Anti-Discrimination Learning: from Association to Causation

Lu Zhang
Xintao Wu
Yongkai Wu
University of Arkansas

SBP-BRiMS 2017 Tutorial, July 5, 2017. Washington DC, USA

Outline

- Part I: Introduction
  - Context
  - Literature Review
- Part II: Causal Modeling Background
- Part III: Anti-Discrimination Learning
- Part IV: Challenges and Directions for Future Research

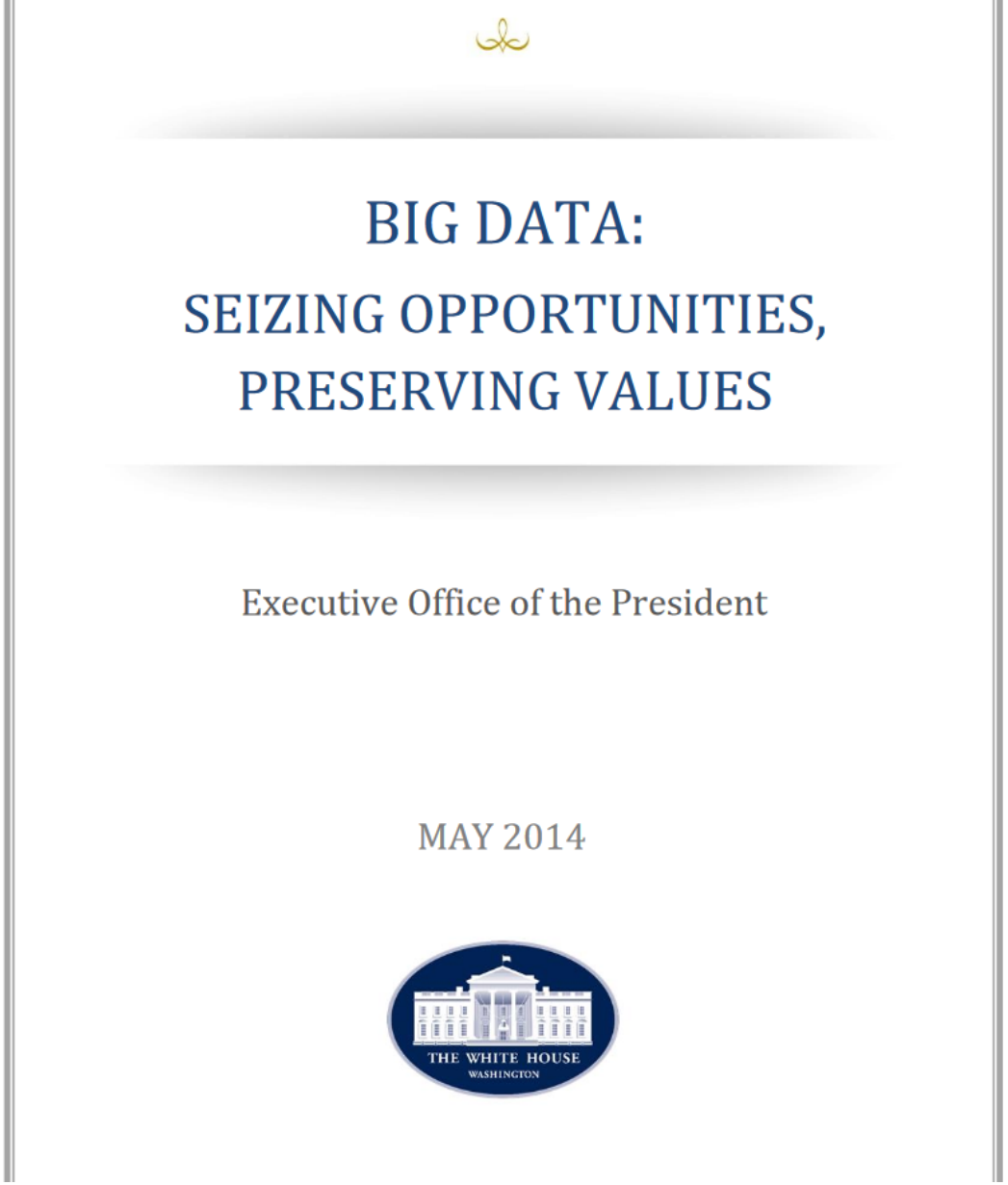
Introduction

- Discrimination refers to unjustified distinctions of individuals based on their membership in a certain group.
- Federal Laws and regulations disallow discrimination on several grounds:
  - Gender, Age, Marital Status, Race, Religion or Belief, Disability or Illness ……
  - These attributes are referred to as the protected attributes.

Introduction

May 2014

Big data technologies can cause societal harms beyond damages to privacy, such as discrimination against individuals and groups.



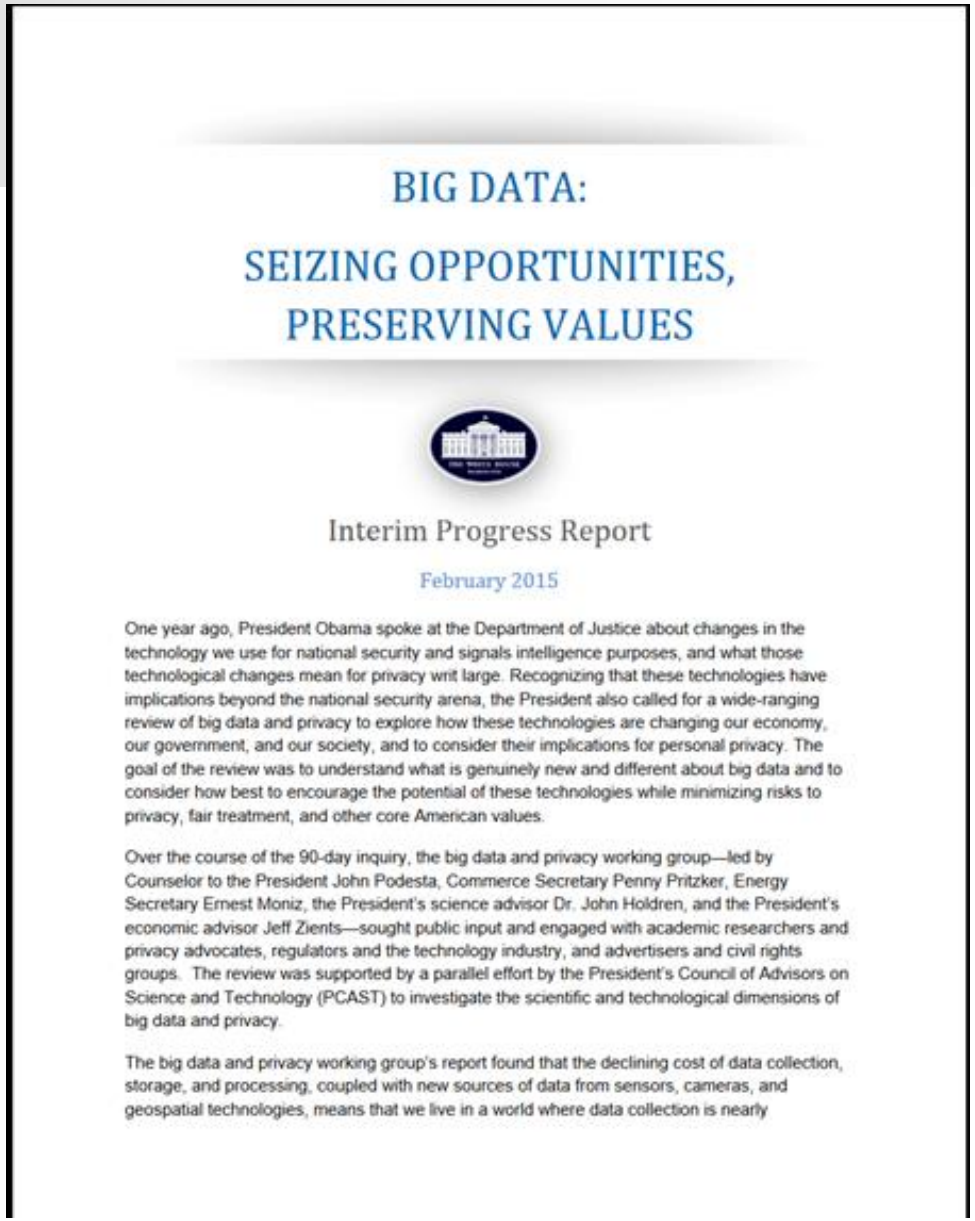
Introduction

February 2015

Pay attention to the potential for big data to facilitate discrimination

Expand technical expertise to stop discrimination

Deepen understanding of differential pricing



BIG DATA:

SEIZING OPPORTUNITIES, PRESERVING VALUES

Interim Progress Report

February 2015

One year ago, President Obama spoke at the Department of Justice about changes in the technology we use for national security and signals intelligence purposes, and what those technological changes mean for privacy writ large. Recognizing that these technologies have implications beyond the national security arena, the President also called for a wide-ranging review of big data and privacy to explore how these technologies are changing our economy, our government, and our society, and to consider their implications for personal privacy. The goal of the review was to understand what is genuinely new and different about big data and to consider how best to encourage the potential of these technologies while minimizing risks to privacy, fair treatment, and other core American values.

Over the course of the 90-day inquiry, the big data and privacy working group—led by Counselor to the President John Podesta, Commerce Secretary Penny Pritzker, Energy Secretary Ernest Moniz, the President's science advisor Dr. John Holdren, and the President's economic advisor Jeff Zients—sought public input and engaged with academic researchers and privacy advocates, regulators and the technology industry, and advertisers and civil rights groups. The review was supported by a parallel effort by the President's Council of Advisors on Science and Technology (PCAST) to investigate the scientific and technological dimensions of big data and privacy.

The big data and privacy working group's report found that the declining cost of data collection, storage, and processing, coupled with new sources of data from sensors, cameras, and geospatial technologies, means that we live in a world where data collection is nearly

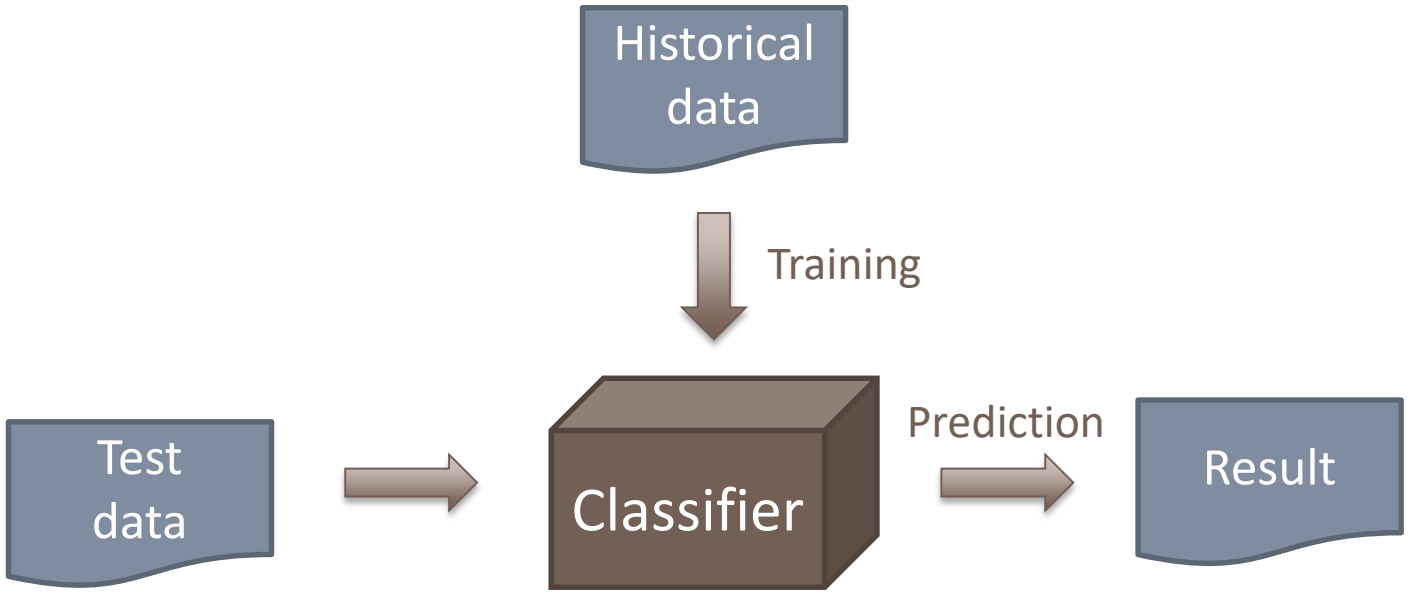
Introduction

May 2016

Support research into mitigating algorithmic discrimination, building systems that support fairness and accountability, and developing strong data ethics frameworks.

Anti-Discrimination Learning

Discover and remove discrimination from the training data



Build discrimination-free classifier

Outline

- Part I: Introduction
  - Context
  - Literature Review
- Part II: Causal Modeling Background
- Part III: Anti-Discrimination Learning
- Part IV: Challenges and Directions for Future Research

Discrimination Categorization

- Category based on two dimensions
  - From the perspective of in what way discrimination occurs, discrimination is legally divided into direct discrimination and indirect discrimination.
  - From the perspective of different level of granularity in studying, discrimination can be divided into system level, group level, and individual level.

Discrimination Categorization

- Category based on two dimensions
  - From the perspective of in what way discrimination occurs, discrimination is legally divided into
    - Direct: explicitly based on the protected attributes.
      - E.g., rejecting a qualified female just because of her gender.
    - Indirect: based on apparently neutral non-protected attributes but still results in unjustified distinctions against individuals from the protected group.
      - E.g., redlining, where the residential Zip Code of an individual is used for making decisions such as granting a loan.
  - From the perspective of different level of granularity in studying, discrimination can be divided into system level, group level, and individual level.

Discrimination Categorization

- Category based on two dimensions
  - From the perspective of in what way discrimination occurs, discrimination is legally divided into direct discrimination and indirect discrimination.
  - From the perspective of different level of granularity in studying, discrimination can be divided into
    - System level: the average discrimination across the whole system, e.g., all applicants to a university.
    - Group level: the discrimination that occurs in one particular subgroup, e.g., the applicants applying for a particular major, or the applicants with a particular score.
    - Individual level: the discrimination that happens to one particular individual, e.g., one particular applicant.

Research Topics

- Discrimination Discovery/Detection
  - Unveil evidence of discriminatory practices by analyzing the historical dataset or the predictive model.

- Discrimination Prevention/Removal
  - Ensure non-discrimination by modifying the biased data (before building predictive models) or twisting the predictive model.

Notations

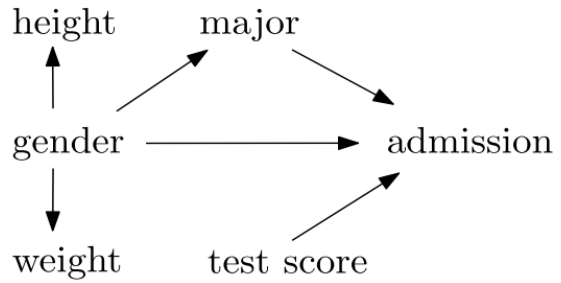
- A binary protected attribute C with values c^-, c^+ (can extend to multiple domain values and multiple protected attributes).
- A binary decision E with values e^-, e^+.
- Non-protected attributes $\boldsymbol{X}$ among which $\boldsymbol{R}$ are redlining attributes.

- Denote an attribute by an uppercase alphabet, e.g., X
- Denote a value of attribute X by x
- Denote a subset of attributes by a bold uppercase alphabet, e.g., $\boldsymbol{X}$
- Denote a value assignment of attributes $\boldsymbol{X}$ by $\boldsymbol{x}$

Illustrative Example

- Gender discrimination in college admission.

No.	gender	major	score	height	weight	ad.
1	F	CS	B	low	low	reject
2	M	CS	B	median	median	admit
3	F	CS	A	low	low	reject
4	M	CS	A	median	median	admit
5	F	CS	C	low	median	reject
6	M	CS	C	median	median	reject
7	M	EE	B	low	low	reject



C is gender, c^- = female, c^+=male.
E is admission, e^-= reject, e^+=admit.

Measuring Discrimination

- Individual fairness
- Statistical parity
- Causal effect

Individual Fairness

- Consistency for individual i
  - $Cons_i = 1 - \frac{1}{k}\sum_{j \in kNN(i)} |e_i - e_j|$
  - Compare the outcome of an individual with its k-nearest neighbors
  - Note that the similar individuals may be from the protected group and all are treated badly.
- Consistency for the whole data
  - $Cons = 1 - \frac{1}{Nk}\sum_i \sum_{j \in kNN(i)} |e_i - e_j|$

Zemel, R.S.,Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In ICML'13. (2013)

Statistical Parity

- Risk Difference (RD), UK law
- Risk Ration (RR), EU Court of Justice
- Relative Chance (RC)
- Odds Ratio (OR)

Protected group vs. unprotected group

- Extended Risk Difference (ED)
- Extended Risk Ratio (ER)
- Extended Chance (EC)

Protected group vs. entire population

	benefit		
group	denied	granted	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

$$p_1 = a/n_1 \quad p_2 = c/n_2 \quad p = m_1/n$$

$$RD = p_1 - p_2 \quad RR = \frac{p_1}{p_2} \quad RC = \frac{1-p_1}{1-p_2} \quad OR = \frac{RR}{RC} = \frac{a/b}{c/d} \quad ED = p_1 - p \quad ER = \frac{p_1}{p} \quad EC = \frac{1-p_1}{1-p}$$

Statistical Parity

- Conditional discrimination
- α-discrimination based on association rules
- Multi-factor interactions
- $belift$ based on Bayesian networks
- kNN-based situation testing
- Disparate impact

Conditional Discrimination

- $diff = P(e^+|c^+) - P(e^+|c^-)$ is a sum of the explainable and the bad discrimination.
  - $D_{all} = D_{exp} + D_{bad} = P(e^+|c^+) - P(e^+|c^-)$
- Explainable Discrimination
  - $D_{exp} = \sum_i P(x_i|c^+)P^*(e^+|x_i) - \sum_i P(x_i|c^-)P^*(e^+|x_i)$
  - $P^*(e^+|x_i) = \frac{P(e^+|x_i, c^+)+P(e^+|x_i, c^-)}{2}$
  - X is an explanatory attribute and x_i is its i-th domain value

Žliobaite, I., Kamiran, F., Calders, T.: Handling conditional discrimination. In: ICDM'11, pp. 992–1001. IEEE, (2011)

Examples

Example 1				
Major	Medicine		Computer	
Gender	female	male	female	male
# of applicants	800	200	200	800
Acceptance rate (%)	20%	20%	40%	40%

$D_{all} = 12\%$
$D_{exp} = 12\%$
$D_{bad} = 0\%$

Example 2				
Major	Medicine		Computer	
Gender	female	male	female	male
# of applicants	800	200	200	800
Acceptance rate (%)	15%	25%	35%	45%

$D_{all} = 22\%$
$D_{exp} = 12\%$
$D_{bad} = 10\%$

α-Discrimination in Association Rules

- Direct Discrimination
  - $C, X \rightarrow E$
    - $elift(C, X \rightarrow E) = \frac{conf(C, X \rightarrow E)}{conf(X \rightarrow E)} \geq \alpha$
    - C is a protected attribute
    - X is a context attribute
    - E is a decision attribute

$$\{Race = black, Income = high\} \longrightarrow Loan = reject$$

- Indirect Discrimination
  - $X_1, X_2 \rightarrow E$
    - X_1, X_2 are both context attributes
    - X_1, X_2 are strongly correlated with C
    - E is a decision attribute

$$\{ZipCode = 70201, Income = high\} \longrightarrow Loan = reject$$

Hajian, S., Domingo-Ferrer, J.: A methodology for direct and indirect discrimination prevention in datamining. IEEE Trans. Knowl. Data Eng. 25(7), 1445–1459 (2013)

Multi-Factor Interaction

- Build a loglinear model from categorical data
- Measure the discrimination based on the strength of interactions among categorical attributes in the fitted model

Data:
A 3-D table (C, E, X) where a cell is denoted as (i, j, k, m_{ijk})

$$\log(m_{ijk}) = \gamma + \gamma_i^C + \gamma_j^E + \gamma_k^X + \gamma_{ij}^{CE} + \gamma_{ik}^{CX} + \gamma_{jk}^{XE} + \gamma_{ijk}^{CEX}$$

$C_i = female$
$X_j = CS$
$E_k = reject$

$$I_{ij|k}^{CE|X} = \gamma_{ik}^{CE} + \gamma_{ijk}^{CEX}$$

$$\log(OR) = I_{ij|k}^{CE|X} + I_{i'j'|k}^{CE|X} - I_{i'j|k}^{CE|X} - I_{ij'|k}^{CE|X}$$

- Extendable to multiple protected/decision attributes

Wu,Y., Wu,X.: Using loglinear model for discrimination discovery and prevention. In: DSAA'16 (2016)

$belift$ Based on Bayesian networks

- $belift = \frac{P(e^+|c_1,c_2,\ldots c_l, x_1,x_2,\ldots x_m, r_1,r_2,\ldots r_n)}{P'(e^+|x_1,x_2,\ldots x_m)}$
  - C_i is a protected attribute
  - X_i is a non-protected attribute
  - R_i is a redlining attribute
  - $belift = 1$: perfect equality
- Two bayesian networks are built from data to calculate conditional probabilities.

Mancuhan, K., Clifton, C.: Combating discrimination using Bayesian networks. Artif. Intell. Law 22(2), 211–238 (2014)

Discrimination discovery using $belift$

- Build a Bayesian network G from training dataset D
- Build a relative Bayesian network G' by removing protected attributes and any attribute directly connected to them in G
- For each instance in D
  - Compute $P(e^+|c_1, c_2, \dots c_l, x_1, x_2, \dots x_m, r_1, r_2, \dots r_n)$ over G
  - Compute $P'(e^+|x_1, x_2, \dots x_m)$ over G'
  - Calculate $belift$ and report discrimination if it exceeds a threshold

Situation Testing

- A legally grounded technique for analyzing the discriminatory treatment on an individual adopted both in the US and the EU.
- In responding to complaint about discrimination:
  1. Pairs of testers who are similar to the individual are sent out to participate in the same decision process (e.g., applying for the same job).
  2. For each pair, the two testers possess the same characteristics except the membership to the protected group.
  3. The distinction of decisions between the protected group and the non-protected group implies discriminatory behavior.

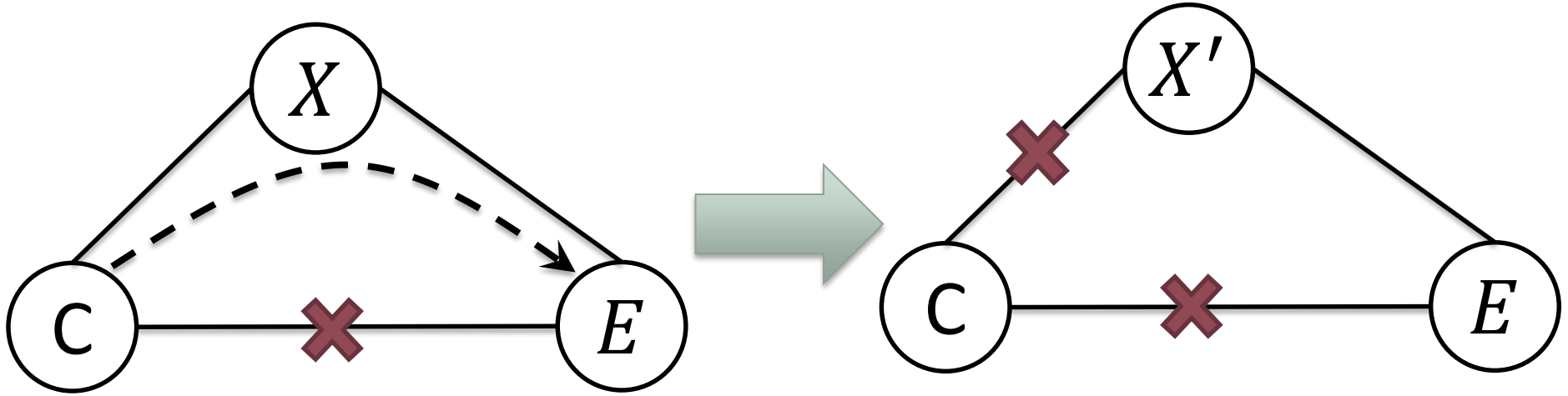
kNN-Based Situation Testing

- Given a individuals tuple t with c^- and e^-;
- Rank all the individuals according to their distances to t;
- Select the individuals that closest to t;
  - individuals with c^+ are added into set $\boldsymbol{S}^+$
  - individuals with c^- are added into set $\boldsymbol{S}^-$;
- If $P(e^+|\boldsymbol{S}^+) - P(e^+|\boldsymbol{S}^-) > \tau$, then t is considered as being discriminated.

Luong, B.T., Ruggieri, S., Turini, F.: k-NN as an implementation of situation testing for discrimination discovery and prevention. In: SIGKDD'11, pp. 502–510. ACM, (2011)

Disparate Impact

- Measured using risk ratio
  - $DI = \frac{1}{LR_+} = \frac{1-specificity}{sensitivity}$
  - LR_+ is the likelihood ratio of the positive class
- Prevention



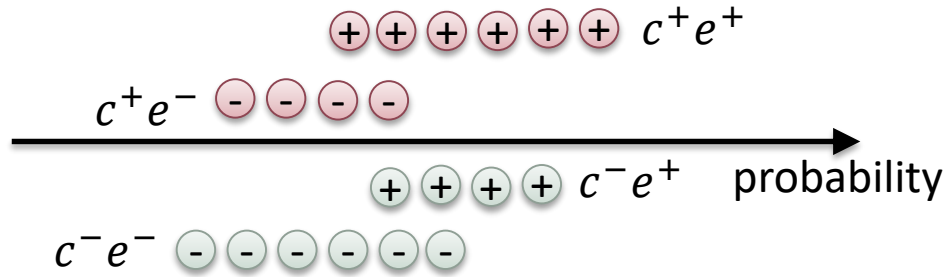
Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C.,Venkata- subramanian, S.: Certifying and removing disparate impact. In: SIGKDD'15, pp. 259–268. ACM, (2015)

Discrimination Prevention

- Data manipulation (Pre-processing)
  - Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. KAIS (2012)
    - Suppression/Massaging/Reweighting/Sampling (uniform vs. preferential sampling)
  - …
- Algorithm tweak (In-processing)
  - Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. DMIN (2010)
  - Kamishima, T., Akaho, S., and Sakuma J.: Fairness-aware Learning through Regularization Approach, ICDMW (2011)
  - …

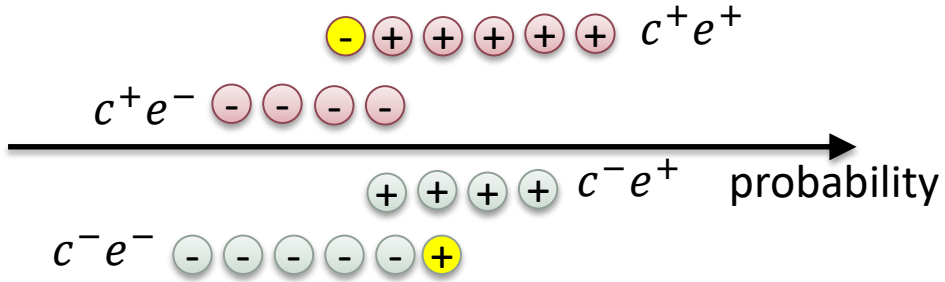
Massaging

- Flip the decision of some individuals according to a ranker
  1. Learn a classifier and estimate the predicted probability of the positive decision of each individual
  2. sort the individuals of four groups according to this probability



$$RD = \frac{6}{10} - \frac{4}{10} = 0.2$$

  3. Flip the decision of individuals that close to the bottom/top

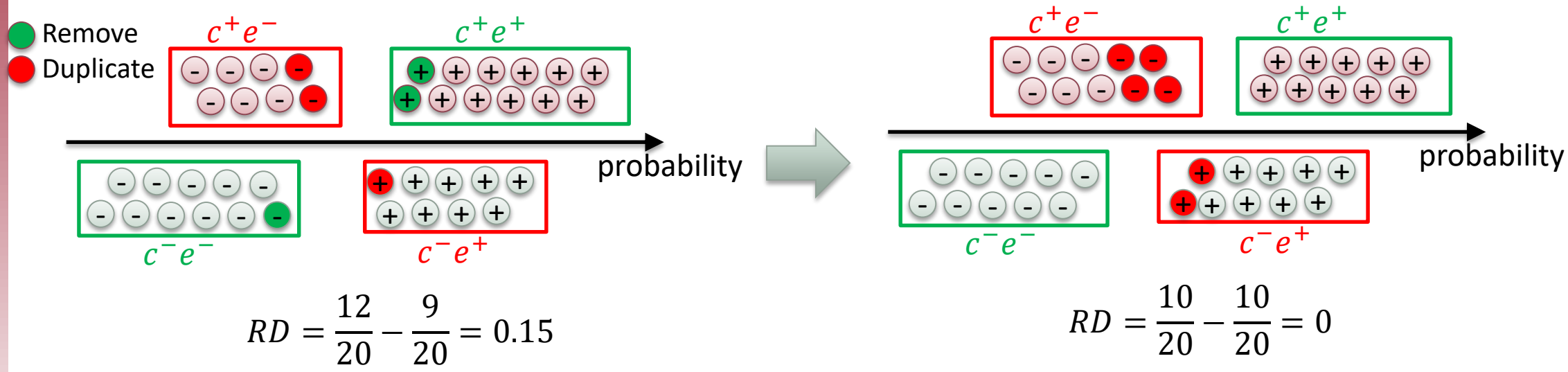


$$RD = \frac{5}{10} - \frac{5}{10} = 0.0$$

Kamiran, F., Calders, T.: Classifying without discriminating. In: IC4'*09 (2019)*

Preferential Sampling

- Partition the data into 4 groups ($c^+e^+, c^-e^-, c^-e^+, c^+e^-$) and two are under-sampled and two over-sampled
- Select and remove/duplicate the individuals close to the top/bottom



$$RD = \frac{12}{20} - \frac{9}{20} = 0.15$$

$$RD = \frac{10}{20} - \frac{10}{20} = 0$$

Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. 33(1), 1–33 (2012)

Summary of Statistical Parity

	system-level	group-level	individual-level
direct discrimination	Kamiran et al. KAIS 2012	Žliobaite et al. ICDM 2011 Hajian et al. TKDE 2013 Mancuhan et al. AIL 2014	Luong et al. KDD 2011
indirect discrimination	Feldman et al. KDD 2015	Hajian et al. TKDE 2013 Mancuhan et al. AIL 2014	

- Lack a unifying framework and a systematic approach for all types of discrimination.
- Gap between association and causation.

Gap Between Association and Causation

- Association does not mean causation, but discrimination is causal.
- The golden rule of causal analysis: no causal claim can be established by a purely statistical method.
- Need causal-aware methods in discovering and preventing discrimination.

Discrimination as Causal Effect

- Probabilistic causation based on Suppes-Bayes Causal Network (Bonchi et al. JDSA 2017)
- Causal modeling-based anti-discrimination framework (Zhang et al. JDSA 2017, Zhang et al. SBP 2017)
  - Zhang, L., Wu, Y., Wu, X.: A causal framework for discovering and removing direct and indirect discrimination. In: IJCAI'17 (2017)
  - Zhang, L., Wu, Y., Wu, X.: Achieving non-discrimination in prediction. arXiv preprint arXiv: 1703.00060 (2017)
  - Zhang, L., Wu, Y., Wu, X.: Achieving non-discrimination in data release. In: SIGKDD'17 (2017)
  - Zhang, L., Wu, Y., Wu, X.: Situation testing-based discrimination discovery: a causal inference approach. In: IJCAI'16 (2016)

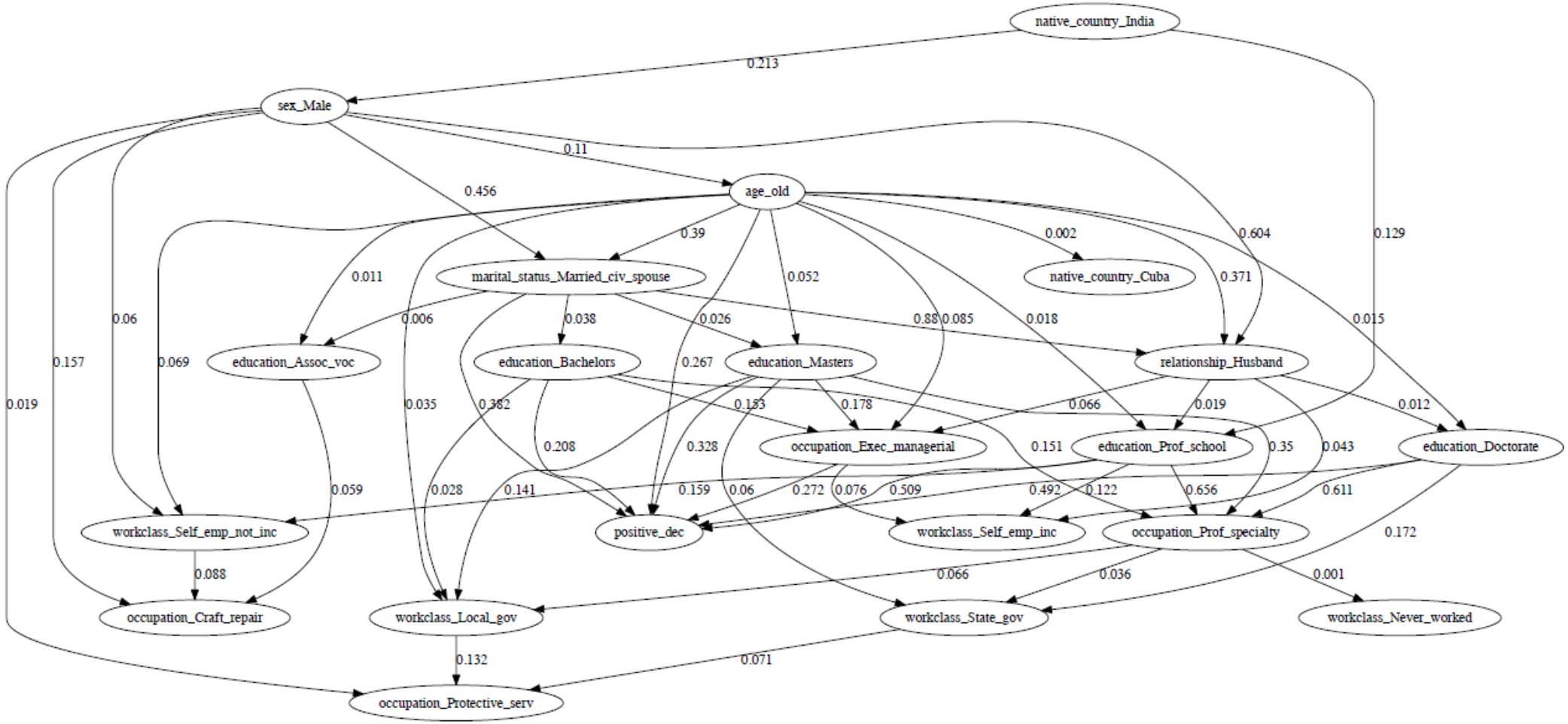
Bonchi, F., Hajian, S., Mishra, B., Ramazzotti, D.: Exposing the probabilistic causal structure of discrimination. Int. J. Data Sci. Anal. 3(1), 1–21 (2017)
Zhang, L., Wu, Y., Wu, X.: On discrimination discovery using causal networks. In: SBP-BRiMS 2016. (2016)
Zhang, L., Wu, X.: Anti-discrimination learning: a causal modeling-based framework. Int. J. Data Sci. Anal. (2017)

Suppes-Bayes Causal Network (SBCN)

- Each node represents an assignment attribute value
- Each arc $v \rightarrow u$ represents the existence of a relation satisfying Suppes' constraints
  - Let v denote cause, u denote effect
  - Temporal priority: $t_v < t_u$
  - Probability raising: $P(u|v) > P(u|\neg v)$
- Each arc is labeled with a positive weight $p(u|v) - p(u|\neg v)$

Bonchi, F., Hajian, S., Mishra, B., Ramazzotti, D.: Exposing the probabilistic causal structure of discrimination. Int. J. Data Sci. Anal. 3(1), 1–21 (2017)

A SBCN Example



Discrimination Score using SBCN

- Discrimination score
  - $ds^{-}(v) = \frac{rw_{v \to e^{-}}}{n}$
  - v is a node of SBCN (e.g. female), e^{-} is the node of negative decision, $rw_{v->e^{-}}$ is the number of random walks from v to e^{-} that earlier than e^{+}, n is the number of random walks from v to e^{+} and from v to e^{-}.
- Generalized score for individual and subgroup discrimination
  - $gds^{-}(v_1, \dots, v_n) = \frac{ppr(e^{-}|v_1, \dots v_n)}{ppr(e^{-}|v_1, \dots v_n) + ppr\left(e^{+}\middle|v_1, \dots v_n\right)}$
  - $ppr(e^{-}|v_1, \dots v_n)$ is output of personalized PageRank.
- Limitations
  - The constructor of SBCN is impractical with large attribute-value pairs.
  - It is unclear how the number of random walks is related to meaningful discrimination metric.

Resources

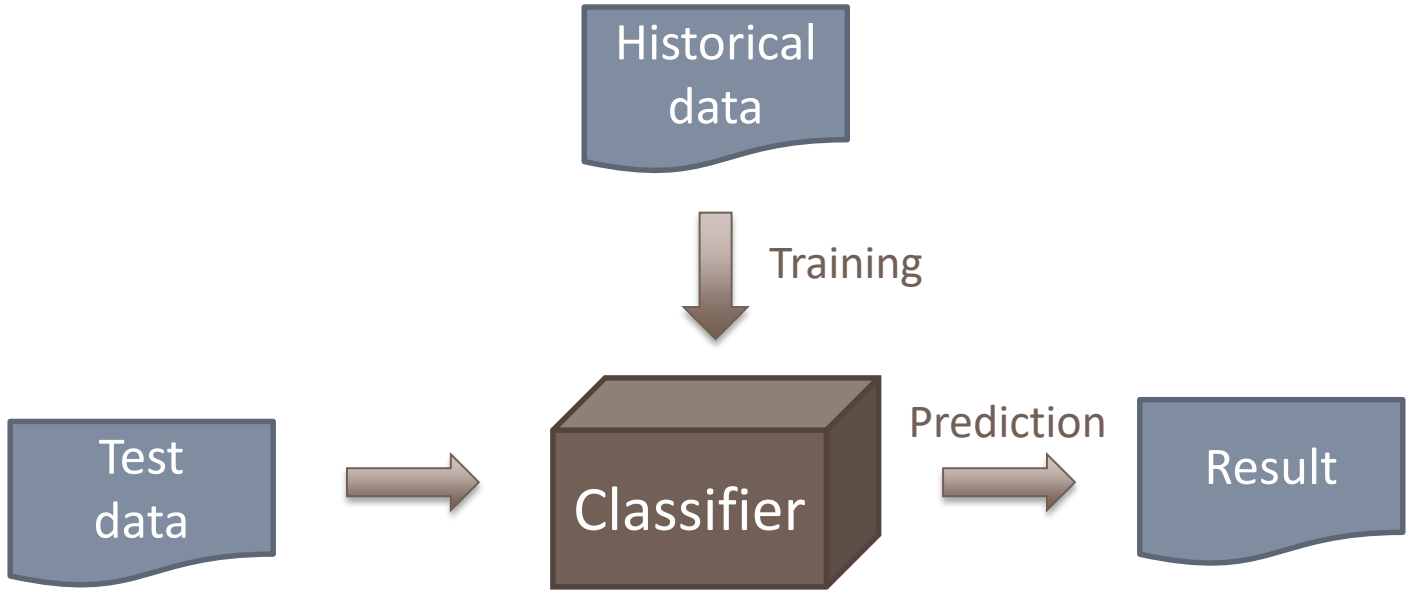
- Tutorials and keynotes
  - Hajian, S., Bonchi, F., & Castillo, C. Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. Tutorial of KDD 2016
  - Abiteboul, S., Miklau, G., & Stoyanovich J. Data Responsibly: Fairness, Neutrality and Transparecy in Data Analysis, Tutorial of EDBT 2016
  - Dwork, C. What's Fair. Keynote of KDD 2017
- Survey papers and books
  - Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review 29(05), 582–638 (2014)
  - Magnani, L., Board, E., Longo, G., Sinha, C., & Thagard, P.: Discrimination and Privacy in the Information Society. Springer (2013)
  - Zhang, L., Wu, X.: Anti-discrimination learning: a causal modeling-based framework. Int. J. Data Sci. Anal. (2017)
- Workshops/Symposiums
  - Fairness, Accountability, and Transparency in Machine Learning (FATML)
  - Machine Learning and The Law

Motivating Examples (ME1)

- How to deal with indirect discrimination due to redlining attributes?
- Assume a bank makes loan decisions based on the areas of residence of the applicants.

Motivating Examples (ME2)

- How to build discrimination-free predictive models?



- Assumption: a classifier learned from a discrimination-free training data will also be discrimination-free.
- Whether and to what extend this assumption holds?

Motivating Examples (ME3)

- How to ensure non-discrimination in data release under all possible scenarios?
- How to identify meaningful partitions?

gender	female	male
admission (%)	37%	47%

$P(e^+|c^+) - P(e^+|c^-) = 0.1$

major	CS				EE			
test score	L		H		L		H	
gender	female	male	female	male	female	male	female	male
admission (%)	20%	20%	50%	50%	40%	40%	70%	70%

$P(e^+|c^+, \{\text{CS}, \text{L}\}) - P(e^+|c^-, \{\text{CS}, \text{L}\}) = 0$

gender	female	male
admission (%)	43%	43%

$P(e^+|c^+) - P(e^+|c^-) = 0$

major	CS				EE			
test score	L		H		L		H	
gender	female	male	female	male	female	male	female	male
admission (%)	30%	36%	50%	40%	40%	45%	60%	50%

$P(e^+|c^+, \{\text{CS}, \text{L}\}) - P(e^+|c^-, \{\text{CS}, \text{L}\}) = 0.06$

Motivating Examples (ME4)

- How to find paired individuals for situation testing in individual discrimination?

No.	gender	major	score	height	weight	ad.
1	F	CS	B	low	low	reject
2	M	CS	B	median	median	admit
3	F	CS	A	low	low	reject
4	M	CS	A	median	median	admit
5	F	CS	C	low	median	reject
6	M	CS	C	median	median	reject
7	M	EE	B	low	low	reject

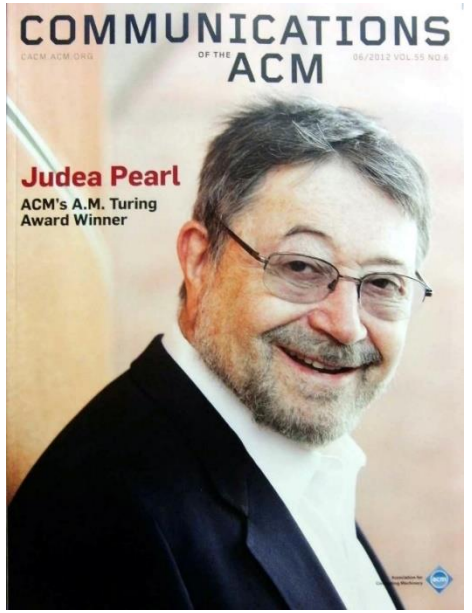
- Which one is closest to 1? 2 or 3 or 7?

Outline

- Part I: Introduction
- Part II: Causal Modeling Background
  - From Statistics to Causal Inference
  - Structural Causal Model and Causal Graph
  - Causal Effect Inference
- Part III: Anti-Discrimination Learning
- Part IV: Challenges and Directions for Future Research

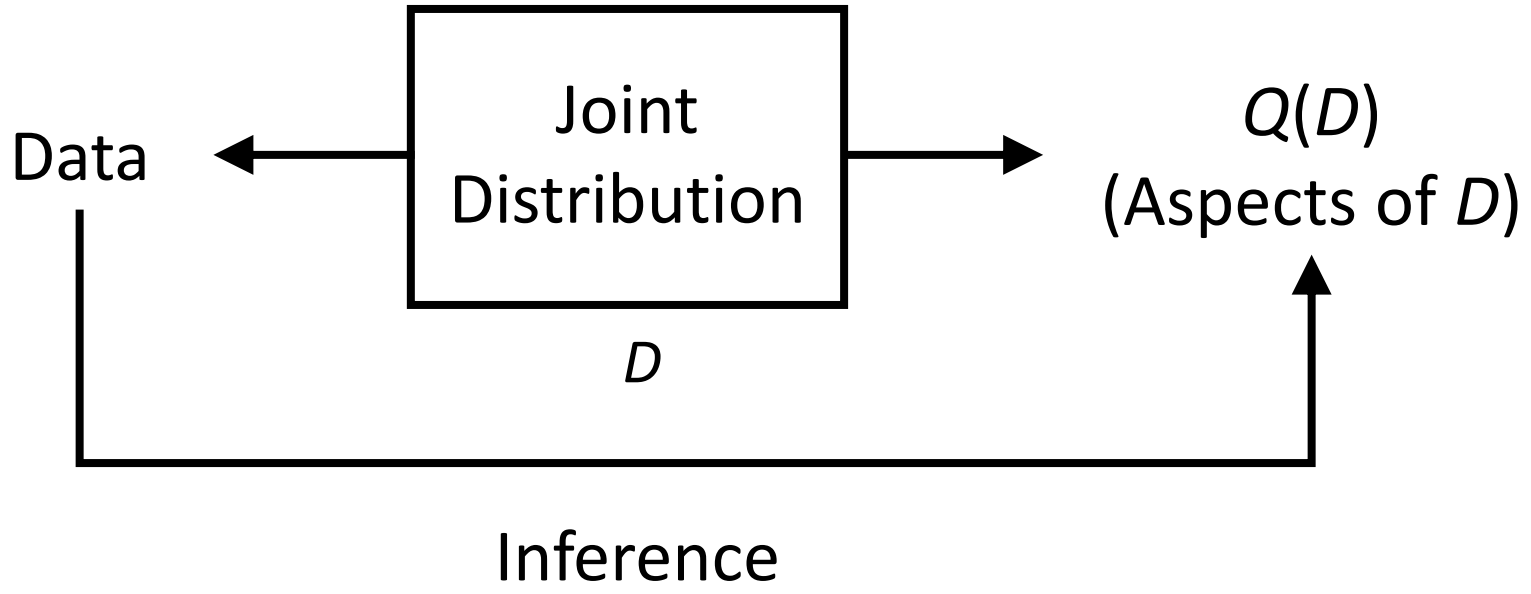
Lessons of Causal Inference (*Pearl*)

1. No causes in — no causes out
2. $\left.\begin{matrix}\text{Data} \\ \text{Causal assumptions/knowledge}\end{matrix}\right\} \Rightarrow$ causal conclusions
3. Causal assumptions/knowledge cannot be expressed in the mathematical language of standard statistics.
4. Need ways of encoding causal assumptions/knowledge mathematically and test their implications.



From Statistics to Causal Inference

- Traditional statistical inference paradigm:



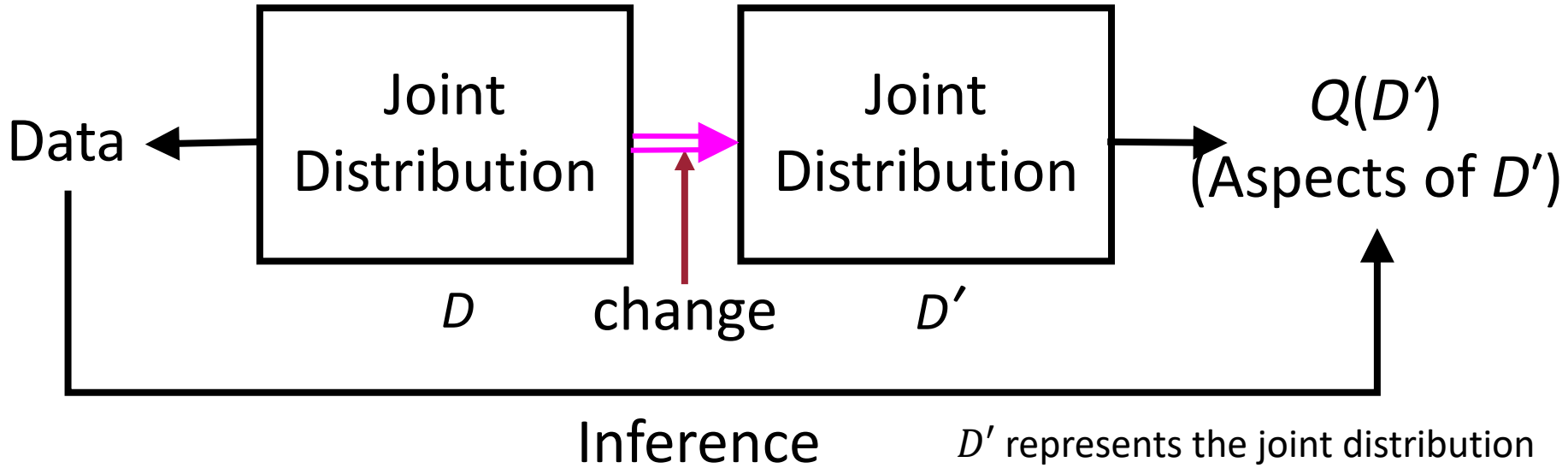
- What is the probability of getting Grade A for the students who study 1 hour each day?

Estimate $Q(D) = P_D(E = \text{'A'} \mid H = 1)$

E (Exam Grade)
H (Hour of Study)
I (Interest)
W (Working Strategy)

From Statistics to Causal Inference

- What is the probability of getting Grade A if a new policy requires all students to study 2 hours each day?
  - The question cannot be solved by statistics.

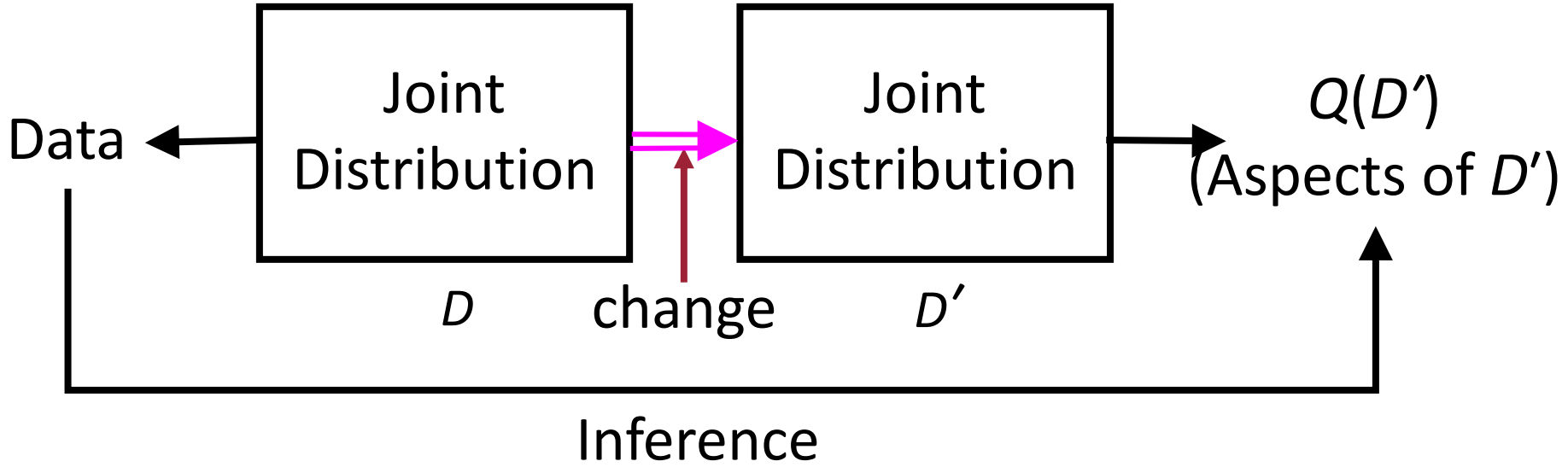


D' represents the joint distribution after adopting the new policy.

Estimate $Q(D') = P_{D'}(E = \text{'}A\text{'})$

From Statistics to Causal Inference

- What is the probability of getting Grade A if a new policy requires all students to study 2 hours each day?
  - The question cannot be solved by statistics.

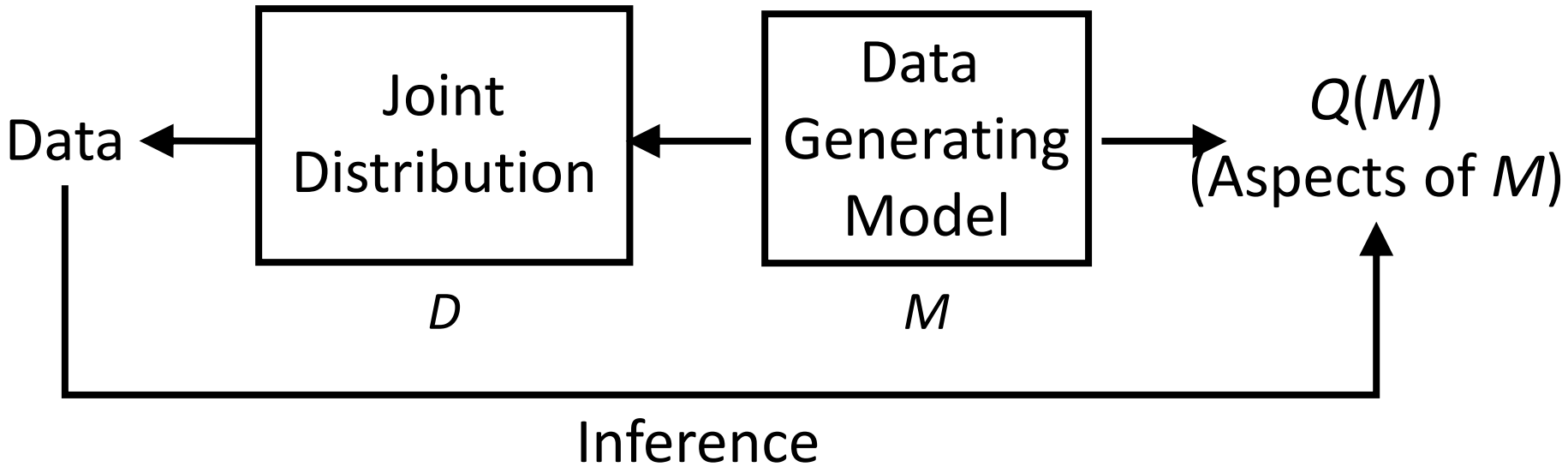


$$P_{D'}(E = \text{‘}A\text{’}) \neq P_D(E = \text{‘}A\text{’} \mid H = 2)$$

The probability of getting Grade A of the students who study 2 hours each day at the first place.

From Statistics to Causal Inference

- Causal inference

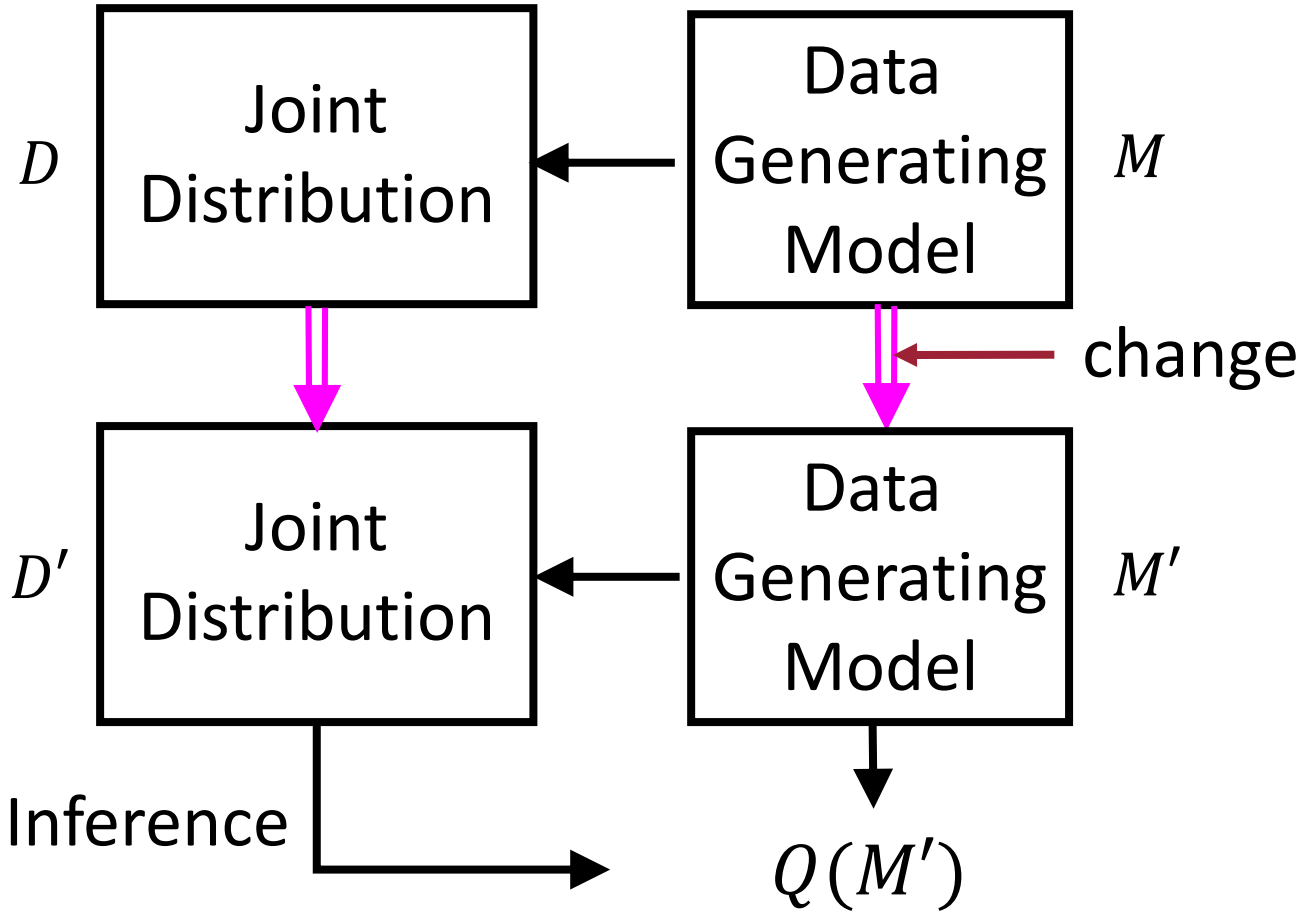


M – Data generation model that encodes the causal assumptions/knowledge.

D – model of data, M – model of reality

From Statistics to Causal Inference

- Causal inference



Outline

- Part I: Introduction
- Part II: Causal Modeling Background
  - From Statistics to Causal Inference
  - Structural Causal Model and Causal Graph
  - Causal Effect Inference
- Part III: Anti-Discrimination Learning
- Part IV: Challenges and Directions for Future Research

Structural Causal Model

- A theory of inferred causation.
- Describe how causal relationships can be inferred from nontemporal statistical data if one makes certain assumptions about the underlying process of data generation.
- Developed since 1988, still growing at an increasing speed.

Structural Causal Model

- A causal model is triple $M =< \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F} >$, where
  - $\boldsymbol{U}$ is a set of exogenous (hidden) variables whose values are determined by factors outside the model;
  - $\boldsymbol{V} = \{X_1, \cdots, X_i, \cdots\}$ is a set of endogenous (observed) variables whose values are determined by factors within the model;
  - $\boldsymbol{F} = \{f_1, \cdots, f_i, \cdots\}$ is a set of deterministic functions where each f_i is a mapping from $\boldsymbol{U} \times (\boldsymbol{V} \setminus X_i)$ to X_i. Symbolically, f_i can be written as

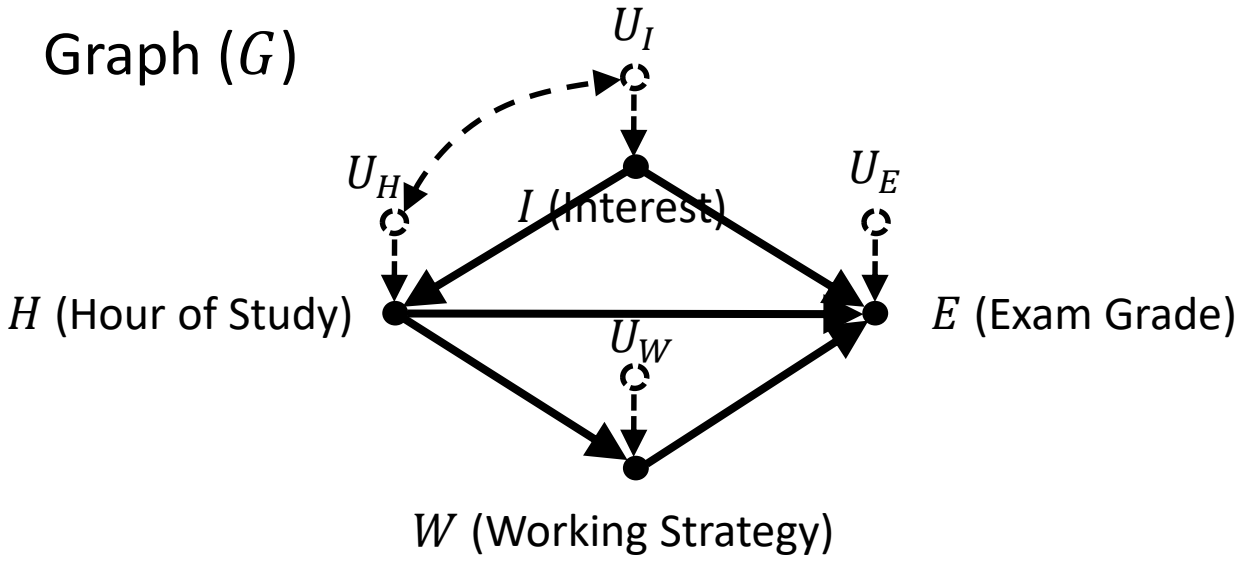
$$x_i = f_i(pa_i, u_i)$$

Causal Graph

- Each causal model M is associated with a direct graph G, where
  - Each node represents a variable.
  - Each direct edge represents the potential direct causal relationship.
  - Absence of direct edge represents zero direct causal relationship.
  - Standard terminology
    - parent, child, ancestor, descendent, path, direct path

A Causal Model and Its Graph

Graph (G)



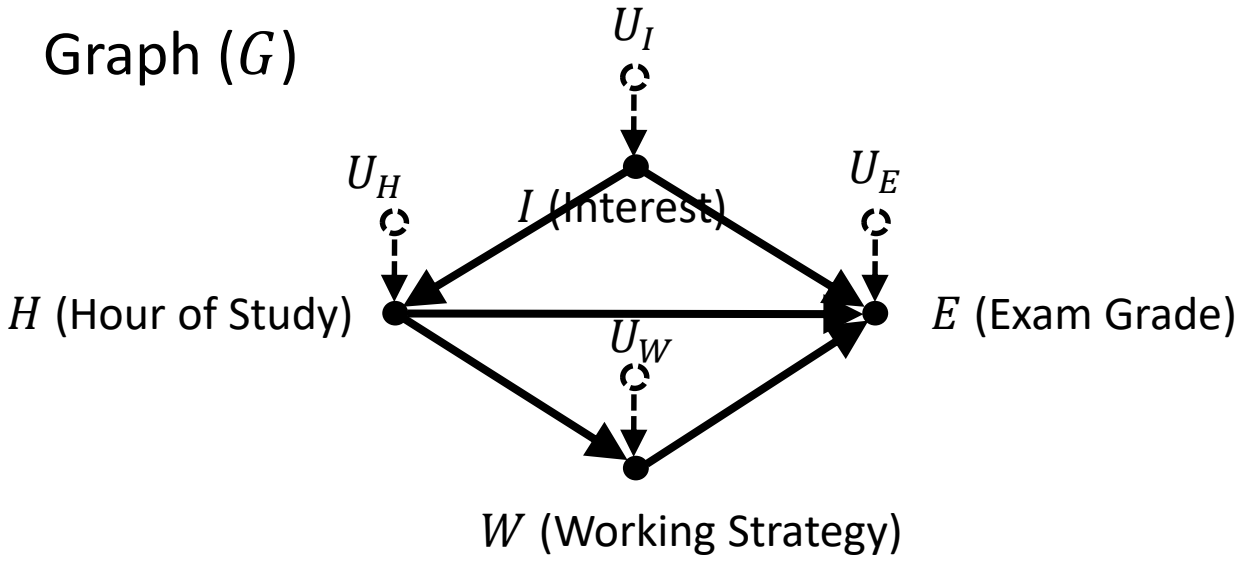
Model (M)

$$i = f_I(u_I)$$
$$h = f_H(i, u_H)$$
$$w = f_W(h, u_W)$$
$$e = f_E(i, h, w, u_E)$$

U_I and U_H are correlated.

A Markovian Model and Its Graph

Graph (G)



Model (M)

$$i = f_I(u_I)$$
$$h = f_H(i, u_H)$$
$$w = f_W(h, u_W)$$
$$e = f_E(i, h, w, u_E)$$

U_I, U_H, U_W, U_E are mutually independent

Markovian Model

- A causal model is Markovian if
  1. The causal graph is acyclic (i.e., DAG);
  2. All variables in $\boldsymbol{U}$ are mutually independent.

Equivalent expression

Each node X is conditionally independent of its non-descendants given its parents $Pa(X)$.

Known as the local Markov condition (e.g., in Bayesian network), or causal Markov condition in the context of causal modeling.

Conditional Independence

- Two random (categorical) variables X and Y are called independent, if for each values of X and Y, x and y,
  - $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$
  - Denoted by $X \perp Y$
- Two random (categorical) variables X and Y are called conditionally independent given Z, if for each values of (X, Y, Z), (x, y, z),
  - $P(X = x, Y = y|Z = z) = P(X = x|Z = z) \cdot P(Y = y|Z = z)$
  - Denoted by $X \perp Y|Z$ or $(X \perp Y|Z)_D$

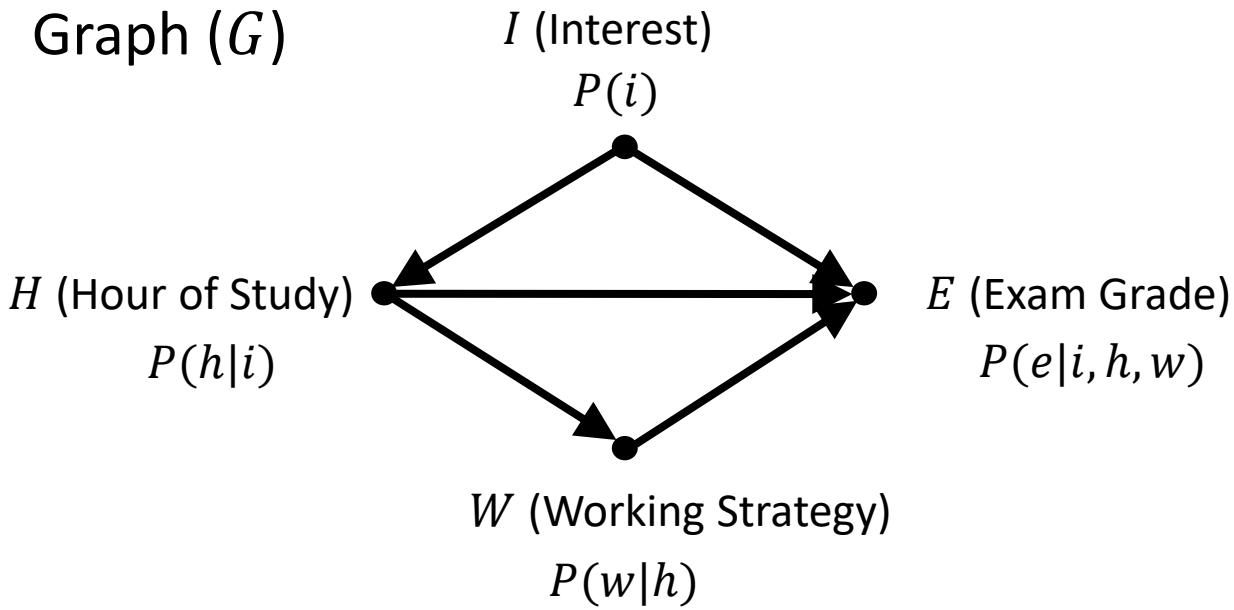
- Note: conditional independence neither implies nor is implied by independence.

Causal Graph of Markovian Model

Each node is associated with a conditional probability table (CPT)

$$P(x_i|pa_i)$$

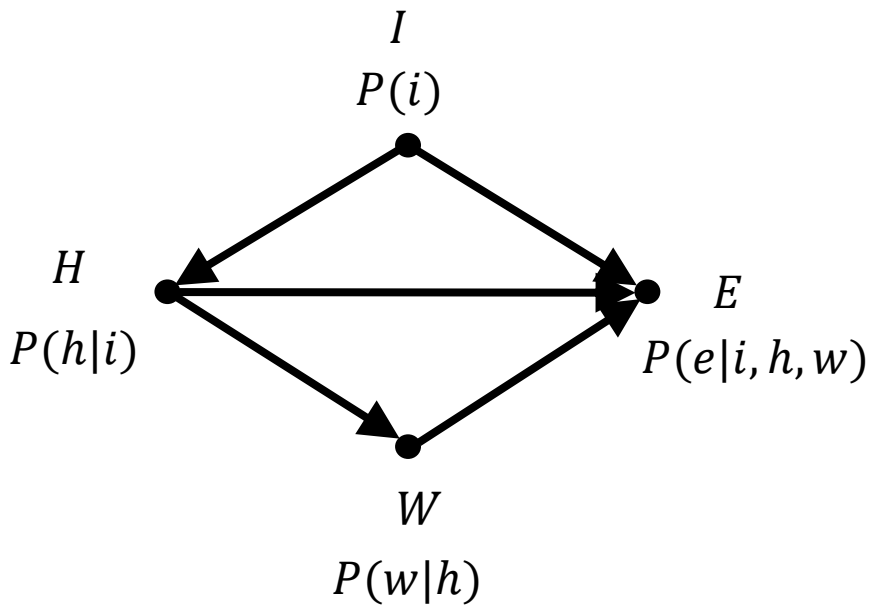
Graph (G)



Factorization Formula

- In a Markovian model, the joint distribution over all attributes can be computed using the factorization formula

$$P(\boldsymbol{v}) = \prod_{X \in \boldsymbol{V}} P(x|Pa(X))$$



$$P(i,h,w,e) = P(i)P(h|i)P(w|h)P(e|i,h,w)$$

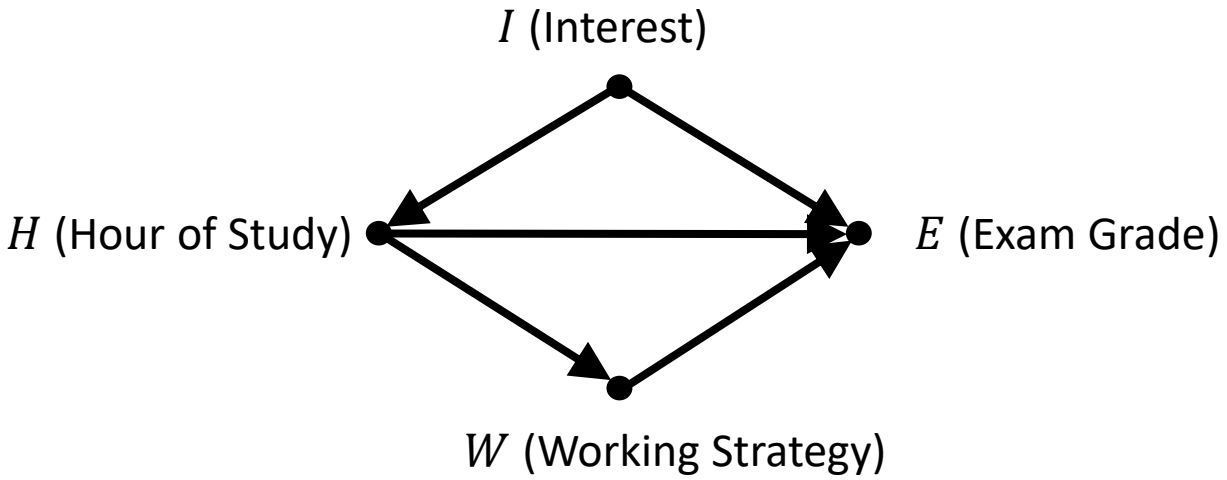
$$P(e) = \sum_{I,H,W} P(i)P(h|i)P(w|h)P(e|i,h,w)$$

Outline

- Part I: Introduction
- Part II: Causal Modeling Background
  - From Statistics to Causal Inference
  - Structural Causal Model and Causal Graph
  - Causal Effect Inference
- Part III: Anti-Discrimination Learning
- Part IV: Challenges and Directions for Future Research

Statistical Inference

- What is the probability of getting grade A if we see that the study hour is 1?



- Find $P(E = \text{'}A\text{'} | H = 1)$

Causal Inference

- What is the probability of getting grade A if we change the study hour to 2?
- The above probability does not equal to $P(E = \text{`}A\text{'} | H = 2)$, i.e., the conditional probability of getting grade A given study hour equals to 2.

Intervention and *do*-Operator

- The basic operation of manipulating a causal model.
  - Simulate the manipulation of the physical mechanisms by some physical interventions or hypothetical assumptions.
- It is treated as a local modification to equations.
- Forces some variables $\boldsymbol{X} \in \boldsymbol{V}$ to take certain constants $\boldsymbol{x}$.
- Mathematically formalized as $do(\boldsymbol{X} = \boldsymbol{x})$ or simply $do(\boldsymbol{x})$.
- The effect of intervention of all other variables $\boldsymbol{Y}$ is mathematically represented by the post-intervention distribution of $\boldsymbol{Y}$, denoted by $P(\boldsymbol{Y} = \boldsymbol{y}|do(\boldsymbol{X} = \boldsymbol{x}))$ or simply $P(\boldsymbol{y}|do(\boldsymbol{x}))$.
  - Sometimes use $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{x})]$.

Intervention and *do*-Operator

- In the Markovian model, the post-intervention distribution $P(\boldsymbol{y}|do(\boldsymbol{x}))$ can be calculated from the CPTs, known as the truncated factorization:

$$P(\boldsymbol{y}|do(\boldsymbol{x})) = \prod_{Y \in \boldsymbol{Y}} P(y|Pa(Y))\delta_{\boldsymbol{X}=\boldsymbol{x}}$$

  - where $\delta_{\boldsymbol{X}=\boldsymbol{x}}$ means assigning attributes in $\boldsymbol{X}$ involved in the term ahead with the corresponding values in $\boldsymbol{x}$.

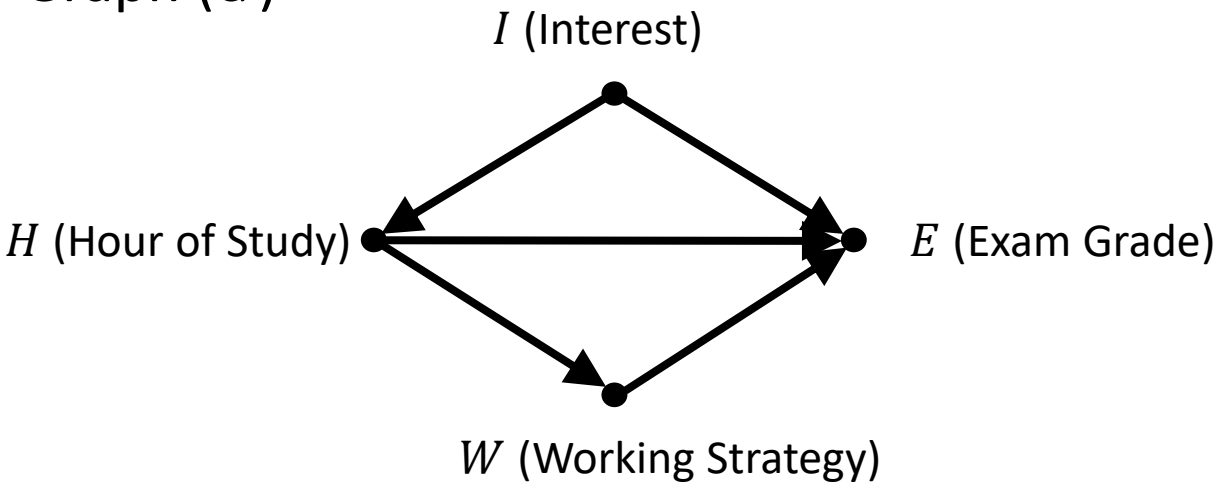
- Specifically, for a single attribute Y given an intervention on a single attribute X,

$$P(y|do(x)) = \sum_{\substack{\boldsymbol{V}\setminus\{X,Y\} \\ Y=y}} \prod_{V \in \boldsymbol{V}\setminus\{X\}} P(v|Pa(V))\delta_{X=x}$$

Intervention and *do*-Operator

- What is the probability of getting grade A if we change the study hour to 2?

Graph (G)



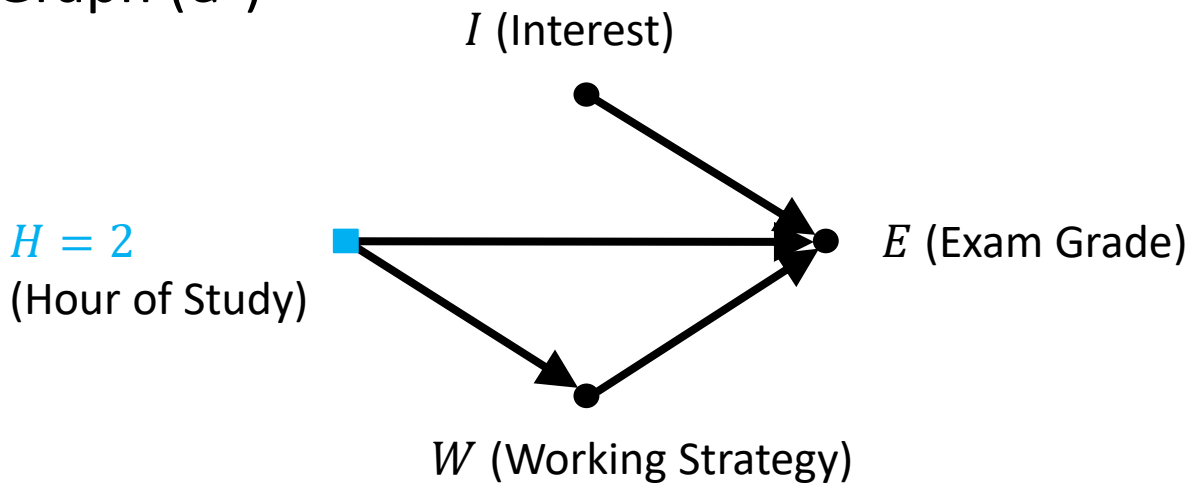
Model (M)

$$i = f_I(u_I)$$
$$h = f_H(i, u_H)$$
$$w = f_W(h, u_W)$$
$$e = f_E(i, h, w, u_E)$$

Intervention and *do*-Operator

- What is the probability of getting grade A if we change the study hour to 2, i.e., $do(H = 2)$?

Graph (G')



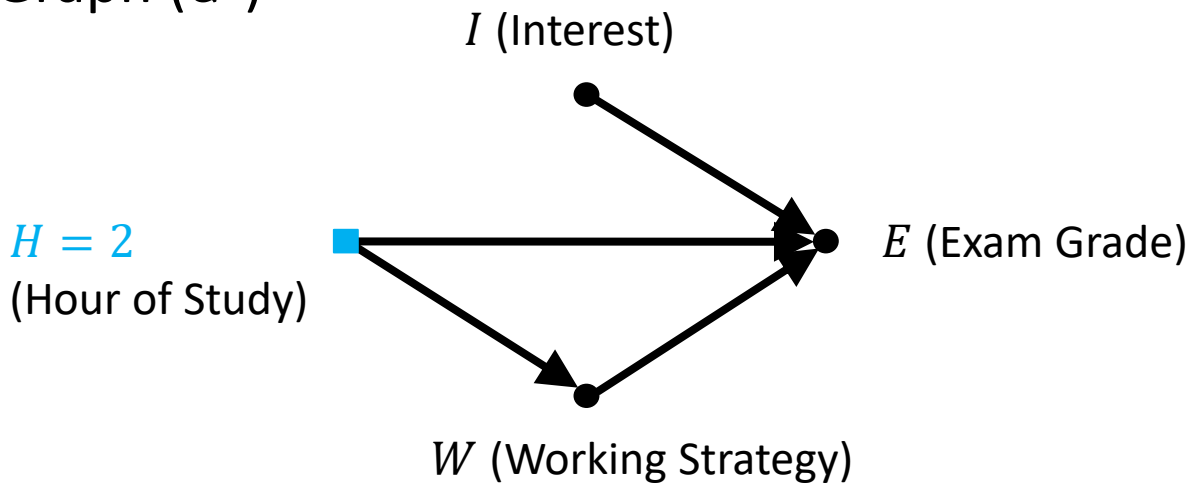
Model (M')

$$i = f_I(u_I)$$
$$h = 2$$
$$w = f_W(h, u_W)$$
$$e = f_E(i, h, w, u_E)$$

- Find $P(E = \text{'}A\text{'} | do(H = 2))$

Intervention and *do*-Operator

Graph (G')



Model (M')

$i = f_I(u_I)$

$h = 2$

$w = f_W(h, u_W)$

$e = f_E(i, h, k, u_E)$

$$P(y|do(x)) = \sum_{\substack{\boldsymbol{V}\setminus\{X,Y\} \\ Y=y}} \prod_{V\in\boldsymbol{V}\setminus\{X\}} P(v|Pa(V))\delta_{X=x}$$

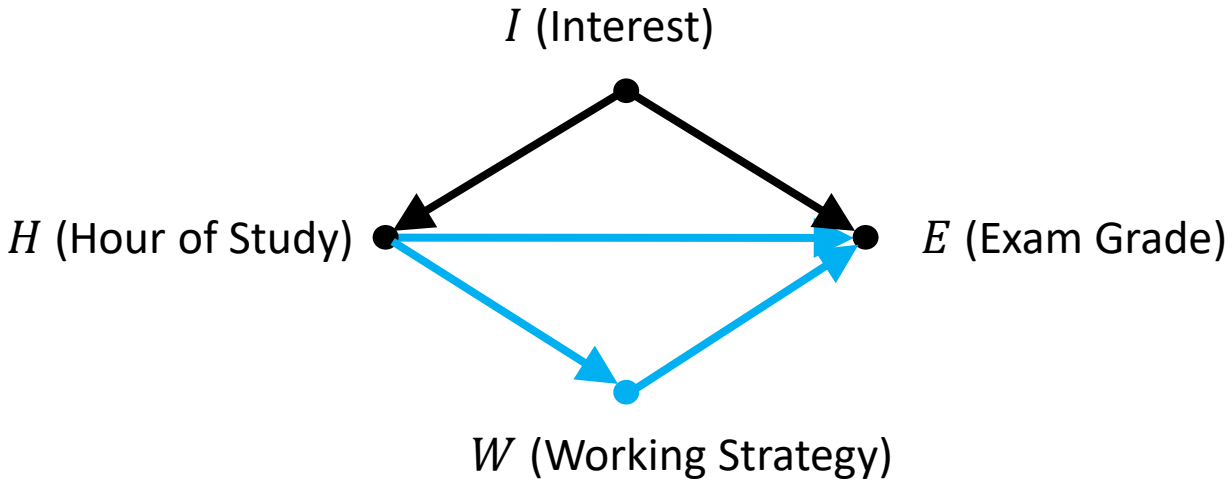
$$P(E = \text{‘}A\text{’}|do(H = 2)) = \sum_{I,W} P(i)P(w|H = 2)P(E = \text{‘}A\text{’}|i, H = 2, w)$$

Total Causal Effect

- A common measure of the causal effect of X on Y is given by

$$TE(x_2, x_1) = P(y|do(x_2)) - P(y|do(x_1))$$

- Called the total causal effect as it measures the causal effect transmitted along all causal paths from X to Y.



$$TE(H = 2, H = 1)$$
$$= P(E = \text{'A'}|do(H = 2)) - P(E = \text{'A'}|do(H = 1))$$
$$= \sum_{I,W} P(i)P(w|H = 2)P(E = \text{'A'}|i, H = 2, w) - \sum_{I,W} P(i)P(w|H = 1)P(E = \text{'A'}|i, H = 1, w)$$

Path-Specific Effect

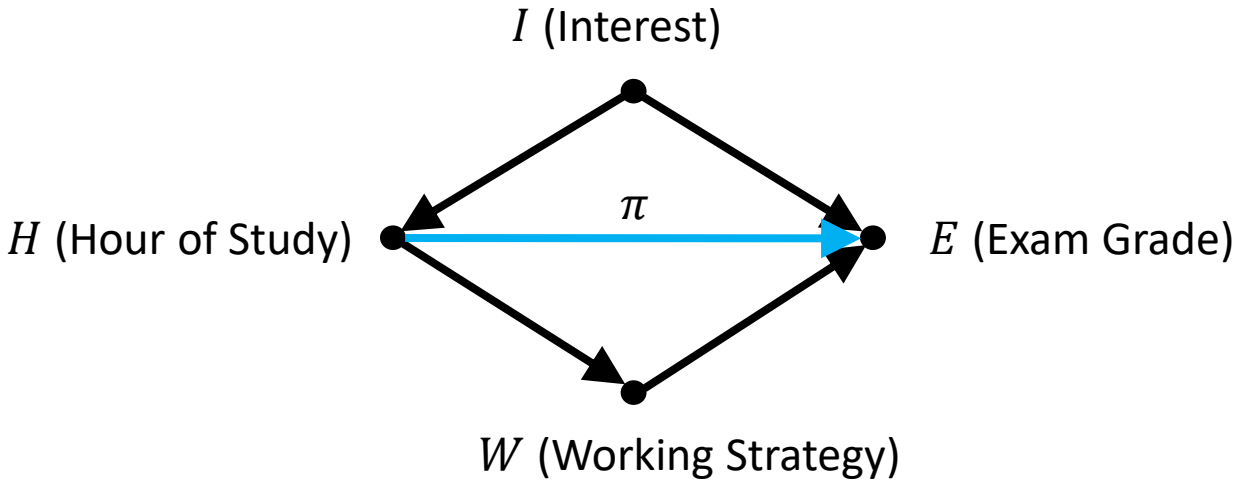
- Path-specific effect measures the causal effect transmitted along certain paths.
- Given a subset of causal paths π, the causal effect of X on Y transmitted along π is denoted by

$$SE_{\pi}(x_2, x_1) = P(y|do(x_2|_{\pi})) - P(y|do(x_1))$$

  - $P(y|do(x_2|_{\pi}))$ denotes the distribution of Y after an intervention of changing X from x_1 to x_2 with the effect transmitted along π.

Path-Specific Effect

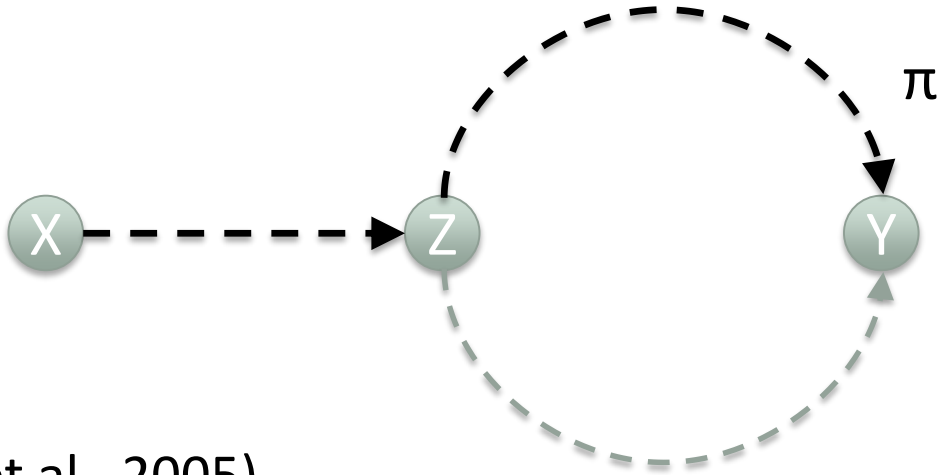
- The causal effect of Study Hour on Exam Grade while keeping the Working Strategy unchanged.
- Measures the causal effect of H on E transmitted along the direct path (π).



$$SE_{\pi}(H=2, H=1)$$
$$= P(E = \mathrm{A}|do(H = 2|_{\pi})) - P(E = \mathrm{A}|do(H = 1))$$
$$= \sum_{I,W} P(i)P(w|H=1)P(E=A|i,H=2,w) - \sum_{I,W} P(i)P(w|H=1)P(E=A|i,H=1,w)$$

Path-Specific Effect

- Identifiability: The path-specific effect can be computed from the observational data if and only if the recanting witness criterion is not satisfied.
- Recanting witness criterion: Given a path set π, let *Z* be a node in the graph such that: 1) there exists a path from *X* to *Z* which is a segment of a path in π; 2) there exists a path from *Z* to *Y* which is a segment of a path in π; 3) there exists another path from *Z* to *Y* which is not a segment of any path in π. Then, the recanting witness criterion for the π-specific effect is satisfied with *Z* as a witness.



- Refer to (Avin et al., 2005).

Avin, C., Shpitser, I., Pearl, J.: Identifiability of path-specific effects. In: IJCAI'05, pp. 357–363. (2005)

Techniques in Causal Modeling

- Markovian model
- Intervention and *do*-operator
- Path-specific effect
- *d*-separation (covered later)
- Semi-Markovian model
- Counterfactual analysis

Outline

- Part I: Introduction
- Part II: Causal Modeling Background
- Part III: Anti-Discrimination Learning
  - Causal Modeling-Based Anti-Discrimination Framework
  - Direct and Indirect Discrimination
  - Non-Discrimination in Prediction
  - Non-Discrimination in Data Release
  - Individual Discrimination
- Part IV: Challenges and Directions for Future Research

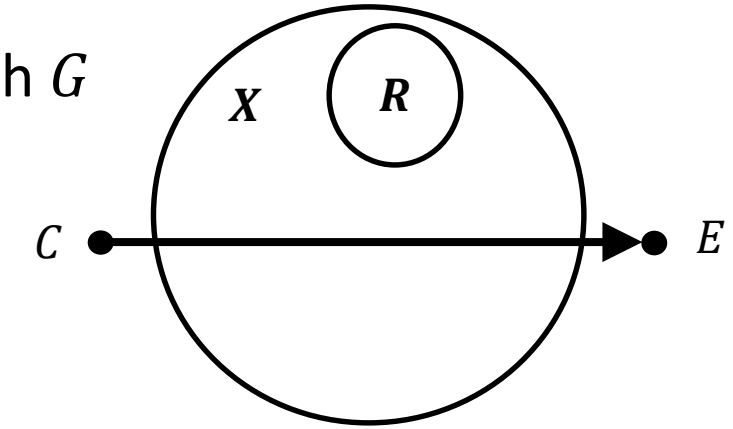
Causal Modeling-Based Anti-Discrimination Framework Assumptions

- A Markovian model M representing the data generation mechanism of the system or population.
  - A dataset D drawn from the population.
  - A classifier h trained by D.
- Two reasonable assumptions to make the framework more concise (not theoretically necessary)
  1. The protected attribute C has no parent;
  2. The decision E has no child.

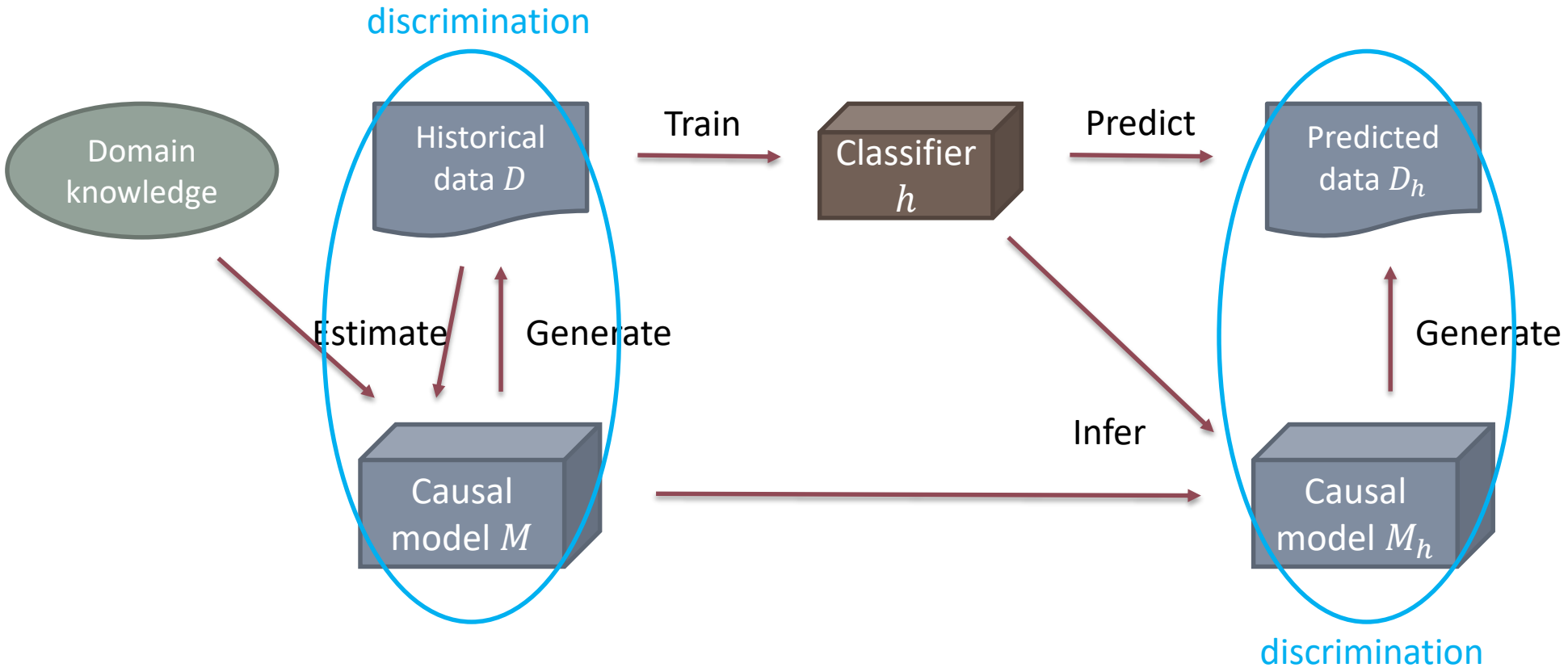
Model M

$$c = f_C(u_C)$$
$$x_i = f_i(pa_i, u_i), i = 1, \cdots, m$$
$$e = f_E(pa_E, u_E)$$

Graph G



Causal Modeling-Based Anti-Discrimination Framework Workflow



Outline

- Part I: Introduction
- Part II: Causal Modeling Background
- Part III: Anti-Discrimination Learning
  - Causal Modeling-Based Anti-Discrimination Framework
  - Direct and Indirect Discrimination
  - Non-Discrimination in Prediction
  - Non-Discrimination in Data Release
  - Individual Discrimination
- Part IV: Challenges and Directions for Future Research

Direct and Indirect Discrimination Discovery and Removal

- Motivating example (ME1): how to deal with indirect discrimination due to redlining attributes?

- Modeling direct and indirect discrimination using the causal model.
- Quantitative discrimination measure and criterion.
- Algorithm for removing direct and indirect discrimination from a dataset.

Zhang, L., Wu, Y., Wu, X.: A causal framework for discovering and removing direct and indirect discrimination. In: IJCAI'17 (2017)

Direct and Indirect Discrimination

- Direct: explicitly based on the protected attribute C.
  - E.g., rejecting a qualified female just because of her gender.
- Indirect: based on apparently neutral non-protected attributes but still results in unjustified distinctions against individuals from the protected group.
  - E.g., redlining, where the residential Zip Code of an individual is used for making decisions such as granting a loan.
  - Redlining attributes $\boldsymbol{R}$: non-protected attributes that can cause indirect discrimination.

Modeling Using CBN

- Direct and indirect discrimination can be captured by the causal effects of C on E transmitted along different paths.
  - Direct discrimination: the causal effect transmitted along the direct path from C to E.
  - Indirect discrimination: the causal effect transmitted along causal paths that pass though the redlining attributes.

Path-Specific Effect

- Given a subset of causal paths π, the path-specific effect is
$$SE_{\pi}(c^{+}, c^{-}) = P(e^{+}|do(c^{+}|_{\pi})) - P(e^{+}|do(c^{-}))$$
  - It is the causal effect transmitted along certain causal paths.
- Measure direct and indirect discrimination based on the path-specific effect.
  - Direct discrimination is measured by $SE_{\pi_d}(c^{+}, c^{-})$ where π_d is the path $C \rightarrow E$.
  - Indirect discrimination is measured by $SE_{\pi_i}(c^{+}, c^{-})$ where π_i contains all the causal paths from C to E through redlining attributes $\boldsymbol{R}$.

Quantitative Measuring

- It is guaranteed SE_{π_d} can be computed

$$SE_{\pi_d}(c^+, c^-) = \sum_{\boldsymbol{V}\backslash\{C,E\}} \left(P(e^+|c^+, Pa(E)\backslash\{C\}) \prod_{V \in \boldsymbol{V}\backslash\{C,E\}} P(v|Pa(V))\delta_{C=c^-} \right) - P(e^+|c^-)$$

- When the recanting witness criterion for SE_{π_i} is not satisfied

$$SE_{\pi_i}(c^+, c^-)$$

$$= \sum_{\boldsymbol{V}\backslash\{C\}} \left(\prod_{G \in \boldsymbol{S}_{\pi_i}} P(g|c^+, Pa(G)\backslash\{C\}) \prod_{H \in \bar{\boldsymbol{S}}_{\pi_i}} P(h|c^-, Pa(H)\backslash\{C\}) \prod_{O \in \boldsymbol{V}\backslash(\{C\}\cup Ch(C))} P(o|Pa(O))\delta_{C=c^-} \right)$$

$$- P(e^+|c^-)$$

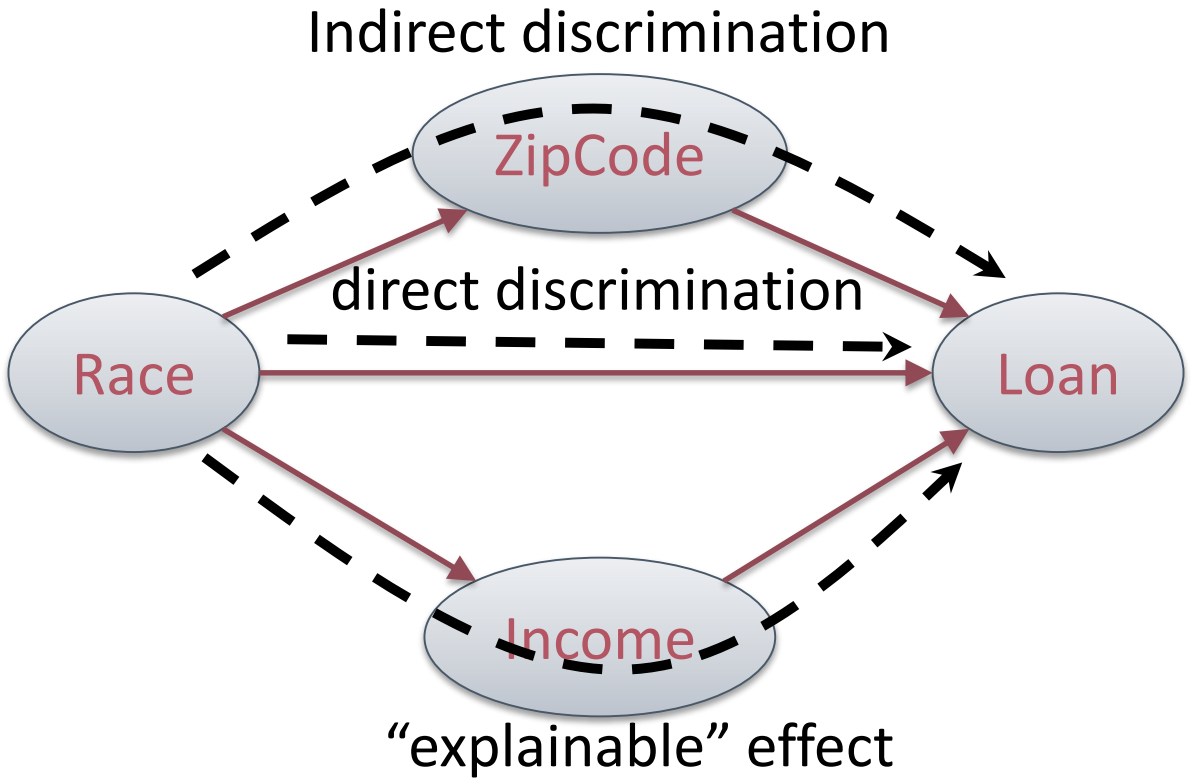
$\boldsymbol{S}_{\pi_i}$: C's children that lie on paths in π_i
$\bar{\boldsymbol{S}}_{\pi_i}$: C's children that don't lie on paths in π_i

- How to deal with the unidentifiable situation is skipped.

Illustrative Example

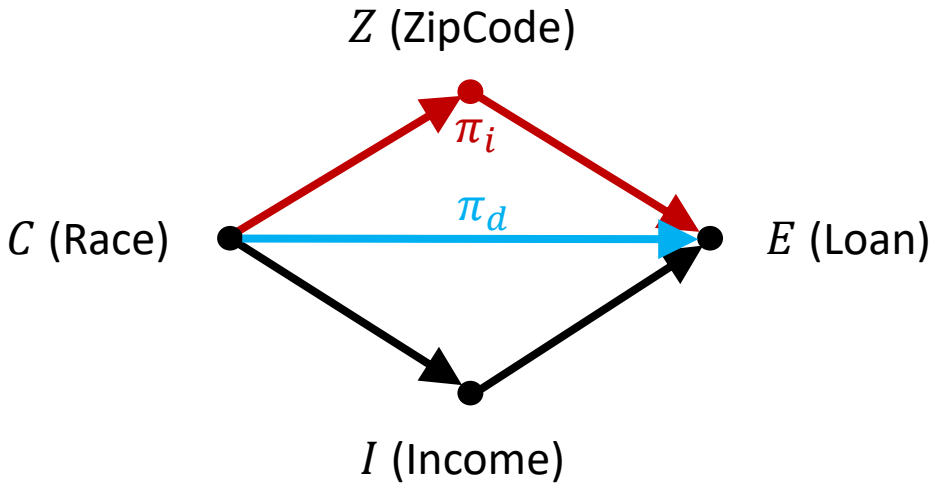
- A bank makes loan decisions based on the Zip Codes, races, and income of the applicants.

- Race: protected attribute
- Loan: decision
- Zip Code: redlining attribute



Quantitative Measuring

- The loan example.



$$SE_{\pi_d}(c^+, c^-) = \sum_{Z,I} \big(P(e^+|c^+, z, i) - P(e^+|c^-, z, i)\big) P(z|c^-) P(i|c^-)$$

$$SE_{\pi_i}(c^+, c^-) = \sum_{Z,I} P(e^+|c^-, z, i) \big(P(z|c^+) - P(z|c^-)\big) P(i|c^-)$$

Discrimination Discovery and Removal

- Path-Specific Effect based Discrimination Discovery (PSE-DD) algorithm
  - Build the causal graph
  - Compute SE_{π_d} and SE_{π_i}
- Path-Specific Effect based Discrimination Removal (PSE-DR) algorithm
  - Modify the CPT of E so that no discrimination exists.
  - Generate a new dataset using the modified graph.
  - Minimize the distance of the joint distributions: quadratic programming.

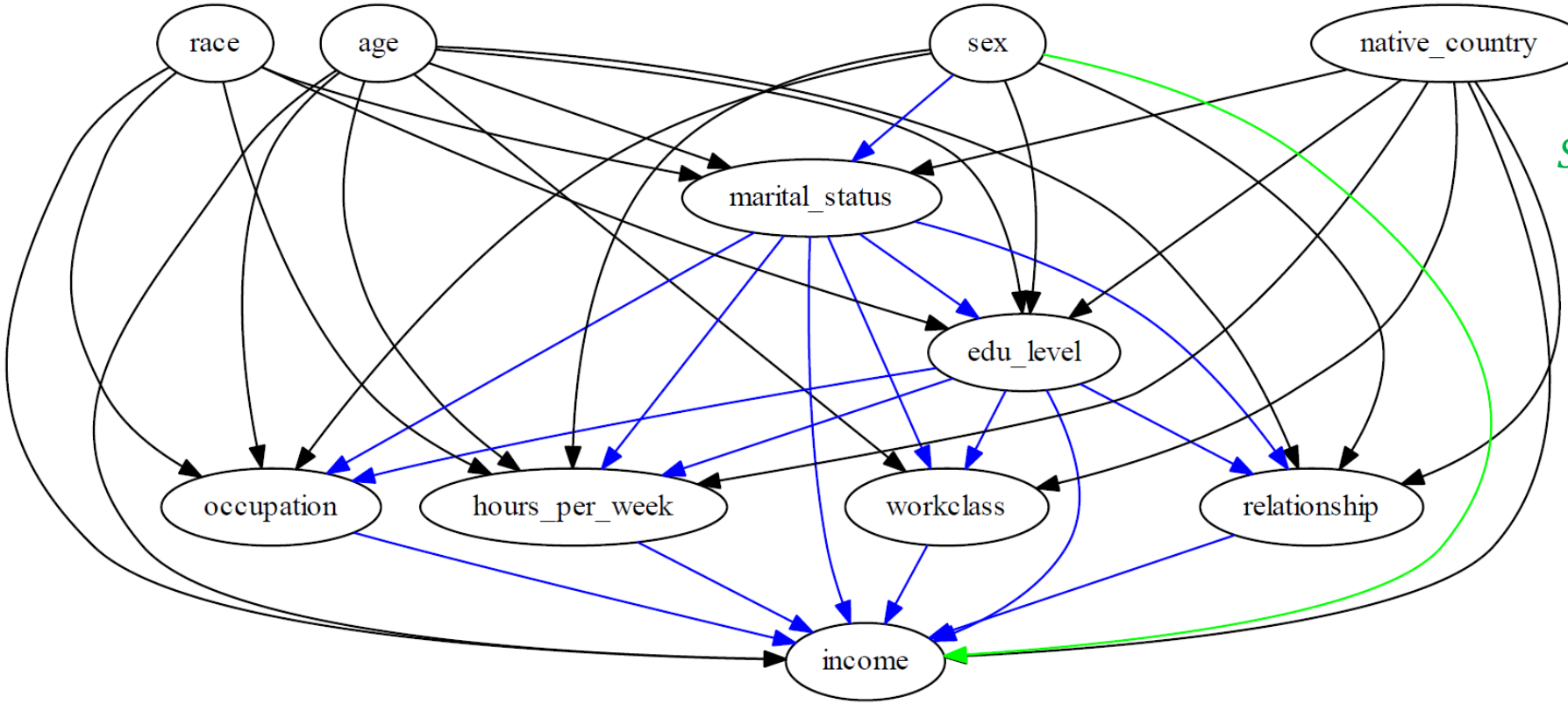
$$\text{minimize} \quad \sum_{\mathbf{V}} \left(P'(\mathbf{v}) - P(\mathbf{v})\right)^2$$

$$\begin{aligned}
\text{subject to} \quad & SE_{\pi_d}(c^+, c^-) \leq \tau, \quad SE_{\pi_d}(c^-, c^+) \leq \tau, \\
& SE_{\pi_i}(c^+, c^-) \leq \tau, \quad SE_{\pi_i}(c^-, c^+) \leq \tau, \\
& \forall Pa(E), \quad P'(e^-|Pa(E)) + P'(e^+|Pa(E)) = 1, \\
& \forall Pa(E), e, \quad Pr'(e|Pa(E)) \geq 0,
\end{aligned}$$

Empirical Evaluation

- Data: Adult dataset

protected attribute: sex decision: income redlining attribute: marital_status



$SE_{\pi_d}(c^+, c^-) = 0.025$

$SE_{\pi_i}(c^+, c^-) = 0.175$

Tool: TETRAD for building the causal graph (using the classic PC algorithm)

Comparison of Different Methods

- Evaluated algorithms:
  - PSE-DD, PSE-DR (Zhang et al. IJCAI 2017)
  - Local massaging (LMSG) and local preferential sampling (LPS) algorithms (Žliobaite et al. ICDM 2011)
  - Disparate impact removal algorithm (DI) (Feldman et al. KDD 2015)
- Local massaging (LMSG) and local preferential sampling (LPS) algorithms still have discrimination.
- Disparate impact removal algorithm (DI) incurs more utility loss.

	Remove Algorithm			
	PSE-DR	*DI*	*LMSG*	*LPS*
Direct	0.013	0.001	-0.142	-0.142
Indirect	0.049	0.050	0.288	0.174
$\chi^2(\times 10^4)$	1.038	4.964	1.924	1.292

Causal Effect vs. Risk Difference

- The total causal effect of C (changing from c^- to c^+) on E is given by

$$TE(c^+, c^-) = P(e^+|do(c^+)) - P(e^+|do(c^-))$$

  - transmitted along all causal paths from C to E.

- Connection with the risk difference

$$TE(c^+, c^-) = P(e^+|c^+) - P(e^+|c^-)$$

Outline

- Part I: Introduction
- Part II: Causal Modeling Background
- Part III: Anti-Discrimination Learning
  - Causal Modeling-Based Anti-Discrimination Framework
  - Direct and Indirect Discrimination
  - Non-Discrimination in Prediction
  - Non-Discrimination in Data Release
  - Individual Discrimination
- Part IV: Challenges and Directions for Future Research

Achieving Non-Discrimination in Prediction

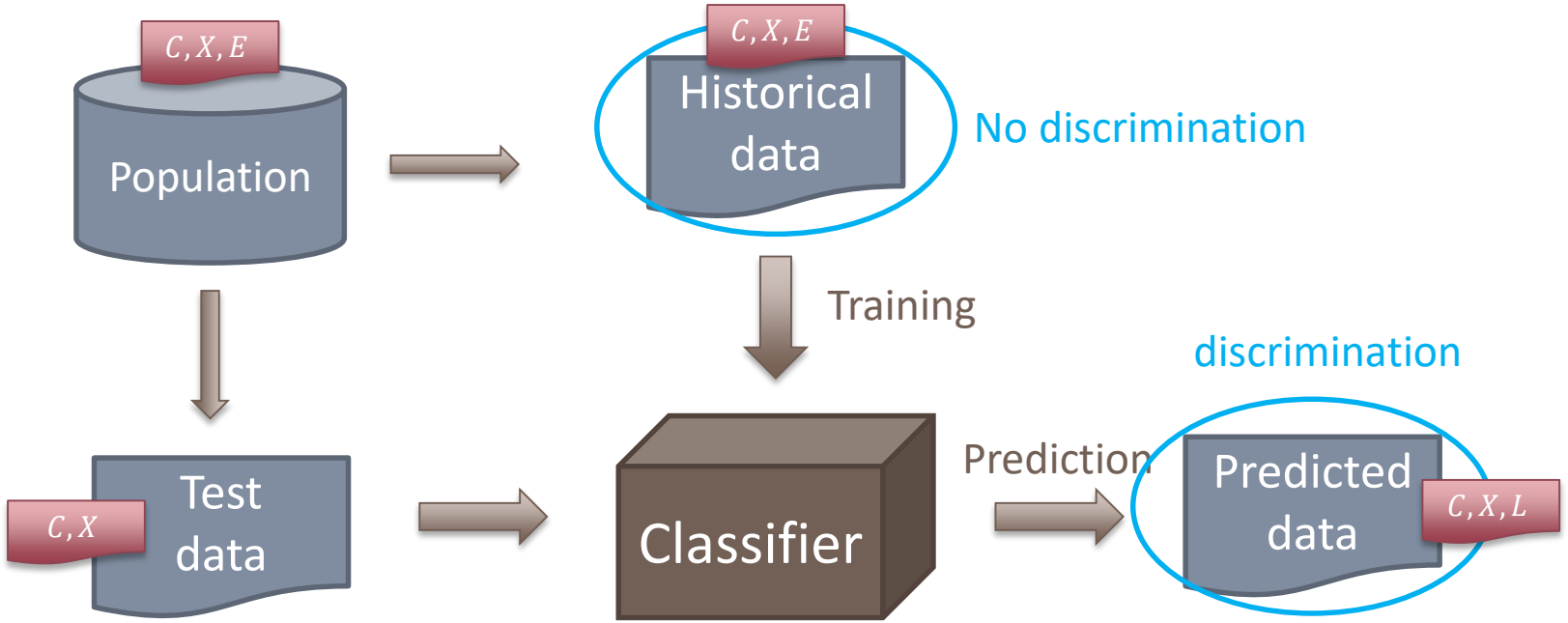
- Motivating example (ME2): will a classifier learned from a discrimination-free training data also be discrimination-free?

- The gap between the discrimination-free training data and the discrimination-free classifier
- Mathematically bound the discrimination in predictions in terms of the training data and the classifier performance.

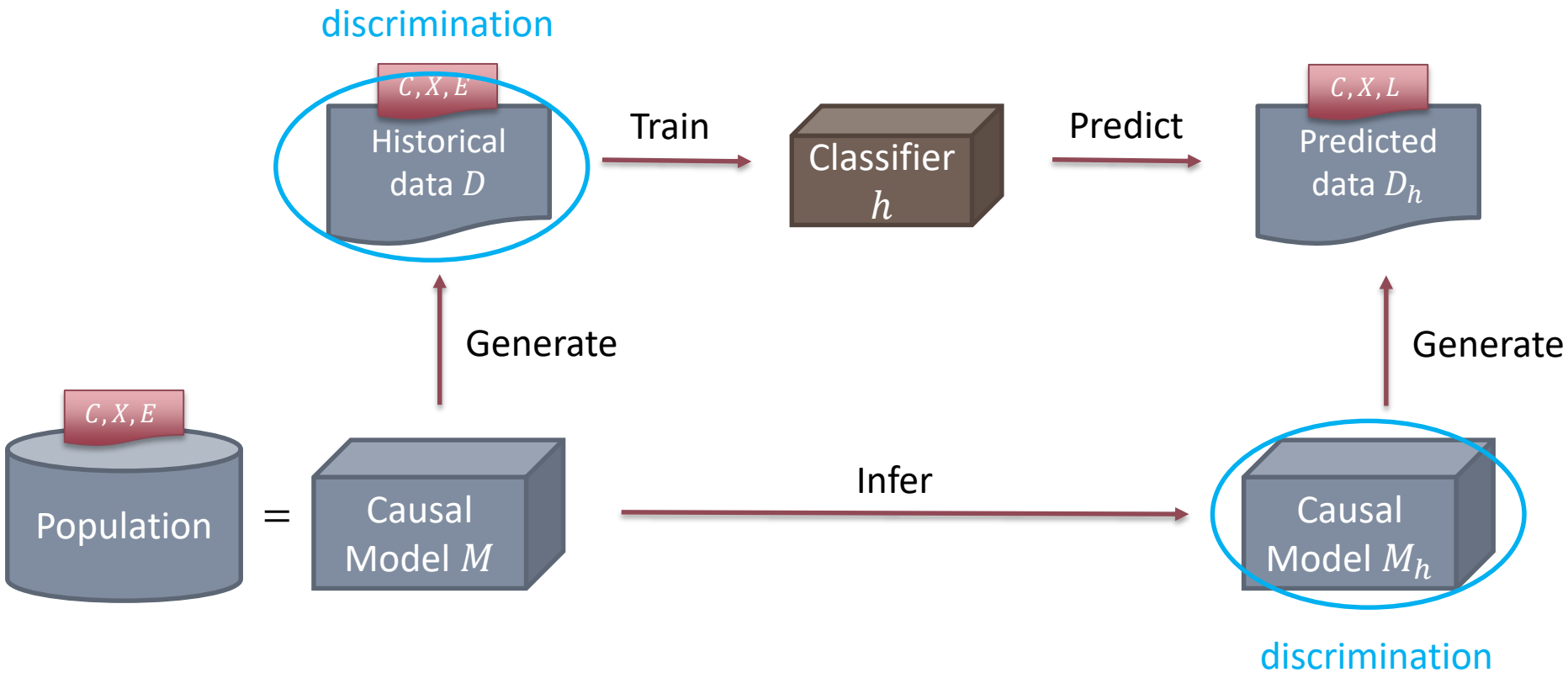
Zhang, L., Wu, Y., Wu, X.: Achieving non-discrimination in prediction. arXiv preprint arXiv:1703.00060 (2017)

Empirical Example

- The training data and test data come from the same population.
- The training data contains no discrimination according to certain discrimination measure.
- The predictions made by the classifier may still contain discrimination.

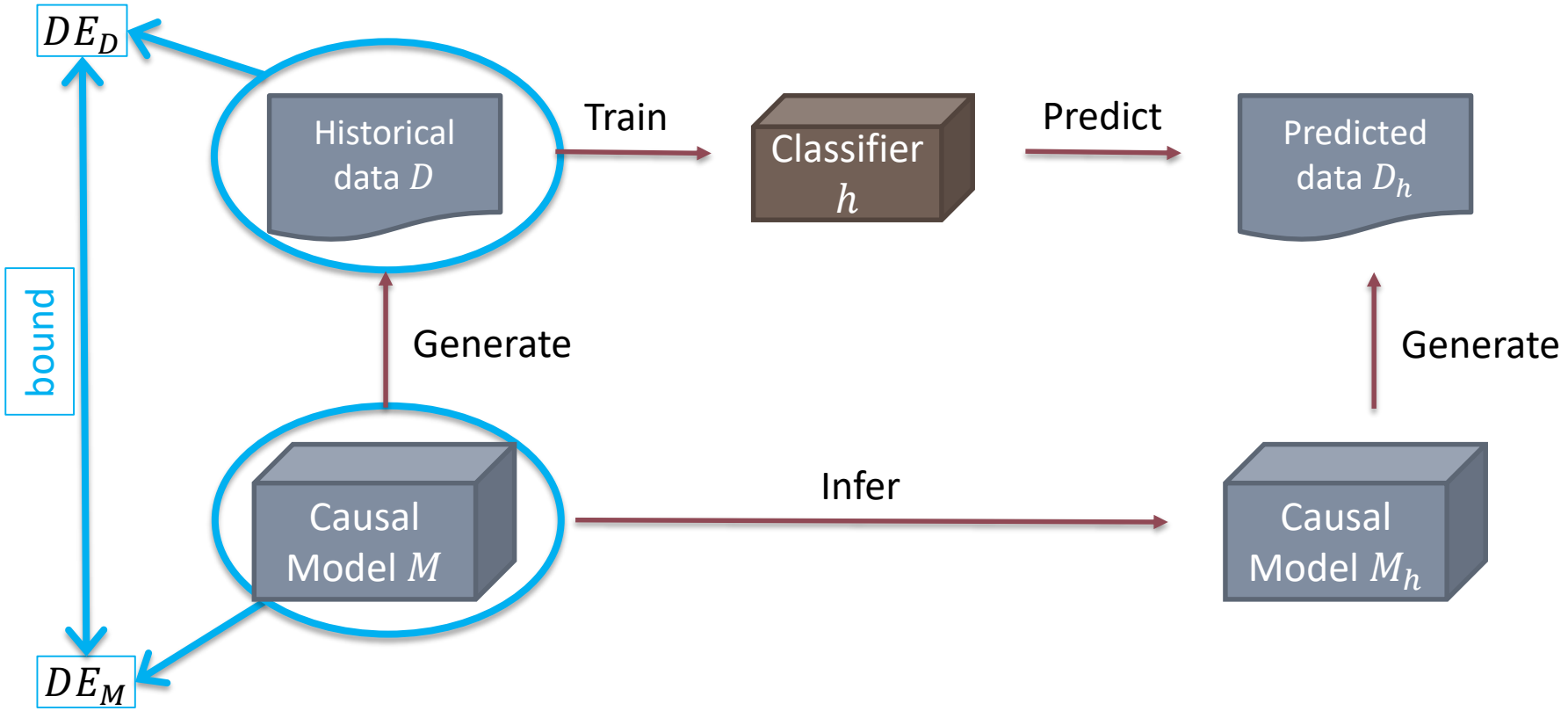


Causal Modeling-Based Anti-Discrimination Framework



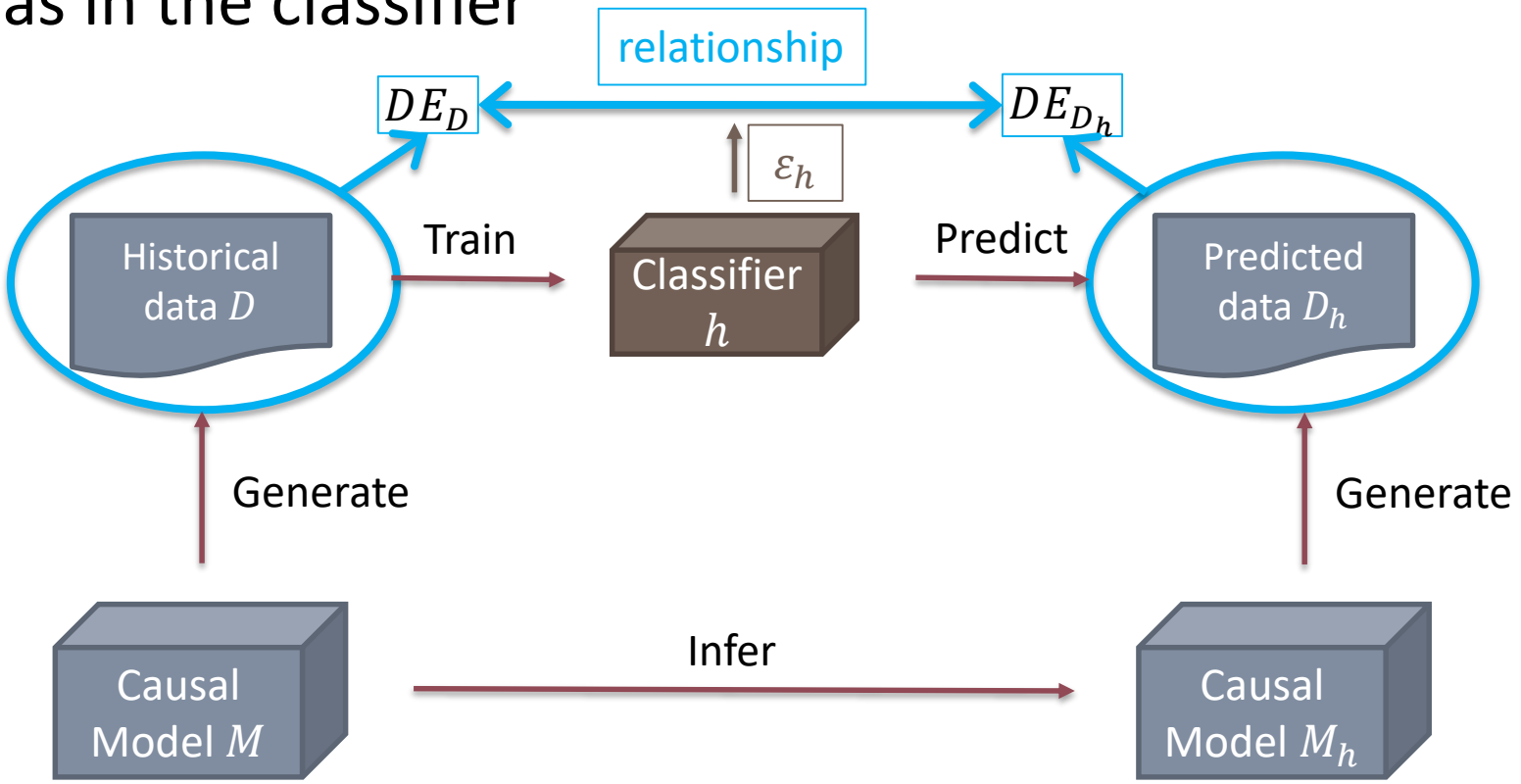
Causal Modeling-Based Anti-Discrimination Framework

- Estimate discrimination in the population.
- DE: discriminatory effect



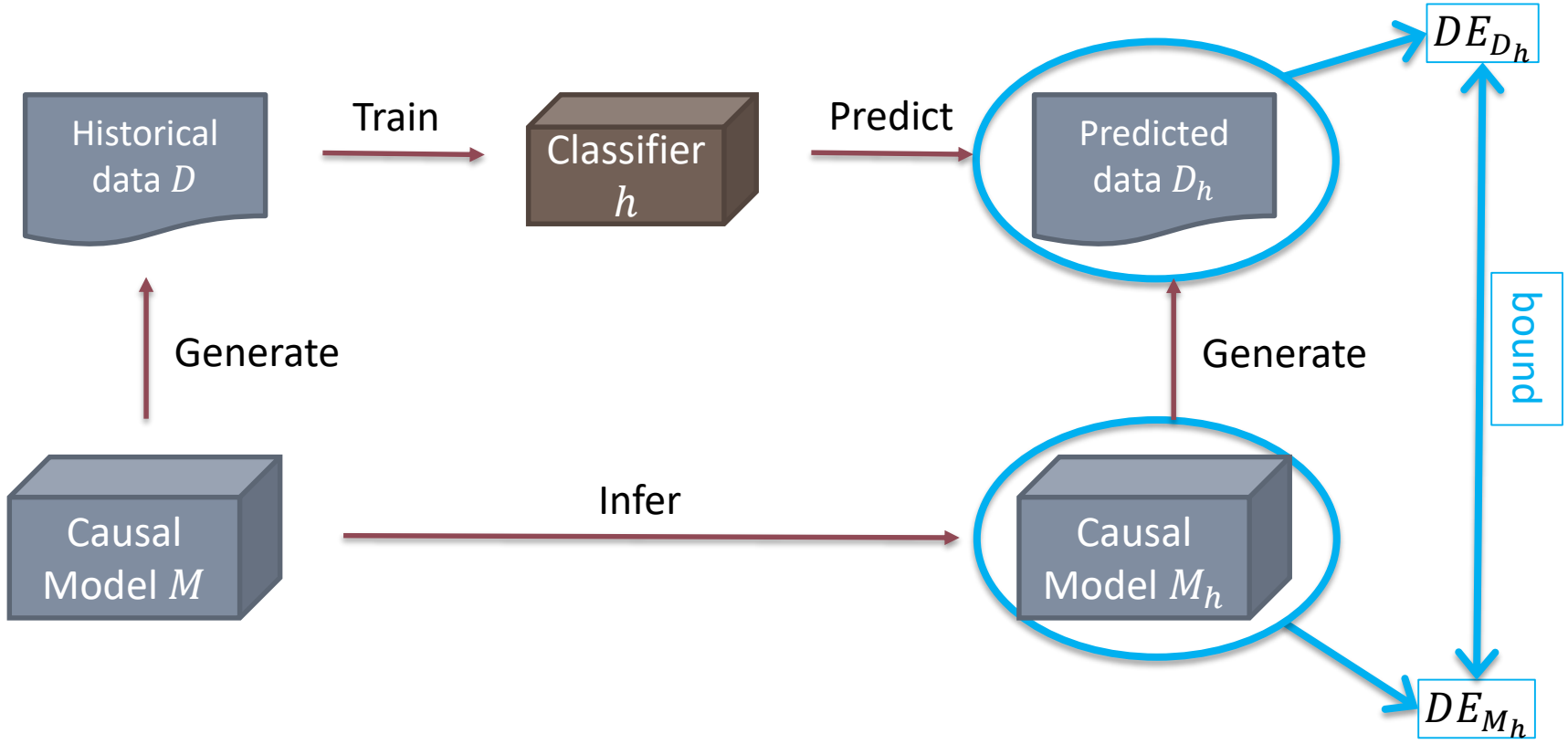
Causal Modeling-Based Anti-Discrimination Framework

- Estimate discrimination in predictions.
- ε_h: bias in the classifier



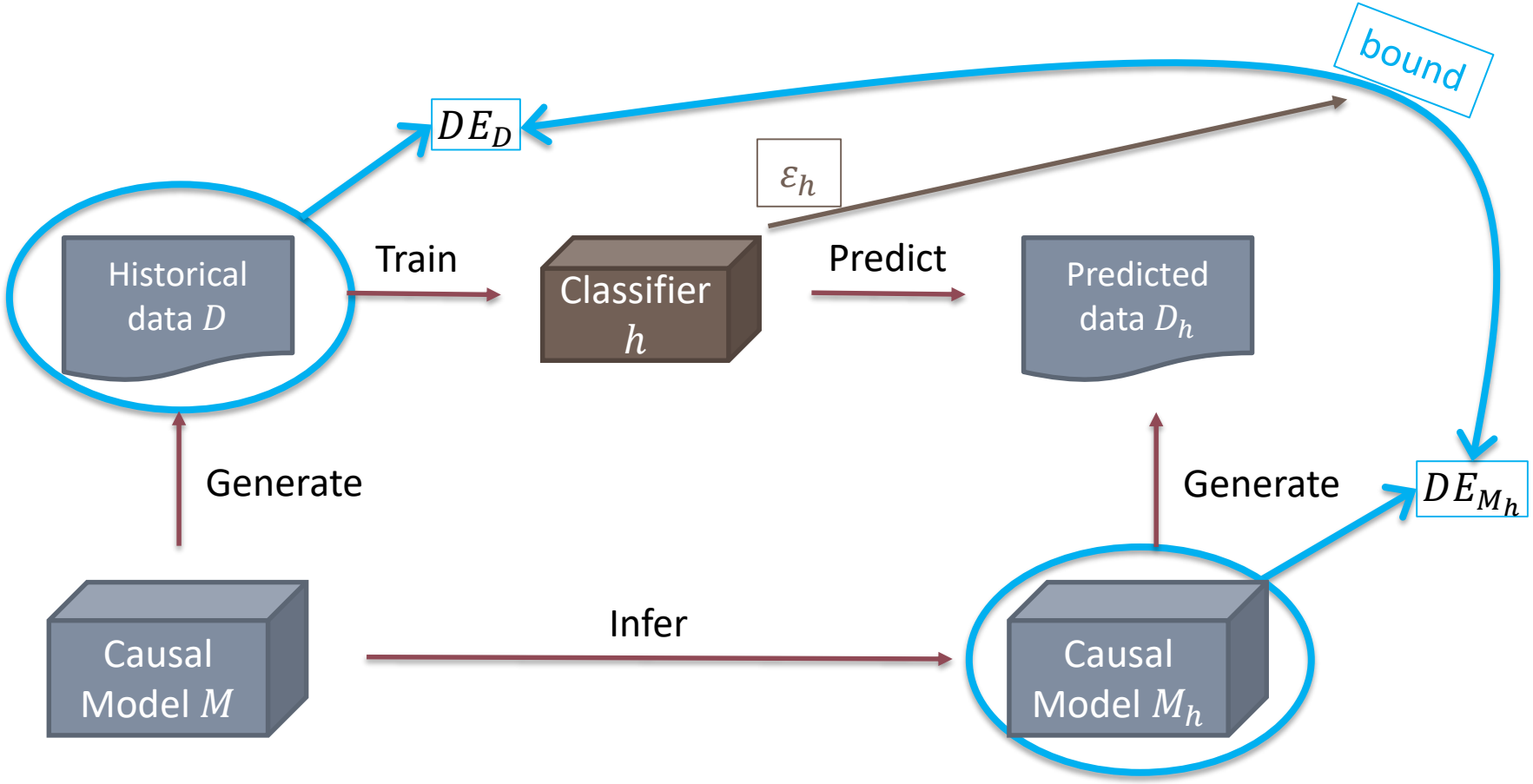
Causal Modeling-Based Anti-Discrimination Framework

- Estimate discrimination in predictions.



Causal Modeling-Based Anti-Discrimination Framework

- Estimate discrimination in predictions.



Ensuring Non-Discrimination in Predictions Results

- Discrimination in predictions (DE_{M_h}) depends on both the discrimination in the training data (DE_D) and the bias in the classifier (ε_h).
- To ensure non-discrimination in predictions:
  1. Remove discrimination from the training data.
     - Need to modify the labels other than the non-protected attributes.
  2. Reduce the bias of the classifier.

- Currently focus on the total causal effect (risk difference).
- Need further investigation.

Zhang, L., Wu, Y., Wu, X.: Achieving non-discrimination in prediction. arXiv preprint arXiv:1703.00060 (2017)

Outline

- Part I: Introduction
- Part II: Causal Modeling Background
- Part III: Anti-Discrimination Learning
  - Causal Modeling-Based Anti-Discrimination Framework
  - Direct and Indirect Discrimination
  - Non-Discrimination in Prediction
  - Non-Discrimination in Data Release
  - Individual Discrimination
- Part IV: Challenges and Directions for Future Research

Achieving Non-Discrimination in Data Release

- Motivating example (ME3): an organization/data-owner aims to achieve a non-discrimination guarantee against all possible lawsuits.
- Risk difference for group discrimination
  - $\Delta P|_{\boldsymbol{s}} = P(e^+|c^+, \boldsymbol{s}) - P(e^+|c^-, \boldsymbol{s})$
  - τ: an user-defined threshold for discrimination detection depending on laws and regulations (e.g., 0.05).
  - If $\Delta P|_{\boldsymbol{s}} < \tau$ holds across all possible partitions and their values $\boldsymbol{s}$, then no discrimination.

Zhang, L., Wu, Y., Wu, X.: Achieving non-discrimination in data release. In: SIGKDD'17 (2017)

Illustration Example

gender	female	male
admission (%)	37%	47%

$$P(e^+|c^+) - P(e^+|c^-) = 0.1$$

test score	L		H	
gender	female	male	female	male
admission (%)	25%	35%	55%	65%

$$P(e^+|c^+, \{\mathrm{L}\}) - P(e^+|c^-, \{\mathrm{L}\}) = 0.1$$

major	CS				EE			
test score	L		H		L		H	
gender	female	male	female	male	female	male	female	male
admission (%)	20%	20%	50%	50%	40%	40%	70%	70%

$$P(e^+|c^+, \{\mathrm{CS}, \mathrm{L}\}) - P(e^+|c^-, \{\mathrm{CS}, \mathrm{L}\}) = 0$$

Ensure Non-Discrimination

- Risk difference for group discrimination
  - Must be based on meaningful partitions.
  - Ensure no bias, if $\Delta P|_{\boldsymbol{b}} < \tau$ holds across all possible meaningful partitions $\boldsymbol{B}$ and their values $\boldsymbol{b}$.
- How to identify meaningful partitions?
- How to ensure no bias over all meaningful partitions?

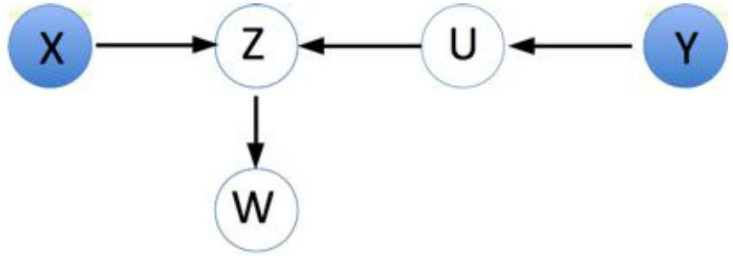
d-Separation

- A graphical criterion which can be used to read off from the graph all the conditional independence relationships encoded in the causal model (graph).
- Definition of *d*-separation
- A path q is said to be blocked by conditioning on a set $\boldsymbol{Z}$ if
  - q contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that the middle node m is in $\boldsymbol{Z}$, or
  - q contains a collider $i \rightarrow m \leftarrow j$ such that the middle node m is not in $\boldsymbol{Z}$ and such that no descendant of m is in $\boldsymbol{Z}$.
- $\boldsymbol{Z}$ is said to *d*-separate nodes X and Y if $\boldsymbol{Z}$ blocks every path from X to Y, denoted by $(X \perp Y|Z)_G$

d-Separation

- Example (blocking of paths)

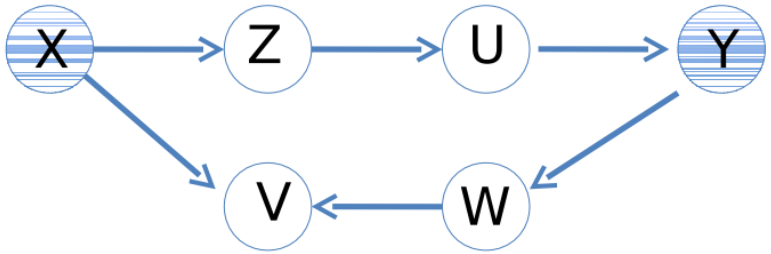
  - Path from X to Y is blocked by conditioning on $\{U\}$ or $\{Z\}$ or both $\{U, Z\}$
- Example (unblocking of paths)



  - Path from X to Y is blocked by $\emptyset$ or $\{U\}$
  - Unblocked by conditioning on $\{Z\}$ or $\{W\}$ or both $\{Z, W\}$

d-Separation

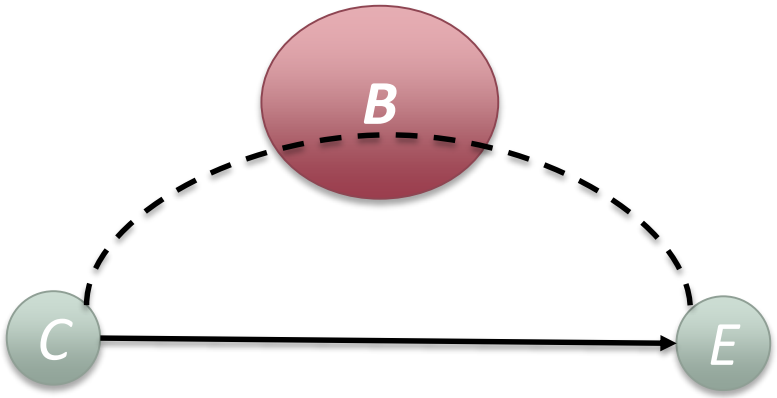
- Examples (*d*-separation)



- We have following *d*-separation relations
  - $(X \perp Y|Z)_G, (X \perp Y|U)_G, (X \perp Y|ZU)_G$
  - $(X \perp Y|ZW)_G, (X \perp Y|UW)_G, (X \perp Y|ZUW)_G$
  - $(X \perp Y|VZUW)_G$
- However we do NOT have
  - $(X \perp Y|VZU)_G$

Identify Meaningful Partition

- A node set $\boldsymbol{B}$ forms a meaningful partition:
  - $(C \perp E|\boldsymbol{B})_{\mathbf{G}'}$
  - None of E's children in $\boldsymbol{B}$
- $\boldsymbol{B}$ is called a block set



- Given a block set $\boldsymbol{B}$, the influence from C to E is only transmitted along $C \rightarrow E$. Hence, $\Delta P|_{\boldsymbol{b}}$ can measure this influence.

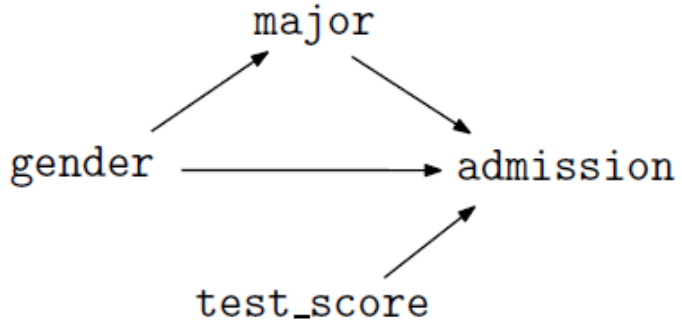
Non-Discrimination Criterion

- For each value assignment $\boldsymbol{b}$ of each block set $\boldsymbol{B}$
  - Partition data by conditioning on attributes of $\boldsymbol{B}$.
  - Use $\Delta P|_{\boldsymbol{b}}$ to measure discrimination.

THEOREM 3.1. *Non-discrimination is claimed for $\mathcal{D}$ if and only if inequality $|\Delta P|_{\mathbf{b}}| < \tau$ holds for each value assignment* $\mathbf{b}$ *of each block set* $\mathbf{B}$.

Motivating Examples (ME3)

- Block sets: {major}, {major, test_score}
- Example 1
  - $\Delta P|_{\{\mathrm{CS}\}} = \Delta P|_{\{\mathrm{EE}\}} = 0$
  - $\Delta P|_{\{\mathrm{CS,L}\}} = \Delta P|_{\{\mathrm{CS,H}\}} = 0$
    $\Delta P|_{\{\mathrm{EE,L}\}} = \Delta P|_{\{\mathrm{EE,H}\}} = 0$
  - Result: no discrimination.
- Example 2
  - $\Delta P|_{\{\mathrm{CS}\}} = \Delta P|_{\{\mathrm{EE}\}} = 0$
  - $\Delta P|_{\{\mathrm{CS,L}\}} = 0.06$
    $\Delta P|_{\{\mathrm{CS,H}\}} = -0.1$
    $\Delta P|_{\{\mathrm{EE,L}\}} = 0.05$
    $\Delta P|_{\{\mathrm{EE,H}\}} = -0.1$
  - Result*: identify discrimination for 4 subgroups.



major	CS				EE			
test score	L		H		L		H	
gender	female	male	female	male	female	male	female	male
admission (%)	20%	20%	50%	50%	40%	40%	70%	70%

Example 1

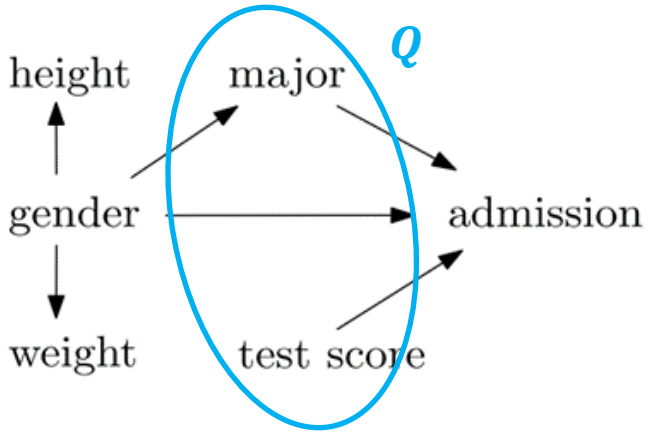
major	CS				EE			
test score	L		H		L		H	
gender	female	male	female	male	female	male	female	male
admission (%)	30%	36%	50%	40%	40%	45%	60%	50%

Example 2

* $\tau = 0.05$

Discrimination Detection

- Examining all block sets
  - Is a brute force algorithm
  - Has an exponential complexity.
- Examining only one set $\boldsymbol{Q} = Pa(E)\backslash\{C\}$
  - If $\left|\Delta P|_{\boldsymbol{q}}\right| < \tau$ holds, it is guaranteed $|\Delta P|_{\boldsymbol{b}}| < \tau$ holds.



Relaxed Non-Discrimination Criterion

- Values of $\Delta P|_{\boldsymbol{b}}$ may vary from one subpopulation to another due to randomness in sampling.

	major 1	major 2	…	major i	…	major n
$\lvert\Delta P\rvert_{\{\text{major}\}}\rvert$	$< \tau$	$< \tau$	$< \tau$	$\geq \tau$	$< \tau$	$< \tau$

  - Discrimination is claimed although majority of values are smaller than τ
- Solution:
  - Treat $\Delta P|_{\boldsymbol{B}}$ as a variable and treat each assignment $\Delta P|_{\boldsymbol{b}}$ as values.
  - If $\mathrm{P}(|\Delta P|_{\boldsymbol{B}}| < \tau) \geq \alpha$, then no discrimination under the partition ***B***.
  - α-non-discrimination is claimed for dataset *D* if it is true for each block set ***B***.

Relaxed Non-Discrimination Criterion

- Use Chebyshev's inequality to estimate $P(|\Delta P|_{\boldsymbol{B}}|<\tau)$.
- The relaxed non-discrimination criterion

THEOREM 4.2. *Given α, α-non-discrimination is claimed if the following inequality holds for each block set* **B**:

$$1-\frac{\sigma_{\mathbf{B}}^2+\mu_{\mathbf{B}}^2}{\tau^2}\geq\alpha,$$

- Examining $\boldsymbol{Q}=Pa(E)\backslash\{C\}$ only instead of examining all block sets.

Discrimination Removal

- Modifying the causal graph (MGraph)
  - Modify the CPT of E so that non-discrimination is achieved over its distribution and graph.
  - Generate a new dataset using the modified graph.
  - Minimize the distance of the joint distributions: quadratic programming.
- Modifying the dataset (MData)
  - If $\Delta P|_{\boldsymbol{q}} \geq \tau$, randomly select a number of individuals from the $\{c^- e^-\}$ group and change decision from e^- to e^+.
  - If $\Delta P|_{\boldsymbol{q}} \leq -\tau$, do the similar modification.
  - As a result, ensure that $|\Delta P|_{\boldsymbol{q}}| \leq \tau$ holds for each $\boldsymbol{q}$.

Comparison of Different Methods

- Data: Adult and Dutch Census
- Evaluated algorithms:
  - MGraph, MData (Zhang et al. KDD 2017)
  - Local massaging (LM) and local preferential sampling (LPS) algorithms (Žliobaite et al. ICDM 2011)
  - Disparate impact removal algorithm (DI) (Feldman et al. KDD 2015)
- Result
  - MGraph and MData totally remove discrimination over all subgroups.
  - LM, LPS, DI still have discriminated groups.
  - MGraph and MData preserve data utility.

Adult	MGraph	MData	Naive	LM	LPS	DI
$d(\times 10^{-3})$	1.18	10.27	39.35	38.65	35.60	60.65
n_T	1114	4122	29944	16048	16366	44582
χ^2	153	8470	18428	26900	10819	99770
Disc.	0	0	64	104	77	128
Dutch	MGraph	MData	Naive	LM	LPS	DI
$d(\times 10^{-3})$	5.68	6.75	13.91	18.00	15.48	14.10
n_T	10422	8838	32516	29288	24648	35728
χ^2	2832	4825	14014	10555	5039	19684
Disc.	0	0	1	9	4	12

Outline

- Part I: Introduction
- Part II: Causal Modeling Background
- Part III: Anti-Discrimination Learning
  - Causal Modeling-Based Anti-Discrimination Framework
  - Direct and Indirect Discrimination
  - Non-Discrimination in Prediction
  - Non-Discrimination in Data Release
  - Individual Discrimination
- Part IV: Challenges and Directions for Future Research

Individual Discrimination Discovery

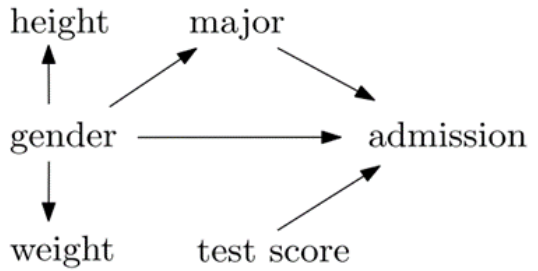
- Individual-level discrimination discovery deals with the discrimination that happens to one particular individual.
- Situation testing-based approach:
  - Select pairs of similar individuals to the target from both the protected (c^-) group and the unprotected (c^+) group.
  - Check whether difference is significant between the decisions of the selected protected and non-protected individuals.

- Motivating example (ME4): how to find similar individuals for situation testing?

Zhang, L., Wu, Y., Wu, X.: Situation testing-based discrimination discovery: a causal inference approach. In: IJCAI'16 (2016)

Illustrative Example

No.	gender	major	Test score	height	weight	ad.
1	F	CS	B	low	low	reject
2	M	CS	B	median	median	admit
3	F	CS	A	low	low	reject
4	M	CS	A	median	median	admit
5	F	CS	C	low	median	reject
6	M	CS	C	median	median	reject
7	M	EE	B	low	low	reject
⋮	⋮	⋮	⋮	⋮	⋮	⋮



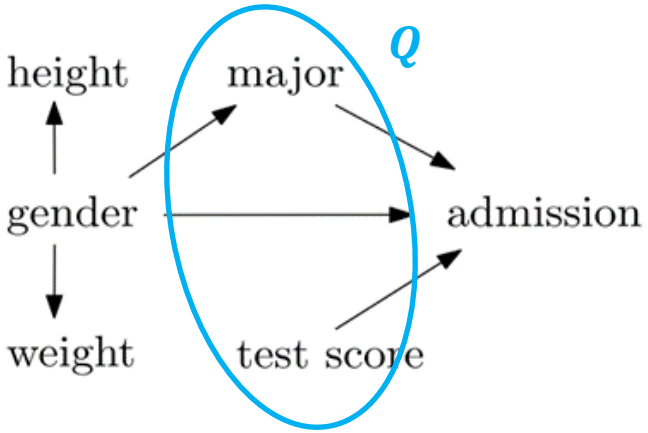
- Which one is the closest to 1, among 2, 3, and 7?
- Is the distance between 2 and 4 same as that between 2 and 6? (is the A-to-B distance the same as the B-to-C distance in regard to test_score?)

Situation Testing

- The key issue:
  - How to determine the similarity/distance between individuals?
- Questions:
  - Which attributes should be used in the distance computation?
  - How to measure the distance between different values of an attribute?

Causal Graph-Based Approach

- Answers:
  - Construct the causal graph.
  - Only the attributes that are the direct causes other than C (denoted by $\boldsymbol{Q} = Pa(E)\backslash\{C\}$) of the decision should be used for measuring distance.
  - Use the combination of causal effect and the distance between to values to calculate the distance of individuals



Discrimination Discovery

- The distance function between two individuals t and t' is defined as:

$$d(t,t') = \sum_{k=1}^{|\boldsymbol{Q}|} |CE(q_k, q'_k) \cdot VD(q_k, q'_k)|$$

- $CE(q_k, q'_k)$ measures the causal effect of each attribute $Q_k \in \boldsymbol{Q}$ on the decision when the value of Q_k changes from q_k to q'_k. Using the *do*-calculus, it is computed with:

$$CE(q_k, q'_k) = P(e^+|do(\boldsymbol{q})) - P(e^+|do(q'_k, \boldsymbol{q}\backslash\{q_k\}))$$

- $VD(q_k, q'_k)$ measures the difference between two values q_k and q'_k of each attribute $Q_k \in \boldsymbol{Q}$.

$$VD(q_k, q'_k) = \begin{cases} Manhattan(q_k, q'_k) & \text{if } Q_k \text{ is ordinal/interval} \\ Overlap(q_k, q'_k) & \text{if } Q_k \text{ is categorical} \end{cases}$$

Comparison of Different Methods

- Data: Dutch Census
- Comparison of Different Methods
  - CBN-based situation testing (CBN-DD) (Zhang et al. IJCAI 2017)
  - KNN-based situation testing (KNN-DD) (Luong et al. KDD 2011)
- Result:
  - KNN-DD and CBN-DD are significantly different.
  - CBN-DD outperforms KNN-DD over the synthetic data.

Accuracy

K	*CBN-DD*		*KNN-DD*	
	TP	TN	TP	TN
10	73.3	63.1	46	66.2
50	85.3	77.6	42.2	76.2
90	81.5	83.9	38.1	81.2

- Add 100 tuples with discrimination to a clean dataset.
- Use these tuples and another 100 tuples without discrimination as the targets.

Outline

- Part I: Introduction
- Part II: Causal Modeling Background
- Part III: Anti-Discrimination Learning
- Part IV: Challenges and Directions for Future Research

Summary

- The causal modeling-based framework for anti-discrimination learning.

Discovering and removing discrimination from dataset:

	system-level	group-level	individual-level
direct discrimination	Zhang et al. IJCAI 2017	Zhang et al. KDD 2017	Zhang et al. IJCAI 2016
indirect discrimination	Zhang et al. IJCAI 2017		

Ensuring non-discrimination in predictions:

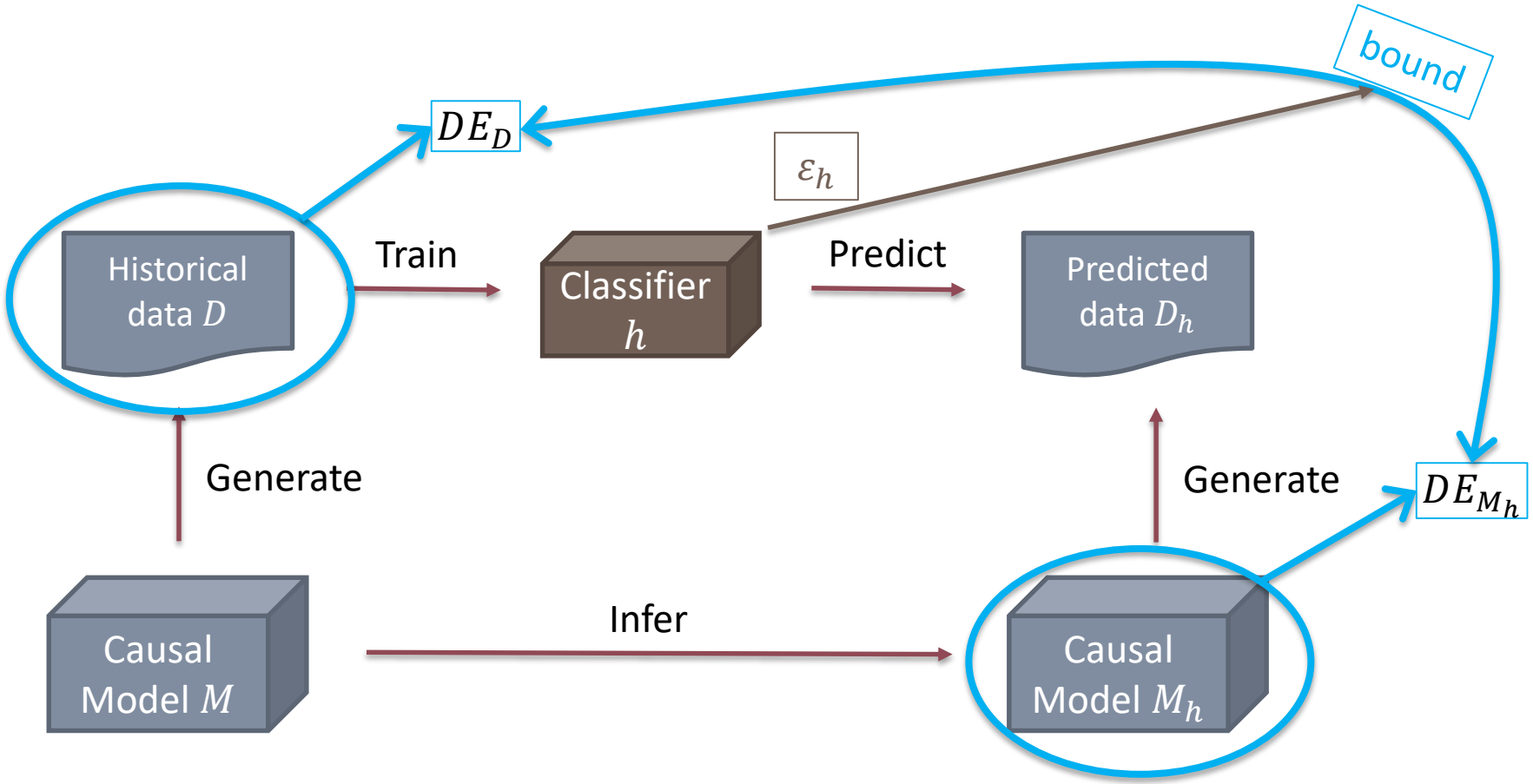
Zhang et al. arXiv 2017 (in term of risk difference)

Group and Individual-Level Indirect Discrimination

- (Zhang et al. IJCAI 2017): direct/indirect discrimination at the system-level.
- (Zhang et al. KDD 2017, Zhang et al. IJCAI 2016): direct discrimination at the group/individual-level.
- How to model, measure, and prevent indirect discrimination at group and individual-level?
- Are the existing techniques still applicable?
  - E.g., the path-specific effect, block set, etc.

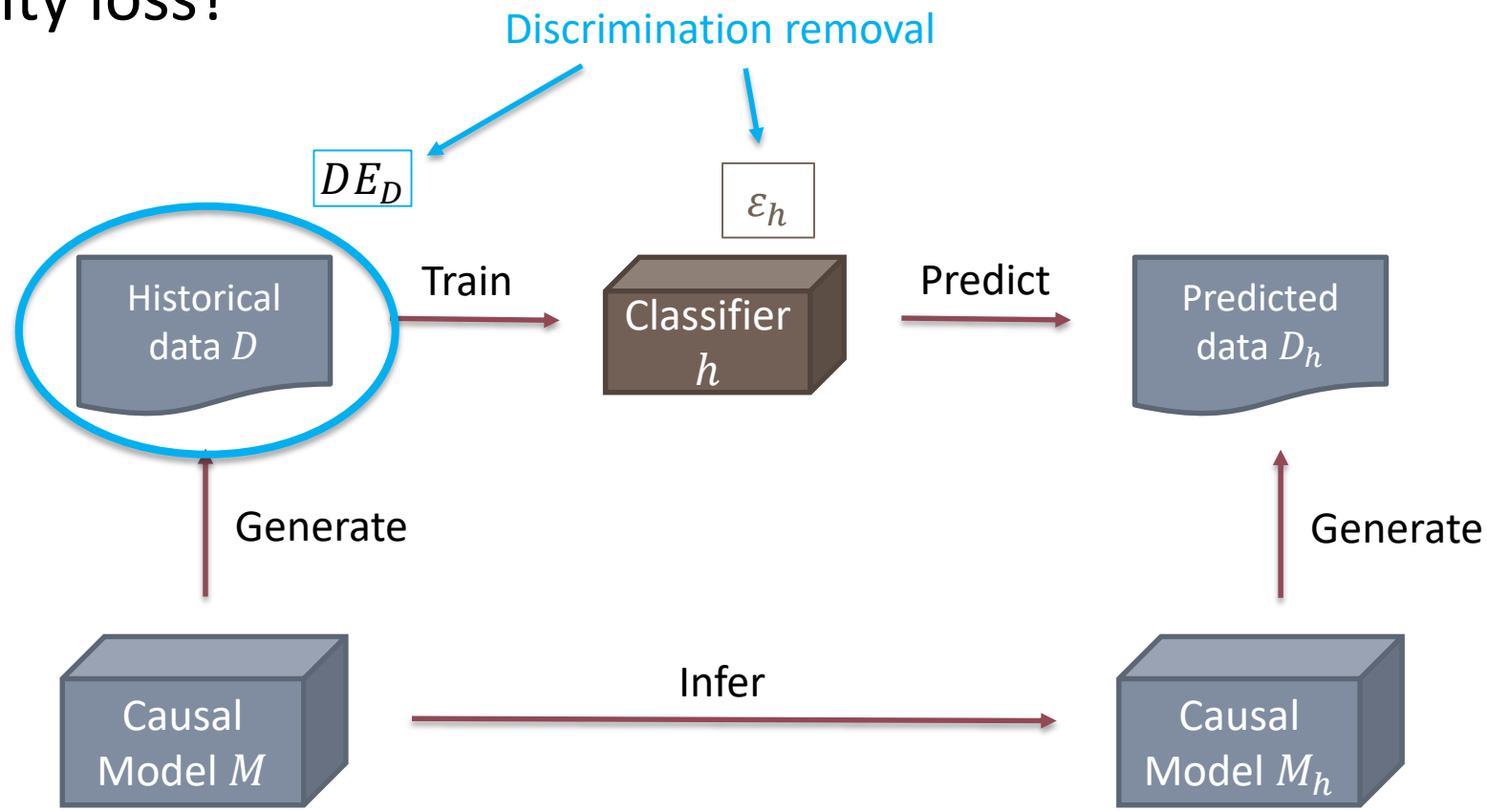
How to Achieve Direct and Indirect Non-Discrimination in Predictions

- Zhang et al. arXiv 2017: target risk difference.



Trade-Off

- How to balance the trade-off between non-discrimination and utility loss?



Relaxing Markovian Assumption

- A causal model is Markovian if
  1. The causal graph is acyclic;
  2. All variables in $\boldsymbol{U}$ are mutually independent.

Relaxing Markovian Assumption

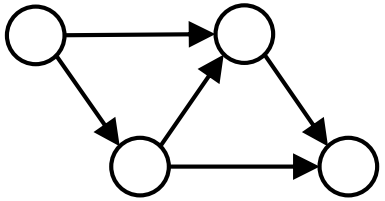
- A causal model is semi-Markovian if
  1. The causal graph is acyclic;
  2. All variables in $\boldsymbol{U}$ are NOT mutually independent.

- Hidden confounders are known to exist in the system.
- Much more complicated than the Markovian model, both in the causal graph construction and causal effect inference.

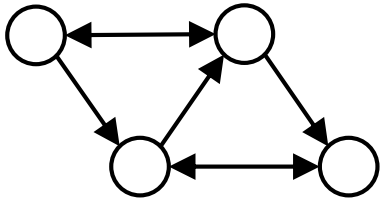
Relaxing Markovian Assumption

- The causal graph of the semi-Markovian model is commonly represented by the acyclic directed mixed graph (ADMG).

DAG



ADMG



- The bidirected arrow ⟷ implies the presence of unobserved confounder(s) between variables.
- How to learn ADMG from data is still under exploration.
  - Some recent advances include the ordinary Markov model and the nested Markov model.

Relaxing Markovian Assumption

- Unlike in the Markovian model, some *do*-operations may not be able to be calculated (identifiable) due to the unobserved confounders.
- Generalize the *d*-separation to *m*-separation.
- The path-specific effect also needs to be generalized in semi-Markovian models.

Any anti-discrimination method designed for semi-Markovian models must be adapted to the differences in the causal inference techniques.

Discrimination in Tasks Beyond Classification

- Currently mainly focus on classification problems.
- Tasks beyond classification:
  - Regression: the decisions are continuous variables
  - Ranking: the outcome is a ranking of candidates
  - Recommendation: the outcome is a list of recommended items
  - …
- Transparency in learning process

References

- Žliobaite, I., Kamiran, F., Calders, T.: Handling conditional discrimination. In: ICDM'11, pp. 992–1001. IEEE, (2011)
- Hajian, S., Domingo-Ferrer, J.: A methodology for direct and indirect discrimination prevention in datamining. IEEE Trans. Knowl. Data Eng. 25(7), 1445–1459 (2013)
- Mancuhan, K., Clifton, C.: Combating discrimination using Bayesian networks. Artif. Intell. Law 22(2), 211–238 (2014)
- Luong, B.T., Ruggieri, S., Turini, F.: k-NN as an implementation of situation testing for discrimination discovery and prevention. In: SIGKDD'11, pp. 502–510. ACM, (2011)
- Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C.,Venkata- subramanian, S.: Certifying and removing disparate impact. In: SIGKDD'15, pp. 259–268. ACM, (2015)
- Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. 33(1), 1–33 (2012)
- Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. Data Min. Knowl. Discov. 21(2), 277–292 (2010)
- Bonchi, F., Hajian, S., Mishra, B., Ramazzotti, D.: Exposing the probabilistic causal structure of discrimination. Int. J. Data Sci. Anal. 3(1), 1–21 (2017)
- Zhang, L., Wu, Y., Wu, X.: On discrimination discovery using causal networks. In: SBP-BRiMS 2016. (2016)
- Zhang, L., Wu, Y., Wu, X.: A causal framework for discovering and removing direct and indirect discrimination. In: IJCAI'17 (2017)
- Zhang, L., Wu, Y., Wu, X.: Achieving non-discrimination in prediction. arXiv preprint arXiv: 1703.00060 (2017)
- Zhang, L., Wu, Y., Wu, X.: Achieving non-discrimination in data release. In: SIGKDD'17 (2017)
- Zhang, L., Wu, Y., Wu, X.: Situation testing-based discrimination discovery: a causal inference approach. In: IJCAI'16 (2016)
- Zhang, L., Wu, X.: Anti-discrimination learning: a causal modeling-based framework. Int. J. Data Sci. Anal. (2017)

References

- Zemel, R.S.,Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In ICML'13. (2013)
- Wu,Y.,Wu,X.: Using loglinear model for discrimination discovery and prevention. In: DSAA'16 (2016)
- Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. KAIS (2012)
- Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review 29(05), 582–638 (2014)
- Magnani, L., Board, E., Longo, G., Sinha, C., & Thagard, P.: Discrimination and Privacy in the Information Society. Springer (2013)

UNIVERSITY OF ARKANSAS

Thank you

Lu Zhang lz006@uark.edu
Xintao Wu yw009@uark.edu
Yongkai Wu xintaowu@uark.edu

This work is supported by NSF 1646654.

Slides will be made available at: http://www.csce.uark.edu/~xintaowu/publ/sbp17.pdf