DEMA: Enhancing Causal Analysis through Data Enrichment and Discovery in Datalakes*

Kayvon Heravi[1,†], Saathvik Dirisala[2,†] and Babak Salimi[3]

[1]University of California, San Diego , 9500 Gilman Dr, La Jolla, CA 92093, United States

Abstract

Causal inference is fundamental to decision-making and policy evaluation, addressing pivotal questions that predictive analysis cannot tackle. Its significance spans various domains such as economics, healthcare, marketing, and e-commerce, enhancing trustworthy machine learning through robustness to distribution shifts, ensuring fairness, interpretability, explainability, and generalizability. However, real-world datasets often collected for operational purposes are diverse and incomplete, lacking key confounding variables and including redundant variables, complicating the analysis. This paper presents an initial framework for systematically identifying and integrating relevant data from diverse sources to facilitate robust causal analysis. Our iterative pipeline addresses challenges such as high-dimensional covariates, missing data, and incomplete joins by curating and ranking features based on their impact, leveraging Double Machine Learning to control for confounding factors. Empirical results demonstrate the framework's capability to uncover meaningful causal relationships in complex, multi-source datasets, thereby enhancing the comprehensiveness and accuracy of data and making machine learning models more dependable and interpretable.

Keywords

Data Discovery, Causal Inference, Data Enrichment, DoubleML

1. Introduction

Causal inference is fundamental to decision-making and policy evaluation, addressing pivotal questions that predictive analysis cannot tackle. Its significance spans economics, healthcare, marketing, and e-commerce, offering critical insights into various domains. Furthermore, its pivotal impact has been recognized for enhancing trustworthy machine learning through robustness to distribution shifts and domain adaptation, ensuring fairness, interpretability, explainability, generalizability, representation learning, and beyond [1, 2, 3].

However, in fields such as economics, epidemiology, and social sciences, causal inference has been successful because the datasets are carefully collected and curated by sophisticated statisticians to test specific hypotheses [4, 5, 6]. These datasets are meticulously designed to ensure completeness and relevance. In the real world, however, data is often collected for various operational purposes rather than for testing specific hypotheses, leading to diverse and often incomplete datasets that are not ready for causal analysis. These real-world datasets can lack key confounding variables or factors that affect an outcome of interest and may include redundant variables that complicate the analysis [7, 8].

Despite these challenges, opportunities provided by data discovery and enrichment are crucial for facilitating causal discovery and data curation for causal inference and causal machine learning. Open datalakes, which aggregate vast amounts of data from multiple sources, can be leveraged to overcome the limitations of real-world data. Advanced data discovery tools can sift through these datalakes to identify relevant datasets, ensuring that all pertinent factors are included in the analysis. By integrating and enriching data from diverse sources, these tools enhance the comprehensiveness and accuracy of the datasets, supporting robust causal inference. This, in turn, improves the reliability of causal effect estimations, making machine learning models more dependable and interpretable.

This paper presents an initial framework DEMA (**D**ata **E**nrichment and **M**erging for Causal **A**nalysis) for data curation for causal inference, aiming to systematically identify and integrate relevant data from diverse sources to facilitate robust causal analysis. This process is challenging due to high-dimensional covariates, missing data, and the issue of incomplete joins in database systems, where attempting a full outer join often results in sparse or empty tables because not all tuples from different datasets have matching keys. Furthermore, aggregating data according to each unit in the base table using pre-defined aggregations can lead to loss of information, introducing biases in the analysis. To address these challenges, we propose an iterative pipeline that curates and ranks features based on their impact, which cannot be explained by other covariates.

To manage high dimensionality, we employ Double Machine Learning (DoubleML), a framework that combines machine learning with econometric techniques

2nd International Workshop on Tabular Data Analysis (TaDA)

† These authors contributed equally.

kheravi@ucsd.edu (K. Heravi); sdirisala@ucsd.edu (S. Dirisala); bsalimi@ucsd.edu (B. Salimi)

to control for confounding factors and ensure robust causal inference [9, 10]. Our method leverages a prebuilt data discovery tool that identifies relevant datasets in a datalake based on a candidate column, facilitating the aggregation of similar data across diverse sources. Once datasets are identified and merged, DoubleML is employed to address potential biases and provide reliable causal effect estimations. The utility of our approach is demonstrated through a series of experiments that highlight its capability to uncover meaningful causal relationships in complex, multi-source datasets, thereby enhancing the comprehensiveness and accuracy of the data and making machine learning models more dependable and interpretable.

2. Background on Causal Inference

The goal of *causal inference* is to estimate the effect of a *treatment variable* T on an outcome variable Y. For instance, in the context of our study, we might want to know the effect of high precipitation (T) on the number of collisions (Y). The gold standard of causal inference is *randomized controlled experiments*, where the population is randomly divided into a *treated* group that receives the treatment (denoted by $do(T = 1)$ for a binary treatment [1]) and a *control* group ($do(T = 0)$). One popular measure of this effect is the *Average Treatment Effect* (ATE). In a randomized experiment, the ATE is the difference in the average outcomes for the treated and control groups [1, 11]:

$$\text{ATE}(T, Y) = \mathbb{E}[Y \mid do(T = 1)] - \mathbb{E}[Y \mid do(T = 0)] \quad (1)$$

Randomized experiments are often infeasible, and in practice, we need to estimate causal effects from observational data, which is collected passively. Business data is inherently observational. *Observational Causal Analysis* offers a reliable method for causal inference with specific assumptions. Controlled trials with randomization address the issue of *confounding factors*—variables influencing both treatment and outcome. One can adjust for these covariates or confounders Z, which should be identified from background knowledge, to achieve unbiased causal inferences from observational data. Two essential assumptions are *Unconfoundedness*: $Y \perp T \mid Z = z$ and *Overlap*: $0 < \Pr(T = 1 \mid Z = z) < 1$. Under these conditions, the average treatment effect (ATE) is expressed as:

$$\text{ATE}(T, Y) = \mathbb{E}_Z\big[\mathbb{E}[Y \mid T = 1, Z = z] - \mathbb{E}[Y \mid T = 0, Z = z]\big] \quad (2)$$

Equation 2 can be estimated from data. There are various methodologies for estimating the ATE in Equation 2. One popular technique is *matching methods* [12], which pair treated and untreated units based on their observed covariates to mitigate confounding bias. However, matching methods often struggle in high-dimensional settings. To address these challenges, both parametric and semi-parametric techniques have been developed, incorporating a range of regression models and advanced machine learning algorithms. These techniques typically estimate the *propensity score* ($m_0(\mathbf{X}) = \mathbb{E}[T = 1 \mid \mathbf{X}]$), which quantifies the probability of treatment given covariates $\mathbf{X}$, and the *prognostic score* ($g_0(\mathbf{X}) = \mathbb{E}[Y \mid T = 0, \mathbf{X}]$), which predicts the expected outcome without treatment. A state-of-the-art methodology in this field is *Double Machine Learning (DML)* [9], which uniquely combines both propensity and prognostic scores. DML ensures robust causal effect estimation by controlling for model misspecification and leveraging the complexity of machine learning models, making it particularly effective for causal inference in high-dimensional data settings.

3. DEMA Methodology

The overall architecture of DEMA is illustrated in Figure 1. The input is a database instance D from a schema $\mathbf{S}(\kappa, \mathbf{X}, T, Y)$, which contains N units of analysis, such as patients, transactions, or events, hence we refer to it as the *unit table*. Each unit is uniquely identified by a key κ_i, for $i \in \{1, 2, \ldots, N\}$, and consists of a set of attributes $\mathbf{X}_i$ and an outcome variable Y_i. The goal of our system is to effectively integrate and analyze data from various sources to curate a dataset that contains a set of features that potentially causally impact the outcome of interest. This curated data is amenable for causal analysis, i.e., it includes a set of relevant features that are correlated with outcomes of interest and such that their correlation cannot be explained by other attributes in the datalake. The data, together with the generated report, can be further used to decide whether additional data collection is necessary or if the current dataset is sufficient for causal analysis based on expert judgment and contextual understanding. The pipeline includes several key components: Data Discovery, Join Viability Assessment, Data Integration and Enrichment, and Impact Analysis that work hand in hand in an iterative process to achieve robust causal inference. Next, we discuss each component in detail.

Data Discovery. The process begins with data discovery, where we explore a datalake to identify relevant datasets that can augment our unit table. Using the *joinable attributes* of D as query columns, i.e., columns that can be used for joining and integration,

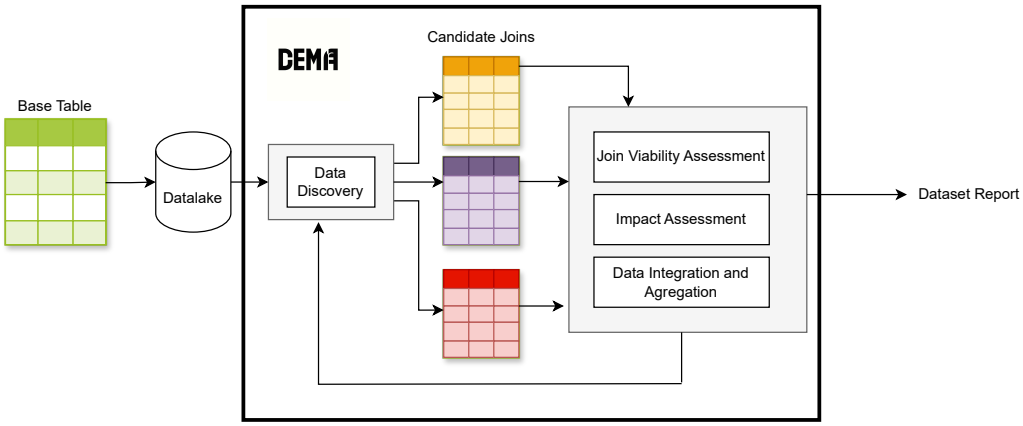

Figure 1: A visual representation of our pipeline Discovery-Enrich-Merge-Analyze (DEMA). We start with a base table fed to a datalake that returns candidate tables and joins. These candidates undergo merging, aggregation, and causal inference using DoubleML, leading to our final results.

we search for candidate tables within the datalake that share high similarities with these columns. Examples of joinable attributes include `patient_id`, `date`, `location`, `zipcode`, `product_id`, `transaction_id`, and `employee_id`, which are common across different tables and can be used to collect more fine-grained information from other tables. Common techniques and existing tools for data discovery could be used here [13, 14, 15, 16, 17]. In this work, DEMA utilizes exact matching as described in [15], where the Jaccard containment between the query column and all other columns in the datalake is computed, providing precise rankings and selecting the most relevant tables for augmentation.

Note that while these joinable features could be treated as both treatments and covariates, data integration based on these features does not bring additional information from an information-theoretic perspective due to functional dependencies [18]. The contributions of data enrichment for causal inference are twofold: 1) Working with joinable features as covariates is often infeasible because these features typically have many distinct values, leading to a violation of overlap. This necessitates the use of one-hot encoding, which results in high dimensionality and poor small-sample properties for causal effect estimators. 2) More fine-grained features provide more interpretable results and allow for more nuanced covariate selection and causal discovery. For instance, in the context of our example, the base table only provides data about collisions and dates. While the attribute date could be featurized into year, time of week, month, season, etc., using it as a joinable variable may not theoretically bring more information about collisions since date functionally determines weather-related information. However, joining it with a weather table allows us to identify which specific weather-related features impact collisions. For instance, if we initially realize that the month extracted from the date highly affects collision data, data discovery enables us to replace the month with more precise weather data, revealing that precipitation, rather than the month, is actually the significant factor.

Join Viability Assessment. Once potential tables are identified, we analyze candidate joins. Given a relevant table R and a joinable variable J, a full outer join often results in sparse or empty tables because not all tuples from different datasets have matching keys. This could be due to several factors, each of which is treated differently and reported: 1) Missing or incomplete data in one or more tables, where we lack information for certain units or their joinable attributes, such as having lab test results only for a subset of patients, or sales data only for certain regions. 2) Data quality issues, such as inconsistencies in data entry, different date formats, or typographical errors in `zipcode` entries. 3) Missingness could be a feature itself; for instance, in a table that has most information but lacks specific data for some units, the absence could indicate a significant underlying factor, like lack of access to services, or unavailability of a product in certain regions. 4) The inherent heterogeneity of data, where units are not homogeneous. For example, products may fundamentally differ, or different types of customers may have varying features, necessitating partitioning the data and analyzing each subpopulation independently.

We capture all these cases using an indicator variable, a binary variable that shows whether a unit in the unit table successfully joins with the relevant table on the joinable attribute. Formally, given a unit i in the unit

table D and relevant table R with a joinable attribute A, the indicator variable $I_{R,A}(i)$ is defined as:

$$I_{R,A}(i) = \begin{cases} 1 & \text{if unit } i \text{ has a matching tuple in } R \text{ on } A, \\ 0 & \text{otherwise.} \end{cases}$$

This variable can be used as a feature itself, and in the Impact Assessment, which we explain below, we analyze its impact on outcomes of interest. The impact could reveal vital information. If the missingness is due to incomplete information, any obtained correlation indicates that missingness is not random, and results obtained after the join could be highly biased. In such cases, careful consideration and possibly additional data collection are required. In the case of heterogeneity and when missingness is a feature itself, this indicator highlights that the type of product or the feature indicated by missingness potentially impacts the outcome. For instance, the lack of access to medical services could significantly affect health outcomes, or the absence of certain products in specific regions could impact sales performance. Thus, DEMA performs a detailed assessment of each join to ensure robust causal inference.

Data Integration and Aggregation. After Join Viability Assessment, the joins are performed, and the data is summarized for many-to-one and many-to-many joins using aggregation. Summarization is needed since integration and enrichment should not change the units of analysis, which is pivotal for causal and statistical analysis. Joining without aggregation distorts the distribution of the unit table and can lead to misleading results. However, since aggregation can lead to loss of information, DEMA performs sensitivity analysis to assess the impact of this loss and ensure that the aggregated data still captures the essential causal relationships.

Impact Assessment. After Data Integration and Aggregation, the next step is to evaluate the impact of the attributes. This involves updating the impact of attributes that were previously present and computing the impact of new attributes obtained through enrichment. The new features can now be used as additional covariates. To achieve this, covariate selection is performed. DEMA uses Large Language Models (LLMs), in particular GPT-4, for covariate selection, which has been shown to be effective in identifying relevant features in high-dimensional data [19, 8, 20]. This selection process is crucial for improving the robustness and validity of the causal inference. We then use DoubleML, which is particularly effective in dealing with very high-dimensional data, a common scenario in our domain where integration involves potentially several tables.

Putting Everything Together. Starting with the base table, DEMA evaluates the existing features and performs feature engineering to extract as many features as possible for attributes with a large number of distinct values, since these values cannot be directly used for causal analysis and are candidates for use as joinable variables for data discovery. Feature engineering helps identify which variables should be used for data discovery and in what order. DEMA uses an impact assessment module to compute the effect of each variable and rank them based on their importance. We then rank all candidate joinable variables accordingly and start discovery with the top-ranked ones using the external datalake. For each candidate table returned by the data discovery tool, DEMA performs a join viability assessment, generates a report, and then performs integration and aggregation for data enrichment if feasible. With these enriched base tables, we run the same analysis as earlier, retaining only attributes that have a significant impact on the outcome and discarding the rest. This process is performed recursively, identifying new sets of joinable variables. We rank features based on their importance and explore the space of all possible joins based on these rankings. This recursive process continues until no more new features are added and the ranking of the most important features becomes stationary.

4. Experiments

This section evaluates the efficacy of using DEMA for data enrichment tailored for causal inference.

4.1. Setup

Data Our datalakes and base data tables are extracted from the NYC Open Data resource [21].

Scenario 1: Taxi Collisions. The base table for analysis consists of taxi collisions, including attributes such as date and number of collisions. The target column is the number of collisions. The primary key for this table is the date, which is initially used to join with various other datasets from New York City.

Scenario 2: School Progress Reports. The base table for analysis consists of school progress reports, which include information on school type, scores in various categories relating to school environment, college readiness, and more. The target column is the school percentile, which scores school quality on a scale from zero to one hundred, based on multiple factors including the school environment and graduation rates. The primary key for this table is DBN (District, Borough, and Number), which is initially used for data discovery and joining school tables across New York City.

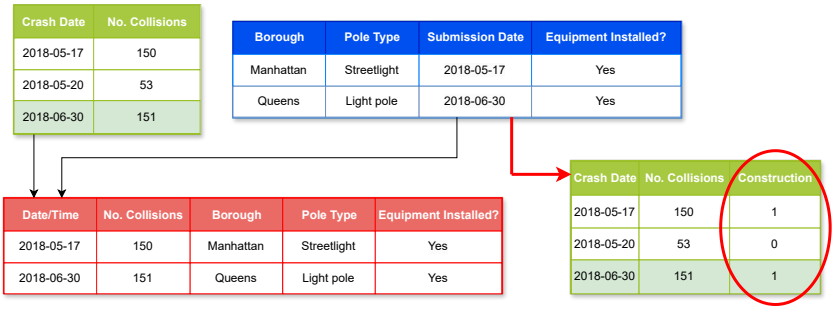

Figure 2: Example illustrating the potential problem of data missingness. The red circle highlights the new column added to show whether the date from the pole construction data appears in our base table.

We used the following two datalakes: 1) New York City: This datalake includes various datasets from New York City, such as weather data, construction records, NYPD reports, and more. 2) Schools: This datalake contains multiple datasets related to New York City schools, including school discipline records, district information, survey results, school performance reports, and more.

Implementation Details For feature engineering and covariate detection, we leveraged GPT-4 to automate these tasks. For impact assessment, we employed DoubleML, using four models: Lasso Regression, XGBoost, Random Forest, and Decision Tree. If at least three of the four models returned significant results (p-value < 0.05), we kept the variable and reported the magnitude of the mean coefficient.

4.2. Results

Taxi Collisions in New York City. The goal of this scenario is to identify impactful factors that contribute to the number of collisions. Firstly, we featurize our base table where our primary key of date is translated to years, months, and weekdays. Initial results from analysis on the base table show that while the year has the highest impact on collisions, its effect is relatively marginal. To gain deeper insights, we aim to use the DEMA pipeline to enrich our base table and analysis by integrating it with various datasets from the New York City datalake.

In each iterative experiment, we generate a report highlighting the top impactful factors and tables. Our analysis reveals that precipitation and the weather table are the most influential factors, significantly affecting collision rates. This is in addition to our featurized attribute of the year, which, though initially having a strong impact, is further contextualized by the additional data.

By recursively joining the tables in order of their impact, we systematically enhance our dataset. This approach allows us to build a more comprehensive and nuanced understanding of the factors contributing to collisions. As shown in Table 1, our final report from the recursive joins provides a detailed picture of the key determinants of collision rates. By the third iteration, we find that precipitation and crash month are the most impactful factors. This highlights that precipitation consistently shows the highest impact on collisions in all our tests, and that certain months play a significant role, possibly due to increased traffic during those times.

In these experiments, it is crucial to acknowledge that the discovered tables may be inherently biased due to the nature of data collection. For instance, a pole construction table that only includes data from days when construction occurred introduces a significant bias. This bias cannot be controlled when joining it with the base table, potentially obscuring true causal relationships and reducing the validity of our findings, as discussed in the Join Viability Assessment of our pipeline. As illustrated in Figure 2, our pole construction table only holds data for days where construction on street poles were ongoing. Consequently, when joining this table with our base table, the resulting dataset only includes days with construction activities. Figure 3 presents the results of this analysis, highlighting the indications of biased tables, with the most significant result being the pole construction table. When we compare our new pole construction results to our original results obtained using the DEMA Pipeline, we observe a higher impact, indicating that days with construction activities correlate with a higher number of collisions.

Our results can help local officials in targeted interventions and improve traffic safety around construction zones, ultimately reducing the risk of accidents and enhancing public safety.

School Performance Using the school datalake, we aim to determine which attributes are indicative of a school's percentile, reflecting the overall quality and performance of the school.

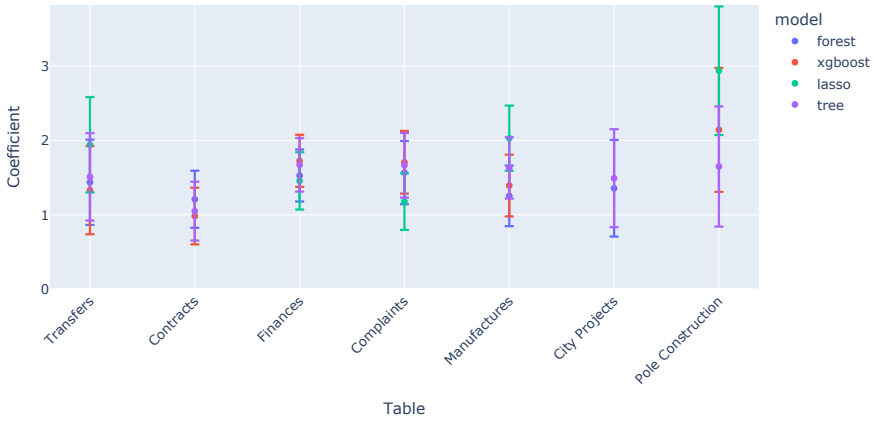

Figure 3: Join Viability Assessment report for tables from the New York City data lake using the base table of taxi collisions

We first featurize DBN into district and borough, and our initial analysis reveals that the district has the highest impact on school percentile. This finding sets a baseline for understanding how geographic and administrative divisions influence school performance. However, to delve deeper, we apply our iterative DEMA pipeline, which refines and expands our analysis.

Upon applying our iterative run, the pipeline returns a table ranking where the two most impactful tables are physical education instructors and math proficiency. Notably, the highest impact within the physical education table is the "ratio of full-time licensed PE teachers to students" indicating that schools with higher teacher-to-student ratios tend to have higher ranking scores. This result suggests that schools with higher teacher-to-student ratios tend to achieve better overall ranking scores more so than the idea that more physical education leads to higher school performance. We enhance our dataset by recursively joining the tables according to our threshold standards which builds a more comprehensive final joined dataset. Combined with the significance of the PE teacher-to-student ratio and the number of disciplinary actions imposed on students, we can illustrate a multifaceted picture of what drives school performance. These final results as shown in Table 2 are intriguing as they underscore the importance of both academic and non-academic factors in evaluating school quality, with the identification of disciplinary actions as a key indicator suggesting a complex interplay between school environment and student outcomes.

Our results obtained through the DEMA pipeline can allow for school administration to implement changes that may benefit the school performance.

Iteration	Table Joined	Most Impactful
1	Weather	Precip, Air Temp
2	Project Status	Precip, Air Temp
3	Manufactures	Precip, Crash month

Table 1
Top 2 factors after every subsequent recursive join for the taxi collisions experiment

Iteration	Table Joined	Most Impactful
1	Math	Level 4 Profficiency
2	PE teachers	Ratio of licensed teachers to students
3	Discipline	Profane, Vulgar language

Table 2
Top 1 factor after every subsequent recursive join for the school ratings experiment

5. Related Works

The field of data discovery has seen significant advancements, particularly in the integration of various data sources such as data lakes. Previous work has focused on the creation, integration, augmentation, searching, and analysis of data within these data lakes [14, 16, 22, 17]. These efforts have laid the groundwork for understanding how to effectively manage and utilize large and diverse datasets for various analytical purposes.

Several frameworks have been developed to integrate external data with causal inference, highlighting the importance of leveraging multiple data sources for more robust causal analysis. Notable contributions include methodologies for causal integration and analysis in complex data environments [7, 8, 23]. Our aim is to build upon these established frameworks to explore the causal analysis of data integration through data discovery. By utilizing enriched datasets from external data lakes, we seek to enhance the accuracy and robustness of causal in-

ferences, addressing gaps and extending the capabilities of existing methodologies.

6. Conclusions and Future Directions

In this paper, we developed an initial framework for data discovery and causal inference using datalakes, demonstrating its effectiveness through practical scenarios involving real-world datasets. Our approach enables users to identify and rank the most causally significant attributes based on their queries, leveraging data discovery tools and advanced causal inference methods. Our empirical results indicated the framework's potential in uncovering significant causal relationships. Our framework took initial steps to address the challenges of data discovery and integration for data curation for causal inference in large, diverse datasets, significantly reducing manual effort and ensuring robust results. This work contributes to the field of data discovery and causal inference, offering a scalable solution adaptable to various domains.

Future work may explore further optimization of the pipeline and its application to additional datasets, enhancing its utility and impact in data-driven decision-making processes. Specific areas for improvement include optimizing the entire pipeline, particularly the recursive processes and the retraining of models given new covariates. There is significant potential for reusing intermediate results and avoiding the need to retrain models from scratch, which would improve efficiency and reduce computational costs. Additionally, novel methods are needed to handle heterogeneous units and address biases induced by data quality issues and missingness, which can significantly affect the reliability of the analysis. Further research could also investigate the development of adaptive algorithms that dynamically adjust the pipeline based on the characteristics of the incoming data, improving the robustness and accuracy of causal inference. Expanding the framework's applicability to various domains, including healthcare, finance, and social sciences, would also demonstrate its versatility and broad impact.

References

[1] J. Pearl, Causal inference in statistics: An overview (2009).

[2] C. Avin, I. Shpitser, J. Pearl, Identifiability of path-specific effects (2005).

[3] S. Galhotra, Y. Brun, A. Meliou, Fairness testing: testing software for discrimination, in: Proceedings of the 2017 11th Joint meeting on foundations of software engineering, 2017, pp. 498–510.

[4] A. Bennett, Causal inference and policy evaluation from case studies using bayesian process tracing, in: Causality in Policy Studies: a Pluralist Toolbox, Springer International Publishing Cham, 2023, pp. 187–215.

[5] F. Kühne, M. Schomaker, I. Stojkov, B. Jahn, A. Conrads-Frank, S. Siebert, G. Sroczynski, S. Puntscher, D. Schmid, P. Schnell-Inderst, et al., Causal evidence in health decision making: methodological approaches of causal inference and health decision science, GMS German Medical Science 20 (2022).

[6] S. Listl, H. Jürges, R. G. Watt, Causal inference from observational data, Community dentistry and oral epidemiology 44 (2016) 409–415.

[7] E. Bareinboim, J. Pearl, Causal inference and the data-fusion problem, Proceedings of the National Academy of Sciences 113 (2016) 7345–7352.

[8] B. Youngmann, M. Cafarella, B. Salimi, A. Zeng, Causal data integration, 2023. arXiv:2305.08741.

[9] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, J. Robins, Double/debiased machine learning for treatment and structural parameters, 2018.

[10] P. Bach, M. S. Kurz, V. Chernozhukov, M. Spindler, S. Klaassen, DoubleML: An object-oriented implementation of double machine learning in R, Journal of Statistical Software 108 (2024) 1–56. doi:10.18637/jss.v108.i03.

[11] D. B. Rubin, Causal inference using potential outcomes: Design, modeling, decisions, Journal of the American Statistical Association 100 (2005) 322–331.

[12] S. L. Morgan, D. J. Harding, Matching estimators of causal effects: Prospects and pitfalls in theory and practice, Sociological methods & research 35 (2006) 3–60.

[13] N. Chepurko, R. Marcus, E. Zgraggen, R. C. Fernandez, T. Kraska, D. Karger, Arda: automatic relational data augmentation for machine learning, Proc. VLDB Endow. 13 (2020) 1373–1387. URL: https://doi.org/10.14778/3397230.3397235. doi:10.14778/3397230.3397235.

[14] R. Castro Fernandez, Z. Abedjan, F. Koko, G. Yuan, S. Madden, M. Stonebraker, Aurum: A data discovery system, in: 2018 IEEE 34th International Conference on Data Engineering (ICDE), 2018, pp. 1001–1012. doi:10.1109/ICDE.2018.00094.

[15] R. Cappuzzo, G. Varoquaux, A. Coelho, P. Papotti, Retrieve, merge, predict: Augmenting tables with data lakes, ArXiv (2024).

[16] M. J. Cafarella, A. Halevy, N. Khoussainova, Data integration for the relational web, Proceedings of

the VLDB Endowment 2 (2009) 1090–1101.
[17] S. Castelo, R. Rampin, A. Santos, A. Bessa, F. Chirigati, J. Freire, Auctus: A dataset search engine for data discovery and augmentation, Proceedings of the VLDB Endowment 14 (2021) 2791–2794.
[18] A. Kumar, J. Naughton, J. M. Patel, X. Zhu, To join or not to join? thinking twice about joins before feature selection, in: Proceedings of the 2016 International Conference on Management of Data, 2016, pp. 19–34.
[19] X. Liu, P. Xu, J. Wu, J. Yuan, Y. Yang, Y. Zhou, F. Liu, T. Guan, H. Wang, T. Yu, et al., Large language models and causal inference in collaboration: A comprehensive survey, arXiv preprint arXiv:2403.09606 (2024).
[20] S. Ashwani, K. Hegde, N. R. Mannuru, M. Jindal, D. S. Sengar, K. C. R. Kathala, D. Banga, V. Jain, A. Chadha, Cause and effect: Can large language models truly understand causality?, arXiv preprint arXiv:2402.18139 (2024).
[21] NYC Open Data, https://opendata.cityofnewyork.us/, 2024.
[22] A. Y. Halevy, F. Korn, N. F. Noy, C. Olston, N. Polyzotis, S. Roy, S. E. Whang, Managing google's data lake: an overview of the goods system., IEEE Data Eng. Bull. 39 (2016) 5–14.
[23] X. Shi, Z. Pan, W. Miao, Data integration in causal inference, Wiley Interdisciplinary Reviews: Computational Statistics 15 (2023) e1581.