$ = predicted, $\boxed{Y^{\circ}}$ = fairized

- $Y{=}1$ advantageous decision / $Y{=}0$ disadvantageous decision

$\boxed{S}$ **sensitive feature**

To ignore the influence to the sensitive feature from a target

Ex: socially sensitive information (gender, race), items' brand

- $S{=}1$ non-protected group / $S{=}0$ protected group
- Specified by a user or an analyst depending on his/her purpose
- It may depend on a target or other features

$\boxed{\mathbf{X}}$ **non-sensitive feature vector**

All features other than a sensitive feature

Other Notations

$\boxed{\mathscr{D} = \{y_i, s_i, \mathbf{x}_i\}_{i=1}^{2}}$ **dataset**

Each datum is a triple of a target value, y_i, a sensitive value, s_i, and non-sensitive feature values, $\mathbf{x}_i$

$\boxed{\mathscr{D}^{(s)} = \{y_i, s_i, x_i\}_{i=1}^{n^{(s)}} \text{ s.t. } s_i = s}$ **sensitive group**

a group consisting of the same sensitive value

If $s_i = 0$ indicates a minority individual to protect, $\mathscr{D}^{(0)}$, is called a **protected group**, and the rest of dataset, $\mathscr{D}^{(1)}$, is called a **non-protected group**

$\boxed{\mathbf{X}^{(e)} \,/\, \mathbf{X}^{(\bar{e})}}$ **explainable / unexplainable non-sensitive feature**

Explainable variables are confounding variables with Y and S, and their influence can be ignored because of legal or other reasons

Type of Formal Fairness

association-based fairness

- defined based on statistical association, namely correlation and independence
- mathematical representation of ethical notions, such as distributive justice

counterfactual fairness

- causal effect of the sensitive information to the outcome
- maintaining a counterfactual situation if the sensitive information was changed

economics-based fairness

- using a notion of a fairness in game theory or econometrics

Accounts of Discrimination

[Lippert-Rasmussen 06]

Why an instance of discrimination is bad?

- **harm-based account:** Discrimination makes the discriminatees worse off
- **disrespect-based account:** Discrimination involves disrespect of the discriminatees and it is morally objectionable
  - An act or practice is morally disrespectful of X

    ⇔ It presupposes that X has a lower moral status than X in fact has

⬇

Techniques of Fairness-Aware Machine Learning
based on the harm-based account

The aim of FAML techniques remedy the harm of discriminatees

Regulations & Laws Related to Association-Based Fairness

[Pedreschi+ 09]

Quantitative restrictions by regulations or laws against discrimination:

Anti-Discrimination Act (Australia, Queensland)

- a person treats, or proposes to treat, a person with an attribute less favorably than another person without the attribute

Racial Equality Directive (EU)

- shall be taken to occur where one person is treated less favorably than another is in a comparable situation on grounds of racial or ethnic origin

Uniform Guidelines on Employee Selection Procedure (US, EEOC)

- a selection rate for any race, sex, or ethnic group which is less than four-fifths (or eighty percent) of the rate for the group with the highest rate will generally be regarded as evidence of adverse impact

Regulations & Laws Related to Association-Based Fairness

[Ishiguro+ 14]

Title VII of the Civil Rights Act of 1964

- Prohibit to discrimination due to race, religion, gender, and ethnicity

Hazelwood School District v. United States, 433 U.S. 299 (1977)

- Evidence of long-lasting and gross disparity between the composition of a workforce and that of the general population thus may be significant even though § 703(j) makes clear that Title VII imposes no requirement that a workforce mirror the general population
- Where gross statistical disparities can be shown, they alone may, in a proper case, constitute *prima facie* proof

Gross Statistical Disparity: Discrimination in employment is determined whether the ratio of protected and non-protected groups of employees is diverged from the corresponding ratio in general population

Baselines in Harm-based Account

[Lippert-Rasmussen 06]

A harm-based account requests a baseline for determining whether the discriminatees have been made worse off

↓

- **Ideal outcome:** the discriminatees are in just, or the morally best
  ➡ **association-based fairness:** letting predictors get ideal outcomes
- **Counterfactual:** the discriminatees had not been subjected to the discrimination
  ➡ **counterfactual fairness:** comparing with the counterfactuals that a status of a sensitive feature was different

Association-Based Fairness: Basics of Associations

Independence

(unconditional) independence

A pair sets of variables, $\mathbf{Y}$ and $\mathbf{S}$, are not influenced from each other

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

conditional independence

$\mathbf{Y}$ and $\mathbf{S}$ are independent, if conditional variables, $\mathbf{X}$, are fixed

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{S} \mid \mathbf{X}$$

✻ **Conditional independence doesn't imply independence, and vice versa**

context-specific independence

$\mathbf{Y}$ and $\mathbf{S}$ are independent, if $\mathbf{X}$ are fixed to specific values, $\mathbf{x}$ [Boutiller+ 96]

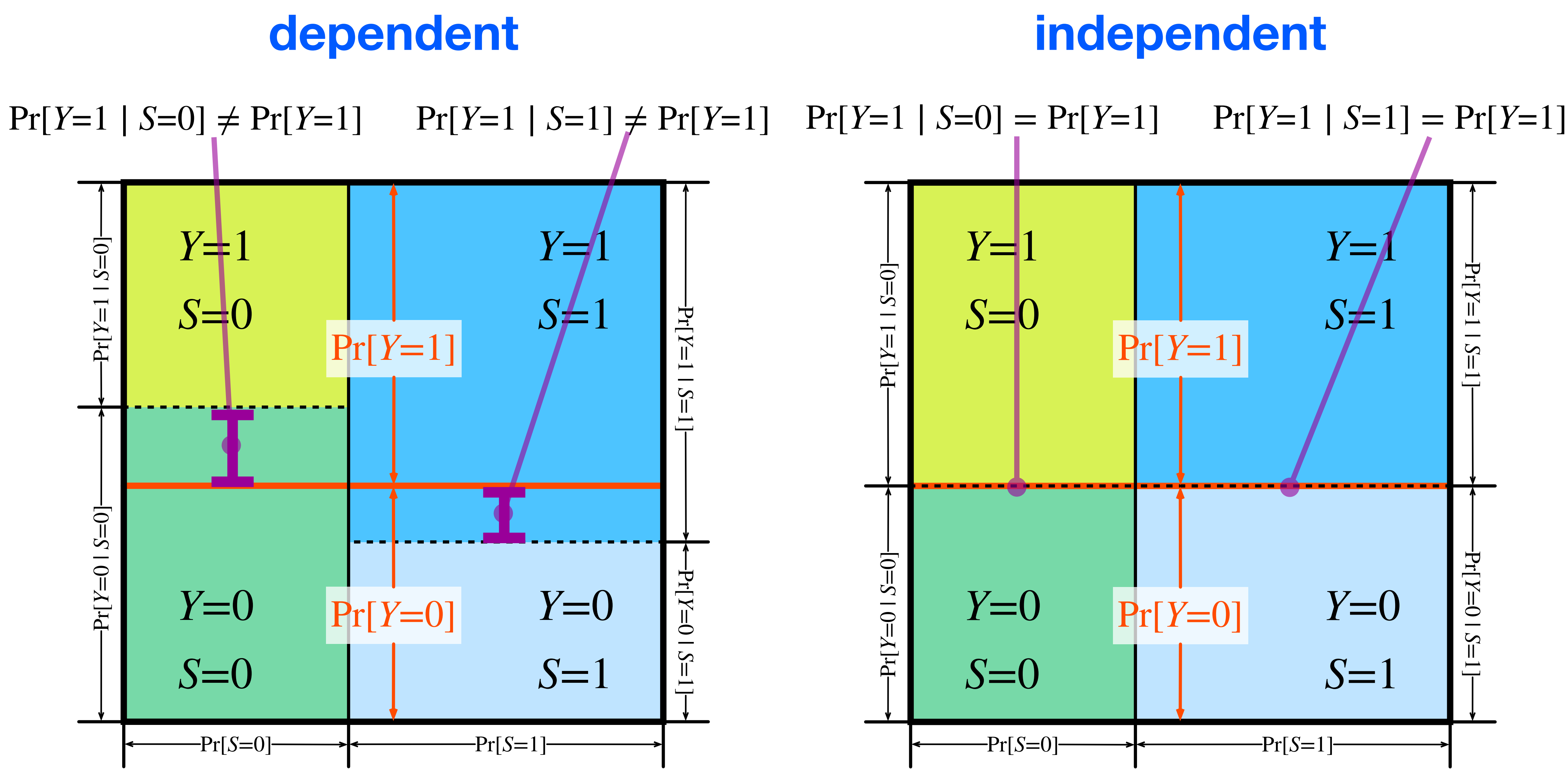
$$\mathbf{Y} \perp\!\!\!\perp \mathbf{S} \mid \mathbf{X}{=}\mathbf{x}$$

✻ Notation with a symbol '⫫' (Unicode 2AEB) is called Dawid's notation

Independence

(Unconditional) Independence: $Y \perp\!\!\!\perp S$

$$\Pr[Y, S] = \Pr[Y]\Pr[S] \iff \Pr[Y \mid S] = \Pr[Y]$$



Conditional Independence

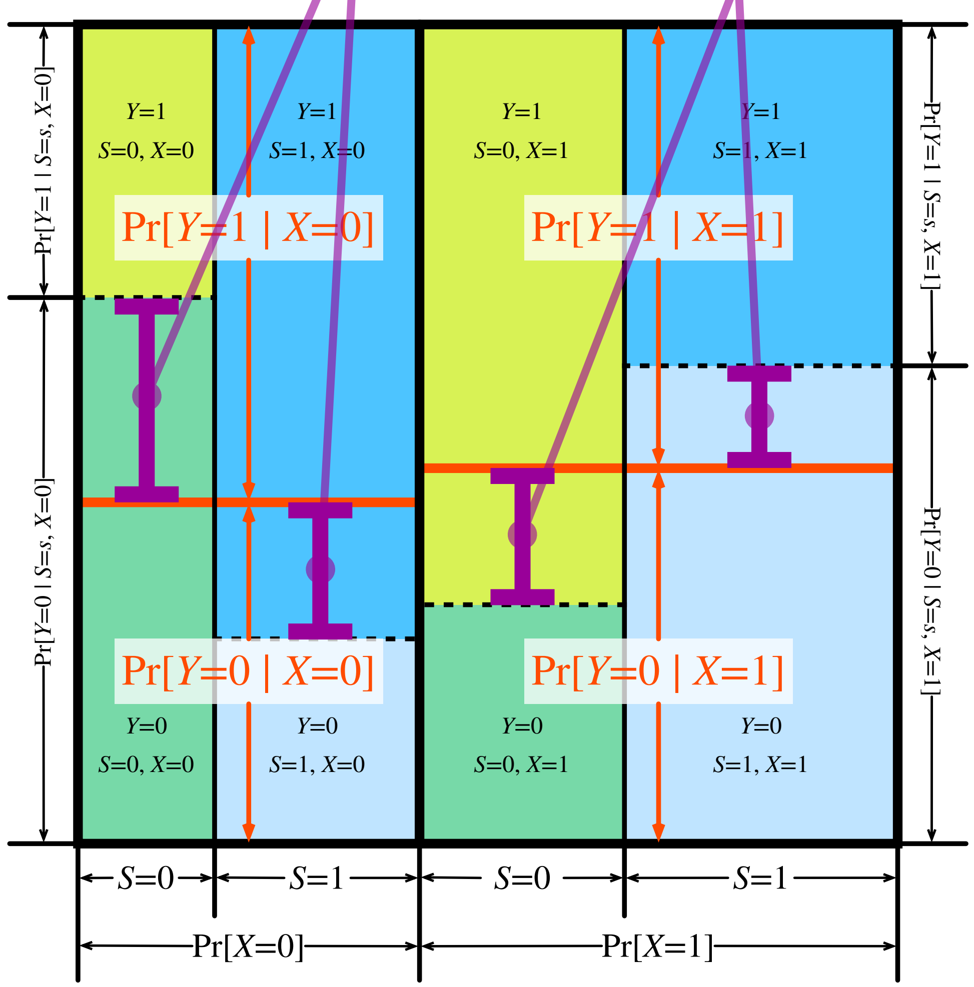
Conditional Independence: $Y \perp\!\!\!\perp S \mid X$

$$\Pr[Y, S \mid X] = \Pr[Y \mid X]\Pr[S \mid X] \iff \Pr[Y \mid S, X] = \Pr[Y \mid X]$$

dependent

$\Pr[Y{=}1 \mid S{=}s, X{=}0] \neq \Pr[Y{=}1 \mid X{=}0]$

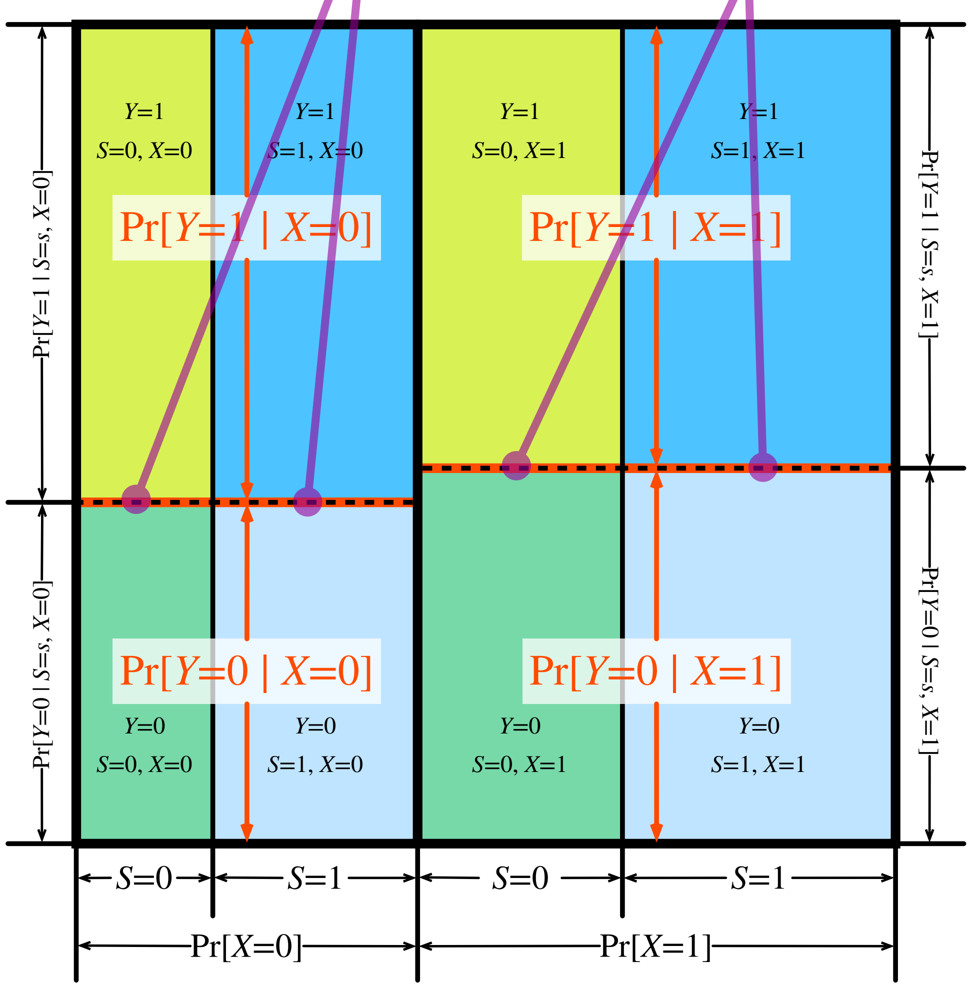
$\Pr[Y{=}1 \mid S{=}s, X{=}1] \neq \Pr[Y{=}1 \mid X{=}1]$



independent

$\Pr[Y{=}1 \mid S{=}s, X{=}0] = \Pr[Y{=}1 \mid X{=}0]$

$\Pr[Y{=}1 \mid S{=}s, X{=}1] = \Pr[Y{=}1 \mid X{=}1]$

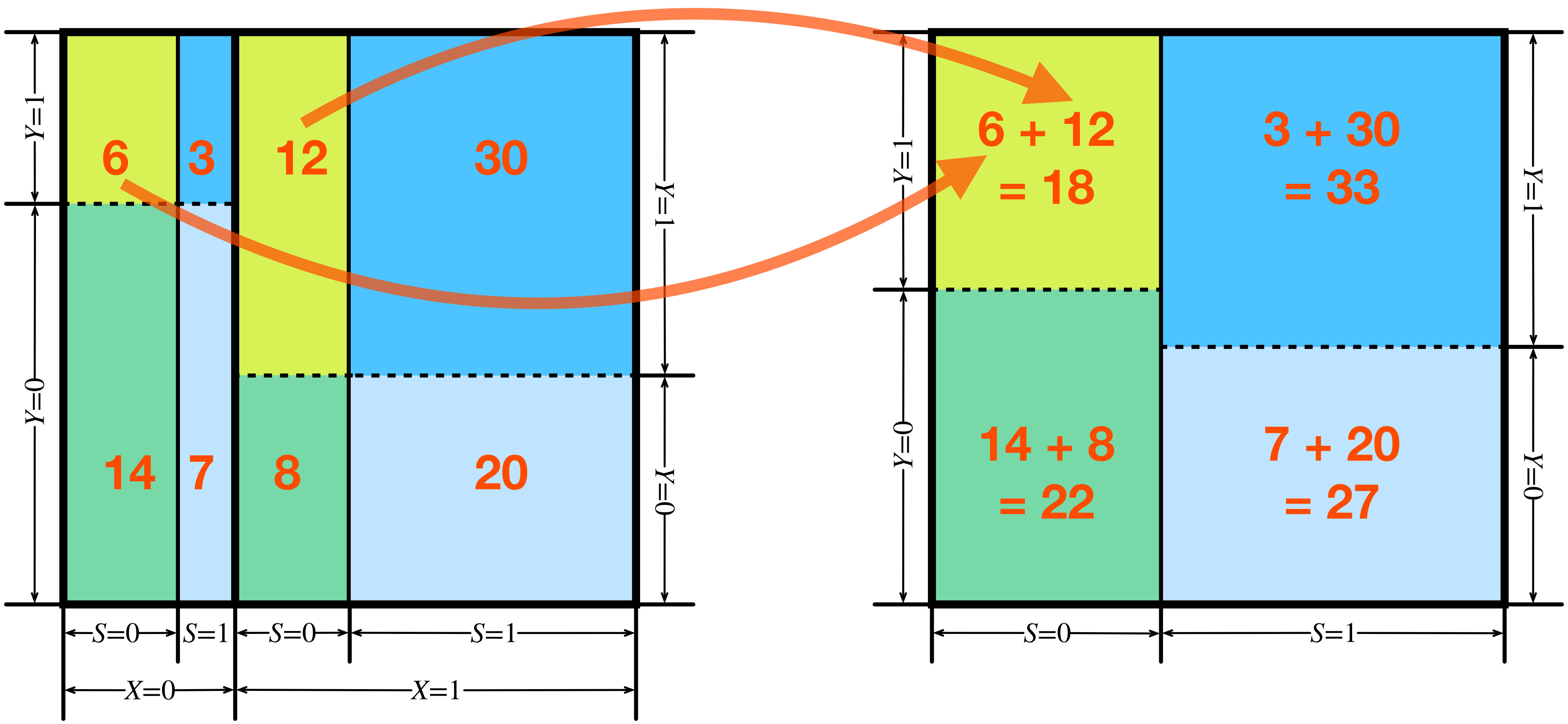


Unconditional & Conditional Independence

Conditional independence does not imply unconditional independence in general

$$S \perp\!\!\!\perp Y \mid X \not\longrightarrow S \perp\!\!\!\perp Y$$

Conditionally Independent **Unconditionally Dependent**

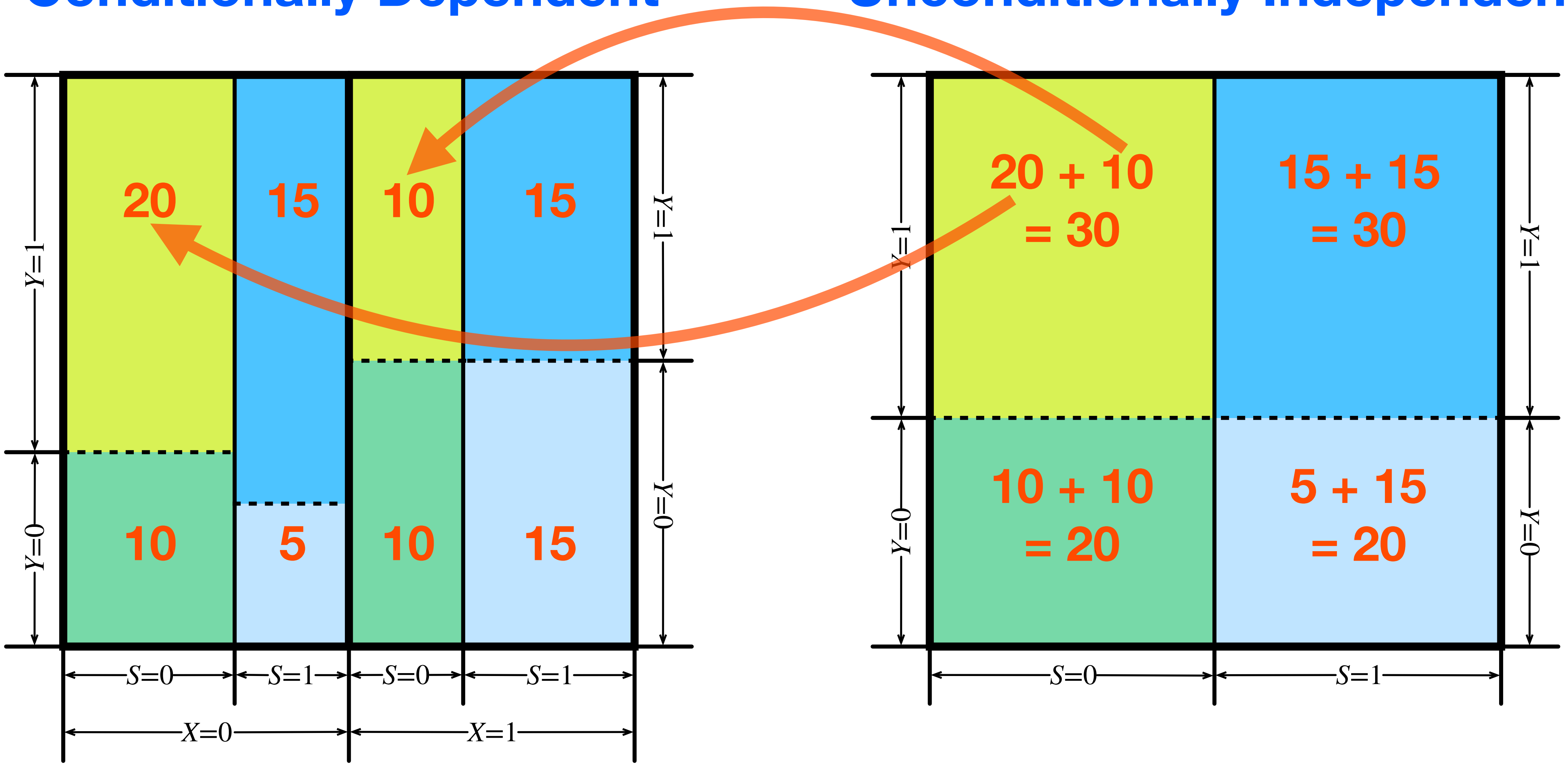


Unconditional & Conditional Independence

Inversely, unconditional independence does not imply conditional independence in general

$$S \perp\!\!\!\perp Y \mid X \;\not\Leftarrow\; S \perp\!\!\!\perp Y$$

Conditionally Dependent | **Unconditionally Independent**



Simpson's Paradox

[Bickel+ 75]

Simpson's Paradox: Numerical facts that the results obtained from a whole dataset is processed are contradicted with the results obtained when a dataset is grouped or stratified

Admission to the Univ. of California, Berkeley, for the fall 1973 quarter

Aggregated data for the campus

- Admission rate: male=44% female=35% ➡ **discriminative**

Grouped by the departments

- Among 85 departments, females are fewer in 4 departments and males are fewer in 6 departments ➡ **non-discriminative**

⬇

This case is not discriminative, because more females were applied to the department whose admission rate was lower

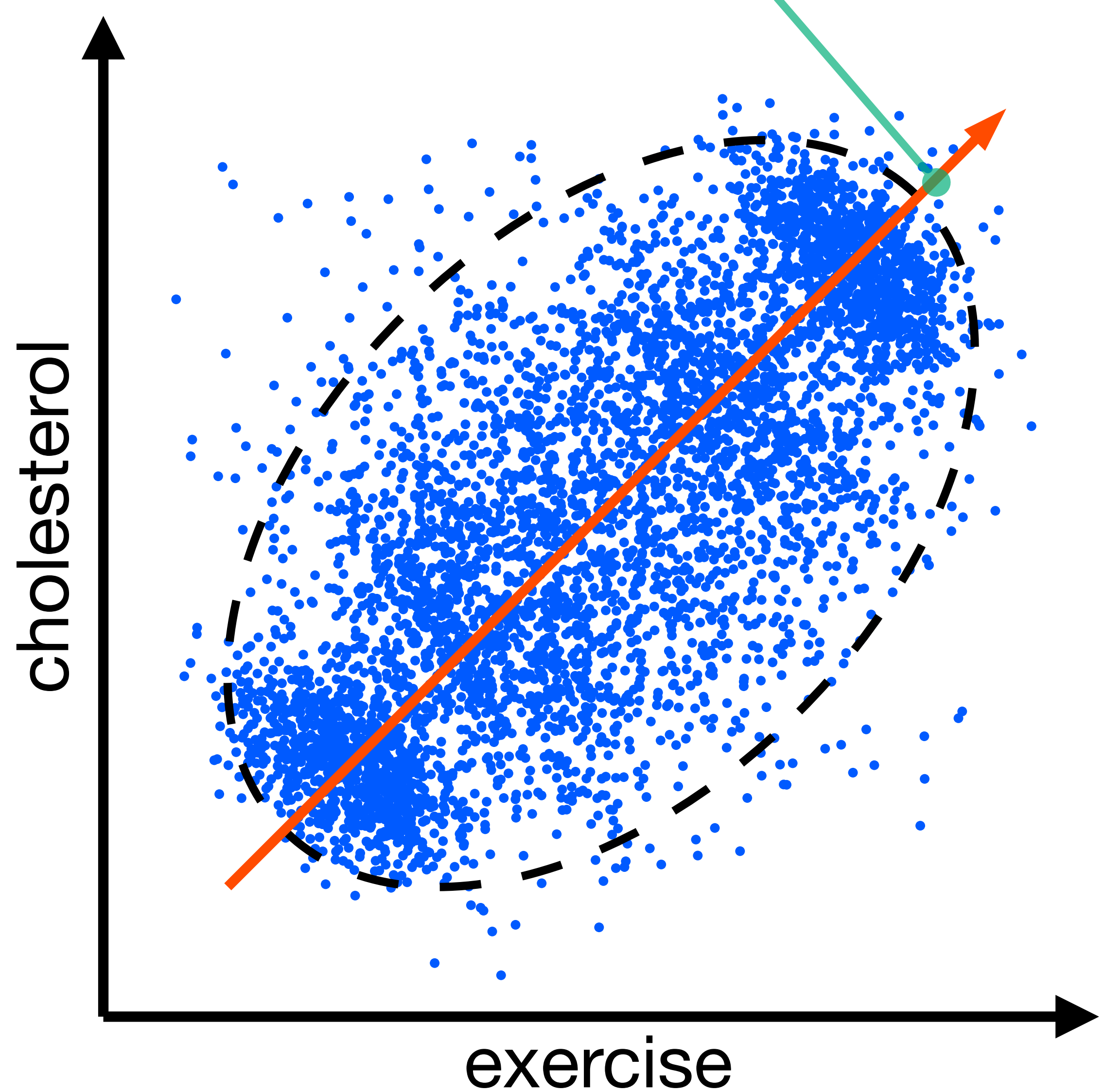
even the naive question could not answered adequately without recourse to sophisticated methodology and careful examination of underlying process

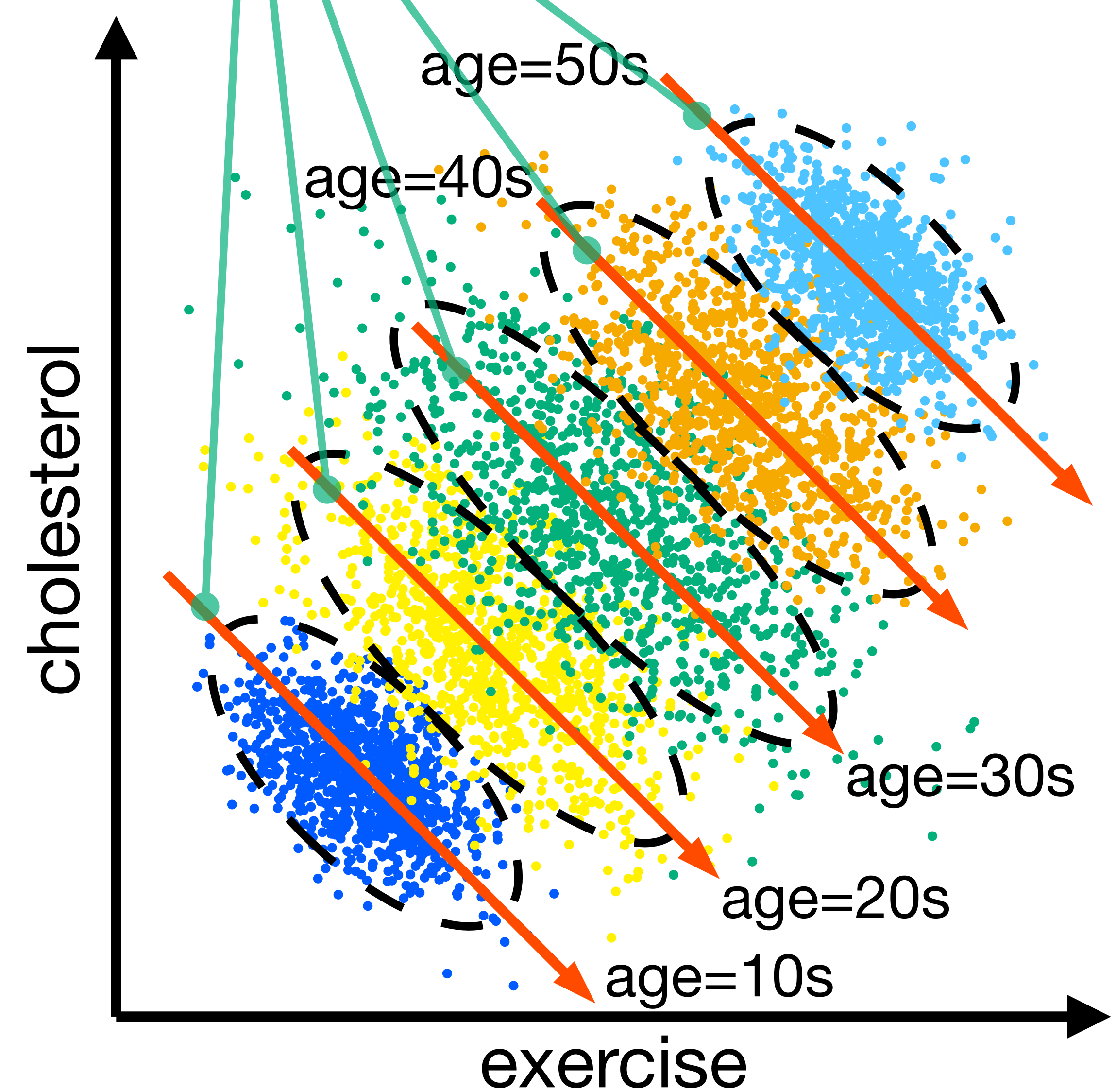
Simpson's Paradox

[Pearl+ 18]

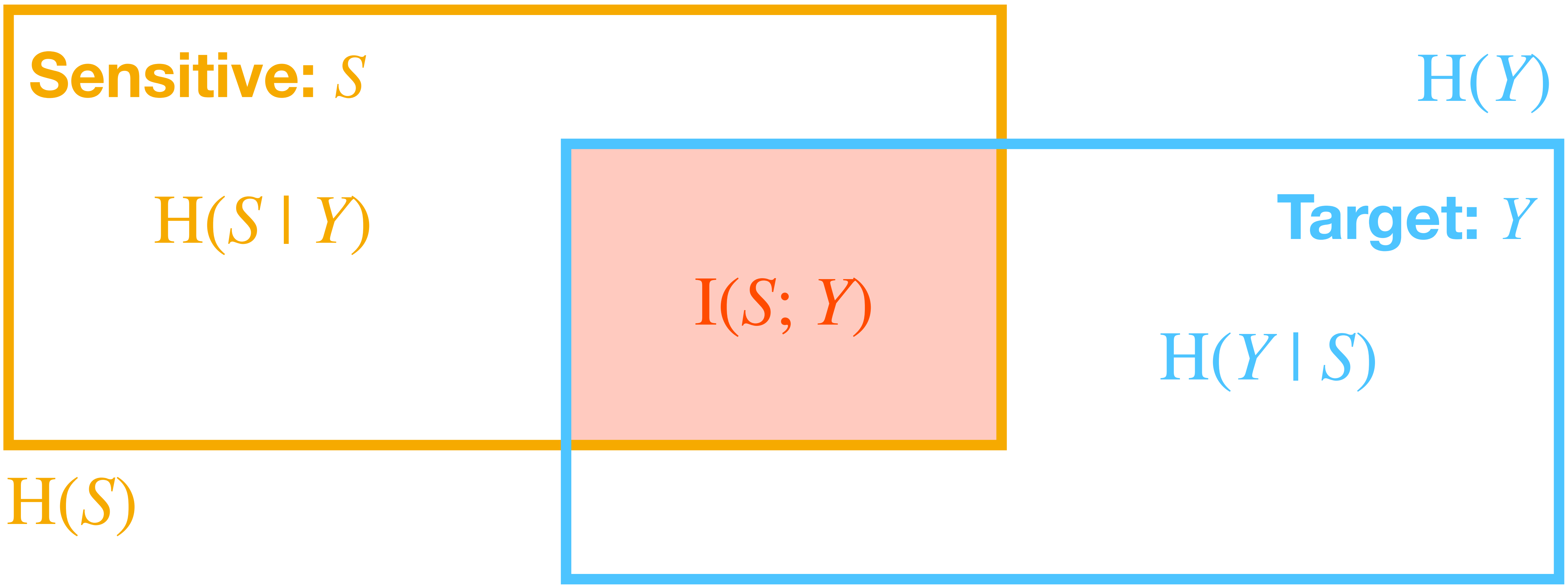
"Cholesterol" and "exercise" are **positively correlated**, if all data are aggregated

If grouped by "age", they are **negatively correlated**, because cholesterol of aged people tends to be higher





Information-Theoretic Interpretation

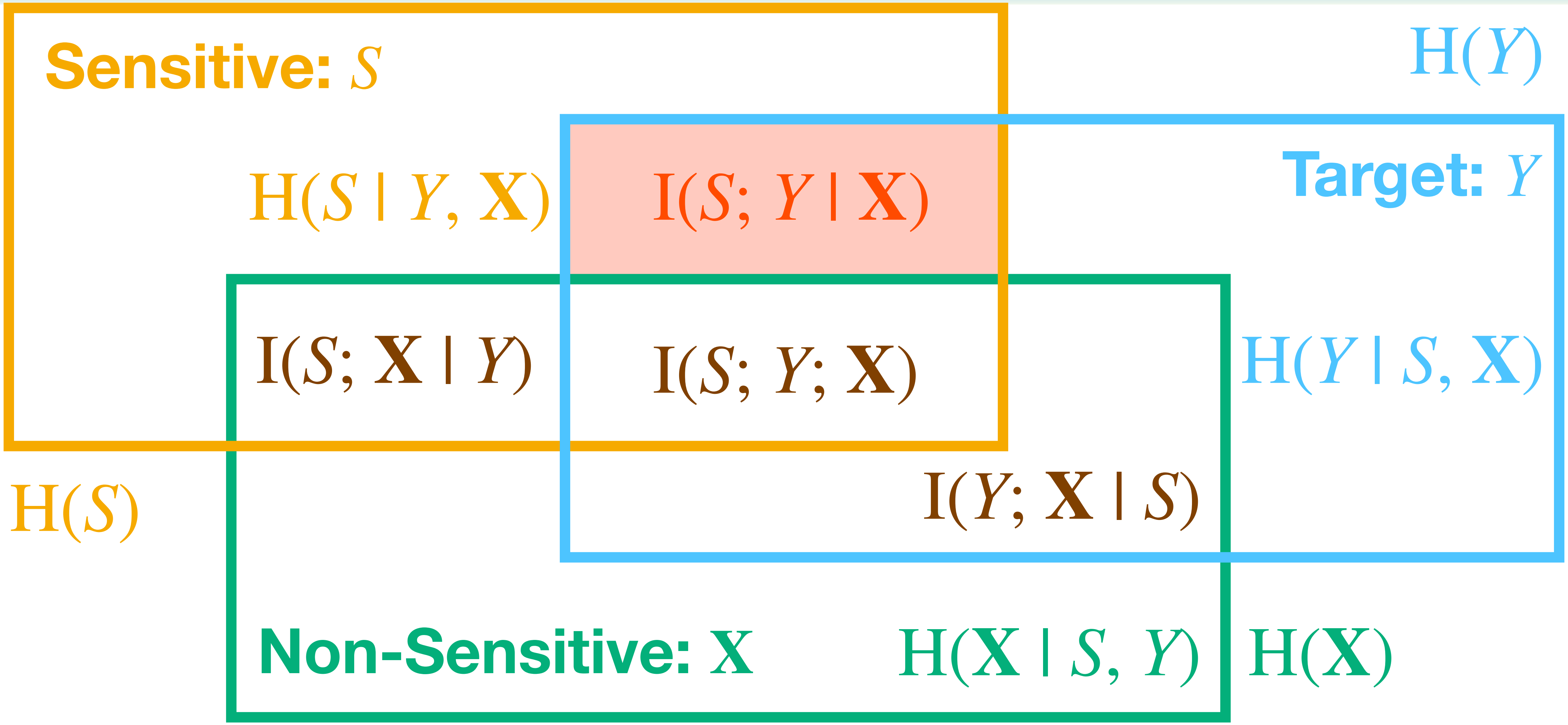


statistical parity, $S \perp\!\!\!\perp Y$, implies zero mutual information: $\mathrm{I}(S; Y) = 0$

⬇

If the information about Y is known, no information about S cannot be gained, and vice versa

Information Theoretic Interpretation



Mutual information, $\mathrm{I}(S; Y \mid \mathbf{X})$, shows the information gained by knowing about Y in the information about S by knowing $\mathbf{X}$ (= $\mathrm{H}(S \mid \mathbf{X})$)

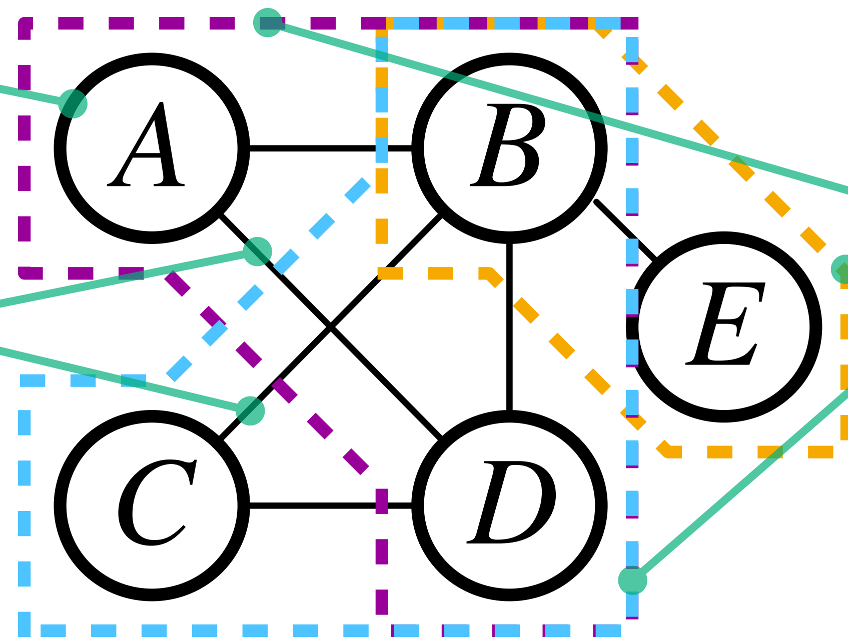
Markov Network

[Bishop 06]

Markov network: undirected graphical model for probabilistic distribution

random variable

dependency



maximal clique
maximal subset of nodes composing a complete graph

potential function
Each corresponds to one clique

standardized constant or partition function

$$\Pr[A,B,C,D,E] = f(A,B,D)f(B,C,D)f(B,E)\big/Z$$

Variables, A and C, are separated by removing B and D

⬇

conditional independence: $A \perp\!\!\!\perp C \mid B, D$

Correlation

Correlation Coefficient

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

* $\bar{x}$ is a sample mean of x. $\mathrm{Var}(X)$ and $\mathrm{Cov}(X, Y)$ are a variance and covariance, respectively.

Independence implies no-correlation, but no-correlation does not generally imply independence

independence ⇒ no-correlation

Continuous Variable

- If X and Y follows Gaussian, no-correlation implies independence

Discrete Variable

- If the rank of a frequency matrix for X and Y is 1, they are independent; If the matrix is singular, They are no-correlation
  ➡ If X and Y are binary, no-correlation implies independence

Partial Correlation

Partial Correlation Coefficient

$$\rho_{xy\cdot z} = \frac{\mathrm{Cov}(\Delta_{xz}, \Delta_{yz})}{\sqrt{\mathrm{Var}(\Delta_{xz})}\sqrt{\mathrm{Var}(\Delta_{yz})}} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{1-\rho_{xz}^2}\sqrt{1-\rho_{yz}^2}}$$

- θ_{xz} : a regression coefficient from z to x.
- $\Delta_{xz}^{(i)} = x_i - \theta_{xz} z_i$
- ρ_{xy} : correlation coefficient between x and y.

- The $\rho_{xy\cdot z}$ (the partial correlation between x and y given z) is the correlation between x and y. while removing the influence of z to x and y, respectively.

Association-Based Fairness: Criteria

Criteria of Association-Based Fairness

Fairness through Unawareness — Fairness through Awareness

- Prohibition to access sensitive information during the process of learning and inference

Group Fairness — Individual Fairness

- Fairness for each group, OR fairness for each individual

Statistical Parity

- Satisfying the equality of outcome

Equalized Odds / Sufficiency

- Equalizing biases of prediction from observed data

Context-Sensitive Independence

- Fairness in Specific Contexts

Correlation-based Fairness

- Sensitive information correlates with a target variable

Association-Based Fairness

	fairness through unawareness $\hat{Y} \perp\!\!\!\perp S \mid \mathbf{X}$	**statistical parity** $\hat{Y} \perp\!\!\!\perp S$	**equalized odds** $\hat{Y} \perp\!\!\!\perp S \mid Y$	**sufficiency** $Y \perp\!\!\!\perp S \mid \hat{Y}$
awareness	unaware	aware		
unit	individual	group		
wordview	WAE		WYSIWYG	
comments	treat like cases alike alias: situation testing	equality of outcomes alias: demographic parity, independence	equality of false positive and false negative rates alias: separation	equality of positive and negative predictive values

Fairness through Unawareness

Fairness through Unawareness: Prohibiting to access individuals' sensitive information during the process of learning and inference

This is a kind of procedural fairness, in which a decision is fair, if it is made by following pre-specified procedure

$\Pr[\,\hat{Y} \mid \mathbf{X}, S\,]$

A **unfair model** is trained from a dataset including sensitive and non-sensitive information

$\Pr[\,\hat{Y} \mid \mathbf{X}\,]$

A **fair model** is trained from a dataset eliminating sensitive information

A unfair model, $\Pr[\,\hat{Y} \mid \mathbf{X}, S]$, is replaced with a fair model, $\Pr[\,\hat{Y} \mid \mathbf{X}\,]$

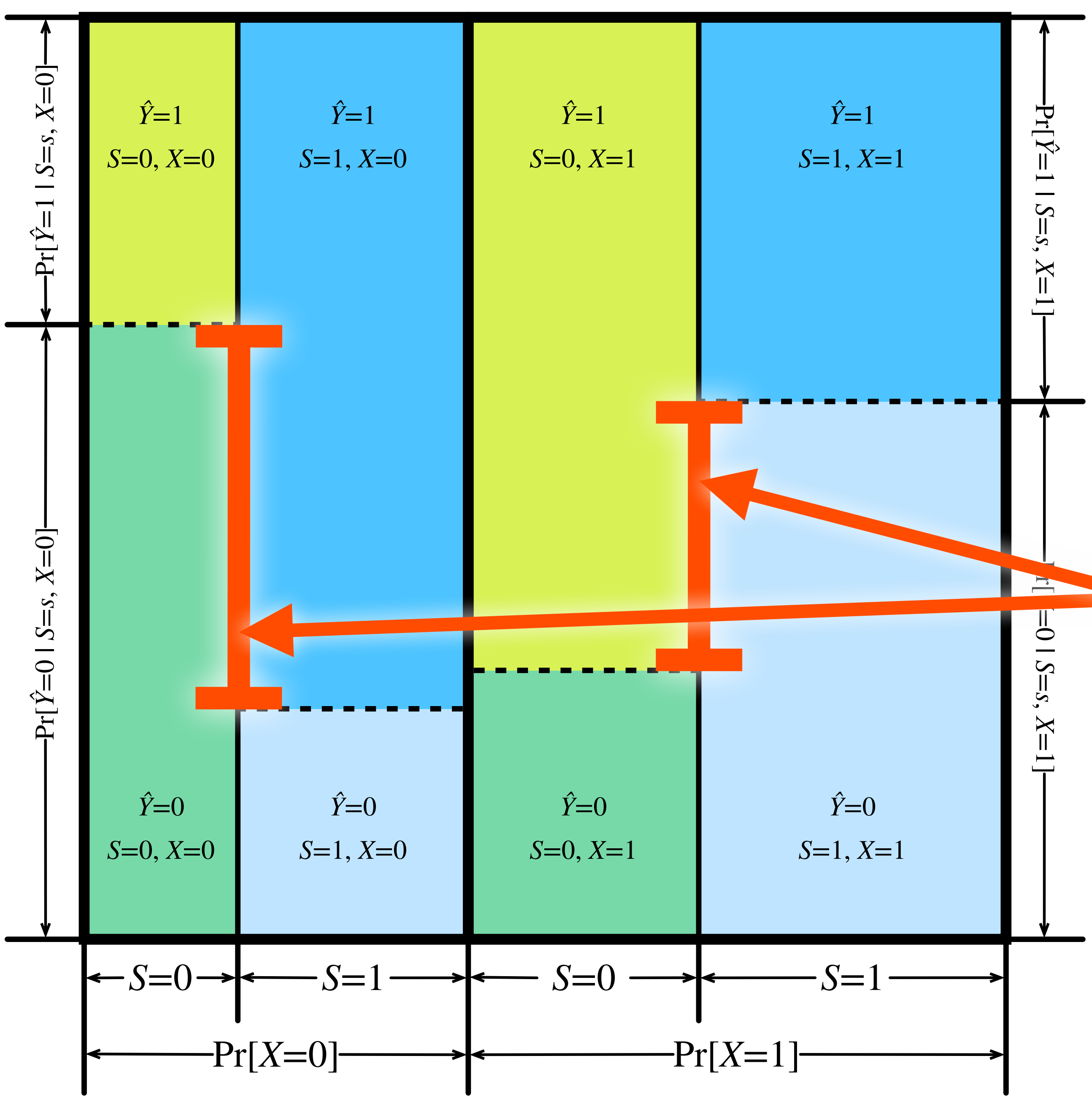
$$\Pr[\,\hat{Y}, \mathbf{X}, S\,] = \Pr[\,\hat{Y} \mid \mathbf{X}, S]\,\Pr[\,S \mid \mathbf{X}\,]\,\mathrm{P}[\,\mathbf{X}\,] \Rightarrow \Pr[\,\hat{Y} \mid \mathbf{X}\,]\,\Pr[\,S \mid \mathbf{X}\,]\,\Pr[\,\mathbf{X}\,]$$

$\Downarrow$

Fairness through Unawareness: $\hat{Y} \perp\!\!\!\perp S \mid \mathbf{X}$

Fairness through Unawareness

a kind of procedural fairness ➜ **Fairness through Unawareness**



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

$$\Updownarrow$$

$$\Pr[\hat{Y}, S \mid X] = \Pr[\hat{Y} \mid X] \Pr[S \mid X]$$

These gaps indicate unfair decision

A learned model directly access sensitive information

Group Fairness / Individual Fairness

Target unit for which a fairness condition is satisfied

Group Fairness

- **Individuals are equally treated as a group**
- Instantiation of the ethical notion "**distributive justice**"
- Implemented by match the aggregated statistics, such as means or errors, between groups
- **Ex:** statistical parity, equalized odds, sufficiency

Individual Fairness

- **Individuals are treated alike regardless of group membership**
- Instantiation of the principle "**treat like cases alike**"
- Implemented by conditioning on individuals, usually represented by $\mathbf{X}$, in a case of association-based fairness
- **Ex:** individual fairness

Group Fairness

Group Fairness: Outcomes of a target variable are equal for all sensitive groups as a whole

- **statistical parity:** equal share between groups

$$\Pr[\hat{Y} \mid S = s] = \Pr[\hat{Y}], \forall s \in \mathrm{Dom}(S) \Rightarrow \hat{Y} \perp\!\!\!\perp S$$

- **equalized odds:** equal errors between group

$$\Pr[\hat{Y} \mid S = s, Y] = \Pr[\hat{Y} \mid Y], \forall s \in \mathrm{Dom}(S) \Rightarrow \hat{Y} \perp\!\!\!\perp S \mid Y$$

Limitations of Group Fairness

- **Individuals are differently treated in each group**

  ➡ some protected individual may receive disadvantageous decision

- **Reverse Tokenism:** justify unfair treatment for members of a protected group by sacrificing a few superior members of a non-protected group [Dwork+ 12]

  ➡ This cannot be prevented by achieving group fairness

Individual Fairness

Individual Fairness: Implementation of the principle of "Treat like cases alike"

Distributions of a target variable are equal for all possible sensitive groups given a specific non-sensitive values

$$\Pr[\, \hat{Y} \mid S, \mathbf{X}{=}\mathbf{x} \,] = \Pr[\hat{Y} \mid \mathbf{X}{=}\mathbf{x}], \forall \mathbf{x} \in \mathrm{Dom}(X) \Rightarrow \hat{Y} \perp\!\!\!\perp S \mid \mathbf{X}$$

Conditioning fairness criteria by $\mathbf{X}$ can be considered as individual fairness

- **Simple individual fairness and fairness through unawareness are the same in a mathematical form, $\hat{Y} \perp\!\!\!\perp S \mid \mathbf{X}$, but not in their semantics**

Ex: To satisfy individual fairness simultaneously with equalized odds, sensitive information must be observed, and this violates a condition of fairness through unawareness

- **Situation Testing:** Legal notion of testing discrimination, comparing individuals having the same non-sensitive values except for their sensitive information

[Luong+ 11]

Detection of Individual Fairness

Probability distributions must be estimated for all non-sensitive values

$$\Pr[\,Y \mid S, \mathbf{X}{=}\mathbf{x}\,] = \Pr[Y \mid \mathbf{X}{=}\mathbf{x}],\ \forall \mathbf{x} \in \mathrm{Dom}(X) \Leftrightarrow Y \perp\!\!\!\perp S \mid \mathbf{X}$$

⬇

To test individual fairness, it is practically impossible to observe data whose non-sensitive values are exactly same

⬇

aggregate information of its neighbors [Luong+ 11]

- A probability distribution, $\Pr[Y \mid S, \mathbf{X}{=}\mathbf{x}]$, is estimated from a dataset composed of the k-nearest neighbor of the point, $\mathbf{x}$

estimate its counterfactual case

- Given a factual case in which $\mathbf{X} = \mathbf{x}$ and $S = s$, its counterfactual case in which $\mathbf{X} = \mathbf{x}$ and $S = s'$ is estimated by assuming the underlying causal relations

Worldview and Bias

[Friedler+ 21]

Worldview is an assumption about mapping from construct space to observed space

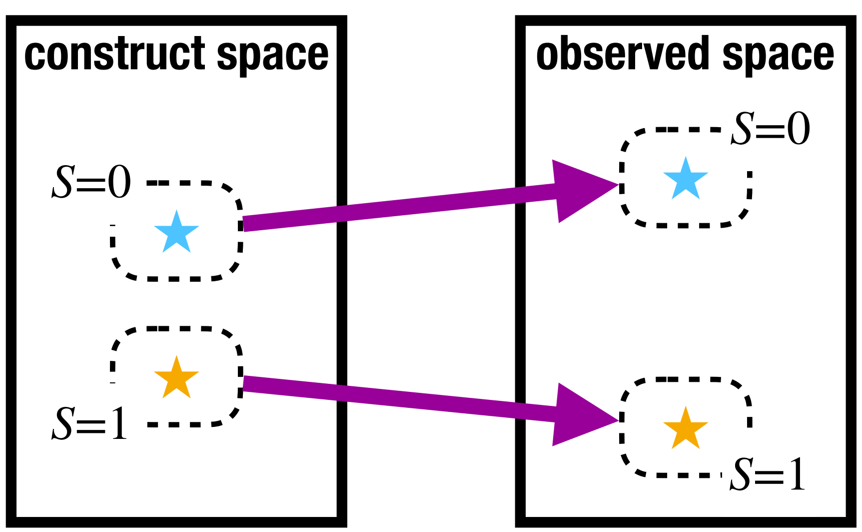
- **construct space:** underlying ideal features and decisions
- **observed space:** observed features and decisions

We're All Equal Worldview

Instances in different groups are mapped differently

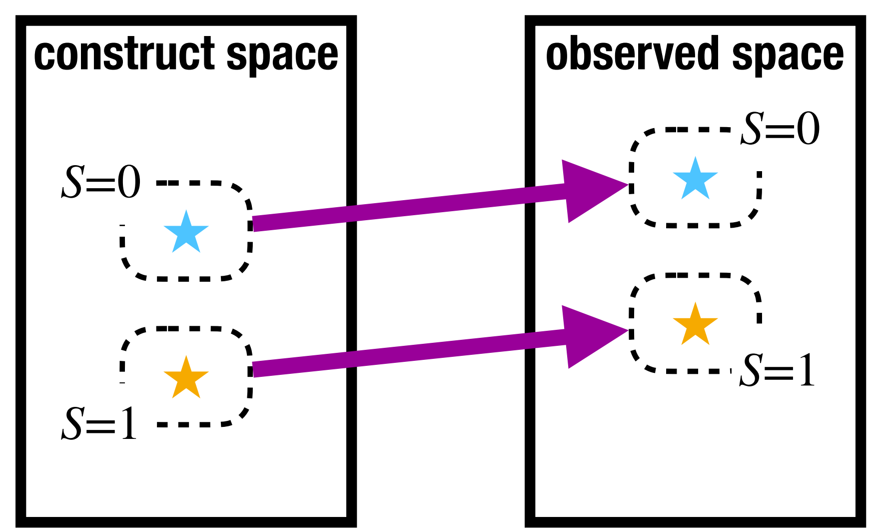
↓

data bias



What You See Is What You Get Worldview

Mapping while keeping relative positions between groups



Statistical Parity / Independence

[Calders+ 10, Dwork+ 12]

equality of outcome: Goods are distributed by following pre-specified procedure

In a context of FAML, the predictions are distributed so as to be proportional to the sizes of sensitive groups

Ratios of predictions are proportional to the sizes of sensitive groups

$$\Pr[Y{=}y_1, S{=}s_1] / \Pr[Y{=}y_2, S{=}s_2] = \Pr[S{=}s_1] / \Pr[S{=}s_2] \quad \forall y_1, y_2 \in \mathrm{Dom}(Y), \forall s_1, s_2 \in \mathrm{Dom}(S)$$

Statistical Parity / Independence: $\hat{Y} \perp\!\!\!\perp S$

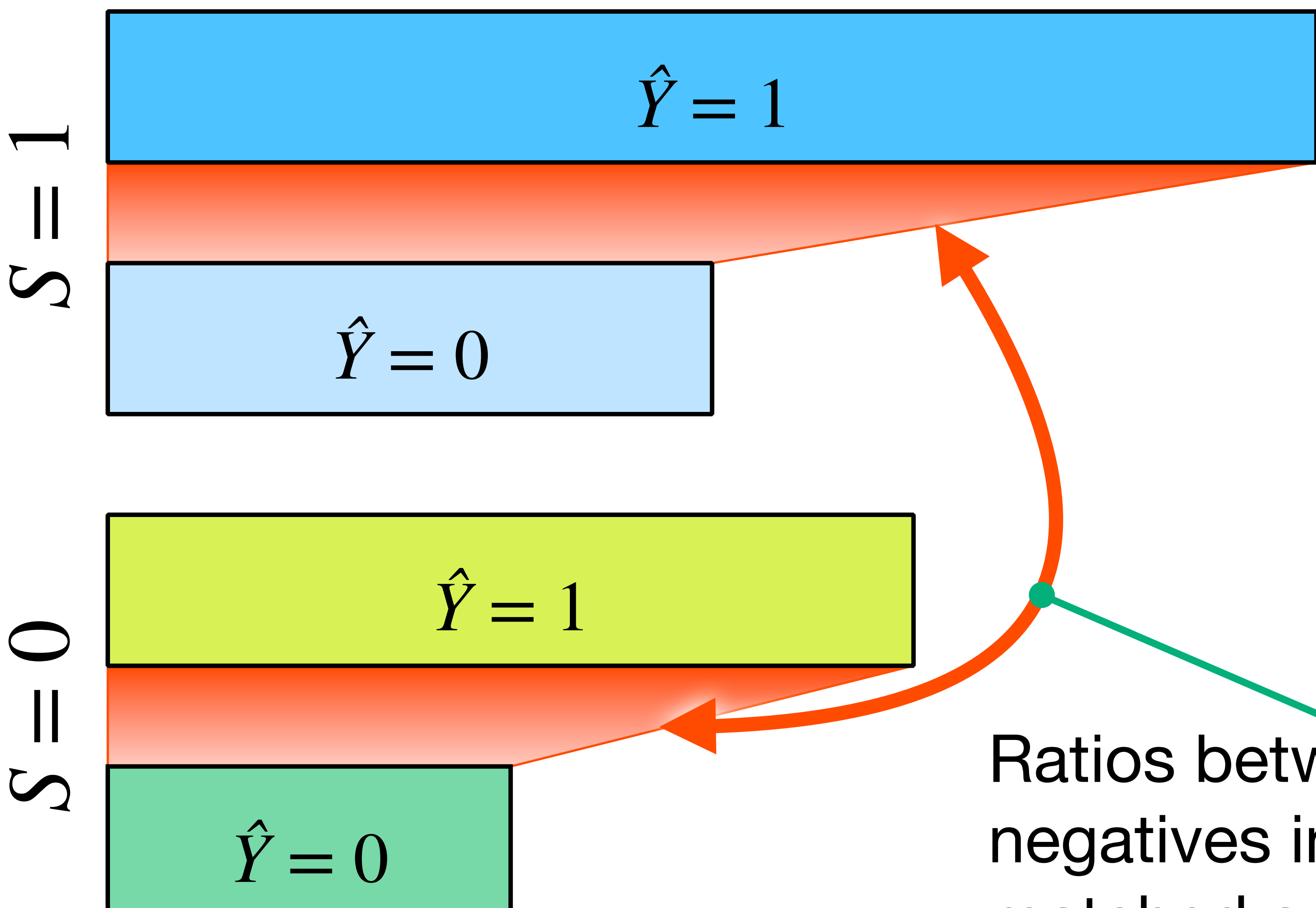
- **Worldview:** “We're All Equal” worldview is assumed, and so it is used for mitigating a data bias
- **Information theoretic view:**

$\hat{Y} \perp\!\!\!\perp S \iff \mathrm{I}(\hat{Y}; S) = 0$ ➡ $\hat{Y}$ has no information about S

Statistical Parity / Independence

[Calders+ 10, Dwork+ 12, Barocas+ 19]

equality of outcome ➡ **Statistical Parity / Independence**



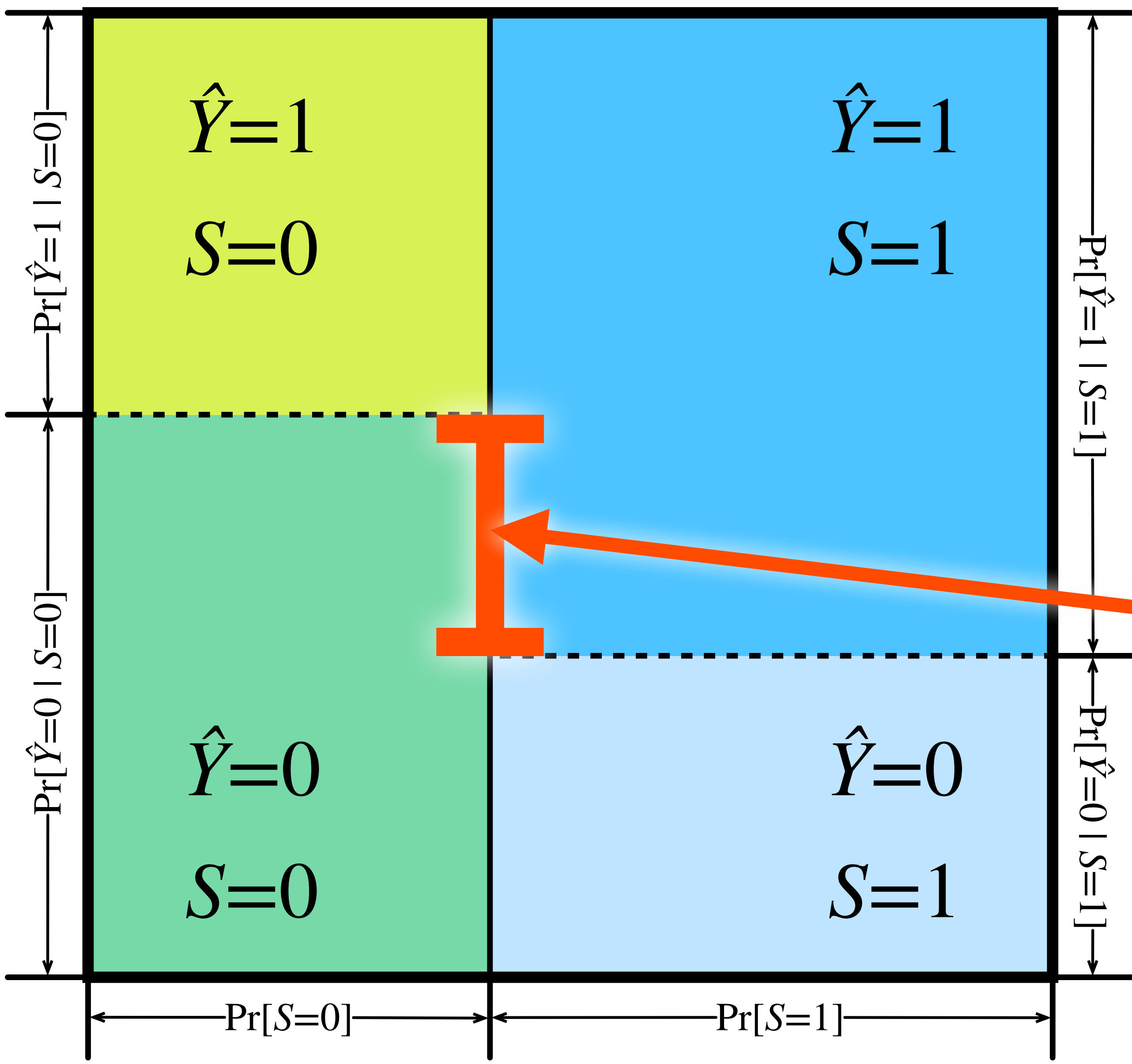
$\hat{Y} \perp\!\!\!\perp S$

Ratios between positives and negatives in prediction should be matched among all sensitive groups

Statistical Parity / Independence

[Calders+ 10, Dwork+ 12, Barocas+ 19]

equality of outcome ➡ **Statistical Parity / Independence**



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

$$\Updownarrow$$

$$\Pr[\hat{Y}, S] = \Pr[\hat{Y}]\,\Pr[S]$$

This gap indicates unfair decision

Ratios between positives and negatives in prediction should be matched among all sensitive groups

Equalized Odds / Separation

[Hardt+ 16, Zafar+ 17]

Removing inductive bias: calibrating inductive errors to observation

- True positive rates should be matched among all sensitive groups

$$\Pr[\hat{Y}{=}1 \mid Y{=}1, S{=}s_1] = \Pr[\hat{Y}{=}1 \mid Y{=}1, S{=}s_2] \quad \forall s_1, s_2 \in \mathrm{Dom}(S)$$

- False positive rates should be matched among all sensitive groups

$$\Pr[\hat{Y}{=}1 \mid Y{=}0, S{=}s_1] = \Pr[\hat{Y}{=}1 \mid Y{=}0, S{=}s_2] \quad \forall s_1, s_2 \in \mathrm{Dom}(S)$$

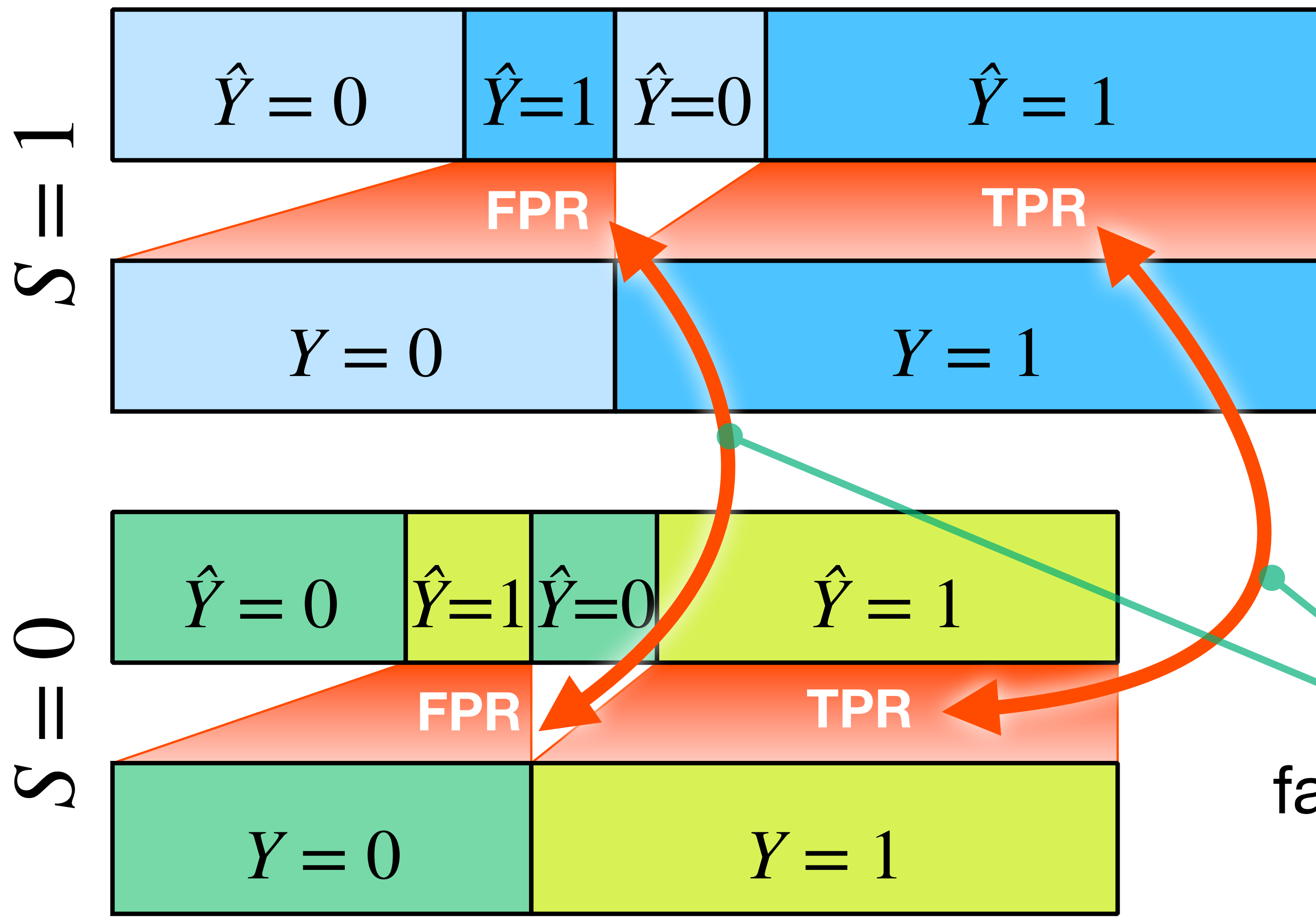
Equalized Odds / Separation: $\hat{Y} \perp\!\!\!\perp S \mid Y$

- **Worldview:** "What You See Is What You Get" worldview is assumed, and so it is used for mitigating an inductive bias

Equalized Odds

[Hardt+ 16, Zafar+ 17, Barocas+ 19]

Removing inductive bias ➜ **Equalized Odds / Separation**



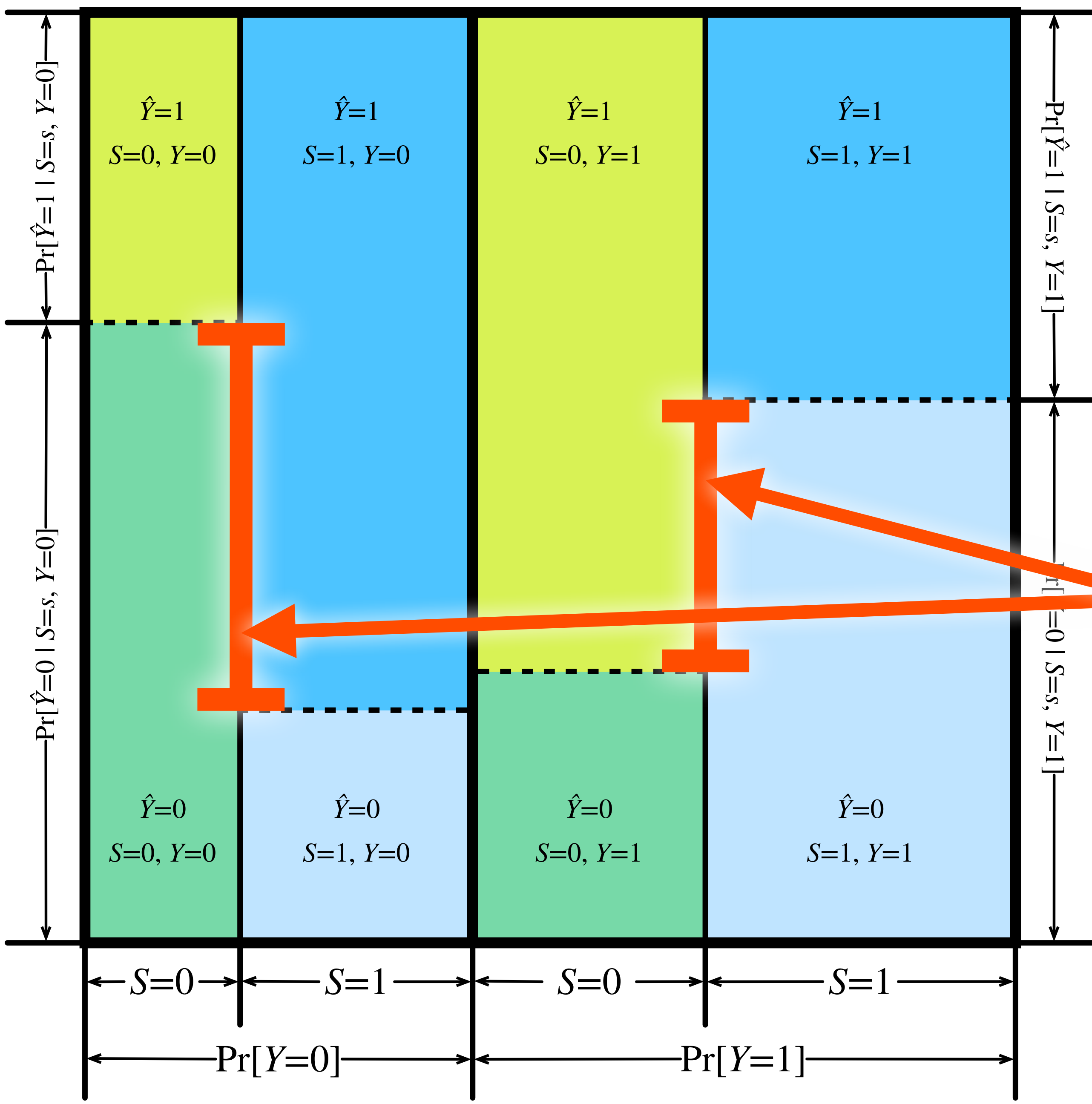
$$\boxed{\hat{Y} \perp\!\!\!\perp S \mid Y}$$

false positive ratio and
true positive ratio
should be matched
among all sensitive groups

Equalized Odds

[Hardt+ 16, Zafar+ 17, Barocas+ 19]

Removing inductive bias ➡ **Equalized Odds / Separation**



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

$$\Updownarrow$$

$$\Pr[\hat{Y}, S \mid Y] = \Pr[\hat{Y} \mid Y]\,\Pr[S \mid Y]$$

These gaps indicate unfair decision

False positive ratio (FPR) and true positive ratio (TPR) should be matched among all sensitive groups

Sufficiency / Calibration

[Flores+ 16, Chouldechova 17, Barocas+ 19]

Removing inductive bias: calibrating inductive errors to observation

- Positive predictive values should be matched between any groups

  $$\Pr[Y{=}1 \mid \hat{Y}{=}1, S{=}s_1] = \Pr[Y{=}1 \mid \hat{Y}{=}1, S{=}s_2] \quad \forall s_1, s_2 \in \mathrm{Dom}(S)$$

- Positive predictive values should be matched between any groups

  $$\Pr[Y{=}0 \mid \hat{Y}{=}0, S{=}s_1] = \Pr[Y{=}0 \mid \hat{Y}{=}0, S{=}s_2] \quad \forall s_1, s_2 \in \mathrm{Dom}(S)$$

Sufficiency / Calibration: $Y \perp\!\!\!\perp S \mid \hat{Y}$

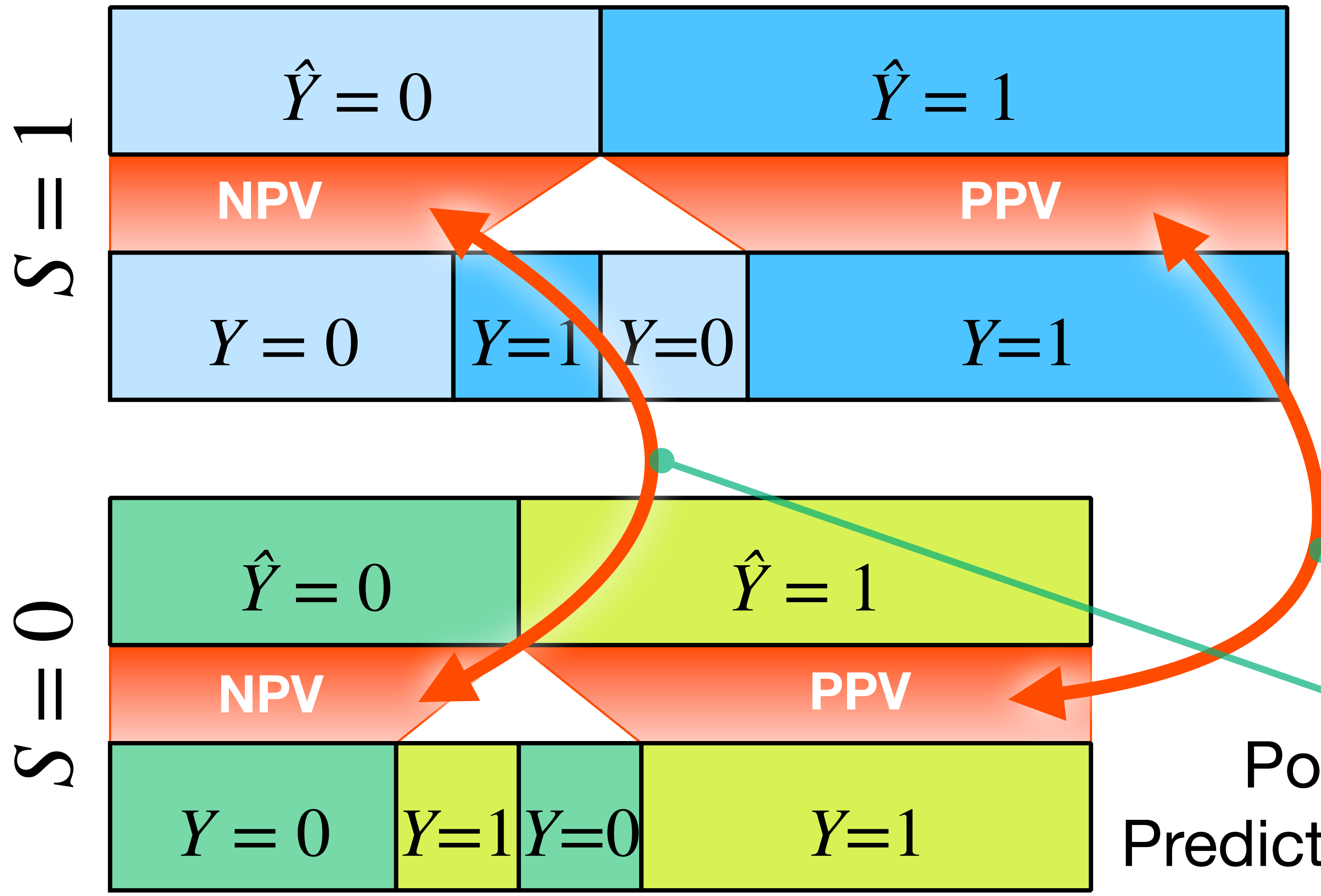
- **Worldview:** "What You See Is What You Get" worldview is assumed, and so it is used for mitigating an inductive bias
- In psychology or education disciplines, this criterion is accepted as a fairness condition

[Chouldechova 17]

Sufficiency

[Flores+ 16, Chouldechova 17, Barocas+ 19]

Removing inductive bias ➡ **Sufficiency / Calibration**



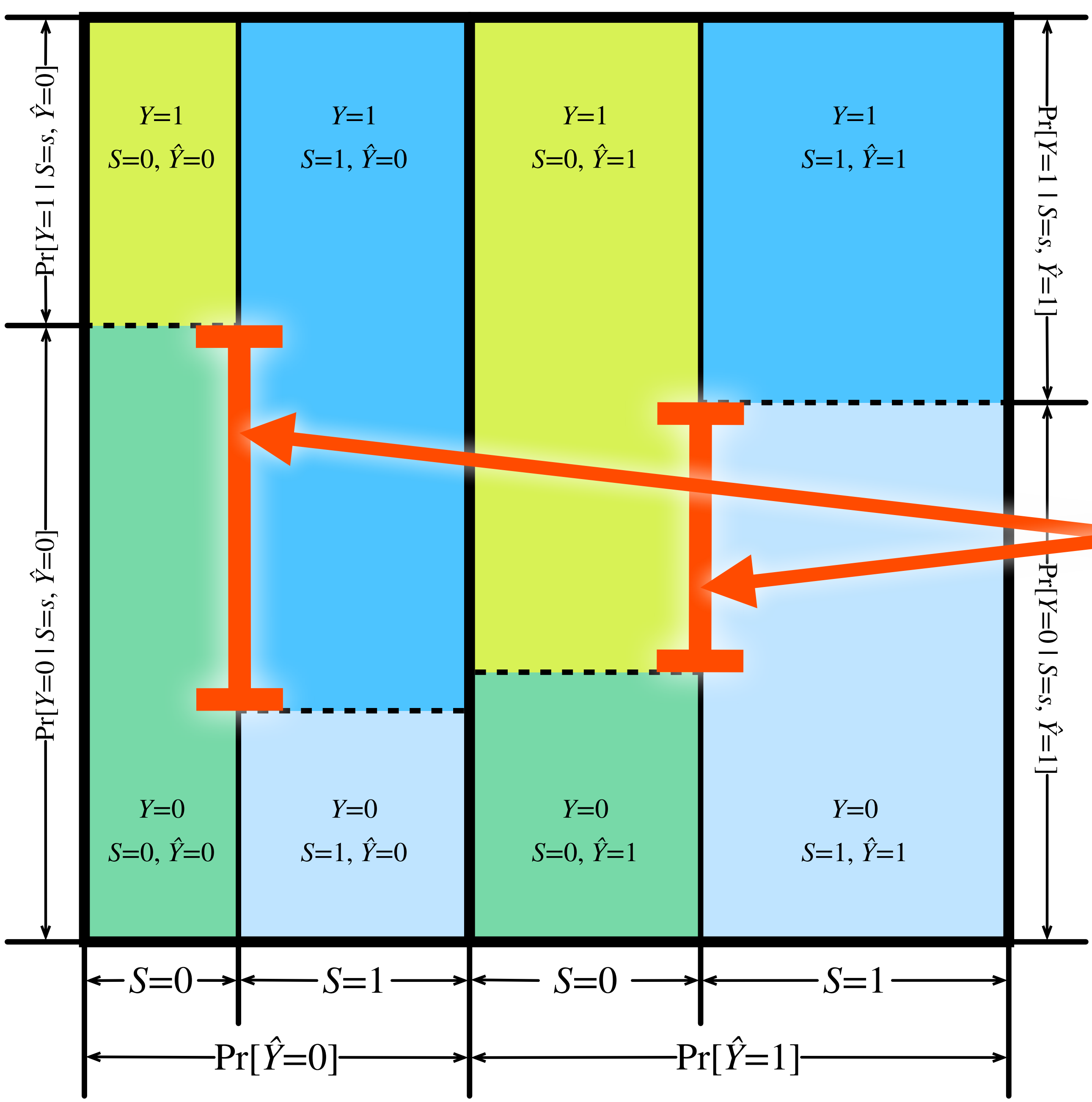
$$\boxed{\hat{Y} \perp\!\!\!\perp S \mid Y}$$

Positive and Negative Predictive Values (PPV & NPV) should be matched among all sensitive groups

Sufficiency

[Flores+ 16, Chouldechova 17, Barocas+ 19]

Removing inductive bias ➡ **Sufficiency / Calibration**



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

$$\Updownarrow$$

$$\Pr[Y, S \mid \hat{Y}] = \Pr[Y \mid \hat{Y}] \Pr[S \mid \hat{Y}]$$

These gaps indicate unfair decision

Precisions for positive and negative classes should be matched among all sensitive groups

* $\hat{Y}$ and Y are exchanged from the separation case

Context-Specific Independence

[Boutiller+ 96]

Context-Specific Independence: Y and S are independent, if $\mathbf{X}$ are fixed to specific values, $\mathbf{x}$

α-protection

[Pedreschi+ 08]

$\Pr[\hat{Y}{=}1 \mid S{=}0, \mathbf{X}{=}\mathbf{x}] \,/\, \Pr[\hat{Y}{=}1 \mid \mathbf{X}{=}\mathbf{x}] \leqq \alpha$

- α-protection is the context-specific independence, $\hat{Y} \perp\!\!\!\perp S \mid \mathbf{X}{=}\mathbf{x}$

Equalized Odds / Equal Opportunity

[Hardt+ 16]

- Equalized odds is conditional independence, $\hat{Y} \perp\!\!\!\perp S \mid Y$
- Equal Opportunity is context-specific independence, $\hat{Y} \perp\!\!\!\perp S \mid Y{=}1$

Sufficiency / Predictive Parity

[Chouldechova 17]

- Sufficiency is conditional independence, $Y \perp\!\!\!\perp S \mid \hat{Y}$
- Predictive Parity is context-specific independence, $Y \perp\!\!\!\perp S \mid \hat{Y}{=}1$

Correlation-Based Fairness

[Hutchinson+ 19]

Fairness in DM/ML has been discussed from 2010s

↑

A statistics literature had discussed fairness criteria in 1960 — 70s after the US Civil Rights Act, 1964

ML / DM		Statistics
Independence	↔	Correlation
Conditional Independence		Partial Correlation
Discovery & Prevention		Discovery only

Statistical Parity / Independence

- Darlington (1971) criterion 4

Equalized Odds / Separation

- Cleary (1968), Darlington (1971) criterion (1), Linn (1973)

Sufficiency / Calibration

- Darlington (1971) criterion (2)

Association-Based Fairness: Properties

Properties of Formal Fairness

Disparate treatment — Disparate Impact

- Groups or individuals are intentionally treated differently, OR
- Unintentional impact on distinct groups or individuals

Direct Discrimination — Indirect Discrimination

- Sensitive information influences targets directly, or indirectly

Type of Biases to Remove

- Fairness criteria are designed to remove a specific type of bias

Relation between Fairness Criteria

- One criterion implies or conflicts with other criterion

Explainable Variable

- Exclusion of the explainable confounding effects between sensitives and targets

Disparate Treatment / Disparate Impact

[Barocas+ 17, Feldman+ 15]

legal notions about fairness

Disparate Treatment		Disparate Impact
equality of opportunity tolerant to unequal outcome	⟷	**equality of outcome** allow reverse discrimination
procedural fairness eliminate sensitive information	⟷	**distributive justice** fair allocation of goods
intended direct or intentional reference of sensitive information	⟷	**unintended** indirect reference of sensitive information

Direct Discrimination & Indirect Discrimination

[Pedreschi+ 08, Žliobaitė+ 16]

technical notions about fairness

Direct Discrimination	↔	**Indirect Discrimination**
discrimination on the basis of sensitive information		discrimination on the basis of other features resulting in direct discrimination

These technical notions are often expressed by legal terms

Disparate Treatment

Strictly speaking, disparate treatment includes intended indirect reference to sensitive information

Disparate Impact

Strictly speaking, whether or not the reference is intended should be cared in a disparate impact case

Red-Lining Effect

[Calders+ 10]

Red-Lining Effect: Simple elimination of a sensitive features from training dataset fails to remove the influence of sensitive information to a target

Eliminating sensitive information is equivalent to replacing an unfair model, $\Pr[Y \mid \mathbf{X}, S]$ with a fair model, $\Pr[Y \mid \mathbf{X}]$

$$\Pr[\, Y, \mathbf{X}, S \,] = \Pr[\, Y \mid \mathbf{X}, S] \Pr[\, S \mid \mathbf{X} \,] \mathrm{P}[\, \mathbf{X} \,] \Rightarrow \Pr[\, Y \mid \mathbf{X} \,] \Pr[\, S \mid \mathbf{X} \,] \Pr[\, \mathbf{X} \,]$$

This corresponds to conditional independence: $\hat{Y} \perp\!\!\!\perp S \mid \mathbf{X}$ (not $\hat{Y} \perp\!\!\!\perp S$)

S still influences Y through $\mathbf{X}$

Red-Lining Effect

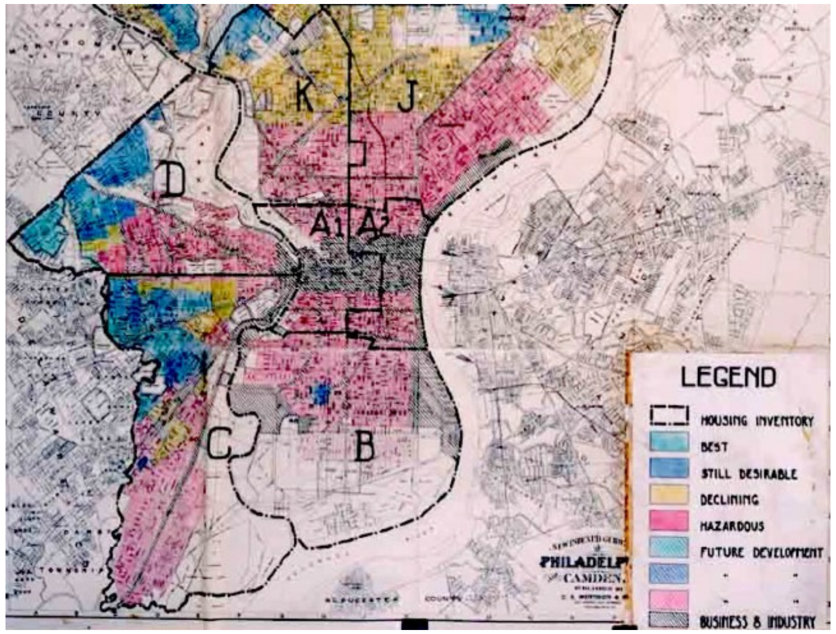
[Calders+ 10]

fairness through unawareness = eliminating a sensitive feature

Red-Lining Effect: Elimination of a sensitive information from training dataset fails to remove the influence of the information to a target

Ex: People of the same race frequently resident in a specific region

Even if their race are not explicitly referred, the information is included in that of their residential region



[Wikipedia]

Distributive justice cannot be satisfied under fairness through unawareness

Types of Bias to Remove

Three sources of biases that undesirably corrupt outcomes

- **Data / Annotation Bias:** unfair labeling by annotators; inappropriately observed feature values
- **Sample Selection Bias:** dataset that is not a representative of population to analyze
- **Inductive Bias:** propensity of ML algorithms caused by assumptions in the algorithms' inductive process

Sources of undesired outcomes depends on problems

Formal fairness have to be selected by considering which type of biases tries to remove

Removing Data Bias

Data / Annotation Bias: Target values or feature values in a training data are biased due to annotator's cognitive bias or inappropriate observation schemes

↓

data are not reliable, and never accessible to a fair dataset

↓

Assumptions about the conditions that values or distributions of target variables and sensitive features should satisfy

↓

Examples of assumptive conditions:

- $\hat{Y} \perp\!\!\!\perp S$: statistical parity
- $\hat{Y} \perp\!\!\!\perp S \mid \mathbf{X}$: fairness through unawareness
- $\hat{Y} \perp\!\!\!\perp S \mid \mathbf{X}{=}\mathbf{x}$: Y and S are context-sensitive independent given $\mathbf{X}{=}\mathbf{x}$

Removing Sample Selection Bias

Sample Selection Bias: Whether a datum is sampled depends on conditions or contents of the datum, and thus an observed dataset is not a representative of population

Batch Learning: Training data violates a condition of random assignment in terms of sensitive information

- incorrectly annotated by an ML algorithm
  ➡ modify an inductive bias of the ML algorithm
- not sampled uniformly at random, as seen in a statistical survey
  ➡ modify data so as to satisfy a condition of random assignment

Online Learning: Selection of data to test is biased in an ML tasks with a feedback loop, e.g., bandits, reinforcement learning, active learning

- biased selection of data to test or investigate
  ➡ select randomly in terms of sensitive information

Removing Inductive Bias

Inductive Bias: a bias caused by an assumption adopted in an inductive machine learning algorithms

⬇

Outcomes in a training dataset, Y, are assumed to be reliable, and the prediction, $\hat{Y}$, might be different from the observed, Y.

⬇

The changes from Y to $\hat{Y}$ should be balanced between sensitive groups defined by S

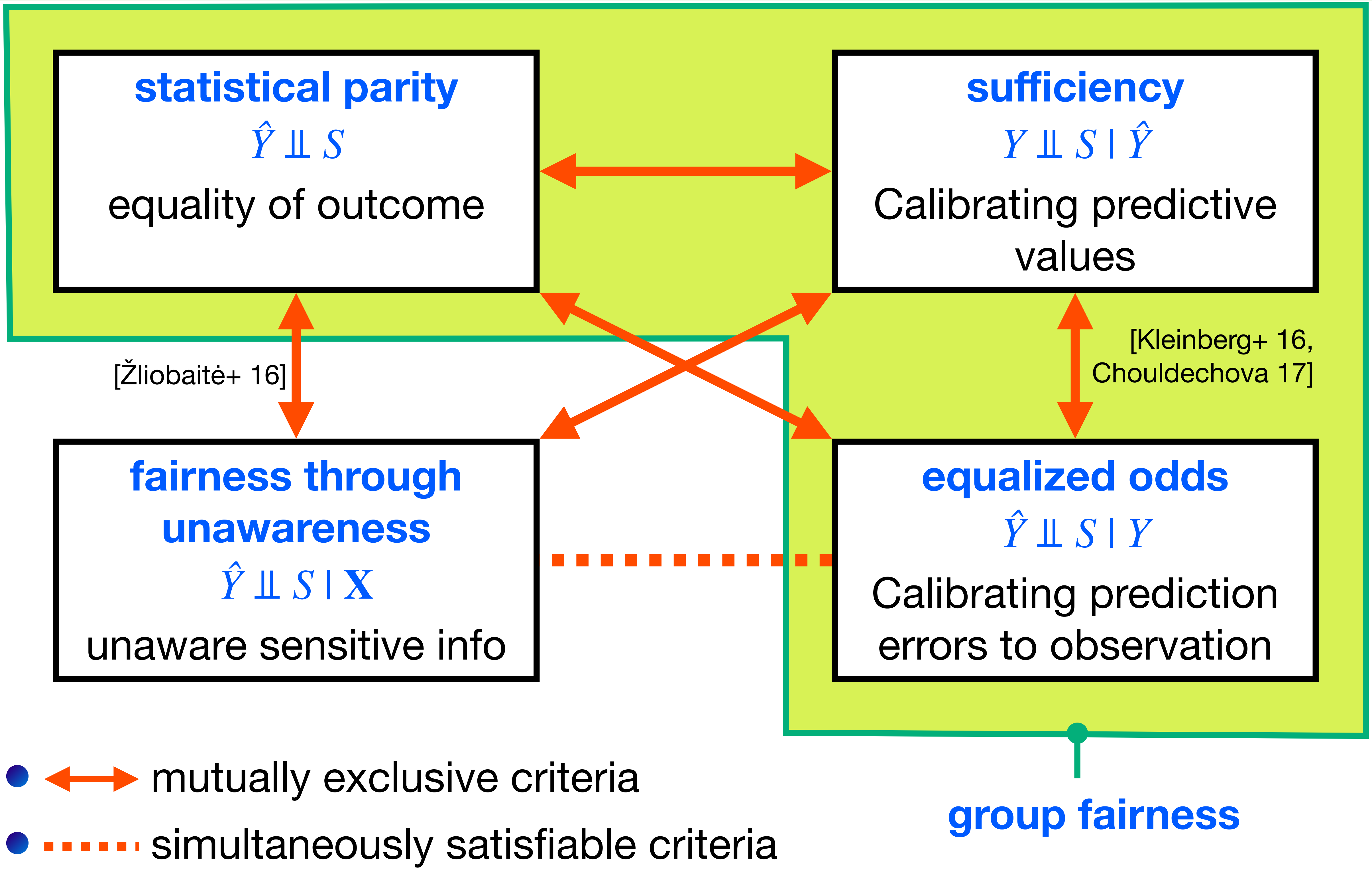
⬇

$\hat{Y} \perp\!\!\!\perp S \mid Y$: Equalized Odds / Separation

II

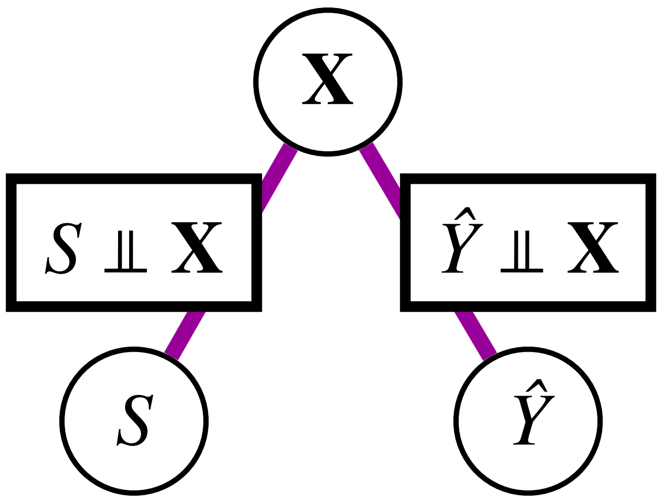
Empirical errors of $\hat{Y}$ over sample outcomes, Y, are equal for all groups consist of the same sensitive values

Satisfiablity between Fairness Criteria



Fairness through Unawareness & Statistical Parity

[Žliobaitė+ 16]



Satisfying **fairness through unawareness**, $S \perp\!\!\!\perp \hat{Y} \mid \mathbf{X}$

⬇

To simultaneously satisfy **statistical parity,** $S \perp\!\!\!\perp \hat{Y}$,
a condition of $S \perp\!\!\!\perp \mathbf{X}$ OR $\hat{Y} \perp\!\!\!\perp \mathbf{X}$ must be satisfied

⬇

$S \perp\!\!\!\perp \mathbf{X}$: a sensitive feature and non-sensitive features are independent

- **unrealistic** ⬅ $\mathbf{X}$ and S are uncontrollable, and $\mathbf{X}$ is high-dimensional

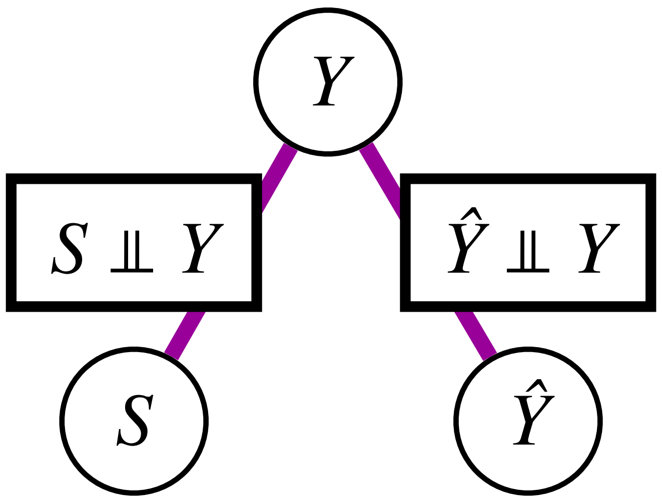
$\hat{Y} \perp\!\!\!\perp \mathbf{X}$: a sensitive feature and a target variable are independent

- **meaningless** ⬅ $\hat{Y}$ must be random guess

⬇

Simultaneous satisfaction of individual fairness and statistical parity is unrealistic or meaningless

Equalized Odds & Statistical Parity



Equalized odds, $S \perp\!\!\!\perp \hat{Y} \mid Y$, is satisfied

↓

To simultaneously satisfy **statistical parity,** $S \perp\!\!\!\perp \hat{Y}$, a condition of $S \perp\!\!\!\perp Y$ OR $\hat{Y} \perp\!\!\!\perp Y$ must be satisfied

↓

$S \perp\!\!\!\perp Y$: a observed class and non-sensitive features are independent

- **violating an assumption** ← observed classes are already fair

$\hat{Y} \perp\!\!\!\perp Y$: a sensitive feature and a target variable are independent

- **meaningless** ← Y depends on $\mathbf{X}$ and $\hat{Y}$ must be random guess

↓

Simultaneously satisfying equalized odds and statistical parity is meaningless

Impossibility between Sufficiency and Equalized Odds

[Kleinberg+ 16]

Well-calibration (= sufficiency):

True class distribution given the prediction is independent from groups

$$Pr[Y \mid \hat{Y} = \hat{y}] = Pr[Y \mid \hat{Y} = \hat{y}, S = s], \forall \hat{y}, s$$

Balance for the positive and negative classes (= equalized odds):

TPR and NPR are equal between sensitive groups

$$\Pr[\hat{Y} = 1 \mid Y = y, s = 0] = \Pr[\hat{Y} = 1 \mid Y = y, s = 1], \forall y$$

$$\Downarrow$$

Perfect prediction: $\Pr[Y = 1 \mid \mathbf{x}] \in \{0,1\}, \forall \mathbf{x} \in \mathrm{Dom}(X)$

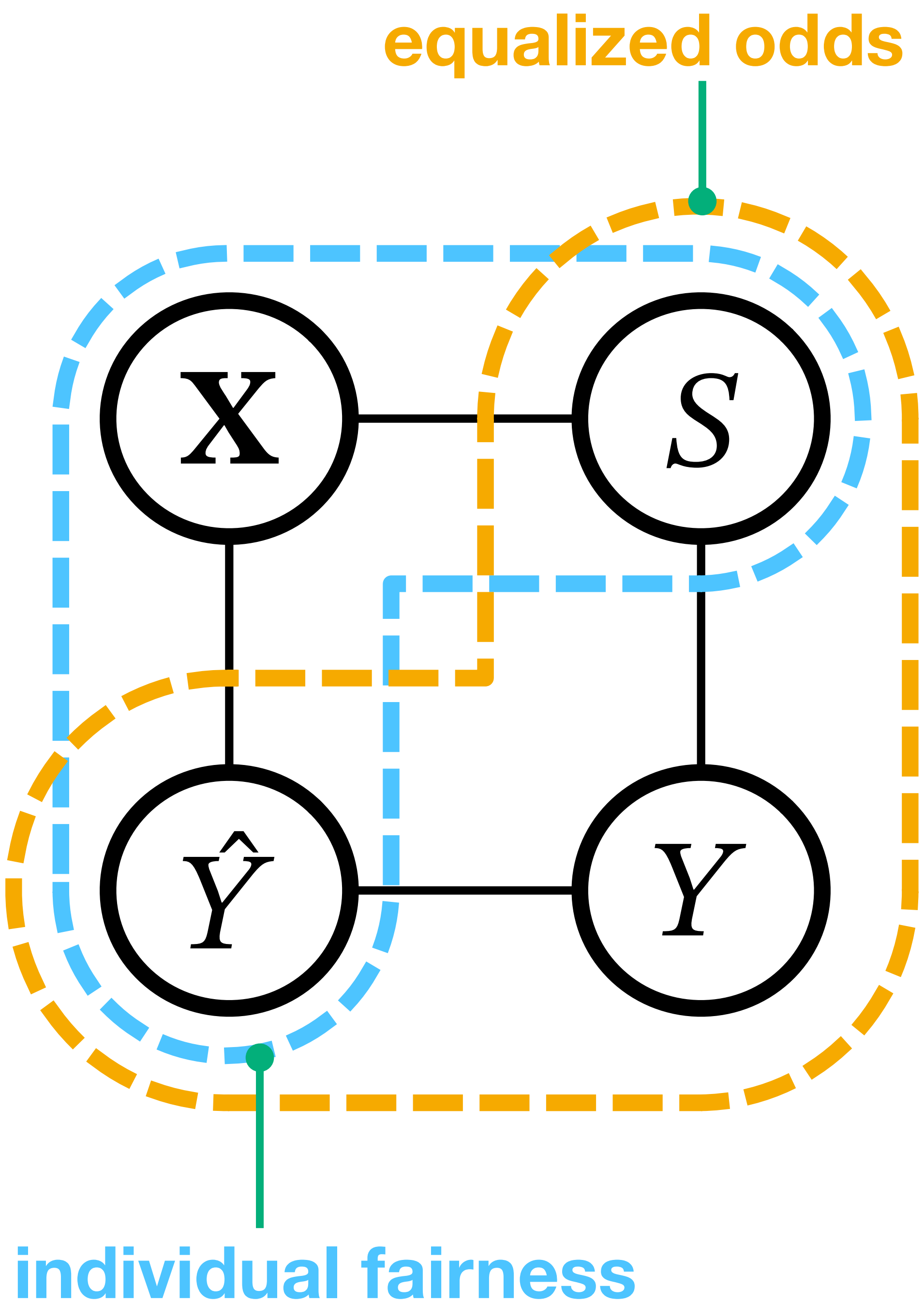
Equal base rates: $\Pr[Y = 1 \mid S = 0] = \Pr[Y = 1 \mid S = 1] \equiv Y \perp\!\!\!\perp S$

Satisfying sufficiency and equalized odds implies distributions of true class must be either perfect prediction or equal base rates

$$\Downarrow$$

Sufficiency and Equalized odds cannot be satisfied simultaneously in general

Individual Fairness & Equalized Odds



- Equalized odds, $\hat{Y} \perp\!\!\!\perp S \mid Y$, and individual fairness, $\hat{Y} \perp\!\!\!\perp S \mid \mathbf{X}$, can be simultaneously satisfiable
- The resultant combined condition is:

$$\Pr[\hat{Y}, Y, S, \mathbf{X}] = \Pr[\hat{Y} \mid \mathbf{X}] \Pr[S \mid \mathbf{X}] \Pr[\mathbf{X}] \Pr[\hat{Y} \mid Y] \Pr[S \mid Y] \Pr[Y]$$

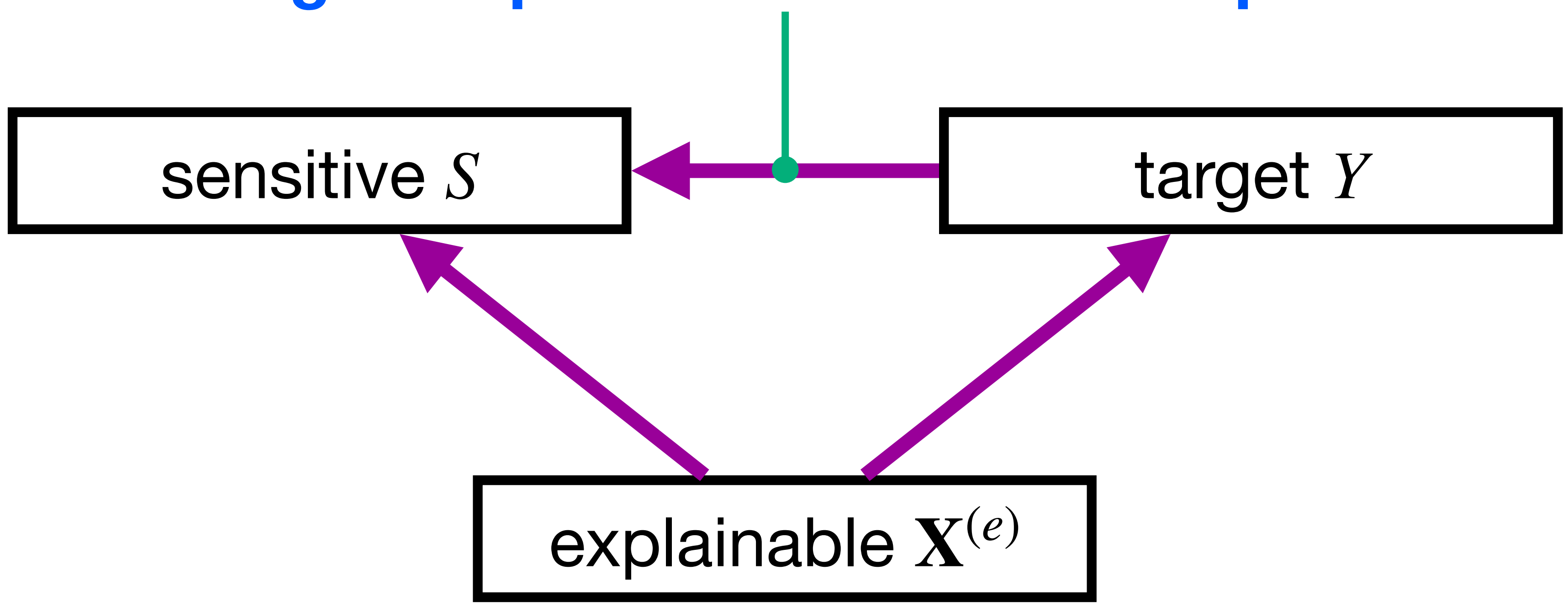
- A condition, $\hat{Y} \perp\!\!\!\perp S \mid \mathbf{X}, Y$, is weaker than the combined condition, but what the two criteria are intended to accomplish is fulfilled

Explainable Variable

[Žliobaitė+ 11, Kamiran+ 13]

Explainable Variable / Legally-grounded Variable: these variables influence both target and sensitive variables, and the influence is not semantically problematic

In FAML, we are interested in the **pure effect** from a sensitive feature to a target **excluding the spurious effect of an explainable variable**



genuine occupational requirement: the nature of he role makes it unsuitable for individuals with a particular sensitive value

Ex: Fashion model for feminine clothes should be female

Fair Determination

[Žliobaitė+ 11, Kamiran+ 13]

Is the target determination fair in terms of a sensitive state

An example of university admission in [Žliobaitė+ 11]

sensitive feature: S
gender
male / female

target variable: Y
admission
admit / not admit

Fair determination: the gender does not influence the acceptance

⬇

(unconditional) independence: $Y \perp\!\!\!\perp S$

Causality with Explainable Features

[Žliobaitė+ 11, Kamiran+ 13]

An example of fair determination
even if S and Y are not independent

sensitive feature: S
gender
male / female

target variable: Y
admission
admit / not admit

female → medicine=high
male → computer=high

medicine → admission=low
computer → admission=high

explainable feature: $\mathbf{X}^{(e)}$
(confounding feature)
department
medicine / computer

Removing the **pure** influence of S to Y, excluding the effect of $\mathbf{X}^{(e)}$

conditional statistical independence: $Y \perp\!\!\!\perp S \mid \mathbf{X}^{(e)}$

Association-Based Fairness: Measures

Difference-based Measures

risk difference / **mean difference** [Calders+ 10, Pedreschi 09]

Difference of receiving advantageous decisions between groups

$$\mathrm{RD} = \Pr[\hat{Y} = 1 | S = 1] - \Pr[\hat{Y} = 1 | S = 0]$$

- RD → 0 ➡ $Y \perp\!\!\!\perp S$
- equivalent to the total causal effect of changing S on $\hat{Y}$

balanced error ratio [Feldman+ 15]

mean of the probability of the disadvantageous decision for a non-protected group and the probability of the advantageous decision for protected group

$$\mathrm{BER} = \frac{\Pr[\hat{Y} = 0 | S = 1] + \Pr[\hat{Y} = 1 | S = 0]}{2} = \frac{1 - \mathrm{RD}}{2}$$

- BER → 1/2 ➡ $Y \perp\!\!\!\perp S$

elift (extended lift)

[Pedreschi+ 08, Ruggieri+ 10]

elift (extended lift) = $$\frac{\mathrm{conf}(\mathbf{X}{=}\mathbf{x}, \boxed{S{=}0} \Rightarrow Y{=}0)}{\mathrm{conf}(\mathbf{X}{=}\mathbf{x} \Rightarrow Y{=}0)}$$

the ratio of the confidence of a rule with a **sensitive condition**,
to that of a rule without the condition

The condition elift = 1 means that no unfair treatments, and it implies

$$\Pr[\ Y{=}0 \mid S{=}0, \mathbf{X}{=}\mathbf{x}\ ] = \Pr[\ Y{=}0 \mid \mathbf{X}{=}\mathbf{x}\ ]$$

when S and Y are additionally binary variables,

This condition is equivalent to the context-sensitive independence:

$$Y \perp\!\!\!\perp S \mid \mathrm{X}{=}\mathrm{x}$$

Useful for finding unfair effects from S to Y under the context of $\mathbf{X}{=}\mathbf{x}$

Measures from Contingency Table

[Pedreschi+ 09, Hajian+ 16, Zhang 18]

	$\hat{Y}=0$	$\hat{Y}=1$
$S=0$	a_1	$n_1 - a_1$
$S=1$	a_2	$n_2 - a_2$

$$p_0 = \Pr[\hat{Y}{=}0 \mid S{=}0] = \frac{a_0}{n_0}$$

$$p_1 = \Pr[\hat{Y}{=}0 \mid S{=}1] = \frac{a_1}{n_1}$$

$$p = \Pr[\hat{Y}{=}0] = \frac{a_0 + a_1}{n_0 + n_1}$$

$p_0 - p_1 =$ risk difference / mean difference / slift$_d$

$p_0 - p =$ extended risk difference / elift$_d$

$p_0 / p_1 =$ risk ratio / relative risk / slift

$(1 - p_0)/(1 - p_1) =$ relative chance

$p_0 / p =$ extended risk ratio / elift

$\dfrac{p_0(1 - p_1)}{p_1(1 - p_0)} =$ odds ratio / olift

Counterfactual Fairness

Counterfactual Fairness: Basics of Causal Inference

Pearl's Ladder of Causation

[Pearl+ 18]

Counterfactuals

Activity: Imaging, Retrospection, Understanding
Questions: What if I had done …? Why?
Examples: Was it the aspirin that stopped my headache?

Intervention

Activity: Doing, Intervening
Questions: What if I do …? How?
Examples: If I take aspirin, will my head ache be cured

Association

Activity: Seeing, Observing
Questions: What if I see …?
Examples: What does a symptom tell me about a disease?

Structural Causal Model

Structural Causal Model: represents causal dependency

Formula Representation

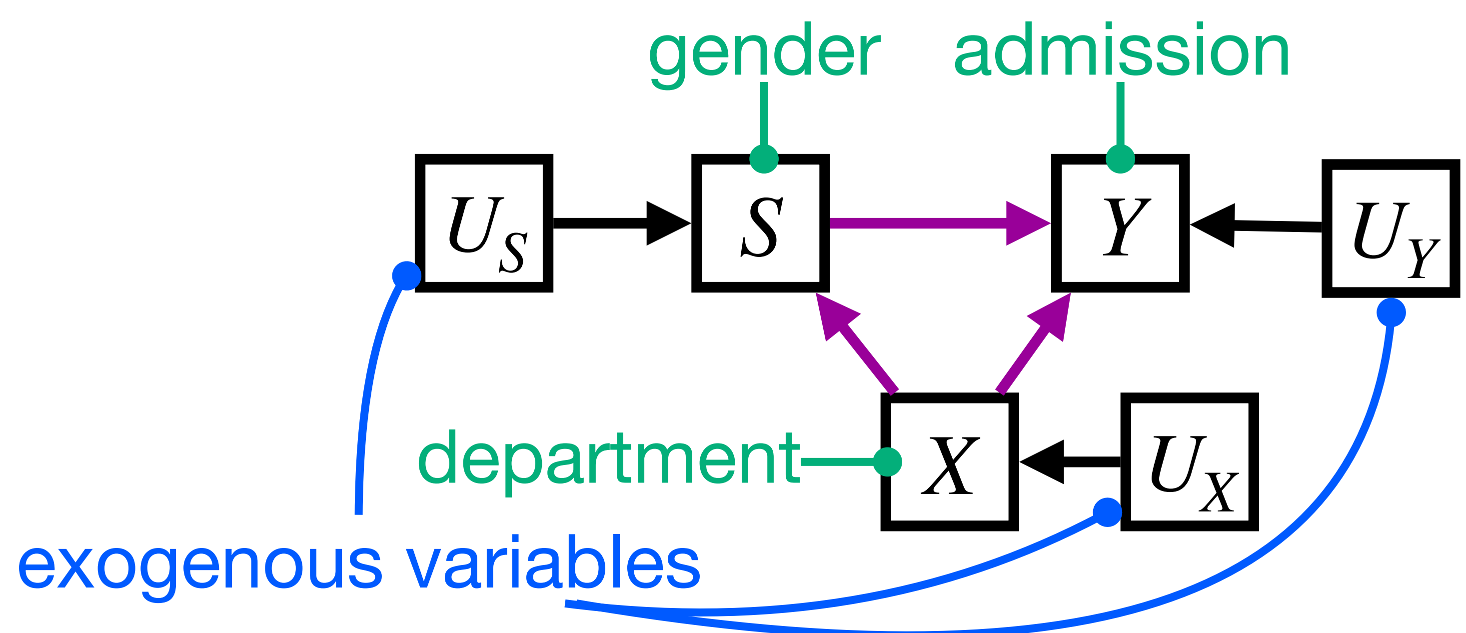
$$X \sim f_X(U_X)$$

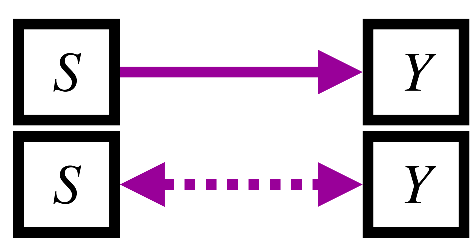
$$S \sim f_S(X, U_S)$$

$$Y \sim f_Y(S, X, U_Y)$$

Graphical Representation

university admission





S is a direct cause of an effect Y

S might cause Y, and vice versa

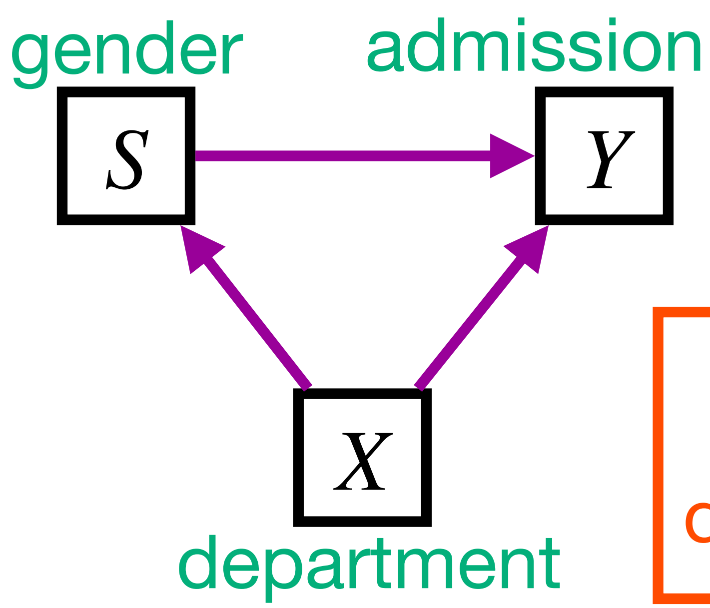
- S, X, Y are observed, and U's are unobserved (usually omitted)
- The SCM is Markovian, if exogenous variables are mutually independent

Intervention

Association

$\Pr[Y \mid S = s]$

Observing Y, where $S = s$

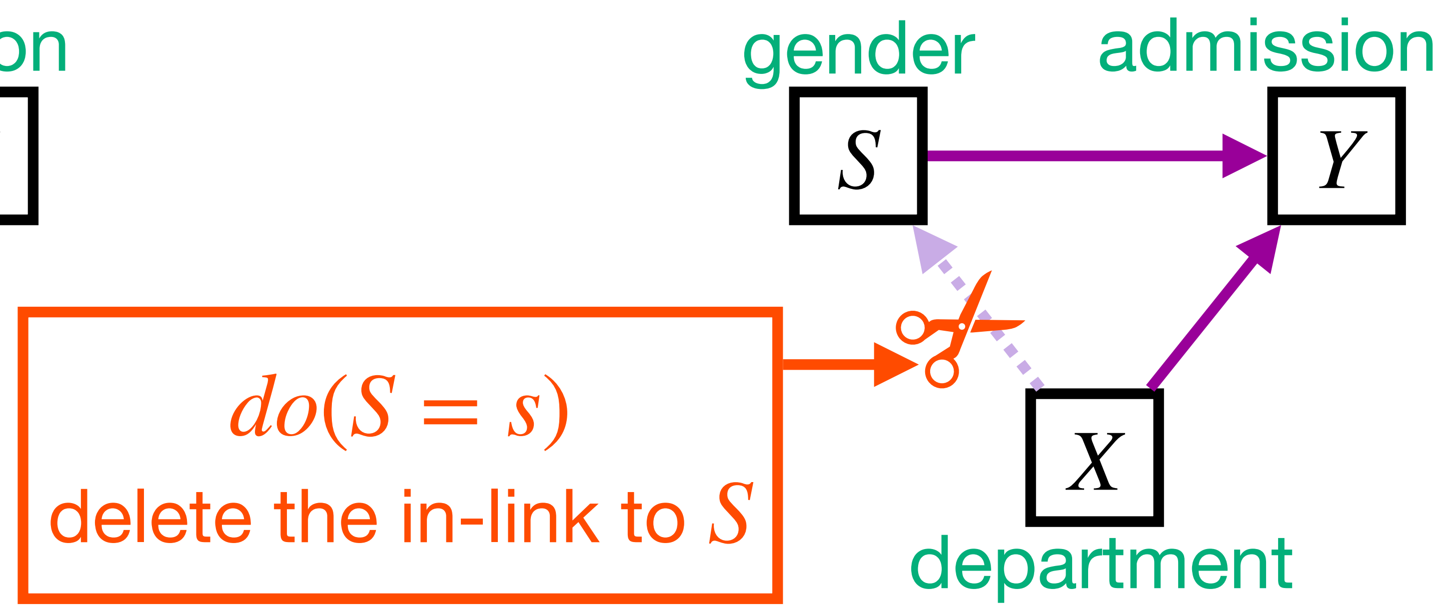


Select the cases $S = s$, without any modifications on the model

Intervention

$\Pr[Y \mid do(S = s)]$

How Y would be changed, If S is changed to s



After deleting all the in-links to the intervened variable, set $S = s$

Association vs. Intervention

[Pearl+ 18]

Department selection of applicants in university admission
Applicants who prefers a philosophy department are talented in history, and those who prefers a computer science are talented in math

Association

assigned to their preferred dep.

↓

talents are unbalanced between deps.

↓

Applicants' talents might influence the outcomes

Intervention

assigned randomly

↓

talents are balanced between deps.

↓

The effects of applicants' talents are balanced and can be ignored

Counterfactual

Notations of counterfactual situations

In the factual world, a person of the minority group having ability $\mathbf{x}$ is declined.
The probability that the person would be declined,
if the person were the majority group?

The outcome in the counterfactual world where the people belonged to the majority group

People of the minority group in the factual world

$$\Pr[Y_{S=1} = 0 \mid \mathbf{X} = \mathbf{x}, S = 0]$$

Admission is declined in the counterfactual world

People whose ability is $\mathbf{x}$ in the factual world

- S is a sensitive variable, 0 ➡ minority, 1 ➡ majority
- Y is an outcome, 0 ➡ declined, 1 ➡ admitted
- $\mathbf{X}$ are non-sensitive variables, indicating personal ability

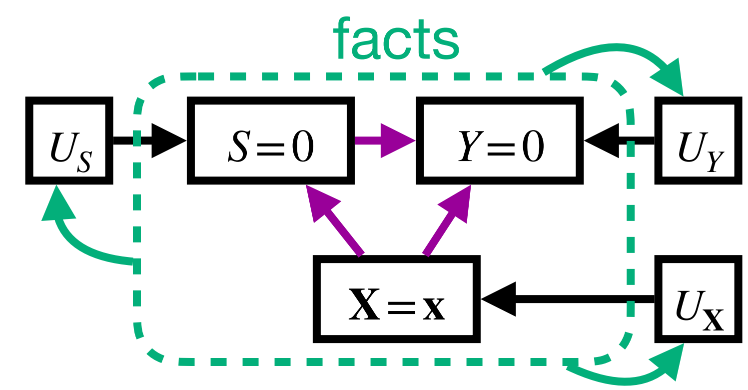
⁎ As a short hand for $Y_{X=0}$, Y_{x_0} or Y_0 is used, if it's apparent from contexts

Counterfactual: Computation Steps

> The person whose personality is $\mathbf{X}=\mathbf{x}$ and group is $S=0$ is declined in admission, $Y=0$, and then, what if the person's group were $S=1$?

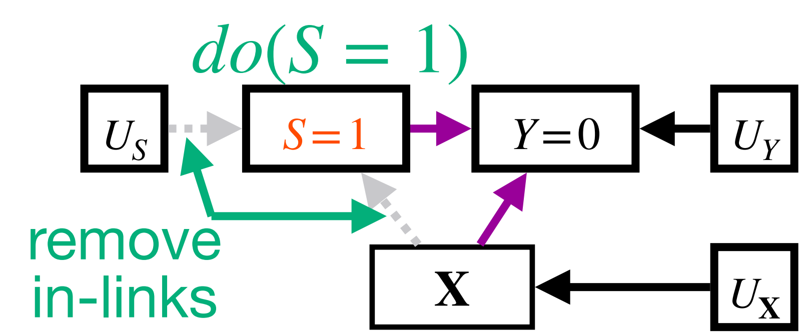
1. Abduction: predict exogenous situation that can cause the observed facts:

$$\Pr[\underbrace{\{U_S, U_{\mathbf{X}}, U_Y\}}_{\text{exogenous variables}} \mid \underbrace{\mathscr{F}=\{S=0, \mathbf{X}=\mathbf{x}, Y=0\}}_{\text{facts}}]$$

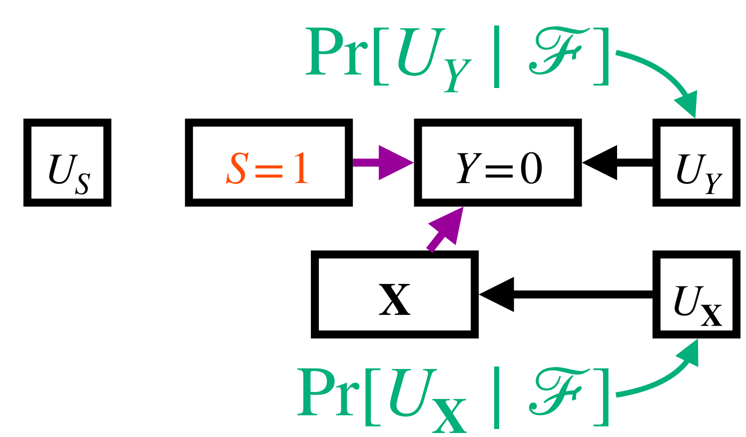


2. Action: intervention:

$$do(S=1)$$



3. Prediction: Predict the expectation of the outcome, Y, given the distribution in step 1 and using the model in step 2



Rubin's Conditions

[Pearl+ 18]

A counterfactual outcome, $Y_{S=s'}$, if a sensitive feature, S, was changed s from s', could be estimated by conditioning by non-sensitive features (confoundings), $\mathbf{X}$

Stable Unit Treatment Value Assumption (SUTVA)

- Each individual will have the same effect of treatment regardless of what treatment the other individuals receive

Consistency

- The same effect that is observed by experimental design will be observed in a real world

Ignorability

- The potential outcome, Y_s, is independent of the treatment actually received, S, given the values of a certain set of confoundings, $\mathbf{X}$

Ignorability

[Pearl+ 18]

Example of the violation of ignorability

An Employee's salary, Y, depends on their "Years of Education", S, and "Years of Experience", X

↓

If an employee had received longer years of education, how much their salary would be

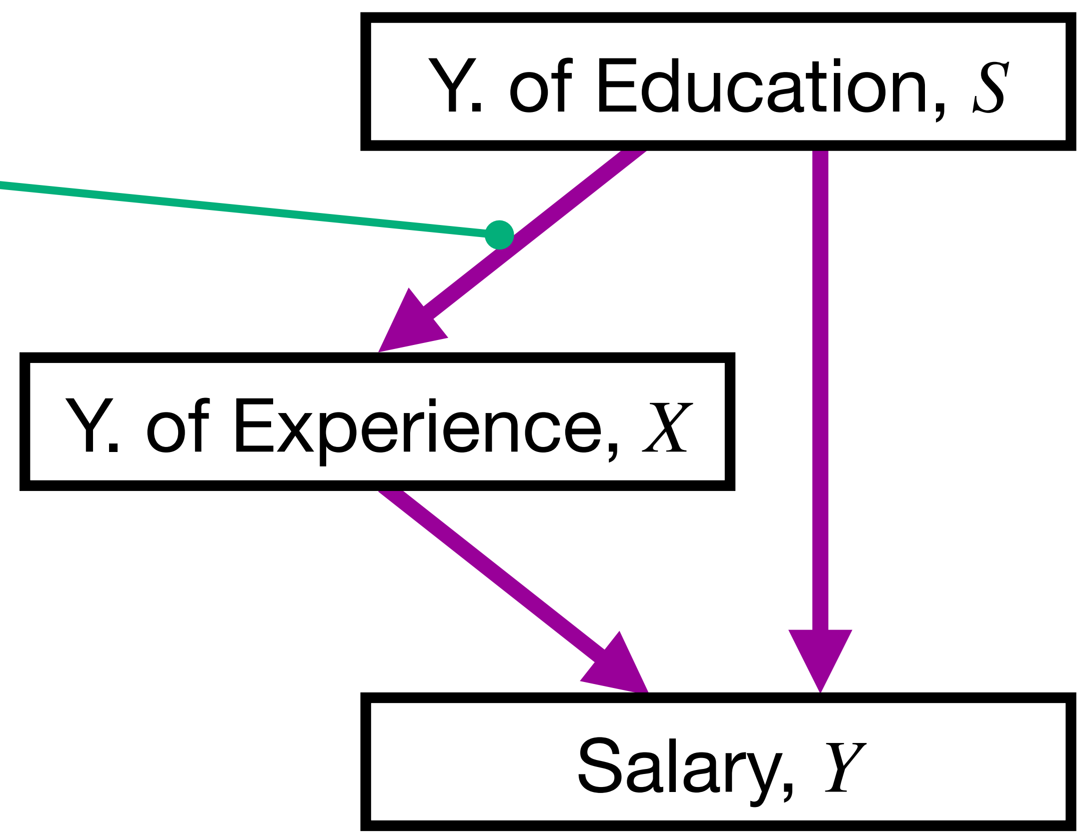
longer years of education
→ shorter years of experience

↓

Due to the conditioning of X, the influence of S on Y through X is distorted

↓

the violation of ignorability



Propensity Score

[Calders+ 13]

Propensity Score: probability to be a protected group given an explainable values, $e(S) = \Pr[\, S{=}0 \mid \mathbf{X}^{(e)} \,]$

propensity score can be used for eliminating the effects of explainable variables due to its **balancing property:** $S \perp\!\!\!\perp \mathbf{X}^{(e)} \mid e(S)$

If S is strongly ignorable given explainable variables, S is strongly ignorable given a propensity score:

$$Y \perp\!\!\!\perp S \mid \mathbf{X}^{(e)} \Rightarrow Y \perp\!\!\!\perp S \mid e(S)$$

The effect of explainable variables is removed by dividing a dataset into strata in which propensity scores are similar

training dataset

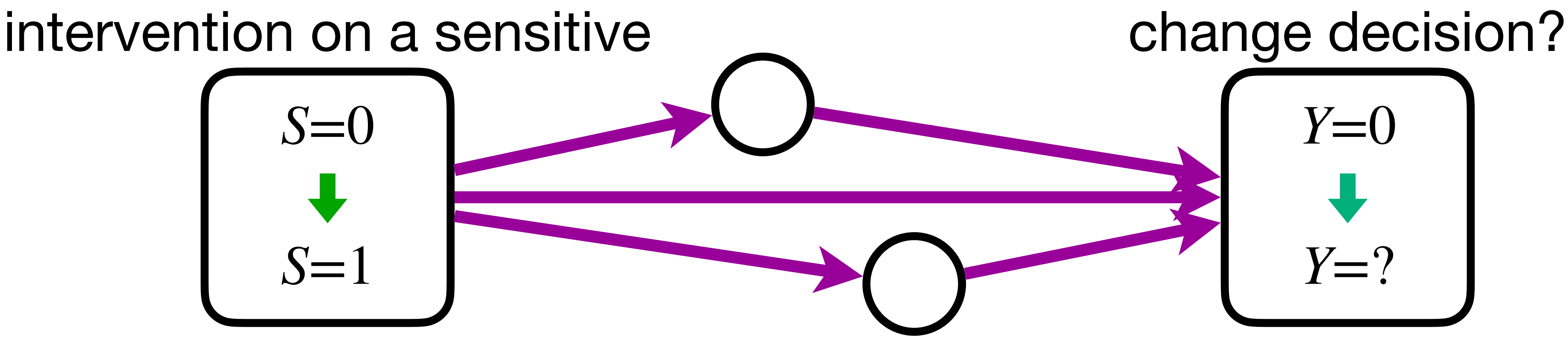
strata 1	$e(S) \in [0, 1/3)$
strata 2	$e(S) \in [1/3, 2/3)$
strata 3	$e(S) \in [2/3, 1]$

Counterfactual Fairness: Total Fairness Criteria

Total Effect and Total Variation

[Zhang+ 18]

total causal effect of changing a sensitive feature, S, on a target, Y



any direct and indirect causal paths are considered

Total Causal Effect

$$\mathrm{TE}(S=0, S=1) = \Pr[Y=1 \mid do(S=1)] - \Pr[Y=1 \mid do(S=0)]$$

Interventional, sensitive values are controlled

↕

Total Variation

$$\mathrm{TV}(S=0, S=1) = \Pr[Y=1 \mid S=1] - \Pr[Y=1 \mid S=0]$$

Observational, equal to TE if a sensitive variable has no in-links

Counterfactual Fairness

[Kusner+ 17]

Observations (Facts): If a sensitive feature is $S = s$ and the corresponding non-sensitive features, $\mathbf{X}_{S=s}$, are given, an outcome, $Y = y$, is observed.

$$\boxed{\mathbf{X}_{S=s}, S = s} \longrightarrow \boxed{Y = y}$$

Counterfactuals: Even if a sensitive feature was changed so that $S = s'$ while holding the non-sensitive features fixed, **it was fair if an outcome of a predictor is unchanged**

$$\boxed{\mathbf{X}_{S=s'}, S = s'} \dashrightarrow \boxed{Y_{S=s'} = y} \quad \textbf{fair}$$

$$\boxed{\mathbf{X}_{S=s'}, S = s'} \dashrightarrow \boxed{Y_{S=s'} = y'} \quad \textbf{unfair}$$

intervention: $S = s \rightarrow S = s'$

Counterfactual Fairness in Law

[Bareinboim+ 21, Pearl+ 18]

Jack Gross, Petitioner, v. FBL Financial Services, US Supreme Court, 2008

- To establish a disparate-treatment claim under this plain language, a plaintiff must prove that age was the but-for cause of the employer's adverse decision
- A plaintiff must prove by a preponderance of the evidence (which may be direct or circumstantial), that age was the but-for cause of the challenged employer decision

⬇

- The but-for cause: After occurring X and Y, if X was not occurred, whether or not Y would be occur?
- In a causal inference context, this is interpreted as **probability of necessity**, that is the probability of the counterfacutual, $Y_{X=0}$, is 0 given facts $X = 1$ and $Y = 1$.

$$\Pr[Y_{X=0} = 0 \mid X = 1, Y = 1]$$

Predictors Enhancing Counterfactual Fairness

[Kusner+ 17]

$\hat{Y}$ is counterfactually fair if it is a function of the non-descendants of S

Learn a predictor from non-descendants of S

$$\hat{Y} \sim f(\mathbf{U}, \mathbf{X}_{\nsucc S})$$

observables that are non-descendants of S

Algorithm for learning a counterfactually fair predictor

1. Data augmentation

   For each training data, $(s_i, \mathbf{x}_I, y_i)$, m data are randomly sampled from $\Pr[\mathbf{U} \mid S, \mathbf{X}]$, which was derived from the causal graph

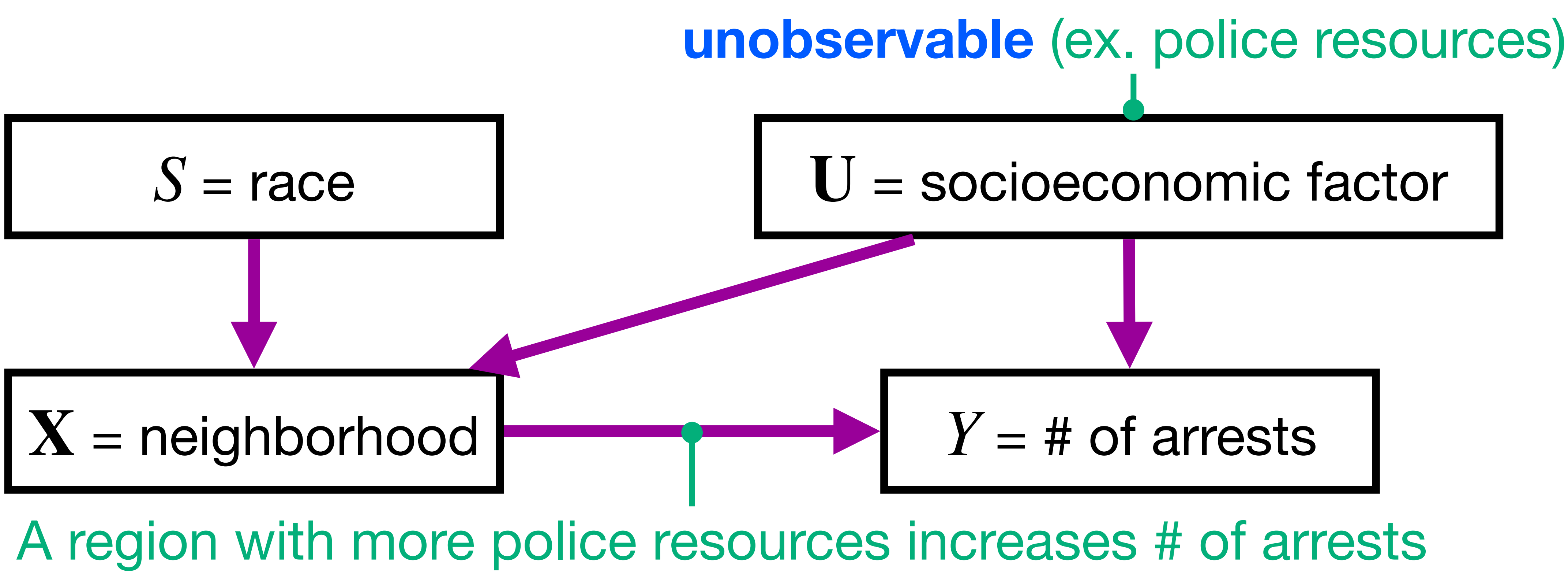
2. Generate a dataset,

   $\mathcal{D}' = \left\{(\mathbf{u}_{1,1}, \mathbf{x}_{\nsucc S,1}, y_1), \ldots, (\mathbf{u}_{1,m}, \mathbf{x}_{\nsucc S,1}, y_1), (\mathbf{u}_{2,1}, \mathbf{x}_{\nsucc S,2}, y_2), \ldots, (\mathbf{u}_{n,m}, \mathbf{x}_{\nsucc S,n}, y_n)\right\}$

3. Learn a predictor, $f(\mathbf{U}, \mathbf{X}_{\nsucc S})$, from $\mathcal{D}'$

Association-based Fairness & Counterfactual Fairness

[Kusner+ 17]



Y depends on $\mathbf{U}$ as well as on S

- **Association-based:** If training data were influenced by $\mathbf{U}$, in other words individuals had not equal opportunity, enhancing equalized odds cannot mitigate unfairness caused by $\mathbf{U}$
- **Counterfactual:** This approach can deal with such unfairness, because it predicts $\mathbf{U}$ and uses the predictions for mitigation

Association-based Fairness & Counterfactual Fairness

Legal Viewpoint

[Ishiguro+ 14, Bareinboim+ 21]

Association-based Fairness	Counterfactual Fairness
Hazelwood School District v. United States, 433 U.S. 299 (1977)	Jack Gross, Petitioner, v. FBL Financial Services, US Supreme Court, 2008
Gross Statistical Parity Outcomes should be equal between groups	**but-for cause** What if the sensitive information had been different?

Ethical Viewpoint

[Lippert-Rasmussen 06]

Association-based Fairness	Counterfactual Fairness

A harm-based account A baseline for determining whether the discriminatees have been made worse off

Ideal outcome	**Counterfactual**

Individual and Group Fairness in Counterfactual Fairness

[Kusner+ 17, Zhang+ 18]

Counterfactual fairness defined by the Kusner et al. is **individual**

The personality of the individual is represented by features, $\mathbf{X}$ and S

This definition targets individuals whose features are $\mathbf{X} = \mathbf{x}$ and $S = s$

$$\Pr[\underbrace{Y_{S=s}}_{= Y} = y \mid \mathbf{X} = \mathbf{x}, S = s] = \Pr[Y_{S=s'} = y \mid \mathbf{X} = \mathbf{x}, S = s]$$

This condition part represents a specific individual

Expectation over individuals so that $S=s$ is considered as criterion of **group fairness**

Effect of Treatment on the Treated

$$\mathrm{ETT}(S=0, S=1) = \Pr[Y_{S=1} = 1 \mid S=1] - \Pr[Y=1 \mid S=0]$$

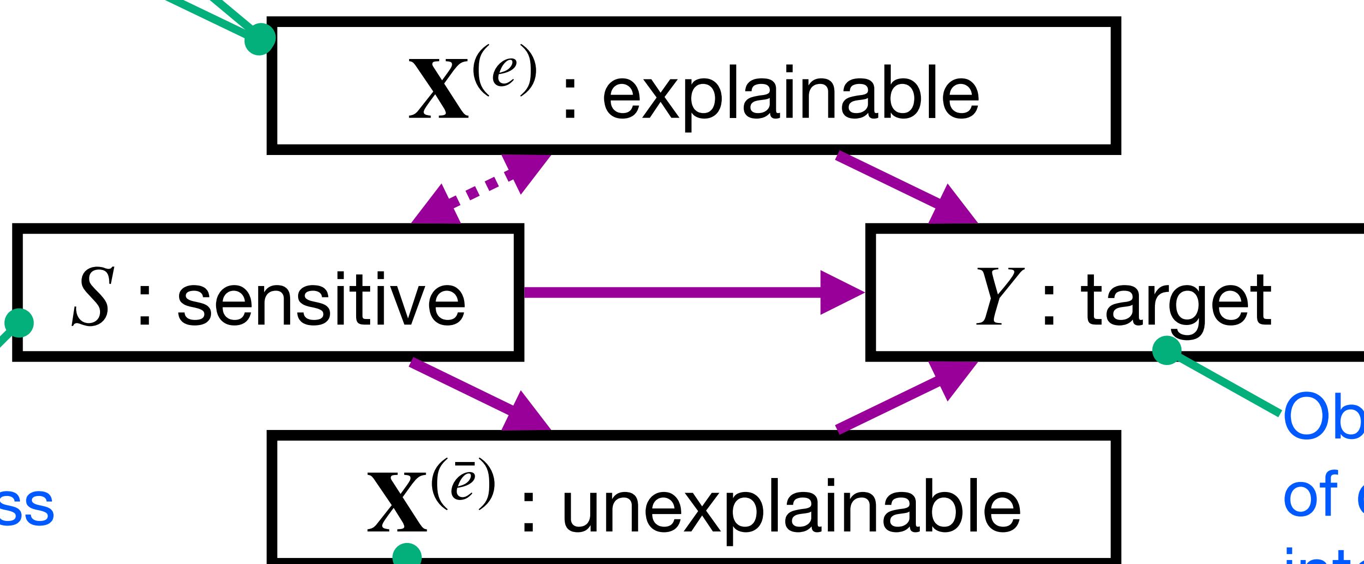
Counterfactual Fairness:
Path-Specific Fairness Criteria

Standard Fairness Model

[Zhang+ 18]

Standard Fairness Model : A basic model to deal with causal fairness based on path-specific analysis

Confounder, producing spurious correlation (ex. a department in the Berkley admission case), OR
Explainable mediator, legally allowed even if it passes the interventional effect (ex. genuine occupational requirement)



Intervened
to test fairness

Observing the change
of outcomes by
intervention

Unexplainable mediator, unfairly passed the interventional effect
(ex. red-lining effect)

＊ The model whose variables are all dependent, it is called *extended standard fairness model*

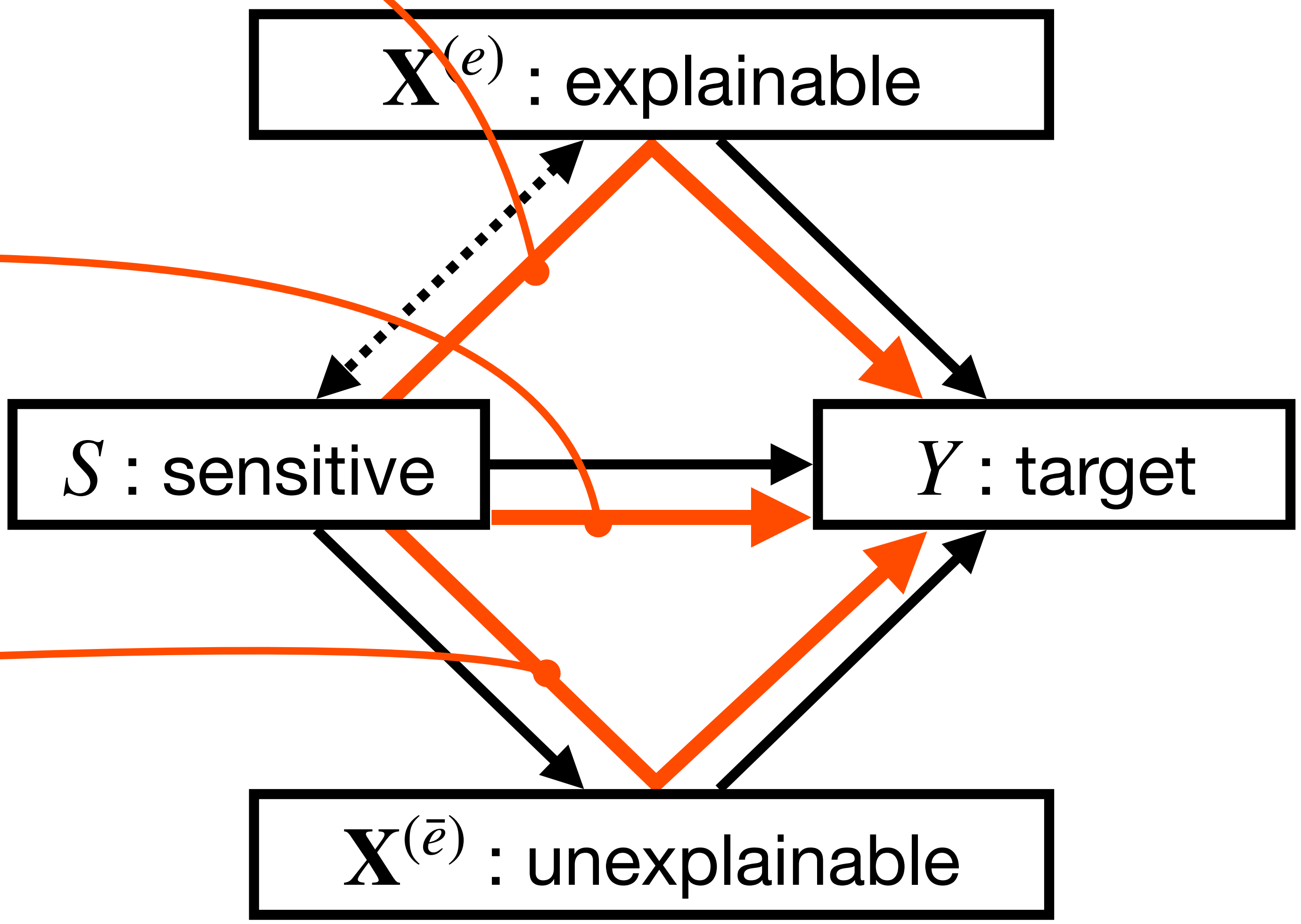
Path-Specific Fairness

Path-Specific Fairness depends on the causal path from a sensitive variable to a target variable

Spurious Effect ➡ Fair
The influences of sensitive information is spurious or legaly-allowed

Direct Effect ➡ Unfair
sensitive information directly influences the target

Indirect Effect ➡ Unfair
sensitive information indirectly influences the target through the unexplainable mediators



Economics-Based Fairness

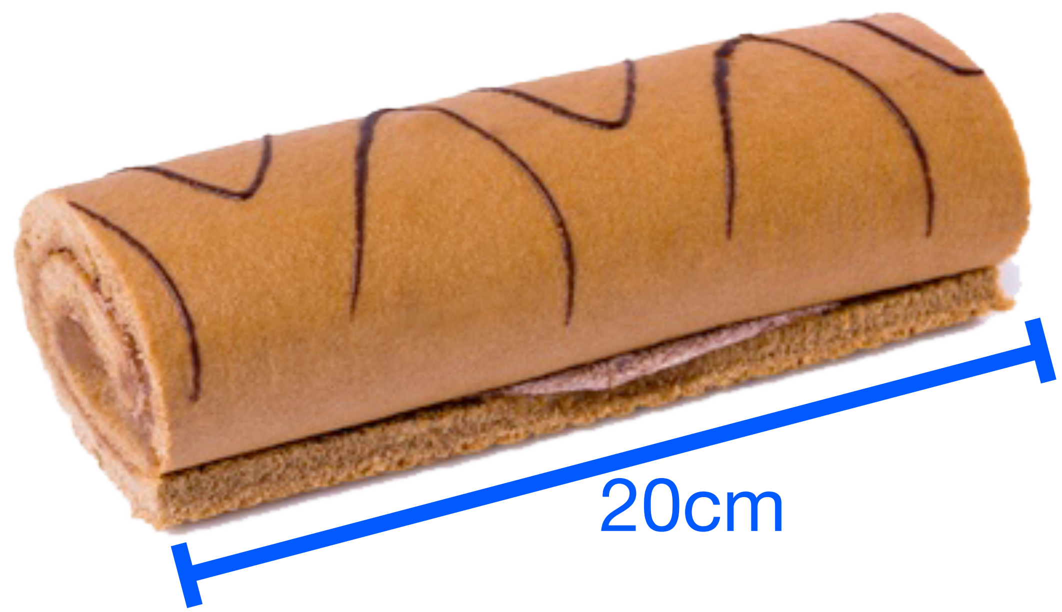
Game Theory: Fair Division

Alice and Bob want to divide this swiss-roll **FAIRLY**

Alice and Bob get half each based on **agreed common measure**

Game Theory: Fair Division

Alice and Bob want to divide this swiss-roll **FAIRLY**

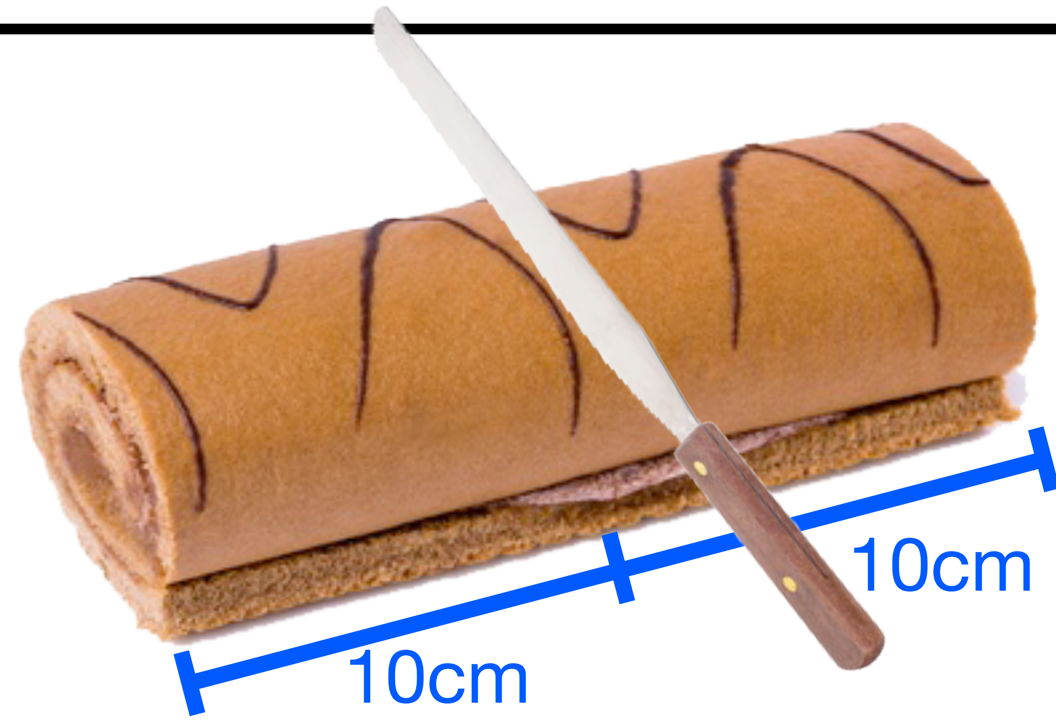


Total length of this swiss-roll is 20cm

Alice and Bob get half each based on **agreed common measure**

Game Theory: Fair Division

Alice and Bob want to divide this swiss-roll **FAIRLY**



divide the swiss-roll into 10cm each

Alice and Bob get half each based on **agreed common measure**

Game Theory: Fair Division

Unfortunately, Alice and Bob don't have a scale

envy-free division: Alice and Bob get a equal or larger piece based on **their own measure**

Game Theory: Fair Division

Unfortunately, Alice and Bob don't have a scale

Alice cuts the swiss-roll exactly in halves based on her own feeling

envy-free division: Alice and Bob get a equal or larger piece based on **their own measure**

Game Theory: Fair Division

Unfortunately, Alice and Bob don't have a scale

Bob picks a larger piece based on his own feeling

envy-free division: Alice and Bob get a equal or larger piece based on **their own measure**

Game Theory: Fair Division

- Every party i has one's own measure $m_i(P_j)$ for each piece P_j
- P_i is the piece selected by the party i, and P_j's are not selected

Fairness in a fair division context

- **Envy-Free Division:** Every party gets a equal or larger piece than other parties' pieces based on one's own measure

$$m_i(P_i) \geq m_i(P_j), \forall i, j$$

- **Proportional Division:** Every party gets an equal or larger piece than $1/n$ based on one's own measure; Envy-free division is proportional division

$$m_i(P_i) \geq 1/n, \forall i$$

- **Exact Division:** Every party gets a equal-sized piece

$$m_i(P_i) = 1/n, \forall i$$

Preferred Treatment

[Zafar+ 17]

Preferred Treatment: A fairness criterion inspired by the notion of envy-freeness. Each group receives more utilities from its own predictor than from any other groups' predictors

$$\mathrm{util}(\Theta_s) \geq \mathrm{util}(\Theta_{s'}),\ \forall s, s' \in \mathcal{S}$$

- As predictor, a linear classifier, $\Theta_s^\top \mathbf{x}$, is adopted
- As $\mathrm{util}(\Theta_s)$, the probability of receiving advantageous decision, $\mathbb{I}(\mathrm{sign}(\Theta_s^\top \mathbf{x}) = 1)$, and then it is convex-relaxed, $\max(\mathrm{sign}(0, \Theta_s^\top \mathbf{x}))$

A learning task while enforcing preferred treatment

$$\min_{\Theta_s} \sum_{(s,\mathbf{x},y)\in\mathcal{D}} \mathrm{loss}(\mathbf{x}, y; \Theta_s) + \lambda_s \mathrm{reg}(\Theta_s)$$

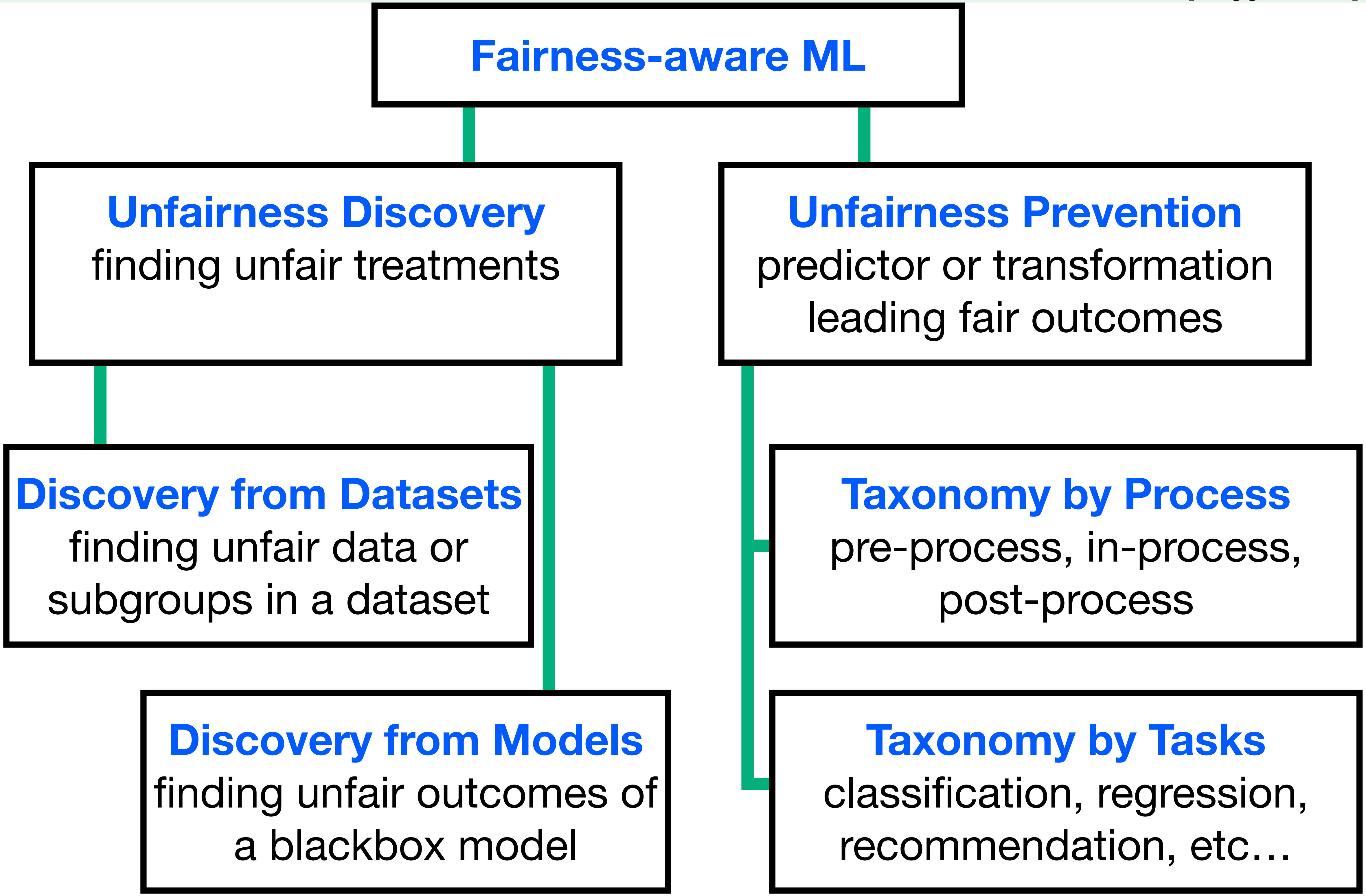
subject to $$\sum_{\mathbf{x}\in\mathcal{D}_s} \max(0, \Theta_s^\top \mathbf{x}) \geq \sum_{\mathbf{x}\in\mathcal{D}'_s} \max(0, \Theta_s^{\prime\top} \mathbf{x}),\ \forall s, s' \in \mathcal{S}$$

Part III
Fairness-Aware Machine Learning

Fairness-Aware Machine Learning: Overview

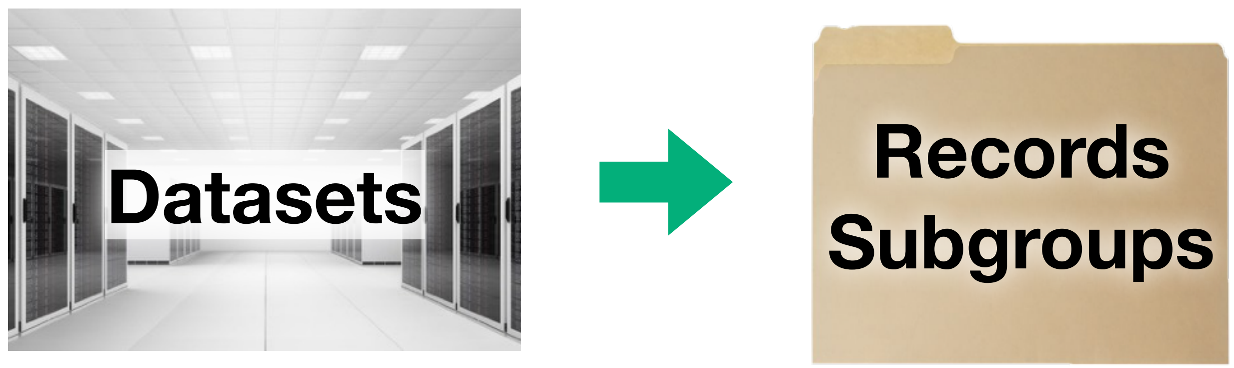
Tasks of Fairness-Aware ML

[Ruggieri+ 10]



Unfairness Discovery from Datasets

Unfairness Discovery from Datasets: Find personal records or subgroups that are unfairly treated from a given dataset



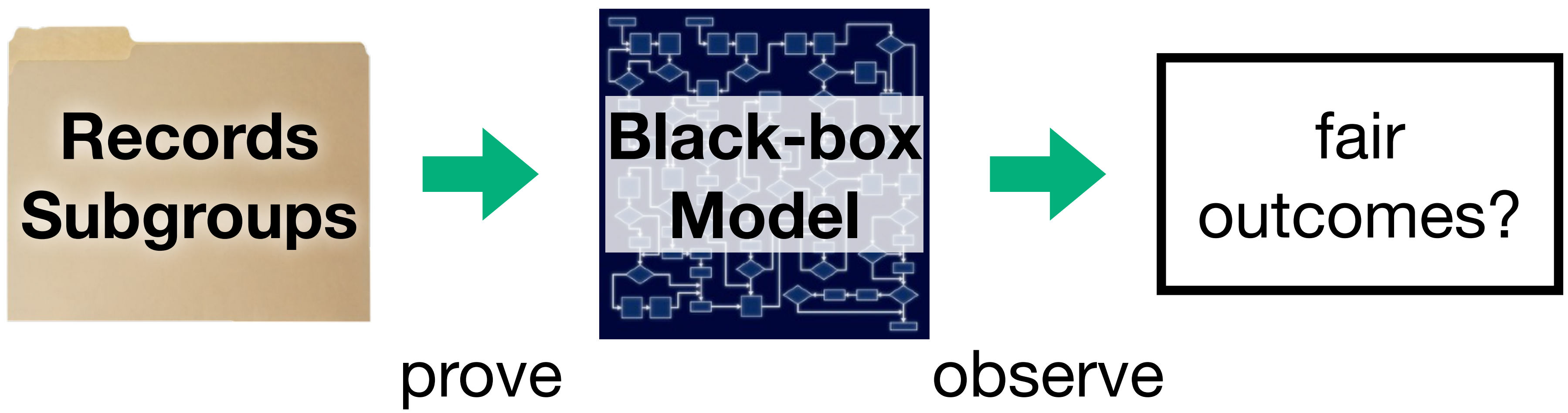
observe

Research Topics

- Definition of unfair records or subgroups in a dataset
- Efficiently searching patterns in the combinations of feature values
- How to deal with explainable variables
- Visualization of discovered records or subgroups

Unfairness Discovery from Models

Unfairness Discovery from Models: When observing outcomes from a specific black-box model for personal records or subgroups, checking fairness of the outcomes



Research Topics

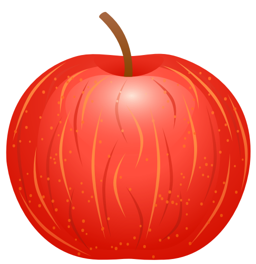
- Definition of unfair records or subgroups in a dataset
- Assumption on a set of black-box models
- How to generate records to test a black-box model

Supervised Learning

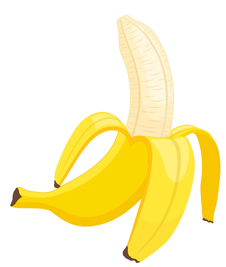
Learning

training data

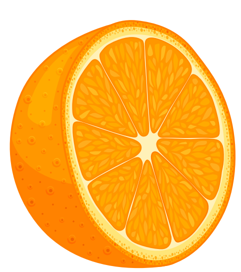
Is this an apple?



Yes



No



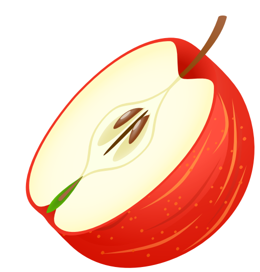
No

classification rule

A pattern between inputs and decisions (labels)

Inference

Is this an apple?



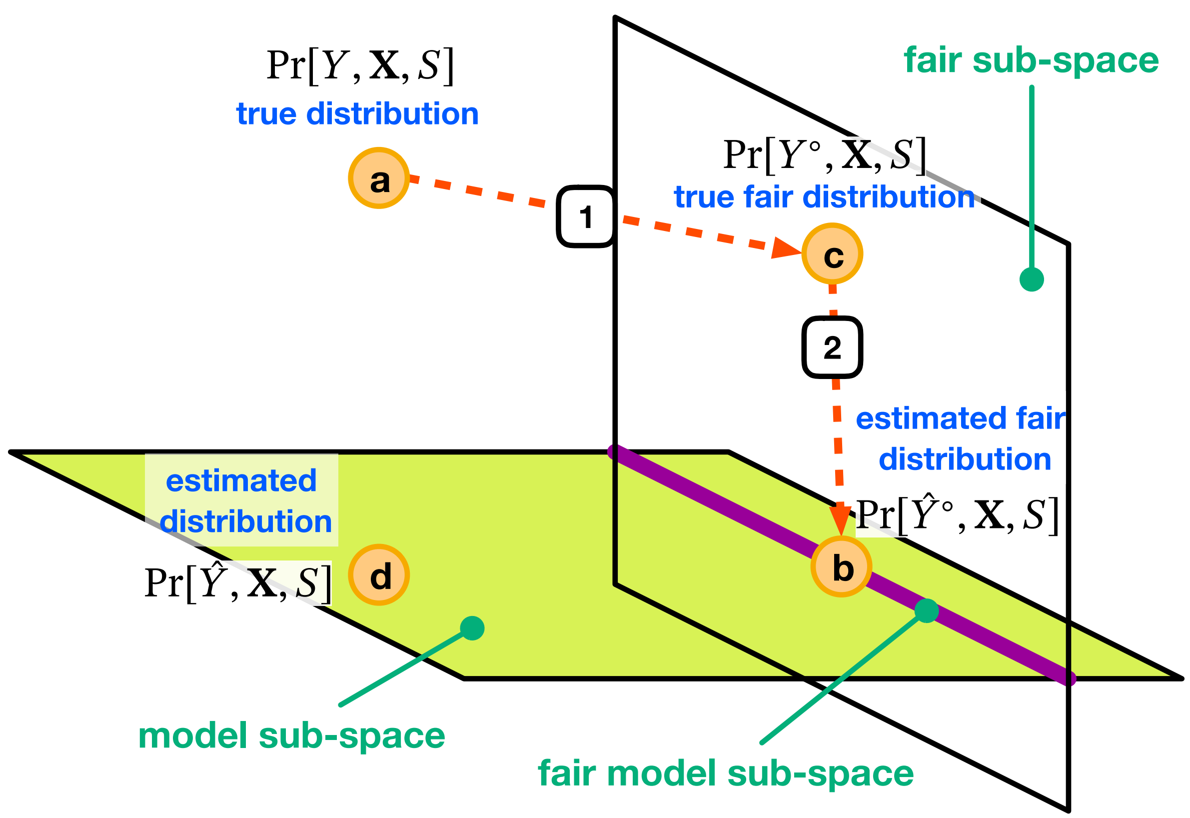
Yes

inference based on the learned rule

Unfairness Prevention: Pre-Process Approach

Pre-Process: potentially unfair data are transformed into fair data ①, and a standard classifier is applied ②

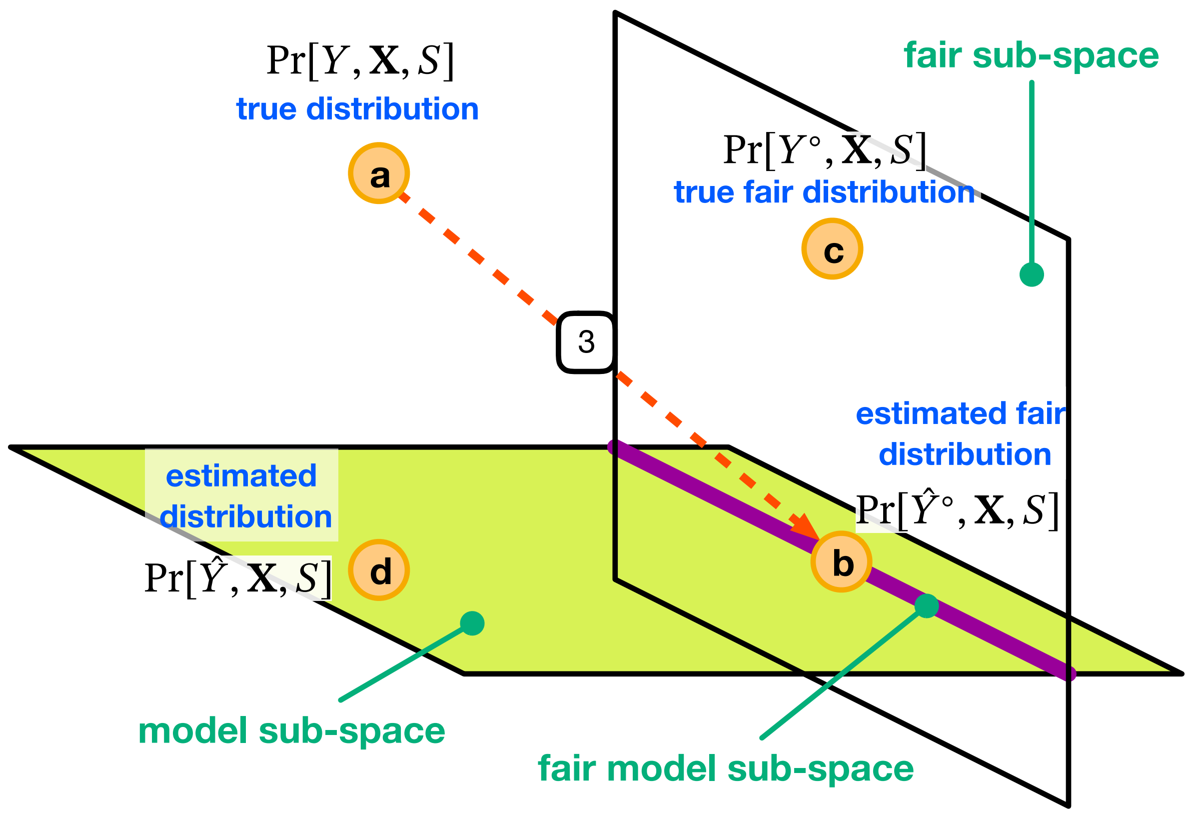
- Any classifier can be used in this approach
- the development of a mapping method might be difficult without making any assumption on a classifier



Unfairness Prevention: In-Process Approach

In-Process: a fair model is learned directly from a potentially unfair dataset ③

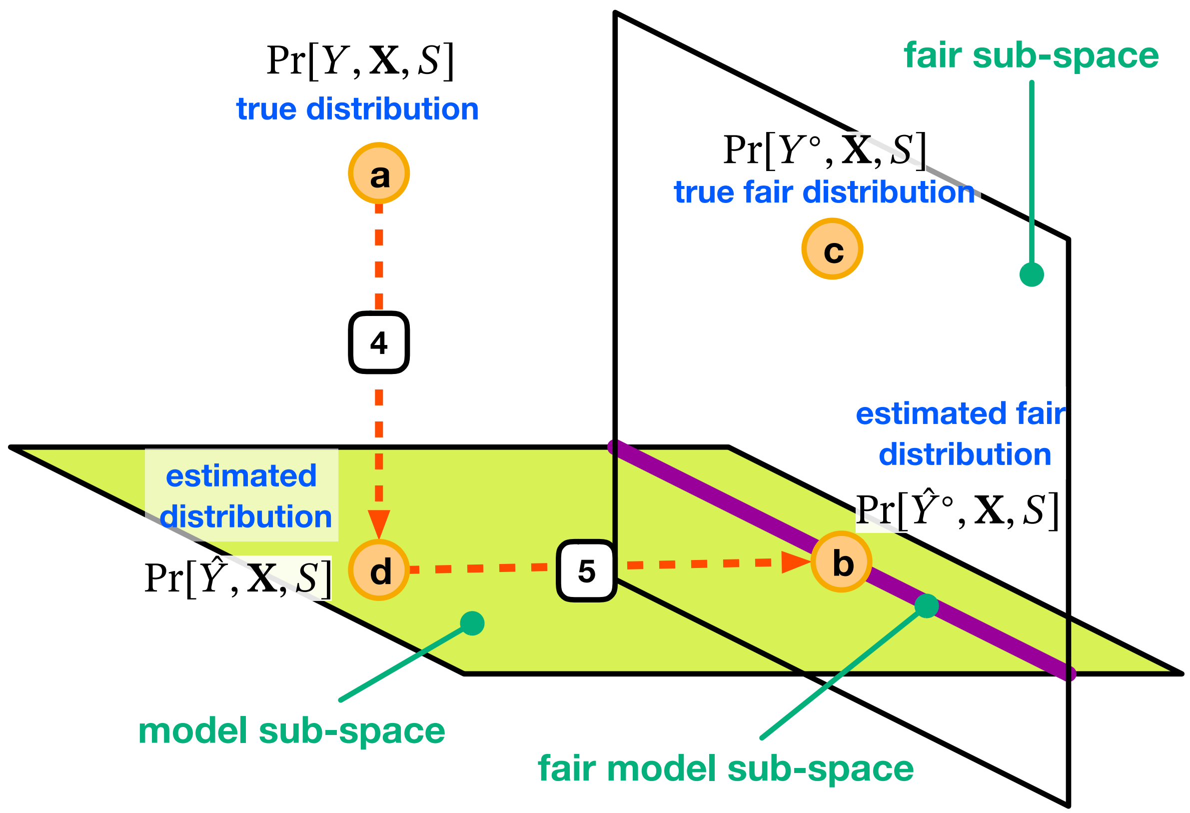
- This approach can potentially achieve better trade-offs, because classifiers can be designed more freely
- It is technically difficult to formalize an objective function, or to optimize the objective function.
- A fair classifier must be developed for each distinct type of classifier



Unfairness Prevention: Post-Process Approach

Post-Process: a standard classifier is first learned ④, and then the learned classifier is modified to satisfy a fairness constraint ⑤

- This approach adopts the rather restrictive assumption, **obliviousness** [Hardt+ 16], under which fair class labels are determined based only on labels of a standard classifier and a sensitive value
- This obliviousness assumption makes the development of a fairness-aware classifier easier



Unfairness Discovery: Discovery from Datasets in Association-Based Fairness Cases

Association Rule

[Agrawal+ 94]

Association Rule

$$\mathbf{X} \Rightarrow Y$$

$\mathbf{X}$: antecedent, Y : consequent

If $\mathbf{X}$ is satisfied, Y is also satisfied with a high probability

Ex:

$(\text{milk} \in Item) \wedge (\text{bread} \in Item) \Rightarrow (\text{egg} \in Item)$

$Item$: a set of simultaneously bought items

A customer who buys milk (= $\mathbf{X}$) and bread simultaneously will buy an egg (= Y) with high probability

Support

$$\text{support}(\mathbf{X}) = \frac{\text{\# of data that satisfy } \mathbf{X}}{\text{total \# of data}} = \Pr[\mathbf{X}]$$

Confidence

$$\text{conf}(\mathbf{X}, Y) = \frac{\text{\# of data that satisfy both } \mathbf{X} \text{ and } Y}{\text{\# of data that satisfy } \mathbf{X}} = \Pr[Y \mid \mathbf{X}]$$

Unfair Association Rules

[Pedreschi+ 08, Ruggieri+ 10]

Association rules extracted from a data set

(a) **city=NYC ⇒ class=bad** (conf=0.25)

0.25 of NY residents are denied their credit application

(b) **city=NYC ∧ race=African ⇒ class=bad** (conf=0.75)

0.75 of NY residents whose race is African are denied their credit application

extended lift (elift)

$$\text{elift} = \frac{\text{conf}(\ A \wedge B \Rightarrow C\ )}{\text{conf}(\ A \Rightarrow C\ )}$$

the ratio of the confidence of a rule with additional condition to the confidence of a base rule

α-protection: considered as unfair if there exists association rules whose elift is larger than α

ex: rule (b) isn't α-protected if a = 2, because elift = conf(b) / conf(a) = 3

Direct Discrimination: a target directly depends on a sensitive feature

Pr[loan=deny | city=NYC, race=African] ≫ Pr[loan=deny | city=NYC]

Unfair Association Rules

[Pedreschi+ 08, Ruggieri+ 10]

Indirect Discrimination: a target depends on a sensitive feature through a non-sensitive feature

A target 'loan' does not directly depends on a sensitive 'race'

$$\Pr[\text{ loan=deny} \mid \text{city=NYC, ZIP=10451 }] \gg \Pr[\text{ loan=deny} \mid \text{city=NYC }]$$

'loan=deny' and 'ZIP=10451' are highly co-occurred

$$\Pr[\text{ race=African} \mid \text{city=NYC, ZIP=10451 }] \sim \text{high}$$
$$\Pr[\text{ ZIP=10451} \mid \text{city=NYC, race=African }] \sim \text{high}$$

a target 'loan' in directly depends on a sensitive 'race'

✻ **Redescription:** the same set of objects are described by two different formulae or descriptions [Miettinen+ 16]

Ex. A literal 'city=NYC $\wedge$ ZIP=10451' is a redescription of 'city=NYC $\wedge$ race=African'

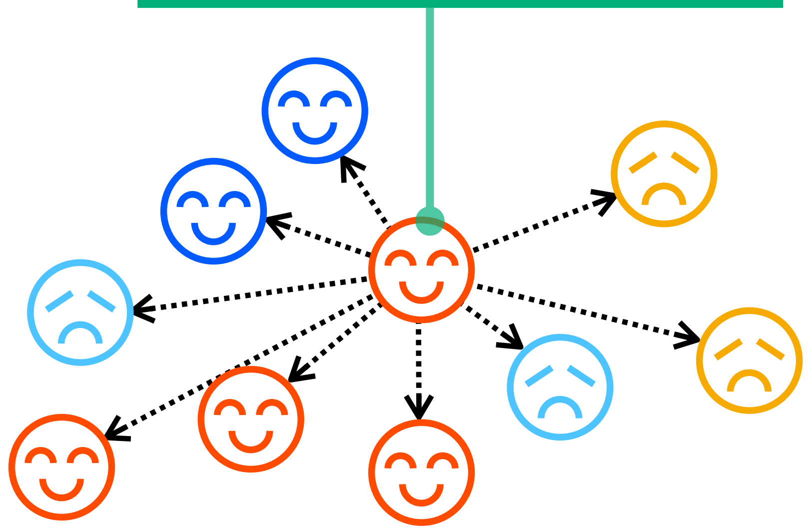
Situation Testing by *k*-NN

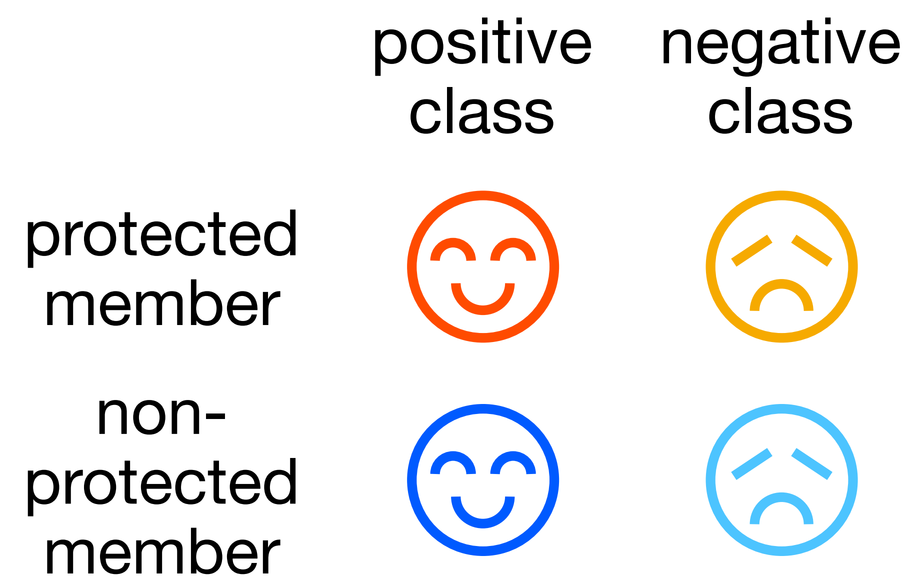
[Luong+ 11]

Situation Testing: When all the conditions are same other than a sensitive condition, people in a protected group are considered as unfairly treated if they received unfavorable decision

***k*-nearest neighbors**

of a protected member





- The statistics of decisions in *k*-nearest neighbors of data points in a protected group
- Condition of situation testing is

$$\Pr[\ Y \mid \mathbf{X}^{(e)}, S{=}0\ ] = \Pr[\ Y \mid \mathbf{X}^{(e)}, S{=}1\ ] \equiv Y \perp\!\!\!\perp S \mid \mathbf{X}^{(e)}$$

Unfairness Discovery: Discovery from Models

Gradient Feature Auditing

[Adler+ 16]

Direct Influence: comparing outputs when changing S

$(\mathbf{X}_i, S)$ original data → Blackbox Model → Y

$(\mathbf{X}_i, S')$ sensitive is perturbed → Blackbox Model → Y'

ignore the influence if a feature in $\mathbf{X}_i$ is correlated with S

Indirect Influence: the influence of features correlated with S

$(\mathbf{X}_i, S)$ original data → Blackbox Model → Y

$(\mathbf{X}'_i, S)$ non-sensitive is perturbed → Blackbox Model → Y'

measure the influence of features in $\mathbf{X}_i$ correlated with S

$\mathbf{X}i$ is perturbed so as not to predict S from the perturbed data $\mathbf{X}'_i$

Unfairness Prevention: Classification (pre-process)

Massaging

[Kamiran+ 12]

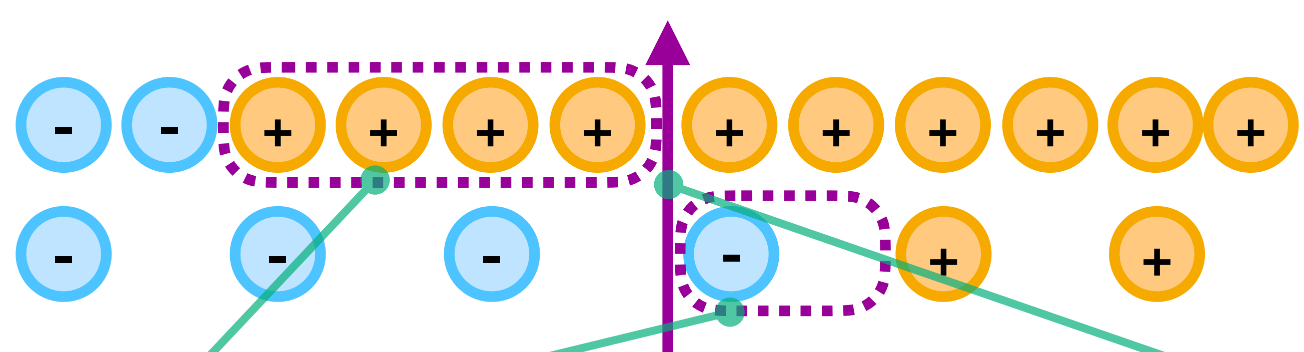
Massaging: Pre-process type method

- A standard classifier is once applied, and class labels are modified so as to be balanced between sensitive groups
- Finally, a standard classifier is trained from the modified dataset

1. A standard classifier is applied, and training data are sorted according to the degree to be a positive class for each sensitive group

non-protected S=1

protected S=0

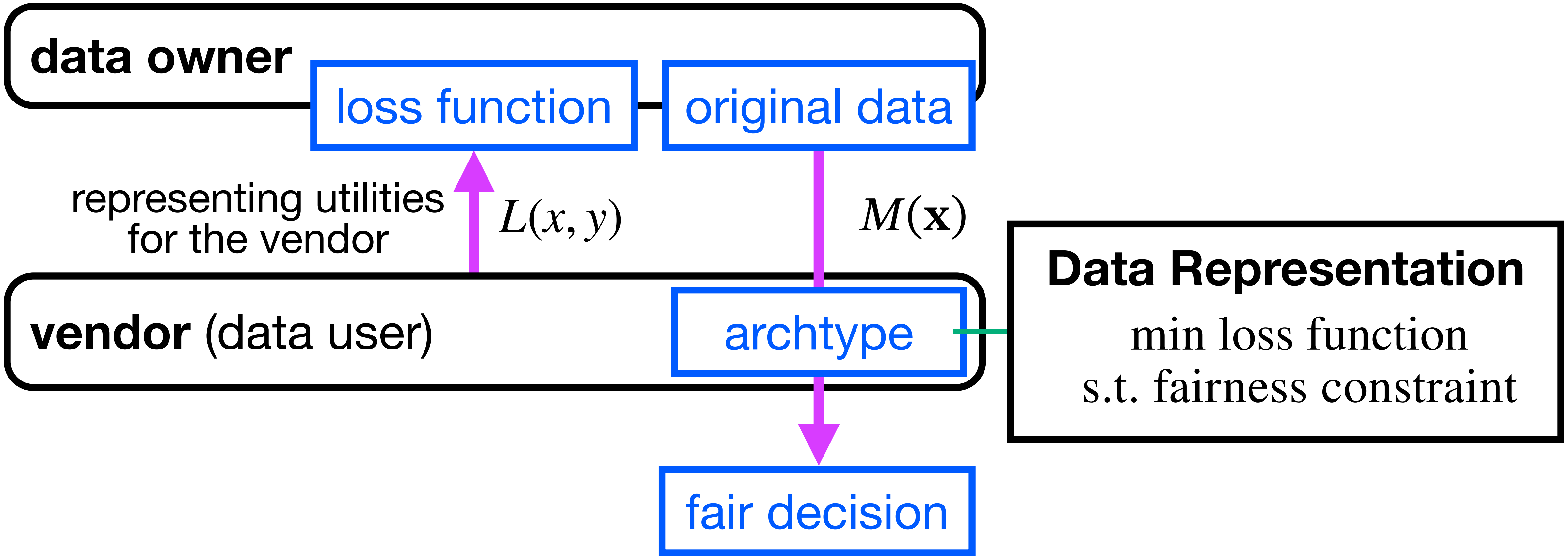


2. class labels are modified so that ratios of a positive class are balanced between sensitive groups

3. A final classifier is trained from the modified training dataset

Dwork's Method (Individual Fairness)

[Dwork+ 12]



Individual Fairness: Treat like cases alike

1. Map original data to archtypes so as to satisfy Lipschitz condition
2. Make prediction referring the mapped architypes

Lipschitz condition: similar data are mapped to similar archtypes

$$\underbrace{D}_{\text{distance between archtypes}}(M(\mathbf{x_1}), M(\mathbf{x_2})) \leq \underbrace{d}_{\text{distance between original data}}(\mathbf{x}_1, \mathbf{x}_2)$$

Dwork's Method (Statistical Parity)

[Dwork+ 12]

Statistical Parity: protected group, S, and non-protected group, $\bar{S}$, are equally treated

⬇

Mean of protected archtypes and mean of non-protected archtypes should be similar

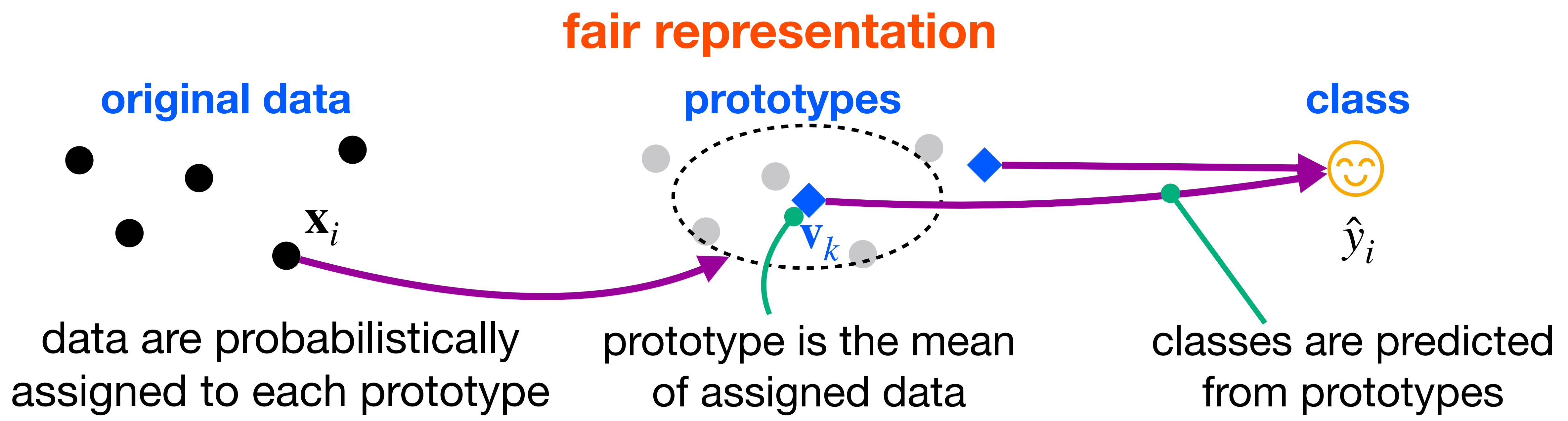
mean of protected archtypes — mean of non-protected archtypes

$$D(\mu_S, \mu_{\bar{S}}) \leq \epsilon$$

- If original distributions of both groups are similar, Lipschitz condition implies statistical parity
- If not, statistical parity and individual fairness cannot be satisfied simultaneously
  - To satisfy statistical parity, protected data are mapped to similar non-protected data while the mapping is as uniform as possible

Learning Fair Representations

[Zemel+ 13]



Requirements for Prototypes

- Probabilities assigned to each prototype is equal between groups

  $$L_z = \sum_k |M_k^{S=0} - M_k^{S=1}|$$

- Original data should be close to the data recovered from prototypes

  $$L_x = \sum_n (\mathbf{x}_n - \hat{\mathbf{x}}_n)^2$$

- Classes predicted from prototypes should close to original classes

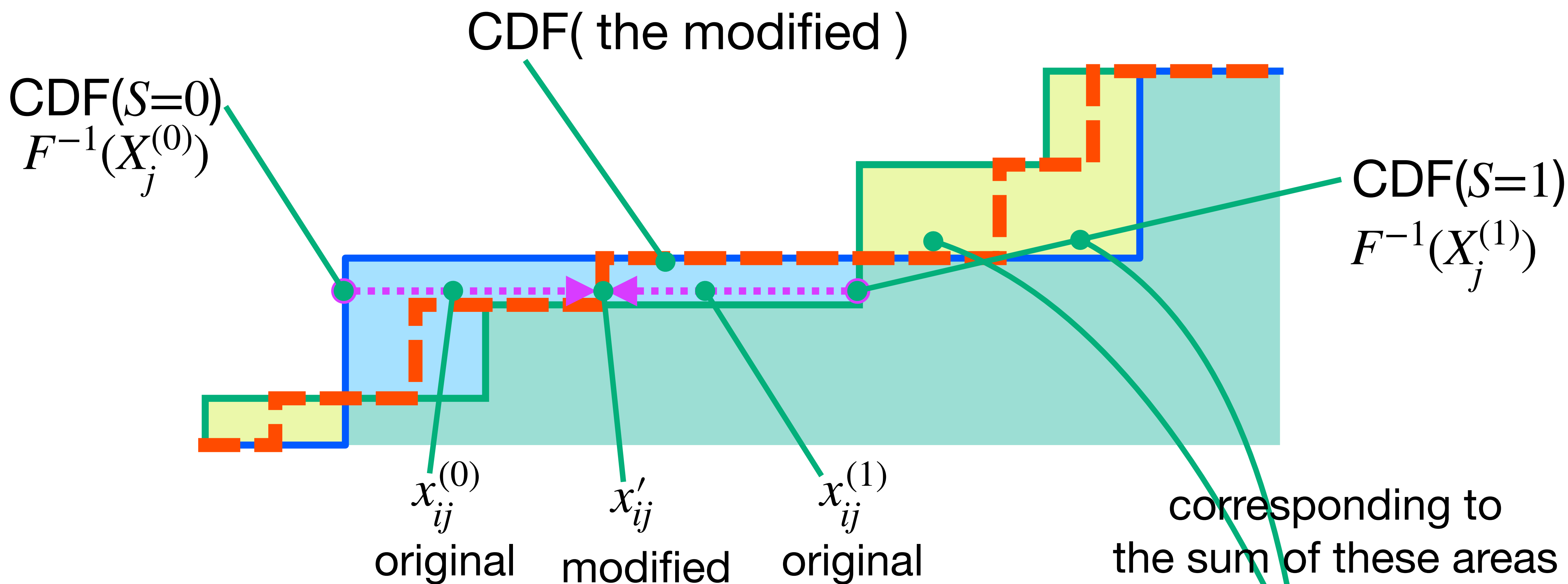
  $$L_y = \sum_n -y_n \log \hat{y}_n - (1 - y_n) \log(1 - \hat{y}_n)$$

Maps to prototypes are learned so as to maximize these requirements

Removing Disparate Impact

[Feldman+ 15]

Distributions of the j-th feature are matched between datasets whose sensitive feature is $S=0$ and $S=1$



Feature values are modified so as to minimize the sum of the L1 distances the modified cumulative distribution function (CDF) from original CDFs

Unfairness Prevention: Classification (in-process)

Prejudice Remover Regularizer

[Kamishima+ 12]

Prejudice Remover: a regularizer to impose a constraint of independence between a target and a sensitive feature, $Y \perp\!\!\!\perp S$

The objective function is composed of classification loss and fairness constraint terms

$$\boxed{-\sum_s \sum_{\mathcal{D}^{(s)}} \ln \Pr[y \mid \mathbf{x}; \Theta^{(s)}] + \frac{\lambda}{2} \sum_s \|\Theta^{(s)}\|} + \eta \, \boxed{\mathrm{I}(Y; S)}$$

fairness parameter to adjust a balance between accuracy and fairness

- A class distribution, $\Pr[Y \mid \mathbf{X}; \Theta^{(s)}]$, is modeled by a set of logistic regression models, each of which corresponds to $s \in \mathrm{Dom}(S)$

$$\Pr[Y = 1 \mid \mathbf{x}; \Theta^{(s)}] = \mathrm{sig}(\mathbf{w}^{(s)\top} \mathbf{x})$$

- As a prejudice remover regularizer, we adopt a mutual information between a target and a sensitive feature, $\mathrm{I}(Y; S)$

Fairness of Actual Class Labels

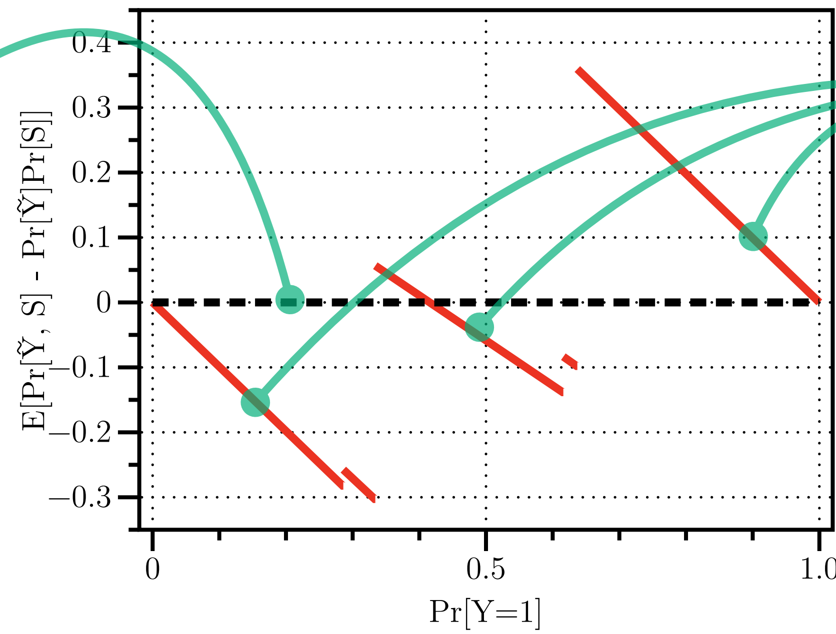
[kamishima+ 18]

Even if Y and S are independent, actual class labels may not satisfy a fairness constraint

deterministic decision rule: Class labels are generated not probabilistically, but deterministically by a decision rule

Difference: $\Pr[Y, S] - \Pr[Y]\Pr[S]$

Always Independent
Labels probabilistically generated according to
$\Pr[Y]\Pr[S]\Pr[\mathbf{X} \mid Y, S]$



Not Independent in general
Bayes optimal Labels are generated by a deterministic decision rule:

$$y^* \leftarrow \arg\max_y \Pr[y|\mathbf{x}, s]$$

model bias: Models doesn't contain true distribution to learn in general

Model-Based & Actual Independence

[Kamishima+ 18]

Model-based Independence: Class labels are assumed to be generated probabilistically

$$\hat{Y}^{\circ} \perp\!\!\!\perp S, \text{where } (\hat{Y}^{\circ}, S) \sim \Pr[\hat{Y}^{\circ}, S]$$

Actual Independence: Class labels are assumed to be deterministically generated by applying a decision rule

$$\tilde{Y}^{\circ} \perp\!\!\!\perp S, \text{where } (\tilde{Y}^{\circ}, S) \sim \Pr[\tilde{Y}^{\circ}, S] = \textstyle\sum_s \Pr[s] \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{D}_s} \Pr[\tilde{Y} \,|\, \mathbf{x}, s]$$

$$\begin{cases} \Pr[\hat{\tilde{y}} = 1 \,|\, \mathbf{x}, s] = 1 & \text{if } \hat{\tilde{y}} = \arg\max_y \Pr[\hat{y} \,|\, \mathbf{x}, s] \\ \Pr[\hat{\tilde{y}} = 0 \,|\, \mathbf{x}, s] = 0 & \text{otherwise} \end{cases}$$

⬇

satisfy actual independence instead of model-based independence

⬇

Fairness in class labels can be greatly improved

Correlation-based Fairness

[Zafar+ 2017]

Quantify unfairness by covariance, proportional to correlation

$$\mathrm{Cov}(Y,S) = \mathrm{E}[YS] - \mathrm{E}[Y]\,\mathrm{E}[S]$$
$$= \mathrm{E}[d_{\boldsymbol{\theta}}(\mathrm{x})(s-\bar{S})] - \mathrm{E}[d_{\boldsymbol{\theta}}(\mathrm{x})]\mathrm{E}[s-\bar{S}]$$
$$= \frac{1}{N}\sum_i^N (s_i - \bar{S})\, d_{\boldsymbol{\theta}}(\mathrm{x})$$

This constraint is convex, helpful for the easy optimization

- $d_{\boldsymbol{\theta}}(\mathbf{x})$ is a signed distance from x to the decision boundary, and is equal to $d_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x}$ in a linear model with a parameter $\boldsymbol{\theta}$

minimize accuracy loss under fairness constraints

$$\min_{\boldsymbol{\theta}} \mathrm{loss}(\boldsymbol{\theta}) \ \mathrm{s.t.}\ |\mathrm{Cov}(Y(\boldsymbol{\theta}), S)| \le \eta$$

accuracy loss
ex. negative log likelihood

trade-off parameter

maximize fairness under accuracy constraints

$$\min_{\boldsymbol{\theta}} \left|\mathrm{Cov}\big(Y(\boldsymbol{\theta})\big)\right| \ \mathrm{s.t.}\ \mathrm{loss}(\boldsymbol{\theta}) \le (1+\eta)\,\mathrm{loss}(\boldsymbol{\theta}^*)$$

optimal loss
without fairness constrains

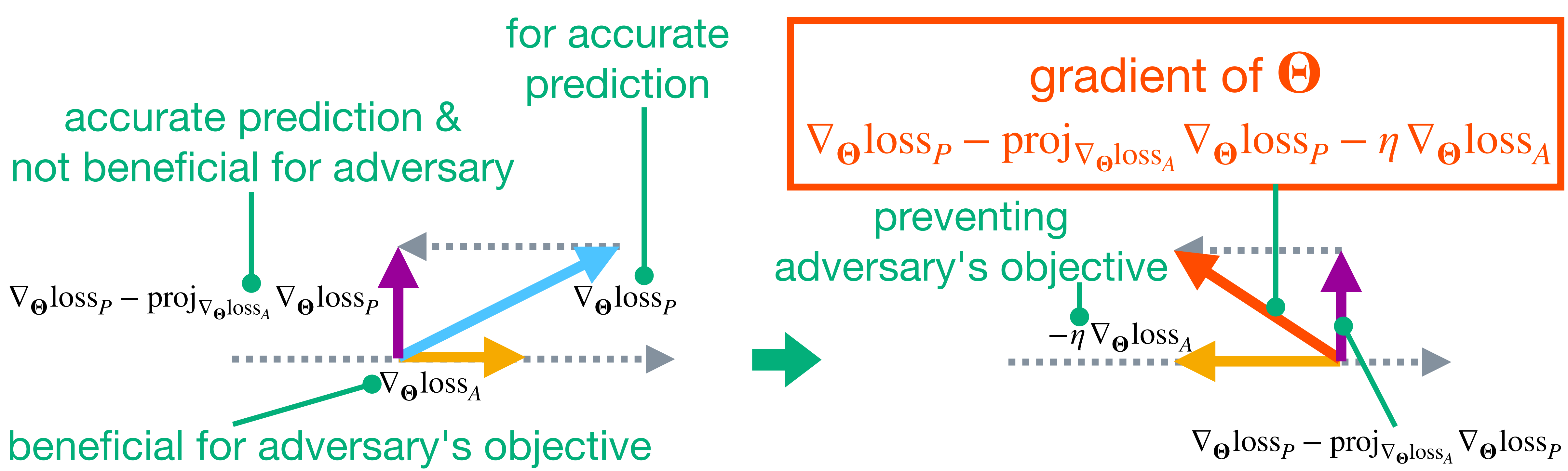
Adversarial Learning

[Zhang+ 18]

gradient-based learner for fairness-aware prediction

$\mathbf{X}$ → **predictor** $\hat{Y} = f_P(\mathbf{X}; \mathbf{\Theta})$ → $\hat{Y}$ → **adversary** $\hat{S} = f_A(\hat{Y}; \mathbf{\Phi})$ → $\hat{S}$

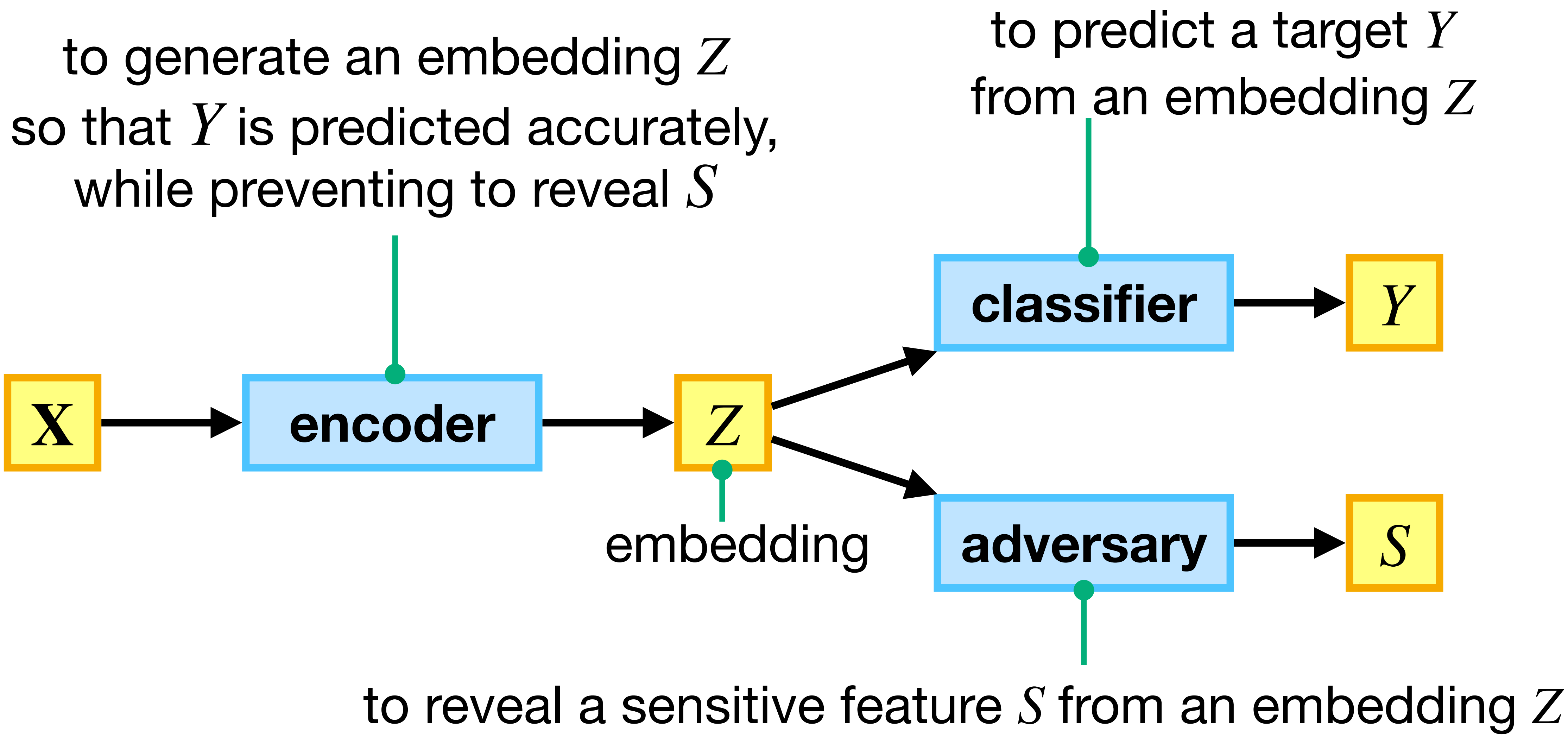
- Predictor minimizes $\mathrm{loss}_P(Y, \hat{Y}; \Theta)$, to predict outputs as accurately as possible while preventing adversary's objective
- Adversary minimizes $\mathrm{loss}_A(S, \hat{S}; W, V)$, to violate fairness condition



Adversarial Learning

[Adel+ 19, Edwards+ 16]

neural network for fairness-aware classification

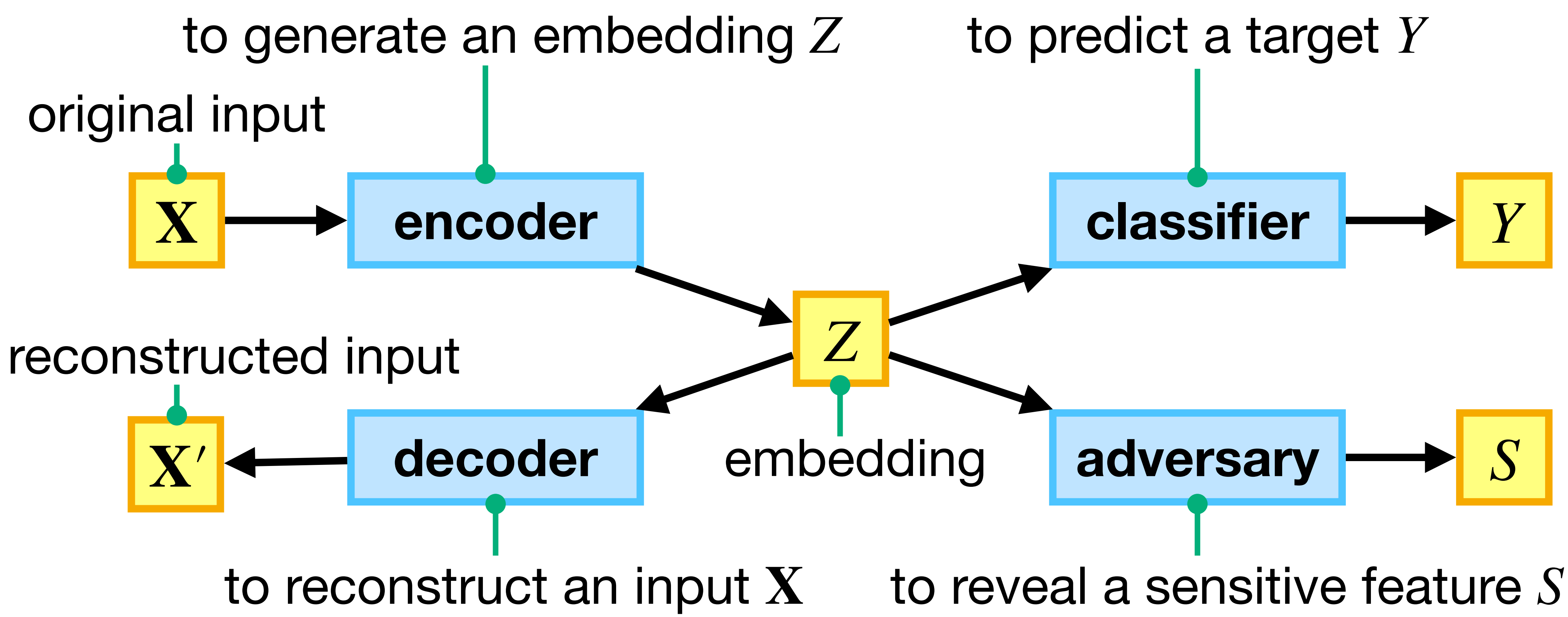


To prevent the prediction of S, gradients from a classifier is propagated straightforward, but those from an adversary is multiplied by -1 in backpropagation

Adversarial Learning

[Edwards+ 16, Madras+ 18]

NN for fair classification and generating fair representation



An embedding Z is generated so that

- minimize the reconstruction error between $\mathbf{X}$ and $\mathbf{X}'$
- minimize the prediction error of the classifier
- maximize the prediction error of the optimized adversary

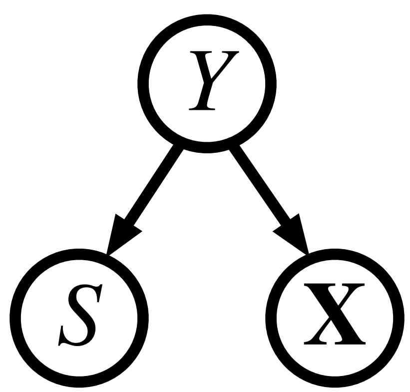
Unfairness Prevention: Classification (post-process)

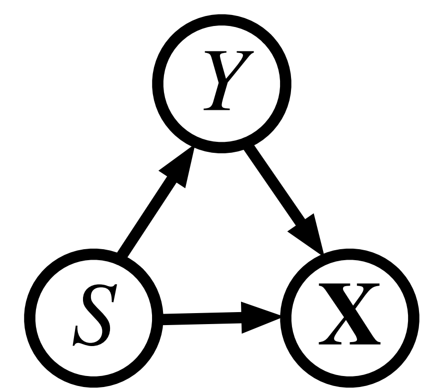
Calders-Verwer's 2-Naive-Bayes

[Calders+ 10]

Unfair decisions are modeled by introducing the dependence of $\mathbf{X}$ on S as well as on Y

Naive Bayes → **Calders-Verwer Two Naive Bayes (CV2NB)**





- S and $\mathbf{X}$ are conditionally independent given Y
- non-sensitive features in $\mathbf{X}$ are conditionally independent given Y and S

＊ It is as if two naive Bayes classifiers are learned depending on each value of the sensitive feature; that is why this method was named by the 2-naive-Bayes

Calders-Verwer's 2-Naive-Bayes

[Calders+ 10]

parameters are initialized by the corresponding sample distributions

$$\Pr[\hat{Y}, \mathbf{X}, S] = \Pr[\hat{Y}, S] \prod_i \Pr[X_i \mid \hat{Y}, S]$$

$\hat{\Pr}[Y, S]$ is modified so as to improve the fairness

estimated model: $\Pr[\hat{Y}, S]$ $\xrightarrow{\text{fairize}}$ fair estimated model: $\Pr[\hat{Y}^{\circ}, S]$

keep the updated marginal distribution close to the $\Pr[\hat{Y}]$

while $\Pr[Y{=}1 \mid S{=}1] - \Pr[Y{=}1 \mid S{=}0] > 0$
 if *# of data classified as "1" < # of "1" samples in original data* **then**
  increase $\Pr[Y{=}1, S{=}0]$, decrease $\Pr[Y{=}0, S{=}0]$
 else
  increase $\Pr[Y{=}0, S{=}1]$, decrease $\Pr[Y{=}1, S{=}1]$
 reclassify samples using updated model $\Pr[Y, S]$

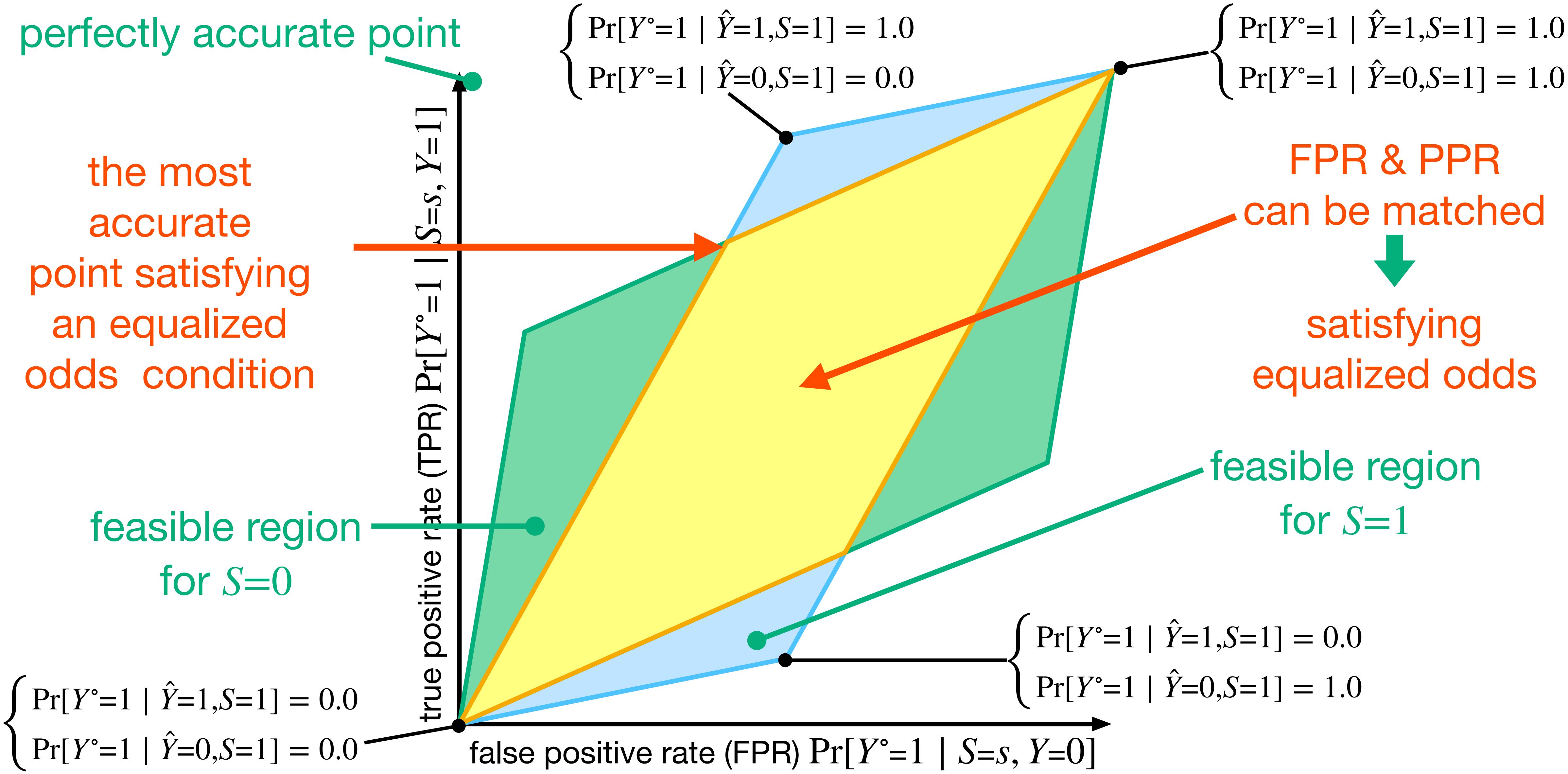
update the joint distribution so that its fairness is enhanced

Hardt's Method

[Hardt+ 16]

Given unfair predicted class, $\hat{Y}$, and a sensitive feature, S, a fair class, Y°, is predicted **maximizing accuracy** under an **equalized odds** condition

✻ True class, Y, cannot be used by this predictor



Unfairness Prevention: Recommendation

Recommender System

[Konstan+ 03]

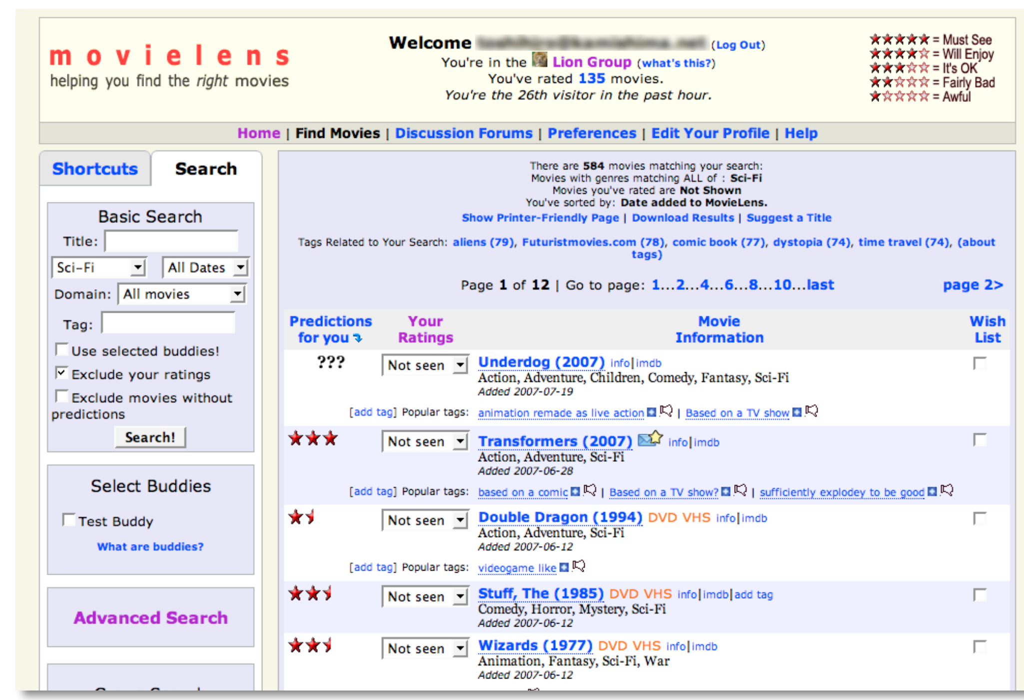
Recommenders: Tools to help identify worthwhile stuff

[Herlocker+ 04, Gunawardana+ 09]

Find Good Items

Ranking items according to users' preference, to help for finding at least one target item

Predicting Ratings



Presenting items with predicted ratings for a user, to help for exploring items

＊ Screen-shots are acquired from Amazon.co.jp and Movielens.org on 2007-07-26

Collaborative Filtering

[Resnick+ 94]

Collaborative filtering is a major approach for predicting users' preference in a word-of-mouth manner

recommending items liked by those who having similar preferences

＊ There are other approaches: content-based filtering or knowledge-based filtering

Adherence to Laws and Regulations

[Sweeney 13]

A recommendation service must be managed while adhering to laws and regulations

suspicious placement in keyword-matching advertisements

Advertisements indicating arrest records were more frequently displayed for names that are more popular among individuals of African descent than those of European descent

↓

Socially discriminative treatments must be avoided

sensitive feature = users' demographic information

↓

Legally or socially sensitive information can be excluded from the inference process of recommendation

Fair Treatment of Content Providers

System managers should fairly treat their content providers

Fair treatment in search engines

[Bloomberg]

The US FTC has investigated Google to determine whether the search engine ranks its own services higher than those of competitors

Fair treatment in recommendation

A hotel booking site should not abuse their position to recommend hotels of its group company

sensitive feature = a content provider of a candidate item

⬇

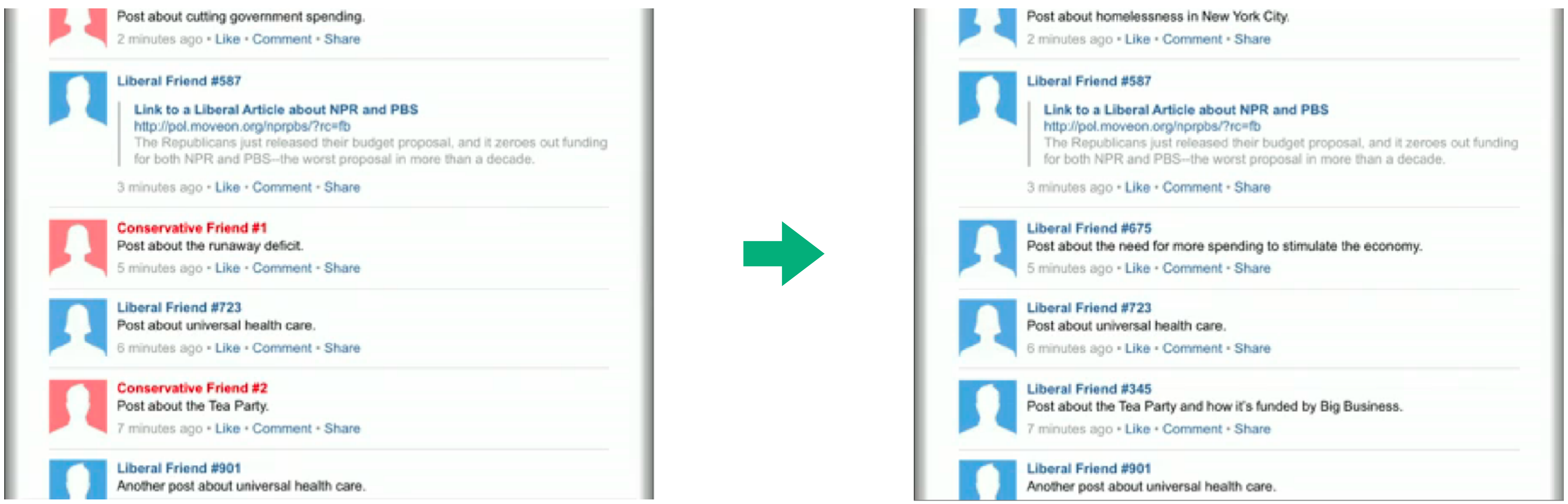
Information about who provides a candidate item can be ignored, and providers are treated fairly

Exclusion of Unwanted Information

[TED Talk by Eli Pariser, http://www.filterbubble.com/]

Information unwanted by a user is excluded from recommendation

Filter Bubble: To fit for Pariser's preference, conservative people are eliminated from his friend recommendation list in Facebook



sensitive feature = a political conviction of a friend candidate

Information about whether a candidate is conservative or progressive can be ignored in a recommendation process

RecSys 2011 Panel on Filter Bubble

[RecSys 2011 Panel on the Filter Bubble]

RecSys 2011 Panel on Filter Bubble

- Are there “filter bubbles?”
- To what degree is personalized filtering a problem?
- What should we as a community do to address the filter bubble issue?

http://acmrecsys.wordpress.com/2011/10/25/panel-on-the-filter-bubble/

Intrinsic trade-off

providing
a diversity of topics ⟷ focusing on
users’ interests

To select something is not to select other things

RecSys 2011 Panel on Filter Bubble

[RecSys 2011 Panel on the Filter Bubble]

Personalized filtering is a necessity

Personalized filtering is a very effective tool
to find interesting things from the flood of information

↓

recipes for alleviating
undesirable influence of personalized filtering

- capture the users' long-term interests
- consider preference of item portfolio, not individual items
- follow the changes of users' preference pattern
- give users to control perspective to see the world through other eyes

Multistakeholder in Recommendation

[Abdollahpouri+ 20]

Utilities of multiple stakeholders

example cases in job recommendation

Consumer: End-users who receive recommendation

- Applicants want to be highly evaluated their own experience or skills

Provider: Entities that supply recommended objects

- Employers should be exposed frequently

System: A platform who manages a recommender system

- Increasing job-matchings is beneficial for the system owner

These fairness constraints might conflict

- Equal exposure of employers

- Employers can be recommended less matched employers frequently
- Less matches reduces the profit of the system owner

Recommendation Independence

[Kamishima+ 12, Kamishima+18]

Recommendation Independence

statistical independence

between a recommendation outcome, R, and a sensitive feature, S

$$\Pr[R \mid S] = \Pr[R] \equiv R \perp\!\!\!\perp S$$

- No information about a sensitive feature influences the outcome
- The status of the sensitive feature is explicitly excluded from the inference of the recommendation outcome

Independence-Enhanced Recommendation

Preferred items are predicted

so as to satisfy a constraint of recommendation independence

Probabilistic Matrix Factorization

[Salakhutdinov+ 08, Koren 09]

Probabilistic Matrix Factorization Model

predict a preference rating of an item y rated by a user x

well-performed and widely used

Prediction Function

global bias

cross effect of users and items

$$\hat{r}(x, y) = \mu + b_x + c_y + \mathbf{p}_x \mathbf{q}_y{}^\top$$

user-dependent bias

item-dependent bias

Objective Function

regularization parameter

$$\underbrace{\sum\nolimits_{\mathscr{D}} \left(r_i - \hat{r}(x_i, y_i)\right)^2}_{\text{squared loss function}} + \lambda \|\boldsymbol{\Theta}\|$$

L_2 regularizer

For a given training dataset, model parameters are learned by minimizing the squared loss function with an L_2 regularizer

Independence Enhanced PMF

[Kamishima+ 12, Kamishima+ 13, Kamishima+ 18]

Prediction Function

a prediction function is selected according to a sensitive value

$$\hat{r}(x, y, s) = \mu^{(s)} + b_x^{(s)} + c_y^{(s)} + \mathbf{p}_x^{(s)} {\mathbf{q}_y^{(s)}}^\top$$

sensitive feature

Objective Function

independence parameter: control the balance between the independence and accuracy

$$\sum_D (r_i - \hat{r}(x_i, y_i))^2 - \eta \, \boxed{\mathrm{indep}(R, S)} + \lambda \, \|\mathbf{\Theta}\|^2$$

independence term: a regularizer to constrain independence

- The larger value indicates that ratings and sensitive values are more independent
- Matching means of predicted ratings for two sensitive values

Independence Terms

Mutual Information with Histogram Models
[Kamishima+ 12]

- computationally inefficient

Mean Matching
[Kamishima+ 13]

$$-\left(\operatorname{mean}\left(\mathcal{D}^{(0)}\right) - \operatorname{mean}\left(\mathcal{D}^{(1)}\right)\right)^2$$

- matching means of predicted ratings for distinct sensitive groups
- improved computational efficiency, but considering only means

Mutual Information with Normal Distributions
[Kamishima+ 18]

$$-\left(\mathrm{H}(R) - \sum_s \Pr[s]\,\mathrm{H}(R|s)\right)$$

Distribution Matching with Bhattacharyya Distance
[Kamishima+ 18]

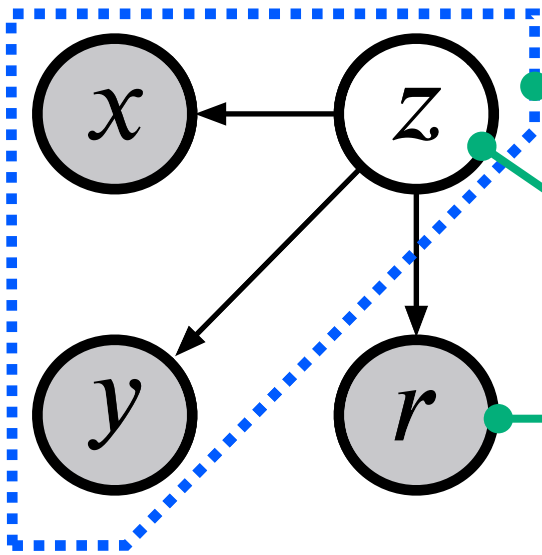
$$-\left(-\ln \int \sqrt{\Pr[r|S{=}0]\Pr[r|S{=}1]}\,dr\right)$$

- These two terms can take both means and variances into account, and are computationally efficient

Latent Class Model

[Hofmann 99]

Latent Class Model: A probabilistic model for collaborative filtering



basic topic model, pLSA

latent topic variable

extended so as to be able to deal with ratings r given by users x to items y

Model parameters can be learned by an EM algorithm

Prediction:

$$\hat{r}(x, y) = \mathrm{E}_{\Pr[r|x,y]}[\mathrm{level}(r)] = \sum_r \Pr[r|x, y]\,\mathrm{level}(r)$$

the r-th rating value

A rating value can be predicted by the expectation of ratings

Independence-Enhanced LCM

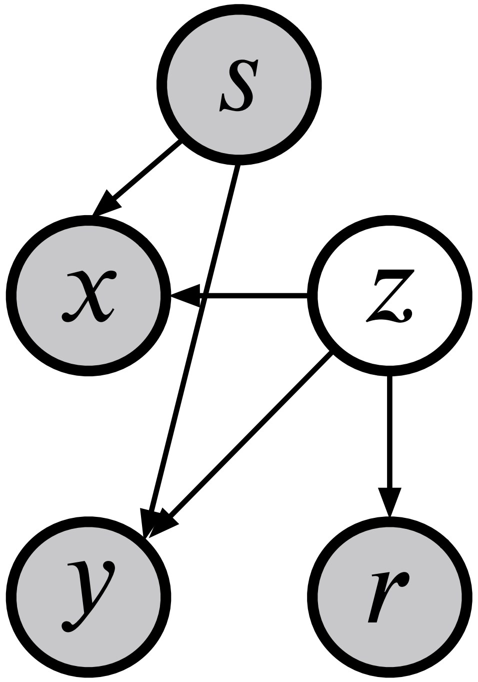
[Kamishima+ 16]

Independence-Enhancement by a Model-based Approach

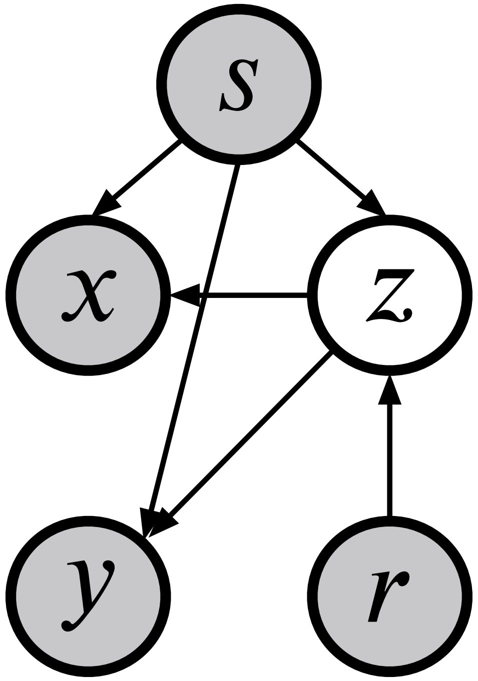
A sensitive variable is embedded into the original LCM

- A rating and a sensitive variable are mutually independent
- A user, an item, and a rating are conditionally independent given Z

Type 1 model



Type 2 model



Experimental results show that the performance of these two models are nearly equal

Unfairness Prevention: Ranking

Ranking

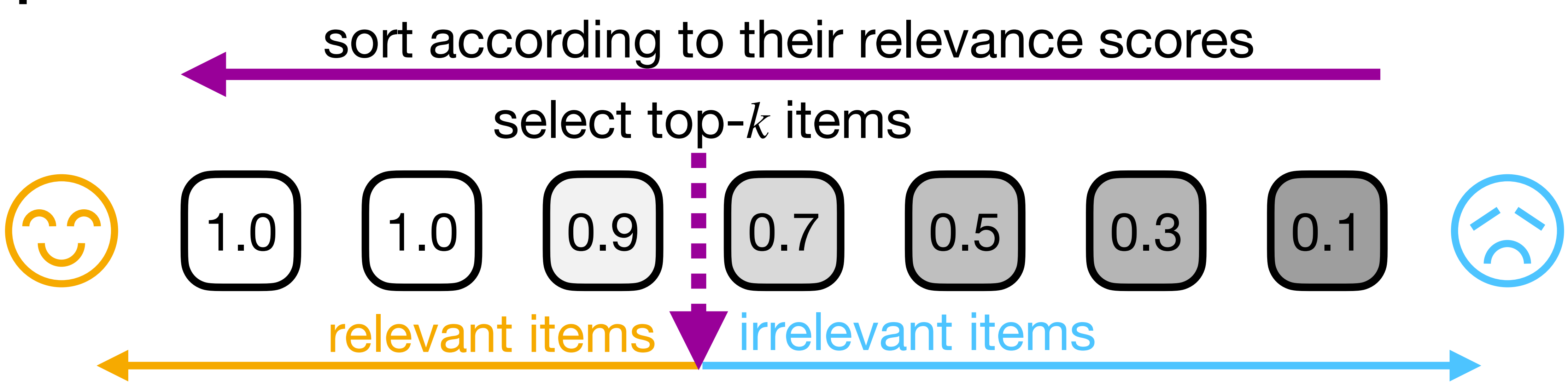
Ranking: select k items and rank them according to the relevance to users' need
A fundamental task for information retrieval and recommendation

Step 1: Calculate Relevance Score

Relevance Score: the degree of relevance to user's need

- **Information Retrieval:** relevance to the user's query
- **Recommendation:** user's preference to the item

Step 2: Rank Items

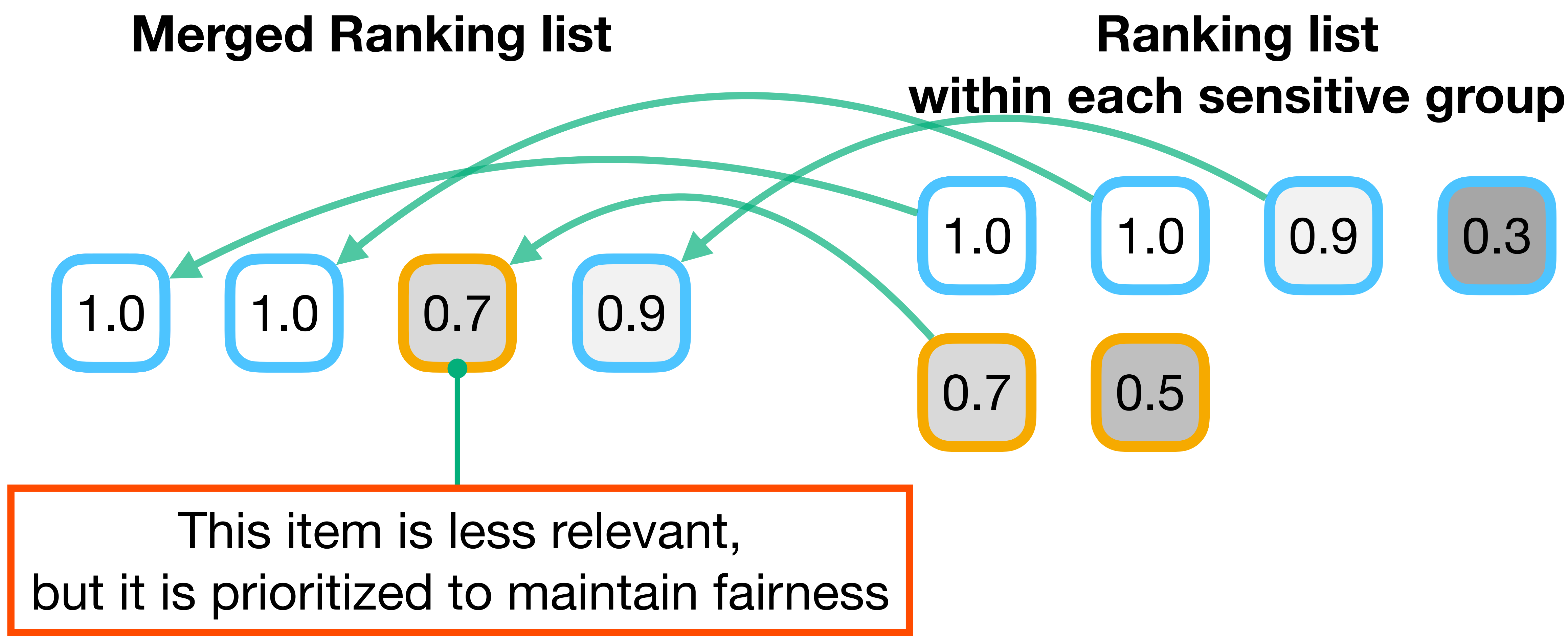


FA*IR

[Zehlike+ 17]

Fair Ranking: for each rank $i = 1, \ldots, k$, the ratio between two sensitive groups must not diverged from the ratio in the entire candidate set

1. Generate ranking lists for each sensitive group
2. Merge two ranking lists so as to the satisfy fair ranking condition



Singh's Method

[Singh+ 2018]

Singh's method is an in-process type ranking algorithm

Step 1: optimize $\mathbf{P}$ by solving the linear programming problem

prob. matrix of $P_{i,j}$

relevance of the document d_i to the query q

$$\min_{\mathbf{P}} \sum_{d_i \in \mathscr{D}} \sum_{j=1}^{N} u(d_i \mid q) P_{i,j} v_j$$

prob. of the document d_i ranked at the j-th position

values of the j-th position

ex. $v_j = 1/\log(j+1)$

subject to: $\mathbf{P}$ satisfies the constraints of probabilities
and **the following fairness constraint (statistical parity)**

the document d_i is a member of the sensitive group 0

the document d_i is a member of the sensitive group 1

$$\sum_{d_i \in \mathscr{D}} \sum_{j=1}^{N} \left(\frac{\mathrm{I}(d_i \in \mathscr{D}_0)}{|\mathscr{D}_0|} - \frac{\mathrm{I}(d_i \in \mathscr{D}_1)}{|\mathscr{D}_1|} \right) P_{i,j} v_j = 0$$

\# of documents in a sensitive group 0

\# of documents in a sensitive group 1

Step 2: By applying the Birkhoff-von Neumann decomposition to $\mathbf{P}$, get probability masses of corresponding rankings

Unfairness Prevention: Other Tasks

Bias in Word Embedding

[Bolukbasi+ 16]

Word Embedding: vector representing semantics of words
The differences of vectors reflect analogy of the corresponding words

he – she = king – queen

Occupational stereotype

Occupational words whose embeddings are the 10 nearest from the word embeddings of **she** or **he**

↓

Word embeddings are unfair due to the gender bias in the training corpus

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

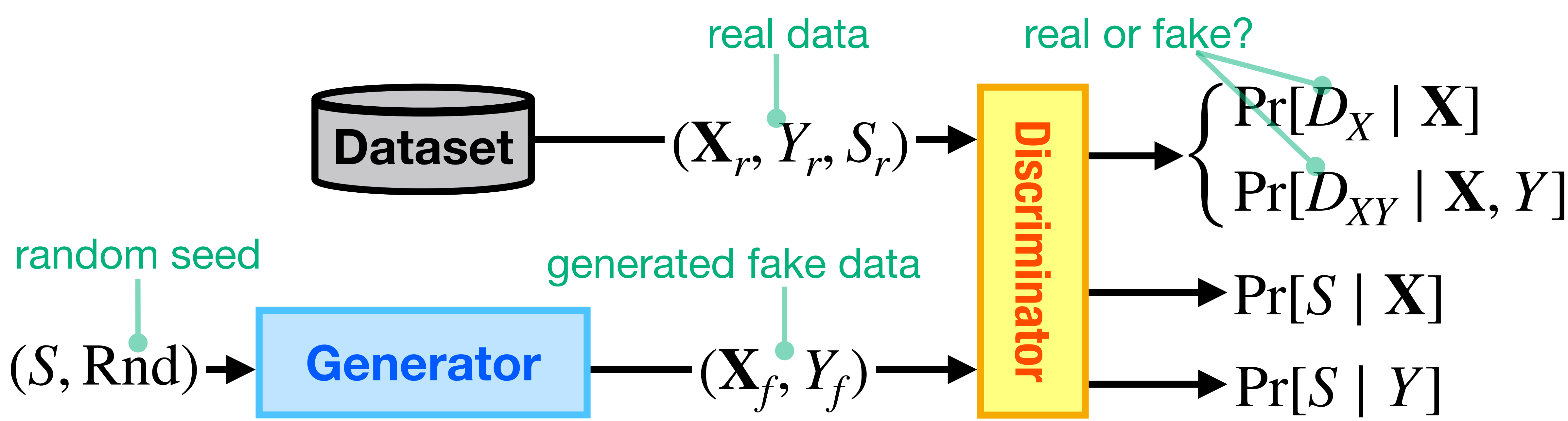
Debiasing Embeddings

- **neutralize:** non-gender words are uncorrelated to gender vector
- **equalize:** equal distance from occupational words to gender words

Fairness GAN: Fair Data Generator

[Sattigeri+ 19]

generative adversarial network for fair data generation



Likelihood to maximize

Discriminator	$\mathscr{L}(D_X \mid \mathbf{X}_{r,f}) + \mathscr{L}(D_{XY} \mid \mathbf{X}_{r,f}, Y_{r,f})$	$+\mathscr{L}(S \mid \mathbf{X}_r)$	$+\mathscr{L}(S \mid Y_r)$
Generator	$-\left(\mathscr{L}(D_X \mid \mathbf{X}_f) + \mathscr{L}(D_{XY} \mid \mathbf{X}_f, Y_f)\right)$	$+\mathscr{L}(S \mid \mathbf{X}_f)$	$-\mathscr{L}(S \mid Y_f)$

Discriminator predicts whether real or fake, but generator prevents it

↓

generating high-quality data

data conditioned on input sensitive value

Preventing to predict S from Y

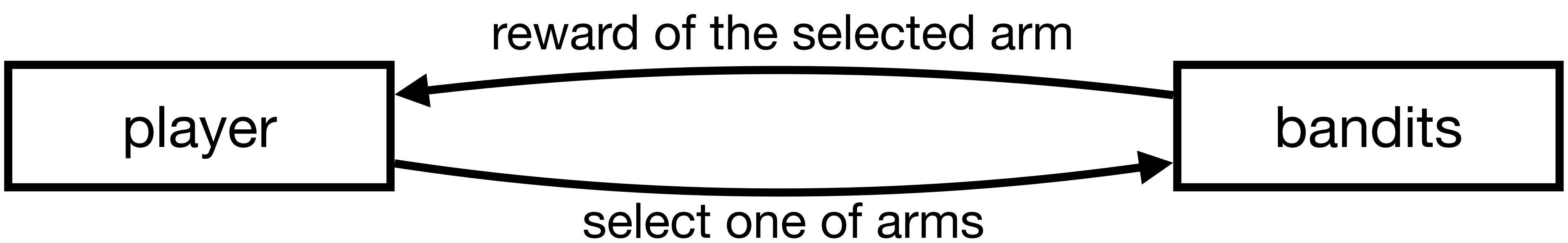
↓

Ensuring statistical parity

Fair Bandit

[Joseph+ 16]

Bandit problem: maximize the cumulative rewards of selected arms



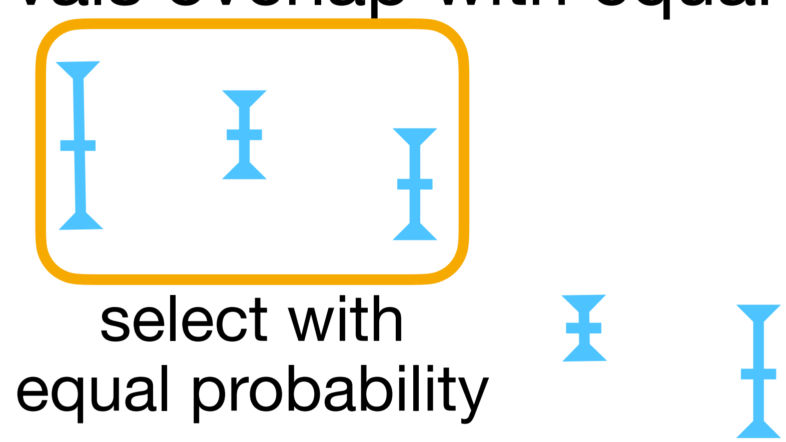
If an arm that is selected initially returns a high-reward by chance, the other arms can be less frequently selected

original UCB

always select the arm whose upper confidence bound is the maximum

fair UCB

select arms whose confidence intervals overlap with equal prob.



Non-Redundant Clustering

[Gondek+ 04]

non-redundant clustering: find clusters that are as independent from a given uninteresting partition as possible

clustering facial images

- A simple clustering method finds two clusters: one contains only faces, and the other contains faces with shoulders
- A data analyst considers this clustering is useless and uninteresting
- By ignoring this uninteresting information, more meaningful female- and male-like clusters could be obtained

The influence of uninteresting information can be ignored

Part IV
Other Topics

Mitigation of a Sample Selection Bias

Zadrony's Theorem

[Zadrozny 04]

$(\mathbf{x}, y)$ is sampled and observed if $z = 1$; it is not sampled if $z = 0$

- $(\mathbf{x}, y) \perp\!\!\!\perp z$: i.i.d. data ➡ **no problem**
- $\mathbf{x} \perp\!\!\!\perp z \mid y$: sampled depending on y ➡ **replacing prior** $\Pr[Y]$
- $y \perp\!\!\!\perp z \mid \mathbf{x}$: sampled depending on $\mathbf{x}$

➡ **assumption of this theory:** The values of $\mathbf{X}$ influence whether or not a datum is observed, but those of y do not

⬇

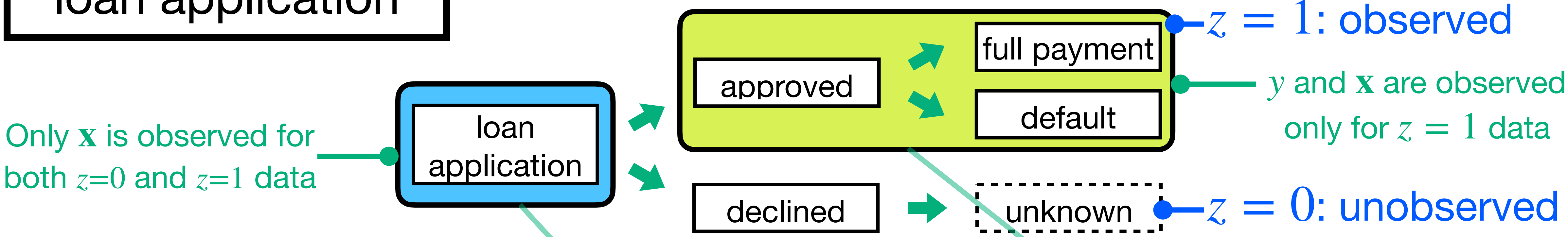
Under the assumption of $y \perp\!\!\!\perp z \mid \mathbf{x}$ and $\Pr[\mathbf{x}] > 0$, local learners are NOT affected by a sample selection bias, but global learners are

- **Local:** the output of learner depends only on $\Pr[y \mid \mathbf{x}]$
  - full Bayes, logistic regression, hard-margin SVM
- **Global:** the output of learner depends on both $\Pr[y \mid \mathbf{x}]$ and $\Pr[\mathbf{x}]$
  - naïve Bayes, decision trees, soft-margin SVM

Zadrony's Theorem

[Zadrozny 04]

loan application



- Under the assumption of $y \perp\!\!\!\perp z \mid \mathbf{x}$ and $\Pr[\mathbf{x}] > 0$, a likelihood function, $\Pr[y \mid \mathbf{x}]$, is unbiased, even if it is learned only from approved data
- $\Pr[z \mid \mathbf{x}]$ can be estimated from all applicants data

A leaner free from a sample selection bias can be trained by maximizing the weighted log-likelihood

$$\max_{\boldsymbol{\Theta}} \sum_{z=1\text{data}} \frac{\Pr[z = 1]}{\Pr[z = 1 \mid \mathbf{x}]} \log \Pr[y \mid \mathbf{x}; \boldsymbol{\Theta}]$$

Covariate Shift

[Shimodaira 00]

Predictors might be applied to data distributed differently from a distribution that it has been trained

⬇

Covariate Shift: $\Pr[\mathbf{X}, S]$ is different between test and training, but $\Pr[y \mid \mathbf{X}, S]$ is same

A distribution of S in training is $\Pr[S]$, and that in test is $\tilde{\Pr}[S]$

⬇

Given a joint distribution of $\mathbf{X}$ and S in training, $\Pr[\mathbf{X} \mid S]$ and , that in test is:

$$\tilde{\Pr}[\mathbf{X}, S] = \sum_s \frac{\tilde{\Pr}[S]}{\Pr[S]} \Pr[\mathbf{X} \mid S]$$

Under the covariate shift assumption, a predictor maximizing the weighted log-likelihood is unbiased

$$\max_{\mathbf{\Theta}} \sum_{\mathbf{x},s,y} \frac{\tilde{\Pr}[\mathbf{x}, s]}{\Pr[\mathbf{x}, s]} \log \Pr[y \mid \mathbf{x}; \mathbf{\Theta}]$$

Disclosure

Misuse of the COMPAS score

[Angwin+ 16]

Paul Zilly heard his COMPAS score, and his lawyer agreed to a plea deal of one year imprisonment, in a court in Barron County, Wisconsin

⬇

Judge James Babler had seen Zilly's high-risk score, and the judge overturned the deal and imposed two year imprisonment

⬇

In an appeal hearing, the developer of the COMPAS, Brennan, testified that the COMPAS was designed not for sentencing

⬇

Zilly’s sentence was reduced to 1.5 years imprisonment

In theory, the COMPAS is designed to determine which defendants are eligible for probation or treatment programs

⬇

Like this case, the disclose of the design intent of the model is important for correcting such a misuse

For a Proper Use of the ML

How to use ML techniques as a tool properly

Quality Control as in Other Industrial Products

- **Design:** datasets, algorithms
- **Test:** performance test, explainable ML
- **Maintenance:** monitoring, model updation

Given a fairness criterion,
an algorithm meats to the criterion can be built

↓

Disclosing which criterion the algorithm is designed to satisfy, and why the criterion is proper for the target task

✻ In a case of the COMPAS, the US court adopts the sufficiency criterion based on the federal Post Conviction Risk Assessment

Model Card

[Mitchell+ 19]

Model Card: standardizing ethical practice and reporting

Model Details
(developper, date, version, …)

Intended Use

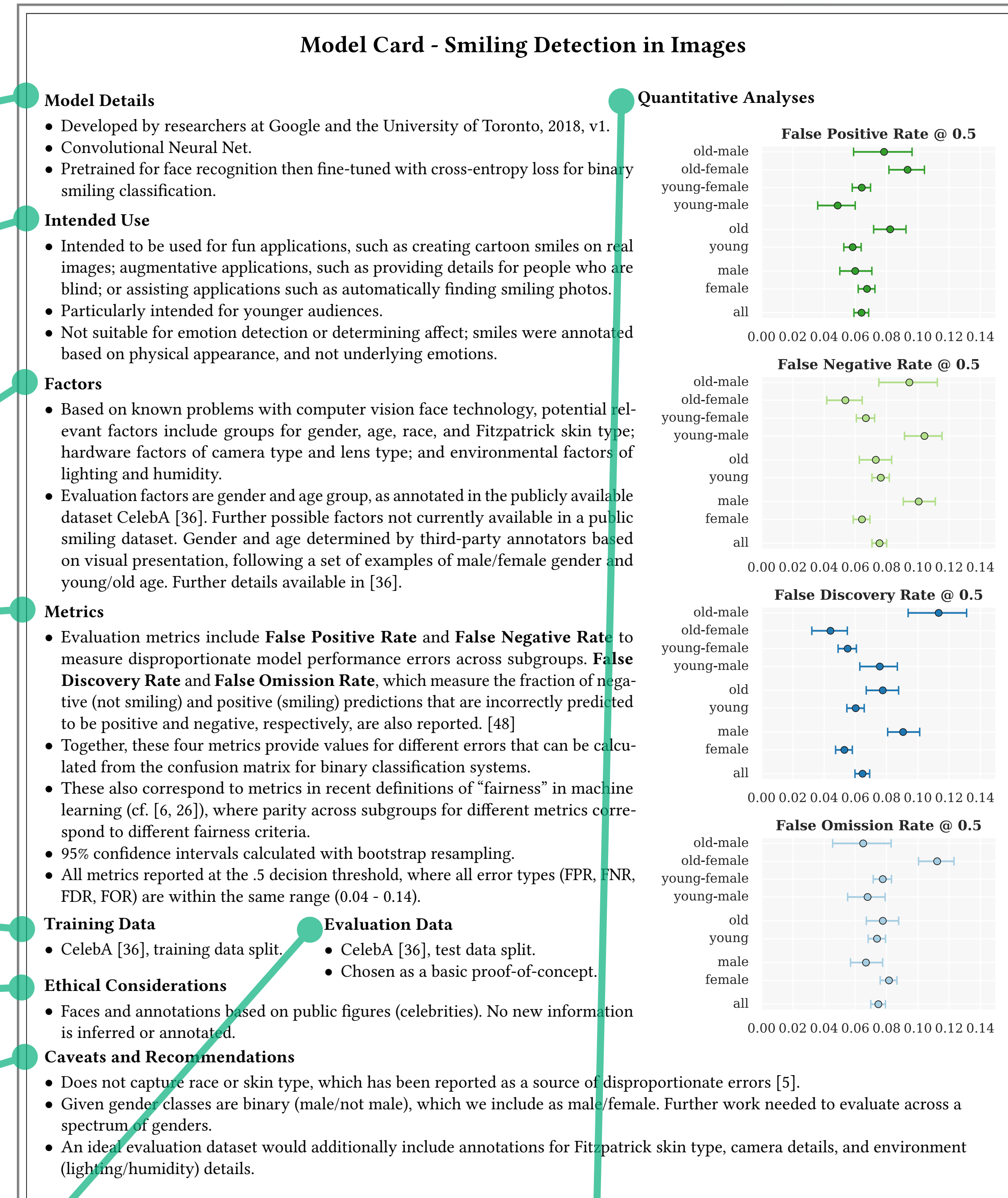
Factors (features, evaluation factors, …)

Metrics
(summary statistics, performance)

Training Data

Ethical Considerations

Caveats and Recommendation



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

Test Data

Quantitative Analysis

Datasheet for Datasets

[Gebru+ 21]

Datasheet for Datasets

- Standardized process for documenting datasets
- Intended to consider potential risks and underlying assumptions

Dataset creators should answer the 57 questions at 7 stages of creating a dataset

motivation	**composition**	**collection process**
purpose, creator, funding	content, size, sampling, features, missing info, splits, noises, external datasets, confidentiality, offensiveness, demographics, identity, sensitive info	method, instruments, sampling, data operators, collection period, ethical review, directly collected, consent, cancel agreement, influence

Uses	**preprocessing / cleaning / labeling**	**distribution**	**maintenance**
use cases, repository, possible use cases, influence of preprocess, prohibited cases	methods, raw data, software	distributer, method, date, license, limitation, regulation	maintainer, contact info, errata, updates, restrictions by subjects, older version, third-party updates

Other
Fairness-Aware Machine Learning
Topics

Bandwagon Effect

Bandwagon Effects in ML

A bias in prediction by ML methods can produce a phenomenon, "richer gets richer"

[Sundar+ 08]

Users' cognitive bias

If others think that something is good, then I should, too

+

[Celma+ 08]

Algorithms' inductive bias

popularity bias: A recommender system tends to select popular items

↓

[Fleder+ 07]

Incorrectly higher-rated items can be more popular, because a recommendation algorithm selects them

↓

A undesirable feedback loop caused by undesired selection

Relation to
Other Machine Learning Topics

Privacy-Preserving Data Mining

Fairness in Machine Learning

the independence between an objective Y and a sensitive feature S

from an information theoretic perspective,

mutual information between Y and S is zero: $\mathrm{I}(Y; S) = 0$

from the viewpoint of privacy-preservation,
protection of sensitive information if an objective is exposed

Difference from PPDM

- introducing randomness is occasionally inappropriate for severe decisions, such as job application
- disclosure of identity isn't problematic in FAML, generally

Cost-Sensitive Learning

[Elkan 01]

Cost-Sensitive Learning: learning classifiers so as to optimize classification costs, instead of maximizing prediction accuracies

FAML can be regarded as a kind of cost-sensitive learning that pays the costs for taking fairness into consideration

Cost matrix $C(i \mid j)$: cost if a true class j is predicted as class i

Total cost to minimize is formally defined as (if class $Y = 1$ or 0):

$$\mathscr{L}(\mathbf{x}, i) = \sum_j \Pr[j \mid \mathbf{x}] C(i \mid j)$$

An object $\mathbf{x}$ is classified into the class i whose cost is minimized

Cost-Sensitive Learning

[Elkan 01]

Theorem 1 in [Elkan 2001]

If negative examples in a data set is over-sampled by the factor of

$$\frac{C(1|0)}{C(0|1)}$$

and a classifier is learned from this samples, a classifier to optimize specified costs is obtained

In a FML case, an over-sampling technique is used for avoiding unfair treatments

A corresponding cost matrix can be computed by this theorem, which connects a cost matrix and the class ratio in training data

* This over-sampling technique is simple and effective for avoiding unfair decisions, but its weak point that it completely ignores non-sensitive features

Other Connected Techniques

Legitimacy / Leakage

- Machine learning models can be deployed in the real world

Independent Component Analysis

- Transformation while maintaining the independence between features

Surrogate Data

- To perform statistical tests, specific information is removed from data sets

Dummy Query

- Dummy queries are inputted for protecting users' demographics into search engines or recommender systems

Visual Anonymization

- To protect identities of persons in images, faces or other information is blurred

Software

Software Frameworks

Non-enterprise Software

- AI Fairness 360 (IBM)
- Fairlearn (Microsoft)
- What-If Tool, ML-fairness-gym (Google)
- **Commercial Packages:** DataRobot, Fiddler AI
- **Non-commercial Packages:** FairTest, Fairness Measures, Aequitas, Fairkit-learn

Enterprise Software

- LinkedIn Fairness Toolkit (LinkedIn)
- Amazon SageMaker (Amazon)

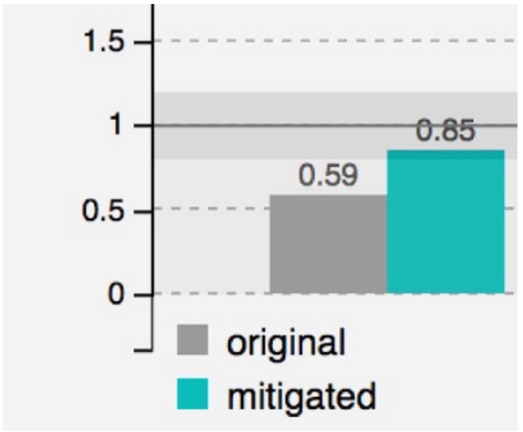
AI Fairness 360

[Bellamy+ 19]

AI Fairness 360 (AIF360): `https://github.com/Trusted-AI/AIF360`

- Software packages for measuring and mitigating fairness
- Developed by IBM, implemented in Python
- **Dataset class:** In addition to the information required for standard ML algorithms, the sensitive information is maintained, and dealing with CSV files or a Pandas DataFrame
- **Metric class:** Evaluate the achievement of the target fairness criteria
- **Explainer class:** Report fairness metric in a text or JSON format, including Web interface

before mitigation



after mitigation

- **Bias Mitigating Algorithms:** 4 pre-processing, 2 in-processing, and 3 post-processing algorithms

* These documented specifications might be updated in the latest version

Fairlearn

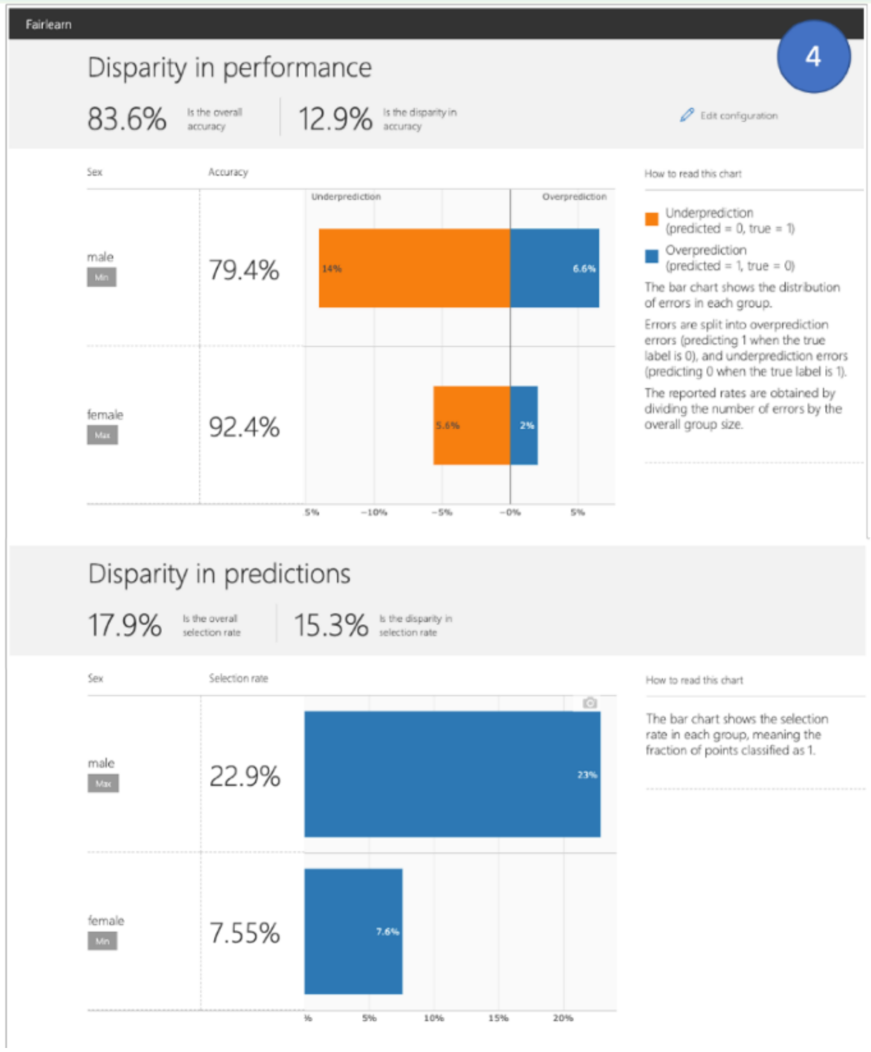
[Bird+ 20]

Fairlearn: https://fairlearn.org/

- Developed by Microsoft, implemented in Python
- Mitigating allocation harms and quality-of-service harms

Interactive visualization dashboard

- Visualize the disparities between sensitive groups



Unfairness mitigation algorithms

- **Hardt's method:** Tuning decision boundaries for each sensitive group to minimize the disparity between the groups
- **Reduction algorithms:** Iterate re-weighting data points and re-training models, to minimize the disparity between sensitive groups

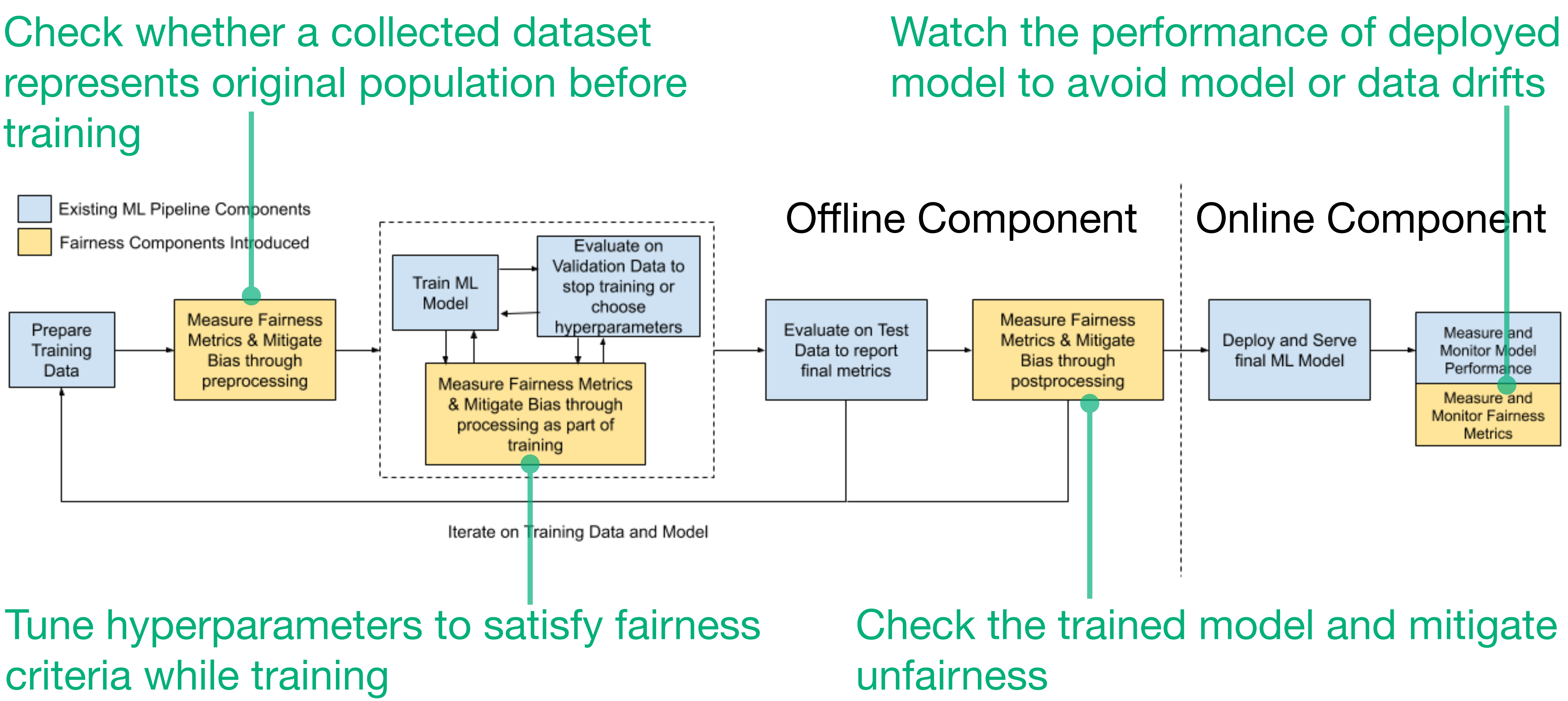
* These documented specifications might be updated in the latest version

LinkedIn Fairness Toolkit

[Vasudevan+ 20]

LinkedIn Fairness Toolkit (Lift): https://github.com/linkedin/LiFT

- Enterprise software for measuring and mitigating fairness
- Developed by LinkedIn
- implemented in Scala, parallel computation using the Apache Spark



Evidence-Based Decision Making

Biased Algorithms Are Easier to Fix Than Biased People

[Mullainathan 19]

Algorithms' biaes are easier to detect than people's biases

People: It takes several months to get one data	↔	**Algorithm:** Massive data can be collected easily

Biased algorithms are easier to fix than biased people

People: The cause of biased decision cannot be cleared up, and evidences showed that training is useless for fixing the biases	↔	**Algorithm:** The cause of biased decision is detectable, and the biases can be fixed

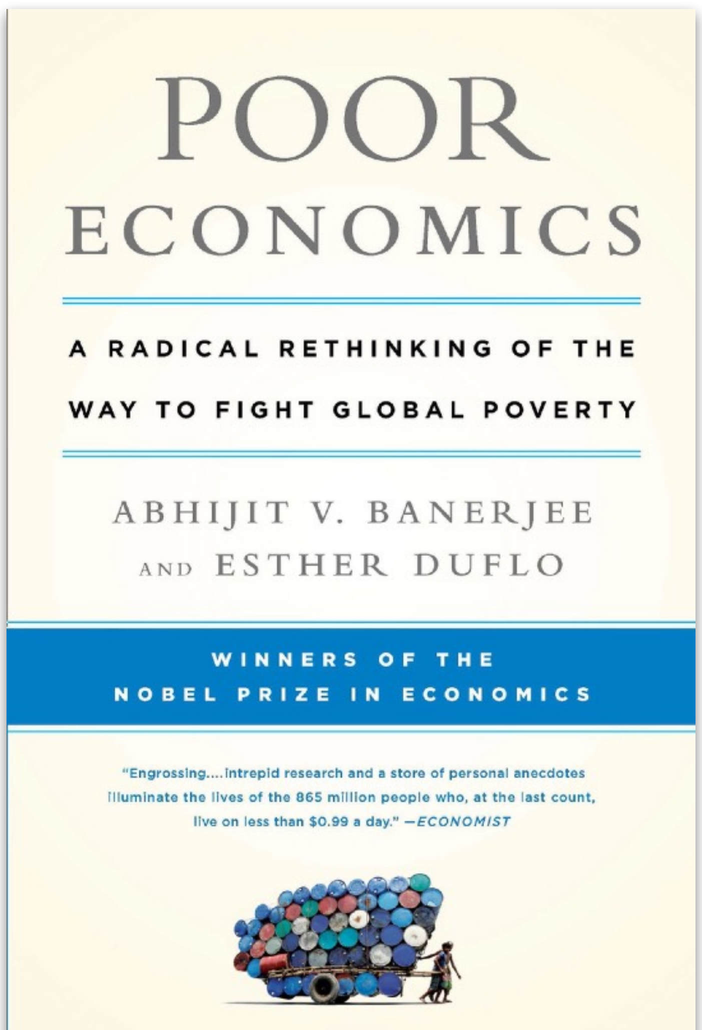
Once proper regulation is in place,
better algorithms can help to ensure equitable treatment in our society

The data for training and test are carefully stored,
and regulatory agency with trained auditors process data

The Three I's Problem

[Banerjee+ 11]

Importance of evidence-based decision making



3I: Ideology, Ignorance, Inertia

why policies fail and why aid does not have the effect it should

The nurses' workload was based on an **ideology** that wants to see nurses as dedicated social workers, designed in **ignorance** of the conditions on the ground, that lives on, mostly just on paper, because of **inertia**.

If we resist the kind of lazy, formulaic thinking that reduces every problem to the same set of general principles; ... if we accept the possibility of error and subject every idea, ..., to rigorous empirical testing, then we will be able not only to construct a toolbox of effective policies but also to better understand why the poor live the way they do.

References

H. Abdollahpouri, G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski, and L. Pizzato. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction*, 30:127–158, 2020. doi: https://doi.org/10.1007/s11257-019-09256-1.

ACMCoE. ACM code of ethics and professional conduct. URL `https://www.acm.org/code-of-ethics`. ⟨`https://www.acm.org/code-of-ethics`⟩.

T. Adel, I. Valera, Z. Ghahramani, and A. Weller. One-network adversarial fairness. In *Proc. of the 33rd AAAI Conf. on Artificial Intelligence*, 2019. doi: https://doi.org/10.1609/aaai.v33i01.33012412.

P. Adler, C. Falk, S. Friedler, G. Rybeck, C. Schedegger, B. Smith, and S. Venkatasubramanian. Auditing black-box models for indirect influence. In *Proc. of the 16th IEEE Int'l Conf. on Data Mining*, pages 1–10, 2016. doi: https://doi.org/10.1109/ICDM.2016.0011. URL `https://doi.ieeecomputersociety.org/10.1109/ICDM.2016.0011`.

R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of the 20th Very Large Database Conf.*, pages 487–499, 1994. URL `http://www.vldb.org/dblp/db/conf/vldb/vldb94-487.html`.

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

S. Athey. Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485, 2017. doi: https://doi.org/10.1126/science.aal4321%20.

R. Baeza-Yates. Bias on the web. *Communications of the ACM*, 61(6):54–61, 2018. doi: https://doi.org/10.1145/3209581.

Abhijit V. Banerjee and Esther Duflo. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. PublicAffairs, 2011.

E. Bareinboim, J. Zhang, and D. Plecko. Causal fairness analysis. The 4th ACM Conference on Fairness, Accountability, and Transparency, Tutorial, 2021.

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

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. URL `https://fairmlbook.org/`.

A. Barr. Google mistakenly tags black people as 'gorillas,' showing limits of algorithms. The Wall Street Journal, 2015. URL `http://on.wsj.com/1CaCNlb`. ⟨`http://on.wsj.com/1CaCNlb`⟩.

R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. of Research and Development*, 2019. doi: https://doi.org/10.1147/JRD.2019.2942287.

B. Berendt and S. Preibusch. Exploring discrimination: A user-centric evaluation of discrimination-aware data mining. In *Proc. of the IEEE Int'l Workshop on Discrimination and Privacy-Aware Data Mining*, pages 344–351, 2012. doi: https://doi.org/10.1109/ICDMW.2012.109. URL `https://doi.ieeecomputersociety.org/10.1109/ICDMW.2012.109`.

B. Berendt and S. Preibusch. Better decision support through exploratory discrimination-aware data mining: Foundations and empirical evidence. *Artificial Intelligence and Law*, 22(2):175–209, 2014. doi: https://doi.org/10.1007/s10506-013-9152-0.

P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175):398–404, 1975. doi: https://doi.org/10.1126/science.187.4175.398.

S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, 2020. URL `https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/`.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.

T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29*, 2016. URL `https://papers.neurips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings`.

C. Boutiller, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in Bayesian networks. In *Uncertainty in Artificial Intelligence 12*, pages 115–123, 1996.

J. Buolamwini and T. Gebru. Gender Shades: Intersectional accuracy disparities in commercial gender classification. In *Proc. of the 1st Conf. on Fairness, Accountability and Transparency*, 2018. URL `https://proceedings.mlr.press/v81/buolamwini18a.html`.

R. Burke, N. Sonboli, and A. Ordonez-Gauger. Balanced neighborhoods for multi-sided fairness in recommendation. In *Proc. of the 1st Conf. on Fairness, Accountability and Transparency*, 2018. URL `http://proceedings.mlr.press/v81/burke18a.html`.

T. Calders. The fairness-accuracy trade-off revisited. ECMLPKDD, Workshop Keynote, 2021.

T. Calders and S. Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21:277–292, 2010. doi: https://doi.org/10.1007/s10618-010-0190-x.

T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang. Controlling attribute effect in linear regression. In *Proc. of the 13th IEEE Int'l Conf. on Data Mining*, pages 71–80, 2013. doi: https://doi.org/10.1109/ICDM.2013.114. URL `https://doi.ieeecomputersociety.org/10.1109/ICDM.2013.114`.

Ò. Celma and P. Cano. From hits to niches?: or how popular artists can bias music recommendation and discovery. In *Proc. of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*, 2008. doi: https://doi.org/10.1145/1722149.1722154.

A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 2017. doi: https://doi.org/10.1089/big.2016.0047.

D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl. Is seeing believing? how recommender interfaces affect users' opnions. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pages 585–592, 2003. doi: https://doi.org/10.1145/642611.642713.

H. Cramer, K. Holstein, J. W. Vaughan, H. Daumé, III, M. Dudík, H. Wallach, S. Reddy, and J. Garcia-Gathright. Challenges of incorporating algorithmic fairness into industry practice. The 2nd ACM Conference on Fairness, Accountability, and Transparency, Tutorial, 2019.

W. Dieterich, C. Mendoza, and T. Brennan. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Northpointe Inc., Research Department, 2016. URL `http://go.volarisgroup.com/rs/430-MBX-989/images/%20ProPublica_Commentary_Final_070616.pdf`. ⟨`http://go.volarisgroup.com/rs/430-MBX-989/images/%20ProPublica_Commentary_Final_070616.pdf`⟩.

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.*, pages 214–226, 2012. doi: https://doi.org/10.1145/2090236.2090255.

EAD. Ethically aligned design (1st edition). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019. URL `https://ethicsinaction.ieee.org/`. ⟨`https://ethicsinaction.ieee.org/`⟩.

H. Edwards and A. Storkey. Censoring representations with an adversary. In *Proc. of the 4th Int'l Conf. on Learning Representations*, 2016. URL `https://arxiv.org/abs/1511.05897`.

M. D. Ekstrand, M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill, and M. S. Pera. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Proc. of the 1st Conf. on Fairness, Accountability and Transparency*, 2018. URL `http://proceedings.mlr.press/v81/ekstrand18b.html`.

M. D. Ekstrand, R. Burke, and F. Diaz. Fairness and discrimination in recommendation and retrieval. The 13th ACM Conf. on Recommender Systems, Tutorial, 2019.

C. Elkan. The foundations of cost-sensitive learning. In *Proc. of the 17th Int'l Joint Conf. on Artificial Intelligence*, pages 973–978, 2001. URL `https://www.ijcai.org/Proceedings/2001-2`.

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*, pages 259–268, 2015. doi: https://doi.org/10.1145/2783258.2783311.

D. Fleder and K. Hosanagar. Recommender systems and their impact on sales diversity. In *ACM Conference on Electronic Commerce*, pages 192–199, 2007. doi: https://doi.org/10.1145/1250910.1250939.

A. W. Flores, K. Bechtel, and C. T. Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.". *Federal Probation Journal*, 80 (2), 2016. URL `https://www.uscourts.gov/federal-probation-journal/2016/09/false-positives-false-negatives-and-false-analyses-rejoinder`.

S. Forden. Google said to face ultimatum from FTC in antitrust talks. Bloomberg, Nov. 13 2012. URL `http://bloom.bg/PPNEaS`. ⟨`http://bloom.bg/PPNEaS`⟩.

S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64, 2021. doi: https://doi.org/10.1145/3433949.

K. Fukuchi and J. Sakuma. Neutralized empirical risk minimization with generalization neutrality bound. In *Proc. of the ECML PKDD 2014, Part I*, pages 418–433, 2014. doi: https://doi.org/10.1007/978-3-642-40991-2_32. [LNCS 8724].

K. Fukuchi, J. Sakuma, and T. Kamishima. Prediction with model-based neutrality. In *Proc. of the ECML PKDD 2013, Part II*, pages 499–514, 2013. doi: https://doi.org/10.1007/978-3-642-40991-2_32. [LNCS 8189].

B. Gao and B. Berendt. Visual data mining for higher-level patterns: Discrimination-aware data mining and beyond. In *In Proc. of 20th Annual Belgian Dutch Conf. on Machine Learning*, pages 45–52, 2011.

GDPR. General data protection regulation. URL `https://eur-lex.europa.eu/eli/reg/2016/679/oj`. ⟨`http://data.europa.eu/eli/reg/2016/679/oj`⟩.

T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: https://doi.org/10.1145/3458723.

D. Gondek and T. Hofmann. Non-redundant data clustering. In *Proc. of the 4th IEEE Int'l Conf. on Data Mining*, pages 75–82, 2004. doi: https://doi.org/10.1109/ICDM.2004.10104. URL `https://doi.ieeecomputersociety.org/10.1109/ICDM.2004.10104`.

D. Gondek and T. Hofmann. Non-redundant clustering with conditional ensembles. In *Proc. of the 11th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 70–77, 2005. doi: https://doi.org/10.1145/1081870.1081882.

A. Gunawardana and G. Shani. A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10:2935–2962, 2009. URL `http://www.jmlr.org/papers/v10/gunawardana09a.html`.

S. Hajian and J. Domingo-Ferrer. A study on the impact of data anonymization on anti-discrimination. In *Proc. of the IEEE Int'l Workshop on Discrimination and Privacy-Aware Data Mining*, pages 352–359, 2012. doi: https://doi.org/10.1109/ICDMW.2012.19. URL `https://doi.ieeecomputersociety.org/10.1109/ICDMW.2012.19`.

S. Hajian and J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE Trans. on Knowledge and Data Engineering*, 25(7):1445–1459, 2013. doi: https://doi.org/10.1109/TKDE.2012.72. URL `https://doi.ieeecomputersociety.org/10.1109/TKDE.2012.72`.

S. Hajian, J. Domingo-Ferrer, and O. Farràs. Generalization-based privacy preservation and discrimination prevention in data publishing and mining. *Data Mining and Knowledge Discovery*, 2014. doi: https://doi.org/10.1007/s10618-014-0346-1.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29*, 2016. URL https://papers.neurips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.

J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47:153–161, 1979. doi: https://doi.org/10.2307/1912352.

J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. on Information Systems*, 22(1):5–53, 2004. doi: https://doi.org/10.1145/963770.963772.

M. Hildebrandt. Rude awakenings from behaviourists dreams. the methodological integrity and the gdpr. The 13th ACM Conf. on Recommender Systems, Keynote, 2019.

T. Hofmann and J. Puzicha. Latent class models for collaborative filtering. In *Proc. of the 16th Int'l Joint Conf. on Artificial Intelligence*, pages 688–693, 1999. URL https://www.ijcai.org/Proceedings/1999-2.

Jason Hon. Is the trolley problem useful for studying autonomous vehicles? BLOG@CACM, May 2019. URL https://cacm.acm.org/blogs/blog-cacm/236606-is-the-trolley-problem-useful-for-studying-autonomous-vehicles/fulltext. ⟨https://cacm.acm.org/blogs/blog-cacm/236606-is-the-trolley-problem-useful-for-studying-autonomous-vehicles/fulltext⟩.

B. Hutchinson and M. Mitchell. 50 years of test (un)fairness: Lessons for machine learning. In *Proc. of the 2nd Conf. on Fairness, Accountability and Transparency*, 2019. doi: https://doi.org/10.1145/3287560.3287600.

IBM. IBM response to "gender shades: Intersectional accuracy disparities in commercial gender classification", 2018. URL http://gendershades.org/docs/ibm.pdf.

IEEEGIoE. The IEEE global initiative on ethics of autonomous and intelligent systems. URL https://ethicsinaction.ieee.org/. ⟨https://ethicsinaction.ieee.org/⟩.

Makio Ishiguro, Motoi Okamoto, Hiroe Tsubaki, Michiko Miyamoto, Masao Yanaga, and Takemi Yanagimoto. *Houteinotameno Toukei Riterashi*. Kindaikagaku-sha, 2014. (in Japanese).

M. Joseph, M. Kearns, J. Morgenstern, and A. Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems 29*, 2016. URL https://papers.neurips.cc/paper/6355-fairness-in-learning-classic-and-contextual-bandits.

F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33:1–33, 2012. doi: https://doi.org/10.1007/s10115-011-0463-8.

F. Kamiran, T. Calders, and M. Pechenizkiy. Discrimination aware decision tree learning. In *Proc. of the 10th IEEE Int'l Conf. on Data Mining*, pages 869–874, 2010. doi: https://doi.org/

10.1109/ICDM.2010.50. URL https://doi.ieeecomputersociety.org/10.1109/ICDM.2010.50.

F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. In *Proc. of the 12th IEEE Int'l Conf. on Data Mining*, pages 924–929, 2012. doi: https://doi.org/10.1109/ICDM.2012.45. URL https://doi.ieeecomputersociety.org/10.1109/ICDM.2012.45.

F. Kamiran, I. Žliobaitė, and T. Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, 35: 613–644, 2013. doi: https://doi.org/10.1007/s10115-012-0584-8.

T. Kamishima and S. Akaho. Considerations on recommendation independence for a find-good-items task. In *Proc.of the FATREC Workshop on Responsible Recommendation*, 2017. doi: https://doi.org/10.18122/B2871W.

T. Kamishima and K. Fukuchi. Future directions of fairness-aware data mining: Recommendation, causality, and theoretical aspects. In *ICML2015 Workshop: Fairness, Accountability, and Transparency in Machine Learning*, 2015. URL http://www.fatml.org/schedule/2015.

T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Enhancement of the neutrality in recommendation. In *The 2nd Workshop on Human Decision Making in Recommender Systems*, 2012a. URL http://ceur-ws.org/Vol-893/.

T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Proc. of the ECML PKDD 2012, Part II*, pages 35–50, 2012b. doi: https://doi.org/10.1007/978-3-642-33486-3_3. [LNCS 7524].

T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Efficiency improvement of neutrality-enhanced recommendation. In *The 3rd Workshop on Human Decision Making in Recommender Systems*, 2013. URL http://ceur-ws.org/Vol-1050/.

T. Kamishima, S. Akaho, H. Asoh, and I. Sato. Model-based approaches for independence-enhanced recommendation. In *Proc. of the IEEE 16th Int'l Conf. on Data Mining Workshops*, pages 860–867, 2016. doi: https://doi.org/10.1109/ICDMW.2016.0127. URL http://doi.ieeecomputersociety.org/10.1109/ICDMW.2016.0127.

T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Model-based and actual independence for fairness-aware classification. *Data Mining and Knowledge Discovery*, 32:258–286, 2018a. doi: https://doi.org/10.1007/s10618-017-0534-x.

T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Recommendation independence. In *Proc of the Conf. on Fairness, Accountability and Transparency*, volume 81 of *PMLR*, pages 187–201, 2018b. URL http://proceedings.mlr.press/v81/kamishima18a.html.

J. Kleinberg, S. Mullainathan, and M. Raghavan:. Inherent trade-offs in the fair determination of risk scores. In *Proc. of 8th Innovations in Theoretical Computer Science Conf.*, 2017. doi: https://doi.org/10.4230/LIPIcs.ITCS.2017.43.

J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133, 2018. doi: https://doi.org/10.1093/qje/qjx032.

J. A. Konstan and J. Riedl. Recommender systems: Collaborating in commerce and communities. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems, Tutorial*, 2003.

Y. Koren. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 426–434, 2008. doi: https://doi.org/10.1145/1401890.1401944.

M. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30*, 2017. URL `https://papers.nips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html`.

K. Lippert-Rasmussen. The badness of discrimination. *Ethical Theory and Moral Practice*, 9: 167–185, 2006. doi: https://doi.org/10.1007/s10677-006-9014-x.

L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt. Delayed impact of fair machine learning. In *Proc. of the 35th Int'l Conf. on Machine Learning*, 2018. URL `http://proceedings.mlr.press/v80/liu18c.html`.

B. T. Luong, S. Ruggieri, and F. Turini. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 502–510, 2011. doi: https://doi.org/10.1145/2020408.2020488.

D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. In *Proc. of the 35th Int'l Conf. on Machine Learning*, 2018. URL `http://proceedings.mlr.press/v80/madras18a.html`.

K. Mancuhan and C. Clifton. Combating discrimination using bayesian networks. *Artificial Intelligence and Law*, 22(2):211–238, 2014. doi: https://doi.org/10.1007/s10506-014-9156-4.

P. Miettinen and E. Galbrun. An introduction to redescription mining. ECMLPKDD, Tutorial, 2016.

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*, 2019. doi: https://doi.org/10.1145/3287560.3287596.

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

E. Pariser. *The Filter Bubble: What The Internet Is Hiding From You*. Viking, 2011.

Judea Pearl and Dana Mackenzie. *The Book of Why — The New Science of Cause and Effect*. Penguin, 2018.

Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, 2016.

D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 560–568, 2008. doi: https://doi.org/10.1145/1401890.1401959.

D. Pedreschi, S. Ruggieri, and F. Turini. Measuring discrimination in socially-sensitive decision records. In *Proc. of the SIAM Int'l Conf. on Data Mining*, pages 581–592, 2009. doi: https://doi.org/10.1137/1.9781611972795.50.

A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muños-Marí, L. Gómez-Chova, and G. Camps-Valls. Fair kernel learning. In *Proc. of the ECML PKDD 2017, Part I*, pages 339–355, 2017. doi: https://doi.org/10.1007/978-3-319-71249-9_21. [LNCS 10534].

A. Rathore, S. Dev, J. Phillips, V. Srikumar, V. Srikumar, and B. Wang. A visual tour of bias mitigation techniques for word representations. In *The 27th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, Tutorial*, 2021.

P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. GroupLens: An open architecture for collaborative filtering of Netnews. In *Proc. of the Conf. on Computer Supported Cooperative Work*, pages 175–186, 1994. doi: https://doi.org/10.1145/192844.192905.

P. Resnick, J. Konstan, and A. Jameson. Panel on the filter bubble. The 5th ACM Conf. on Recommender Systems, 2011. URL `http://acmrecsys.wordpress.com/2011/10/25/panel-on-the-filter-bubble/`. ⟨`http://acmrecsys.wordpress.com/2011/10/25/panel-on-the-filter-bubble/`⟩.

G. Ristanoski, W. Liu, and J. Bailey. Discrimination-aware classification for imbalanced datasets. In *Proc. of the 22nd ACM Conf. on Information and Knowledge Management*, 2013. doi: https://doi.org/10.1145/2505515.2507836.

S. Ruggieri. Data anonymity meets non-discrimination. In *Proc. of the 4th IEEE Int'l Workshop on Privacy Aspects of Data Mining*, pages 875–882, 2013. doi: https://doi.org/10.1109/ICDMW.2013.56. URL `https://doi.ieeecomputersociety.org/10.1109/ICDMW.2013.56`.

S. Ruggieri, D. Pedreschi, and F. Turini. Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data*, 4(2), 2010a. doi: https://doi.org/10.1145/1754428.1754432.

S. Ruggieri, D. Pedreschi, and F. Turini. DCUBE: Discrimination discovery in databases. In *Proc of The ACM SIGMOD Int'l Conf. on Management of Data*, pages 1127–1130, 2010b. doi: https://doi.org/10.1145/1807167.1807298.

S. Ruggieri, S. Hajian, F. Kamiran, and X. Zhang. Anti-discrimination analysis using privacy attack strategies. In *Proc. of the ECML PKDD 2014, Part II*, pages 694–710, 2014. doi: https://doi.org/10.1007/978-3-662-44851-9_44. [LNCS 8725].

R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems 20*, pages 1257–1264, 2008. URL `https://papers.neurips.cc/paper/3208-probabilistic-matrix-factorization`.

P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63, 2019. doi: https://doi.org/10.1147/JRD.2019.2945519.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90:227–244, 2000. doi: https://doi.org/10.1016/S0378-3758(00)00115-4.

A. Singh and T. Joachims. Fairness of exposure in rankings. In *Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, 2018. doi: https://doi.org/10.1145/3219819.3220088. URL http://www.kdd.org/kdd2018/accepted-papers/view/fairness-of-exposure-in-rankings.

T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, and M. B. Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, 2018. doi: https://doi.org/10.1145/3219819.3220046. URL https://www.kdd.org/kdd2018/accepted-papers/view/a-unified-approach-to-quantifying-algorithmic-unfairness-measuring-individu.

E. Steel and J. Angwin. On the web's cutting edge, anonymity in name only. The Wall Street Journal, 2010. URL http://on.wsj.com/1zD2BQP. ⟨http://on.wsj.com/aimKCw⟩.

S. S. Sundar, A. Oeldorf-Hirsch, and Q. Xu. The bandwagon effect of collaborative filtering technology. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, 2008. doi: https://doi.org/10.1145/1358628.1358873.

L. Sweeney. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54, 2013. doi: https://doi.org/10.1145/2447976.2447990.

S. Vasudevan and K. Kenthapadi. LiFT: A scalable framework for measuring fairness in ML applications. In *Proc. of the 29th ACM Conf. on Information and Knowledge Management*, 2020. doi: https://doi.org/10.1145/3340531.3412705.

H. Wallach. Moving beyond prediction: Big data, transparency, and accountability. In *NIPS2014 Workshop: Fairness, Accountability, and Transparency in Machine Learning*, 2014. URL https://hannawallach.medium.com/big-data-machine-learning-and-the-social-sciences-927a8e20460d.

M. Wick, S. Panda, and J.-B. Tristan. Unlocking fairness: a trade-off revisited. In *Advances in Neural Information Processing Systems 32*, 2019. URL https://papers.nips.cc/paper/2019/hash/373e4c5d8edfa8b74fd4b6791d0cf6dc-Abstract.html.

S. Yao and B. Huang. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems 30*, 2017. URL https://papers.neurips.cc/paper/6885-beyond-parity-fairness-objectives-for-collaborative-filtering.

B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proc. of the 21st Int'l Conf. on Machine Learning*, pages 903–910, 2004. doi: https://doi.org/10.1145/1015330.1015425.

M. B. Zafar, I. Valera, M. Rodriguez, K. Gummadi, and A. Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems 30*, 2017a. URL https://papers.neurips.cc/paper/6627-from-parity-to-preference-based-notions-of-fairness-in-classification.

M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proc. of the 20th International Conference on Artificial Intelligence* *and Statistics*, volume 54 of *PMLR*, pages 962–970, 2017b. URL http://proceedings.mlr.press/v54/zafar17a.html.

M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proc. of the 26th Int'l Conf. on World Wide Web*, pages 1171–1180, 2017c. doi: https://doi.org/10.1145/3038912.3052660.

M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. Fa∗ir: A fair top-k ranking algorithm. In *Proc. of the 25th ACM Conf. on Information and Knowledge Management*, 2017. doi: https://doi.org/10.1145/3132847.3132938.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *Proc. of the 30th Int'l Conf. on Machine Learning*, pages 325–333, 2013. URL http://jmlr.org/proceedings/papers/v28/zemel13.html.

B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proc. of the 2018 AAAI/ACM Conf. on AI, Ethics, and Society*, 2018a. doi: https://doi.org/10.1145/3278721.3278779.

J. Zhang and E. Bareinboim. Fairness in decision-making — the causal explanation formula. In *Proc. of the 32nd AAAI Conf. on Artificial Intelligence*, 2018. doi: https://doi.org/10.1609/aaai.v32i1.11564.

L. Zhang, Y. Wu, and X. Wu. Situation testing-based discrimination discovery: A causal inference approach. In *Proc. of the 25th Int'l Joint Conf. on Artificial Intelligence*, pages 2718–2724, 2016. URL http://www.ijcai.org/Abstract/16/386.

L. Zhang, Y. Wu, and X. Wu. Anti-discrimination learning: From association to causation. The 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, Tutorial, 2018b. URL http://csce.uark.edu/~xintaowu/kdd18-tutorial/.

I. Žliobaitė. On the relation between accuracy and fairness in binary classification. In *ICML2015 Workshop: Fairness, Accountability, and Transparency in Machine Learning*, 2015.

I. Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 2017. doi: https://doi.org/10.1007/s10618-017-0506-1.

I. Žliobaitė and B. Custers. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, 24:183–201, 2016. doi: https://doi.org/10.1007/s10506-016-9182-5.

I. Žliobaitė, F. Kamiran, and T. Calders. Handling conditional discrimination. In *Proc. of the 11th IEEE Int'l Conf. on Data Mining*, 2011. doi: https://doi.org/10.1109/ICDM.2011.72. URL https://doi.ieeecomputersociety.org/10.1109/ICDM.2011.72.