ExDis: Causal Explanations for Disparate Trends

Tal Blau
Ben-Gurion University of the Negev
tbl@post.bgu.ac.il

Brit Youngmann
Technion
Haifa
brity@technion.ac.il

Anna Fariha
University of Utah
Salt Lake City
afariha@cs.utah.edu

Yuval Moskovitch
Ben-Gurion University of the Negev
yuvalmos@bgu.ac.il

ABSTRACT

In today's data-driven world, insights collected from the data and trends observed in the data significantly contribute to decision making. However, users are often perplexed by certain surprising data trends, especially the *disparate* ones. For example, upon observing a disparate trend that "men are more likely to have a heart-attack than women", a health-care professional wonders, "is there a certain demographic where the trend is more pronounced or even reversed?", "what factors further exacerbate or alleviate such disparity?". To this end, we introduce ExDis, a system for automatically identifying data regions where an observed Disparity is pronounced (or reversed) and Explaining the associated causes that exacerbate (or alleviate) the disparity. ExDis equips policy-makers to recognize the factors that causally contribute to certain disparities and implement targeted corrective measures.

Instead of generating the top-k explanations, ExDis takes a holistic approach to generate *a set of explanations*, where each explanation has reasonably high utility, while ensuring diversity among the explanations. ExDis allows the users to provide their required number of explanations (budget k) and desired balance between utility and diversity, and automatically finds an explanation set according to the user's preferences. Under the hood, ExDis employs a greedy algorithm that exploits the sub-modularity property of the objective function for the underlying problem and can efficiently produce the desired explanation set. We demonstrate how ExDis can discover interesting data regions and meaningful, often surprising, explanations effectively and efficiently over high-dimensional datasets.

PVLDB Reference Format:
Tal Blau, Brit Youngmann, Anna Fariha, and Yuval Moskovitch. ExDis: Causal Explanations for Disparate Trends. PVLDB, 14(1): XXX-XXX, 2025.
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/TalBl/CausalExplanationforDisparity.

1 INTRODUCTION

Drawing conclusions from data based on observed trends is common practice. However, when analyzing large, high-dimensional datasets, these trends often require deeper exploration or *explanations* to help the analyst gain a better understanding of the data. For instance, after identifying a notable disparity between two groups in the data, an analyst might be interested in identifying subpopulations in the data where the disparity is more pronounced or even reversed. Uncovering the causal reasons behind the observed disparity can further enhance data understanding.

Example 1. *The Medical Expenditure Panel Survey (MEPS) dataset provides detailed information on healthcare utilization, expenditures, insurance coverage, and demographic characteristics of individuals in the United States. Based on this dataset, in general, males have a lower likelihood (37%) of feeling nervous frequently than non-males (45%). Soha, an analyst analyzing the data, is interested in finding subpopulations where a reverse trend exists, i.e., males have a higher likelihood of feeling nervous than non-males. Indeed, a closer look at the data can reveal such cases. For instance, one such subpopulation is "divorced people with age between 51–63 who have a recommendation to exercise from doctor", where males have a higher likelihood (47%) of feeling nervous than non-males (43%). Interestingly, within this subpopulation, "not currently smoking" exacerbates the situation for males (increases the likelihood of feeling nervous by 21%) but improves the situation for non-males (decreases the likelihood of feeling nervous by 14%). Discovering such a reverse trend requires examining all possible subpopulations, which is tedious.*

We aim to *explain* disparities observed in the average of a given outcome for two groups[1] of interest in the data. An explanation should pinpoint *data regions* showing interesting facets of the disparity, either emphasizing or contradicting the observed trend and provide *causal reasons* behind disparities, explaining the associated causes that exacerbate (or alleviate) the disparity.

A single causal explanation is often insufficient to explain the observed disparity for the entire population. In fact, in different subpopulations, the reasons behind the disparity between the two groups may vary, with some contributing more to the disparity than others. Therefore, we aim to discover *high-utility* subpopulations, for which a strong causal explanation exists for the observed disparity. Preferring explanations consisting of subpopulations with high utility in terms of having a strong causal factor may result in small subpopulations with low support with respect to the entire data, which is undesirable as insights drawn from such a small subpopulation are not statistically significant. To avoid reporting small subpopulations, we consider only subpopulations with *high support* (data coverage) as reasonable explanations. Finally, reporting multiple high-utility and high-support subpopulations where the disparity is most pronounced may result in redundancy. E.g., "never married" and "people under the age of 18" may comprise the same individuals as most people under the age of 18 never married. Therefore, beyond finding high-utility and high-support subpopulations, we aim to minimize the overlap among the reported subpopulations, and, thus, ensure *high diversity* among the explanations.

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

[1] Groups may overlap, e.g., one group can be defined as the entire data.

Manual exploration of the data to discover explanations (data regions and causes) of an observed disparity in a dataset can be complex and tedious, particularly when the dataset is large and high-dimensional. To this end, we introduce ExDis (Explaining Disparate trends), a system that automatically identifies data regions where an observed disparity is pronounced (or reversed) and associates specific factors that causally contribute to the disparity. Given a dataset and two groups of interest, ExDis generates a set of explanations, where each explanation has reasonably high utility while ensuring diversity among the explanations. Users can provide their required number of explanations (budget parameter k) and desired balance (α) between utility and diversity.

Related work. Previous work introduced methods to identify intriguing data subsets for exploration [4, 7] and detect subsets responsible for fairness violations in classifier outcomes [8]. We focus on identifying subpopulations with substantial disparities between two, possibly overlapping, groups and providing causal explanations for these disparities. To achieve this, we build upon the DivExplorer algorithm [4], modifying it to suit our setting, enabling an effective identification of subpopulations.

Recent works [3, 9] leveraged causal inference to explain aggregate query results. CauSumX [9] focuses on causal explanations for group-by-average queries, identifying influential factors (patterns) that drive outcomes. While our goals differ, we adapt CauSumX's treatment mining algorithm to find localized causal explanations for the subpopulations with significant disparity. XInsight [3] explains group disparities in aggregate query results by identifying both causal and non-causal patterns. However, unlike XInsight, which does not support overlapping groups, we provide localized causal insights rather than a single explanation for the entire dataset. We argue that disparities are best understood through precise, subpopulation-specific causal reasoning.

Demonstration. We will demonstrate the usefulness of ExDis in different scenarios using three real-life datasets where participants will be able to interact with the system to explore the datasets and ask ExDis to explain disparities among selected groups.

2 SOLUTION SKETCH

We provide an overview of our theoretical foundations of the development of ExDis and sketch our solution.

Background on Causal Inference. We use Pearl's model for *observational causal analysis* [5]. The broad goal of causal inference is to estimate the effect of a *treatment variable* T on an outcome O. One common measure of causal estimate is *Average Treatment Effect* (ATE), defined as the difference in the average outcomes of the treated and control groups:

$$ATE(T, O) = \mathbb{E}_Z \left[\mathbb{E}[O \mid T = 1, Z = z] - \mathbb{E}[O \mid T = 0, Z = z]\right] \quad (1)$$

Since ATE is computed over observational data, the treatment and control groups may not be assigned randomly. Therefore, to mitigate the effect of *confounding factors* (i.e., attributes that can affect both the treatment and outcome), we must control for *confounding variables* [5] (Z in Eq. (1)). A sufficient set of confounders can be determined by applying graphical criteria [5], which can be evaluated against a *causal Directed Acyclic Graph (DAG)*. A causal DAG represents potential direct causal relationships between variables in a given dataset [5]. It can be constructed by a domain expert, or by using existing causal discovery algorithms [1].

In ExDis, where the explanation of the disparity between groups may vary among different subpopulations, we are interested in computing the *Conditional Average Treatment Effect* (CATE), which measures the effect of a treatment on an outcome within *a subpopulation of interest*. Given a subpopulation defined by a predicate $B = b$, we compute $CATE(T, O \mid B = b)$ by adding this predicate to the conditioning sets in Eq. 1.

Disparity Explanations. We consider a database D associated with a causal DAG $\mathcal{G}$. A *pattern* [9] is a conjunction of predicates (attribute-value assignments). An example pattern is {Gender = Female ∧ Race = Asian}. In this work, we only consider equality or inequality predicates for enhanced interpretability. The two groups of interest, g_1 and g_2, are defined by the patterns ψ_{g_1} and ψ_{g_2}, respectively.

Given an outcome attribute O, we aim to discover explanations for an observed disparity in the average value of O between g_1 and g_2. Our building blocks are *disparity explanations* that identify *where* the average outcomes for g_1 and g_2 differ significantly and *why*. Restricting to average is typical for causal explanations [6], as causal effects estimate the expected difference between groups.

We assume the dataset attributes are partitioned into two disjoint sets: *actionable* attributes that can be used to define what affects the outcome (e.g., currently smokes, exercises) and *immutable* attributes, which are inherent and cannot be changed (e.g., race, age), that can be used to identify where the disparity is significant. This categorization ensures that treatments consist solely of mutable attributes that can imply corrective measures to reduce the disparity.

Given a database D with an outcome variable O and two groups g_1 and g_2, a disparity explanation ϕ is defined as a pair of patterns (ψ_g, ψ_e) where: ψ_g is defined by immutable attributes, describing a subpopulation with significant disparity between g_1 and g_2 in terms of AVG(O), and ψ_e is defined by mutable attributes, indicating a treatment that explains the disparity between g_1 and g_2 within the subpopulation defined by $\psi_g(D)$.

To assess the impact of the treatment ψ_e on the outcome O within the subpopulation $\psi_g(D)$, we compare the causal effect of ψ_e on O within the two subpopulations: $(\psi_g \wedge \psi_{g_1})(D)$ and $(\psi_g \wedge \psi_{g_2})(D)$.

Example 2. *Continuing with our example, where g_1 is* males *and g_2 is* non-males*, an example disparity explanation is: Among "divorced people with age between 51—63 who have a recommendation to exercise from doctor", the treatment "not currently smoking" increases the* Likelihood of Feeling Nervous *for* males*, while it decreases for* non-males*. Here, the subpopulation pattern ψ_g is defined by* MaritalStatus=Divorced ∧ Age=[51 − 63] ∧ DoctorRecommendsExercise=True *and the treatment pattern ψ_e is* SmokesCurrently=False.

Problem Formulation. Our goal is to find a bounded-sized set of disparity explanations Φ to identify subpopulations of the data that (1) provide insights into the disparity between g_1 and g_2, and (2) avoid redundancy across different subpopulations to cover different data regions. To this end, we define $IScore(\Phi)$ to measure the *utility* of Φ in explaining the disparity, and $NOverlap(\Phi)$ to measure its *diversity* (lack of redundancy).

IScore: *NDScore* of a disparity explanation $\phi = (\psi_g, \psi_e)$ measures the normalized difference between two CATE values over the subpopulations $(\psi_g \wedge \psi_{g_1})(D)$ and $(\psi_g \wedge \psi_{g_1})(D)$:

$$NDScore(\phi){=}Norm(\left|CATE_{\mathcal{G}_D}(\psi_e, O|\psi_g \wedge \psi_{g_1}){-}CATE_{\mathcal{G}_D}(\psi_e, O|\psi_g \wedge \psi_{g_2})\right|)$$

To prioritize disparity explanations that cover more data, we consider their *support*. The support of a disparity explanation $\phi = (\psi_g, \psi_e)$ is defined by the fraction of tuples $\in D$ that take part in the explanation: $support(\phi) = \frac{|\psi_{g \wedge g_1}(D) \cup \psi_{g \wedge g_2}(D)|}{|D|}$.

We define the *IScore* of a set of disparity explanations Φ as the weighted average NDScore of each disparity explanation within Φ to measure its utility:

$$IScore(\Phi) = \sum_{\phi \in \Phi} NDScore(\phi) \times support(\phi)$$

NOverlap: Given two disparity explanations $\phi_1{=}(\psi_g^1, \psi_e^1)$ and $\phi_2{=}(\psi_g^2, \psi_e^2)$, $Overlap(\phi_1, \phi_2)$ quantifies the number of tuples shared between $\psi_g^1(D)$ and $\psi_g^2(D)$. Since we prefer explanations with low overlap, following [2], we define the notion of *non-overlap* to quantify the degree of exclusiveness (diversity) among the disparity explanations within Φ. $NOverlap(\Phi)$ is defined as:

$$NOverlap(\Phi) = \frac{|D| \cdot l^2 - \sum_{\phi_i, \phi_j \in \Phi} overlap(\phi_i, \phi_j)}{|D| \cdot l^2}$$

where l is the number of candidate disparity explanations. Higher overlap between explanation pairs results in lower NOverlap.

Optimization problem: ExDis aims to select a disparity explanation set $\Phi \subseteq \{\phi_1, \ldots, \phi_l\}$, such that: **(i) (size constraint)** $|\Phi| \leq k$, and **(ii) (objective)** $\alpha \cdot IScore(\Phi) + (1 - \alpha) \cdot NOverlap(\Phi)$ is maximized. The corresponding decision problem is NP-hard; however, the above objective function is submodular, which justifies our choice of a greedy approach for ExDis (described next).

Algorithms. Figure 1 provides an overview of ExDis. Given a database D and two groups g_1 and g_2, the number of possible disparity explanations to explain the disparity among g_1 and g_2 may be exponential in the number of attributes in D. ExDis avoids generating all possible disparity explanations and instead, generates only the promising ones. It first mines subpopulations with significant disparity using the subpopulation-miner module. It then identifies a causal explanation for each subpopulation through the explanation miner module. Then, it applies a greedy approach to select a subset of disparity explanations of size k as the solution, which is then presented to the user.

Mining subpopulations: ExDis first detects data regions (subpopulations) with significant disparity between g_1 and g_2. We adapt DivExplorer [4], which analyzes the divergence of learning models, to measure divergence as the difference in average outcome values. This identifies subpopulations where the outcome gap is significant. To generate candidate subpopulation patterns, we restrict DivExplorer to immutable attributes.

Mining candidate explanations: Next, ExDis searches for an explanation pattern for each subpopulation. We adapt the treatment-mining step of the CauSumX [9], which provides causal explanations for aggregate queries, to maximize the difference between the two CATE values. Unlike CauSumX, which finds treatment patterns with high CATE values, we estimate the *IScore* of candidate treatment patterns. We apply additional optimizations such as parallelism, sampling, and caching to improve ExDis's runtime.

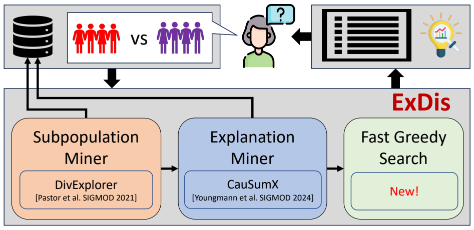


Figure 1: The ExDis architecture. The user provides a database and two groups of interest to ExDis, which returns a set of causal explanations that explains the locations and causes of the observed disparity between the groups.

Fast greedy search: Given the set of candidate disparity explanations obtained in the previous step, ExDis identifies a set of k disparity explanations using a greedy approach. At each iteration, it selects the next best disparity explanation that maximizes our objective function (weight average of *IScore* and NOverlap), then outputs the generated k-size solution to the user.

3 DEMONSTRATION SCENARIO

ExDis is built to support general datasets. For the demonstration, we will showcase its capabilities using the MEPS dataset.[2] and will make two additional datasets available: Stack Overflow Developer Survey[3] and ACS[4]. Below we provide a demonstration scenario over the MEPS dataset, which contains information on healthcare utilization, expenditures, insurance coverage, and demographic characteristics of individuals in the United States. We will guide the users through 8 steps (annotated in Figure 2) impersonating Soha, who is interested in finding a reverse trend within the dataset subpopulations.

Step Ⓐ. The user provides a database, which may be accompanied by a causal DAG. If no causal DAG is provided, ExDis utilizes a causal discovery method [1] to obtain one. For our guided demonstration, the user uploads the MEPS dataset and a causal DAG.

Step Ⓑ. In this step, the user specifies the two groups of interest. The user chooses `Males` for Group A (g_1) and `Non-males` for Group B (g_2). Note that the groups may overlap. If the user doesn't specify a condition, then the entire dataset is considered as a group.

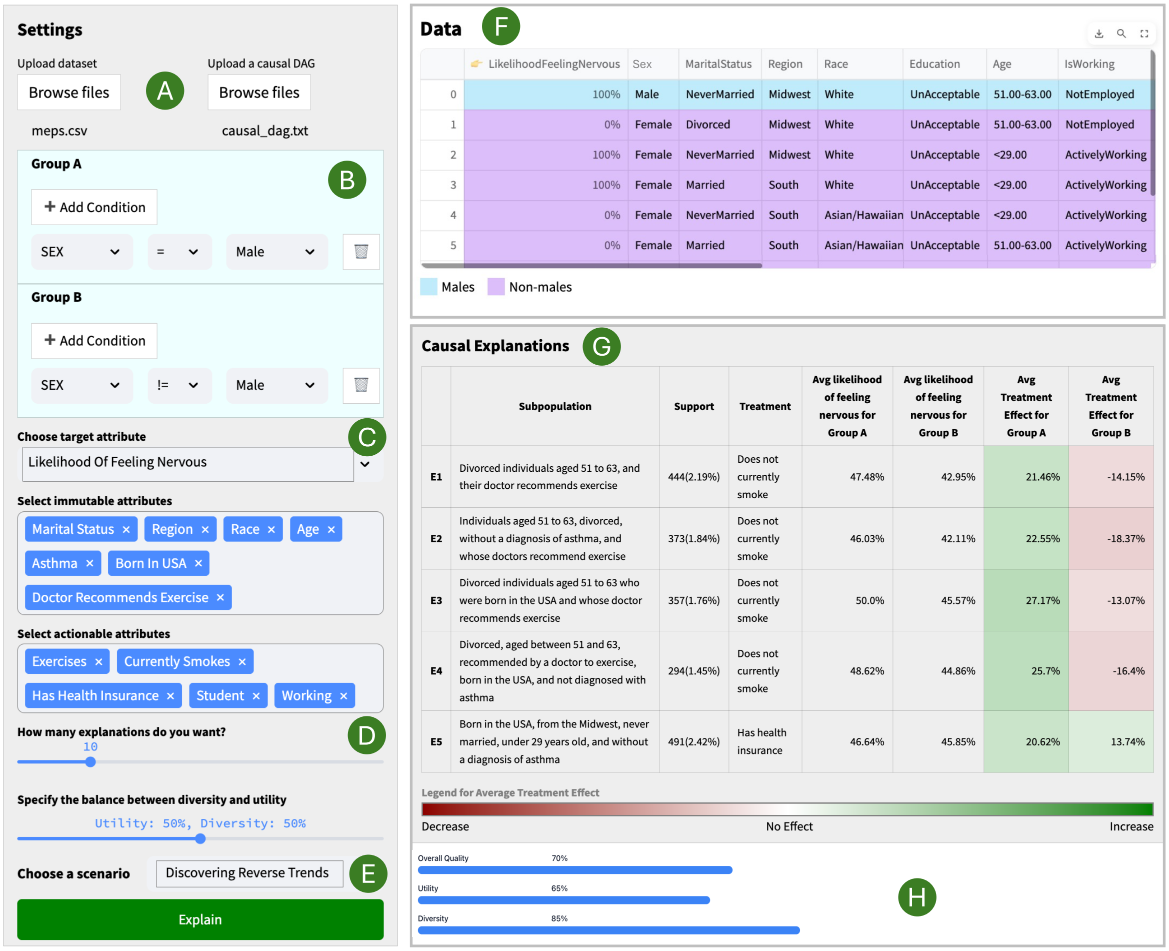
Step Ⓒ. In this step, the user chooses `Likelihood of Feeling Nervous` as the target attribute since they are interested in observing disparity in this attribute between `Males` and `Non-males`. Furthermore, the user chooses several attributes as *immutable* (such as `Marital Status`, `Race`, `Age`, etc.) that cannot be altered to define the subpopulations with disparity. The user also selects a few attributes that they seem actionable (e.g., `Exercises`, `Currently smokes`, etc.) to be considered as causal explanations.

Step Ⓓ. The user selects 10 as the desired number of disparity explanations ($k = 10$) in the causal explanation set and sets equal weights (50%) to both utility and diversity, which implicitly sets the balance parameter $\alpha = 0.5$.

[2] https://www.meps.ahrq.gov/mepsweb/

[3] https://survey.stackoverflow.co/2021

[4] https://www.census.gov/programs-surveys/acs



Settings

Upload dataset: meps.csv

Upload a causal DAG: causal_dag.txt

Group A

SEX = Male

Group B

SEX != Male

Choose target attribute

Likelihood Of Feeling Nervous

Select immutable attributes

Marital Status × Region × Race × Age × Asthma × Born In USA × Doctor Recommends Exercise ×

Select actionable attributes

Exercises × Currently Smokes × Has Health Insurance × Student × Working ×

How many explanations do you want?

10

Specify the balance between diversity and utility

Utility: 50%, Diversity: 50%

Choose a scenario Discovering Reverse Trends

Data

	LikelihoodFeelingNervous	Sex	MaritalStatus	Region	Race	Education	Age	IsWorking
0	100%	Male	NeverMarried	Midwest	White	UnAcceptable	51.00-63.00	NotEmployed
1	0%	Female	Divorced	Midwest	White	UnAcceptable	51.00-63.00	NotEmployed
2	100%	Female	NeverMarried	Midwest	White	UnAcceptable	<29.00	ActivelyWorking
3	100%	Female	Married	South	White	UnAcceptable	<29.00	ActivelyWorking
4	0%	Female	NeverMarried	South	Asian/Hawaiian	UnAcceptable	<29.00	ActivelyWorking
5	0%	Female	Married	South	Asian/Hawaiian	UnAcceptable	51.00-63.00	ActivelyWorking

🟦 Males 🟪 Non-males

Causal Explanations

	Subpopulation	Support	Treatment	Avg likelihood of feeling nervous for Group A	Avg likelihood of feeling nervous for Group B	Avg Treatment Effect for Group A	Avg Treatment Effect for Group B
E1	Divorced individuals aged 51 to 63, and their doctor recommends exercise	444(2.19%)	Does not currently smoke	47.48%	42.95%	21.46%	-14.15%
E2	Individuals aged 51 to 63, divorced, without a diagnosis of asthma, and whose doctors recommend exercise	373(1.84%)	Does not currently smoke	46.03%	42.11%	22.55%	-18.37%
E3	Divorced individuals aged 51 to 63 who were born in the USA and whose doctor recommends exercise	357(1.76%)	Does not currently smoke	50.0%	45.57%	27.17%	-13.07%
E4	Divorced, aged between 51 and 63, recommended by a doctor to exercise, born in the USA, and not diagnosed with asthma	294(1.45%)	Does not currently smoke	48.62%	44.86%	25.7%	-16.4%
E5	Born in the USA, from the Midwest, never married, under 29 years old, and without a diagnosis of asthma	491(2.42%)	Has health insurance	46.64%	45.85%	20.62%	13.74%

Legend for Average Treatment Effect

Decrease No Effect Increase

Overall Quality 70%

Utility 65%

Diversity 85%

Figure 2: The ExDis demo: Ⓐ: upload data and causal DAG, Ⓑ: specify groups of interest, Ⓒ: specify target, immutable, and actionable attributes, Ⓓ: specify the budget k and balance parameter α, Ⓔ: choose a scenario and request explanation, Ⓕ: the dataset, Ⓖ: causal explanation set generated by ExDis, Ⓗ: view quality of the generated explanation set.

Step Ⓔ. For our demonstration, the user chooses "Discovering reverse trends" as the scenario. In this scenario, the goal is to find opposite trends (one increasing and the other decreasing) between the two groups. The other options are: "Investigating disparate trends"—where the difference between the two groups matters the most, regardless of their direction—and "Debugging bias"—where the user wants to find causes of discrimination against one specific subgroup with respect to the entire dataset.

Step Ⓕ. In this step, the user views the dataset (partially shown), color-coded by the two groups of interest. The first row is `Male` and the rest are `Non-males`.

Step Ⓖ. ExDis generates a set of 10 explanations (only the first 5 are shown here), balancing the utility and diversity equally. The first 4 explanations show opposite trends for the subpopulations, where the first explanation corresponds to the one described in Example 1. Any shade of red implies that a decrease in the target attribute is expected after treatment and any shade of green indicates an increase in the target attribute is expected after treatment.

Step Ⓗ. The user observes the overall quality (70%) of the disparity explanation set, along with utility (65%) and diversity (85%) scores. If the user can go back to step Ⓔ and tune the balance between utility and diversity.

After the guided demonstration, participants may use ExDis to explore their own datasets. The main goal of this demonstration is to showcase how ExDis enables users to explore various scenarios to gain insights regarding disparity, according to their preferences.

REFERENCES

[1] C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[2] H. Lakkaraju, S. H. Bach, and J. Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *SIGKDD*, 2016.

[3] P. Ma, R. Ding, S. Wang, S. Han, and D. Zhang. XInsight: EXplainable Data Analysis Through The Lens of Causality. *Proc. ACM Manag. Data*, 2023.

[4] E. Pastor, L. De Alfaro, and E. Baralis. Looking for trouble: Analyzing classifier behavior via pattern divergence. In *SIGMOD*, 2021.

[5] J. Pearl. *Causality*. Cambridge university press, 2009.

[6] B. Salimi, J. Gehrke, and D. Suciu. Bias in olap queries: Detection, explanation, and removal. In *SIGMOD*, 2018.

[7] G. Sathe and S. Sarawagi. Intelligent rollups in multidimensional olap data. In *VLDB*, pages 307–316, 2001.

[8] T. Surve and R. Pradhan. Example-based explanations for random forests using machine unlearning. *CoRR*, abs/2402.05007, 2024.

[9] B. Youngmann, M. Cafarella, A. Gilad, and S. Roy. Summarized causal explanations for aggregate views. *SIGMOD*, 2(1), 2024.