DEMA: Enhancing Causal Analysis through Data Enrichment and Discovery in Data Lakes

Kayvon Heravi
University of California, San Diego
San Diego, United States
kheravi@ucsd.edu

Saathvik Dirsala
University of California, San Diego
San Diego, United States
sdirisala@ucsd.edu

Babak Salimi
University of California, San Diego
San Diego, United States
bsalimi@ucsd.edu

ABSTRACT

Understanding causal relationships is crucial in fields like economics, healthcare, marketing, and e-commerce for effective decision-making. Unlike predictive analysis, causal inference provides deeper insights into outcomes. However, real-world datasets often lack key variables and contain redundancies, complicating analysis. This paper introduces a framework that integrates relevant data from varied sources to facilitate robust causal analysis. Our iterative pipeline addresses high-dimensional covariates, missing data, and incomplete joins using Double Machine Learning to control for confounding factors. Empirical results show the framework's ability to uncover meaningful causal relationships, enhancing data accuracy and improving the reliability of machine learning models.

VLDB Workshop Reference Format:
Kayvon Heravi, Saathvik Dirsala, and Babak Salimi. DEMA: Enhancing Causal Analysis through Data Enrichment and Discovery in Data Lakes. VLDB 2024 Workshop: Tabular Data Analysis Workshop (TaDA).

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 [2, 12, 21].

In fields like economics, epidemiology, and social sciences, causal inference succeeds because datasets are carefully collected and curated to test specific hypotheses [5, 14, 16]. These datasets ensure completeness and relevance. In contrast, real-world data, often collected for operational purposes, is diverse and incomplete, lacking key confounding variables and including redundant ones, complicating causal analysis [4, 24].

Despite these challenges, data discovery and enrichment are essential for effective causal analysis. Open data lakes can mitigate real-world data limitations by aggregating diverse data sources. Advanced data discovery tools identify relevant datasets, ensuring comprehensive and accurate causal inference. This integration enhances dataset quality, improving the reliability and interpretability of causal effect estimations and machine learning models.

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 use Double Machine Learning (DoubleML), which combines machine learning with econometric techniques to control confounding factors and ensure robust causal inference [3, 11]. A data discovery tool identifies and merges relevant datasets in a data lake, allowing DoubleML to address biases and provide reliable causal effect estimations. Experiments demonstrate our approach's ability to uncover meaningful causal relationships in complex datasets, enhancing data accuracy 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 [21]) 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 [21, 22]:

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

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

Proceedings of the VLDB Endowment. ISSN 2150-8097.

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:

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

Equation 2 can be estimated from data. There are various methodologies for estimating the ATE in Equation 2. One popular technique is *matching methods* [19], 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)* [11], 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 shown in Figure 1. The input is a database instance D from a schema $\mathbf{S}(\kappa, \mathbf{X}, T, Y)$, containing N units of analysis, such as patients, transactions, or events, referred to as the *unit table*. Each unit is identified by a key κ_i, with attributes $\mathbf{X}_i$ and an outcome variable Y_i. The system integrates and analyzes data from various sources to curate a dataset with features that causally impact the outcome. This curated data is suitable for causal analysis, ensuring relevant features are correlated with outcomes without being explained by other attributes. The data and generated report help determine if additional data collection is needed. The pipeline includes Data Discovery, Join Viability Assessment, Data Integration and Enrichment, and Impact Analysis, working iteratively to achieve robust causal inference. Next, we detail each component.

Data Discovery. The process begins with data discovery, where we explore a data lake 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, we search for candidate tables within the data lake 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 [6–10]. In this work, DEMA utilizes exact matching as described in [7], where the Jaccard containment between the query column and all other columns in the data lake is computed in order to select relevant tables for augmentation.

While joinable features can be treated as treatments and covariates, data integration based on these features does not add information due to functional dependencies [15]. Data enrichment for causal inference has two main contributions: 1) Using joinable features as covariates is often infeasible due to many distinct values, leading to high dimensionality and poor properties for causal effect estimators. 2) Fine-grained features provide more interpretable results and better covariate selection. For example, instead of using the date from collision data, joining with a weather table can reveal specific weather-related factors like precipitation as significant, rather than just the month.

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 due to several factors: 1) Missing or incomplete data in some tables, lacking information for certain units or their joinable attributes, like lab test results for only a subset of patients. 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, indicating significant underlying factors like lack of access to services. 4) The inherent heterogeneity of data, requiring partitioning and independent analysis of different subpopulations, as products or customer types may fundamentally differ.

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, we analyze its effect on outcomes. If missingness is due to incomplete information, any correlation indicates non-random missingness, potentially biasing results and necessitating additional data collection. When missingness is a feature itself, it highlights how factors like lack of access to services or product unavailability impact outcomes. 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

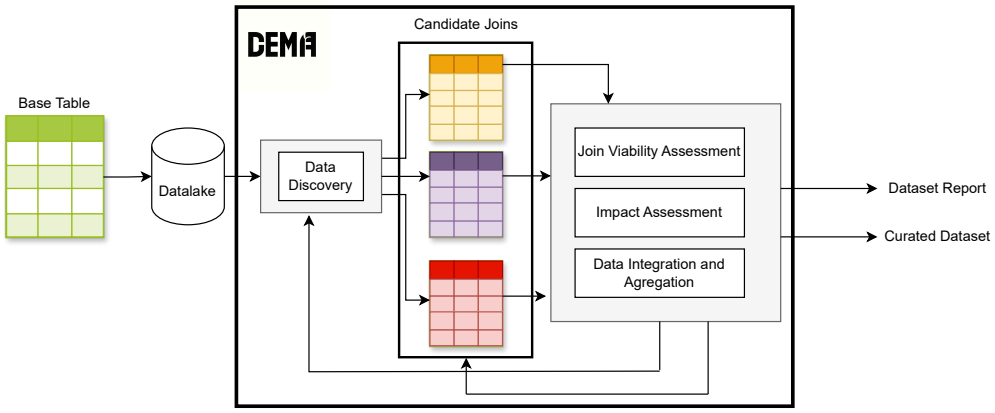

Figure 1: A visual representation of our pipeline Discovery-Enrich-Merge-Analyze (DEMA). We start with a base table fed to a data lake that returns candidate tables and joins. These candidates undergo merging, aggregation, and causal inference using DoubleML, leading to our final results.

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 [1, 18, 24]. 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 existing features and performs feature engineering to extract as many features as possible, identifying which variables should be used for data discovery. The impact assessment module computes and ranks the importance of each variable. We begin discovery with top-ranked joinable variables from the external data lake. For each candidate table, DEMA assesses join viability, generates a report, and performs integration and aggregation if feasible. The enriched base tables are analyzed, retaining only significant attributes. This recursive process continues, identifying new joinable variables and exploring all possible joins until no new features are added and the ranking stabilizes.

4 EXPERIMENTS

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

4.1 Setup

Data. Our data lakes and base data tables are extracted from the NYC Open Data resource [20]. The size of these tables ranges from 100 rows to several million rows. The taxi collision data lake comprises thirty tables, while the school data lake includes twenty tables.

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.

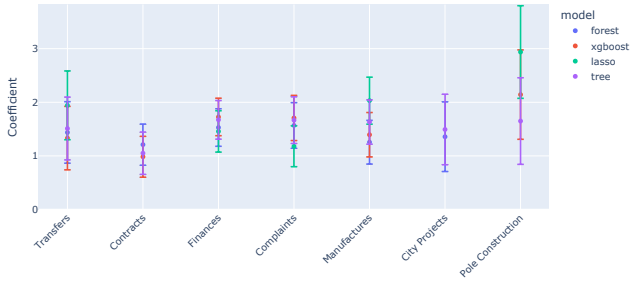

Figure 2: Join Viability Assessment report for tables from the New York City data lake using the base table of taxi collisions

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.

We used the following two data lakes: 1) New York City, with datasets like weather, construction records, and NYPD reports; 2) Schools, with datasets on discipline records, district information, survey results, and performance reports.

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 factors contributing to the number of collisions. We first featurize our base table by translating the date to years, months, and weekdays. Initial results show the year has a marginal impact on collisions. Using the DEMA pipeline, we integrate various datasets from the New York City data lake for deeper insights.

Each iterative experiment generates a report highlighting the most impactful factors and tables. Our analysis reveals that precipitation and weather data are the most influential, significantly affecting collision rates. After our iterative approach, we recursively join tables based on their impact, systematically enhancing our dataset, providing a nuanced understanding of collision determinants. As shown in Table 1, by the final iteration, precipitation and crash month are the most impactful factors.

In these experiments, it's crucial to acknowledge potential biases in discovered tables due to data collection methods, as discussed in the Join Viability Assessment section. For instance, a pole construction table only including data from construction days introduces significant bias, potentially obscuring true causal relationships and reducing the validity of our findings. When joining this table with our base table, the dataset only includes construction days. Figure 2 shows that this bias results in a higher observed impact, indicating that construction days correlate with more collisions.

Our results can help local officials target interventions and improve traffic safety around construction zones, reducing accidents and enhancing public safety.

School Performance. Using the school data lake, we aim to identify attributes indicative of a school's percentile, reflecting overall quality and performance.

We first featurize DBN (District, Burough, Number) into district and borough, finding that the district has the highest impact on school percentile, setting a baseline for understanding geographic influences on performance. To delve deeper, we apply our iterative DEMA pipeline.

The pipeline ranks tables by impact, revealing physical education instructors and math proficiency as the most impactful. The "ratio of full-time licensed PE teachers to students" shows that schools with higher teacher-to-student ratios tend to have higher scores. This suggests that teacher-student ratios are more significant than the amount of physical education. By recursively joining tables, we enhance our dataset, showing that both academic and non-academic factors, including disciplinary actions, are key indicators of school performance. As shown in Table 2, with our second table of Math being added, students with high math proficiency and ratio of teachers to students from the PE Teachers table is our highest impact. The final iteration shows that Math proficiency and school discipline have the highest effect compared to our original findings of ratio of teacher to student ratio with our more comprehensive final dataset.

These results, shown in Table 2, highlight the complex interplay between school environment and student outcomes. The DEMA pipeline's findings can help school administration implement beneficial changes to improve performance.

Iteration	Table Joined	Most Impactful
0	Base Table	Year, Month
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
0	Base Table	District, School level
1	+ PE Teachers	Ratio teachers to students, District
2	+ Math	Level 4, Ratio teachers to students
3	+ Discipline	Level 4, Profane language

Table 2: Top 2 factors after every subsequent recursive join for the school ratings experiment

Sensitivity And Runtime Analysis. In our experimentation, we find that our results are consistent across different aggregation methods. While the impact values vary, the top-ranking results remain the same when using mean, median, max, and min aggregations. The runtime for the NYC Taxi collision experiment, which involves 30 tables with a large number of rows, is approximately five minutes. This demonstrates the efficiency of our approach in handling extensive datasets.

5 RELATED WORKS

Data discovery has advanced significantly, especially in integrating data sources like data lakes [6, 8, 9, 13]. These efforts have laid the groundwork for managing and utilizing large, diverse datasets.

Feature augmentation, a key aspect, involves generating new features from existing data to enhance model performance and uncover hidden patterns, improving predictive accuracy and robustness [10, 17].

Several frameworks integrate external data with causal inference, leveraging multiple data sources for robust analysis [4, 23, 24]. Our aim is to build on these frameworks, exploring causal analysis through data discovery and feature augmentation. Utilizing enriched datasets and augmented features enhances the accuracy and robustness of causal inferences, extending existing methodologies.

6 CONCLUSIONS AND FUTURE DIRECTIONS

We developed an initial framework for data discovery and causal inference using data lakes, demonstrating its effectiveness with real-world datasets. Our approach helps users identify and rank causally significant attributes using data discovery tools and advanced causal inference methods. Empirical results highlight its potential in uncovering significant causal relationships, reducing manual effort, and ensuring robust results. This work offers a scalable solution, adaptable to various domains.

Future work may optimize the pipeline, enhance efficiency, and apply it to more datasets. Improvements include handling heterogeneous units, addressing biases, and developing adaptive algorithms to adjust based on data characteristics. Expanding the framework to healthcare, finance, and social sciences would demonstrate its versatility and broad impact.

REFERENCES

[1] Swagata Ashwani, Kshiteesh Hegde, Nishith Reddy Mannuru, Mayank Jindal, Dushyant Singh Sengar, Krishna Chaitanya Rao Kathala, Dishant Banga, Vinija Jain, and Aman Chadha. 2024. Cause and Effect: Can Large Language Models Truly Understand Causality? *arXiv preprint arXiv:2402.18139* (2024).

[2] Chen Avin, Ilya Shpitser, and Judea Pearl. 2005. Identifiability of path-specific effects. (2005).

[3] Philipp Bach, Malte S. Kurz, Victor Chernozhukov, Martin Spindler, and Sven Klaassen. 2024. DoubleML: An Object-Oriented Implementation of Double Machine Learning in R. *Journal of Statistical Software* 108, 3 (2024), 1–56. https://doi.org/10.18637/jss.v108.i03

[4] Elias Bareinboim and Judea Pearl. 2016. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7345–7352.

[5] Andrew Bennett. 2023. Causal Inference and Policy Evaluation from Case Studies Using Bayesian Process Tracing. In *Causality in Policy Studies: a Pluralist Toolbox*. Springer International Publishing Cham, 187–215.

[6] Michael J Cafarella, Alon Halevy, and Nodira Khoussainova. 2009. Data integration for the relational web. *Proceedings of the VLDB Endowment* 2, 1 (2009), 1090–1101.

[7] Riccardo Cappuzzo, Gaël Varoquaux, Aimee Coelho, and Paolo Papotti. 2024. Retrieve, Merge, Predict: Augmenting Tables with Data Lakes. *ArXiv* (2024).

[8] Sonia Castelo, Rémi Rampin, Aécio Santos, Aline Bessa, Fernando Chirigati, and Juliana Freire. 2021. Auctus: A dataset search engine for data discovery and augmentation. *Proceedings of the VLDB Endowment* 14, 12 (2021), 2791–2794.

[9] Raul Castro Fernandez, Ziawasch Abedjan, Famien Koko, Gina Yuan, Samuel Madden, and Michael Stonebraker. 2018. Aurum: A Data Discovery System. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. 1001–1012. https://doi.org/10.1109/ICDE.2018.00094

[10] Nadiia Chepurko, Ryan Marcus, Emanuel Zgraggen, Raul Castro Fernandez, Tim Kraska, and David Karger. 2020. ARDA: automatic relational data augmentation for machine learning. *Proc. VLDB Endow.* 13, 9 (may 2020), 1373–1387. https://doi.org/10.14778/3397230.3397235

[11] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters.

[12] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*. 498–510.

[13] Alon Y Halevy, Flip Korn, Natalya Fridman Noy, Christopher Olston, Neoklis Polyzotis, Sudip Roy, and Steven Euijong Whang. 2016. Managing Google's data lake: an overview of the Goods system. *IEEE Data Eng. Bull.* 39, 3 (2016), 5–14.

[14] Felicitas Kühne, Michael Schomaker, Igor Stojkov, Beate Jahn, Annette Conrads-Frank, Silke Siebert, Gaby Sroczynski, Sibylle Puntscher, Daniela Schmid, Petra Schnell-Inderst, et al. 2022. Causal evidence in health decision making: methodological approaches of causal inference and health decision science. *GMS German Medical Science* 20 (2022).

[15] Arun Kumar, Jeffrey Naughton, Jignesh M Patel, and Xiaojin Zhu. 2016. To join or not to join? thinking twice about joins before feature selection. In *Proceedings of the 2016 International Conference on Management of Data*. 19–34.

[16] Stefan Listl, Hendrik Jürges, and Richard G Watt. 2016. Causal inference from observational data. *Community dentistry and oral epidemiology* 44, 5 (2016), 409–415.

[17] Jiabin Liu, Chengliang Chai, Yuyu Luo, Yin Lou, Jianhua Feng, and Nan Tang. 2022. Feature augmentation with reinforcement learning. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE, 3360–3372.

[18] Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, et al. 2024. Large language models and causal inference in collaboration: A comprehensive survey. *arXiv preprint arXiv:2403.09606* (2024).

[19] Stephen L Morgan and David J Harding. 2006. Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological methods & research* 35, 1 (2006), 3–60.

[20] NYC Open Data. 2024. https://opendata.cityofnewyork.us/.

[21] Judea Pearl. 2009. Causal inference in statistics: An overview. (2009).

[22] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.

[23] Xu Shi, Ziyang Pan, and Wang Miao. 2023. Data integration in causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics* 15, 1 (2023), e1581.

[24] Brit Youngmann, Michael Cafarella, Babak Salimi, and Anna Zeng. 2023. Causal Data Integration. arXiv:2305.08741 [cs.DB]