Digging Up Threats to Validity: A Data Marshalling Approach to Sensitivity Analysis

Digging Up Threats to Validity: A Data Marshalling Approach to Sensitivity Analysis

Anna Zeng
annazeng@mit.edu
MIT CSAIL
Cambridge, MA, USA

Michael Cafarella
michjc@csail.mit.edu
MIT CSAIL
Cambridge, MA, USA

ABSTRACT

Causal inference remains a cornerstone for scientific discovery in the natural and social sciences; however, the accuracy of such causal discoveries is susceptible to unobserved confounding bias, the "Achilles heel of most non-experimental studies". Our principal objective is to bolster the validity of reported causal findings by marshalling pertinent data to corroborate or refute them. In this workshop submission, we describe how data marshalling can turbocharge sensitivity analysis, detail technical challenges, and illustrate a case study as a proof of concept. Our aim in this work is to gather feedback from the audience, gauge interest in solving this open problem relevant to the responsible AI and data management community, and continue iterating on systems that advance the trustworthiness and reproducibility of scientific discoveries.

CCS CONCEPTS

• **Computing methodologies** → **Causal reasoning and diagnostics**; • **Information systems** → **Information retrieval**; *Incomplete data*; *Data analytics*; *Information integration*; *Information retrieval query processing*.

KEYWORDS

Causality, Sensitivity Analysis, Potential Confounder Search

ACM Reference Format:
Anna Zeng and Michael Cafarella. 2024. Digging Up Threats to Validity: A Data Marshalling Approach to Sensitivity Analysis. In *Governance, Understanding and Integration of Data for Effective and Responsible AI (GUIDE-AI '24), June 09–15, 2024, Santiago, AA, Chile.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3665601.3669850

1 INTRODUCTION

Causal inference remains a cornerstone for scientific discovery, helping us answer questions like "Does social media cause teenage mental illness?", or "Does cured meat consumption cause colon cancer?". While manual intervention and randomized control trials are considered gold standard, often such approaches are not feasible or ethically permitted. Causal inference methods give scientists rigorous means to estimate answers from observational data, guiding experiments toward real-world discoveries.

GUIDE-AI '24, June 09–15, 2024, Santiago, AA, Chile

ACM ISBN 979-8-4007-0694-3/24/06
https://doi.org/10.1145/3665601.3669850

However, the accuracy of such causal discoveries can depend on the dataset that was used, as datasets can be susceptible to bias in unexpected ways. One such bias is unobserved confounding bias, where the existence or magnitude of some causal effect is misattributed by unobserved, associated *confounding* factors; when unaddressed, such bias can lead to erroneous conclusions due to the presence of spurious correlations [17]. For example, children sleeping with a light on leading to child myopia [27] was later found to be confounded by parental myopia [37] and no causal effect has been found after adjusting for that factor. Prior literature addresses this "Achilles heel of most non-experimental studies" through sensitivity analysis [22], which allow one to test the robustness of one's conclusions against possible unforeseen or unobserved threats to validity. Sensitivity analysis featured prominently in establishing the link between smoking and lung cancer when Cornfield et al. [8] established that an unobserved confounder would need to be "at least nine-fold more prevalent" among smokers compared to non-smokers to explain away the smoking-lung cancer link; no such confounder was found in the study and the authors asserted such a confounder does not exist. Uses of these analysis methods generally follow this same argument structure [15]; however, the inclusion of new data can dramatically revise confidence in conclusions believed to be strong and defensible [9, 12, 33].

Our principal objective is to automatically find possible unobserved confounders in causal mechanisms by building automatic data marshalling systems. While demand for reproducibility is strong [3], the manual effort required to verify causal claims can be prohibitively high. To address this, we propose constructing systems which can refute causal claims by marshalling pertinent data to corroborate or refute these causal assertions.

In this submission, we describe how data marshalling can turbocharge sensitivity analysis, detail technical challenges, provide a solution sketch, and illustrate a small case study as a proof of concept. We aim to gather feedback from the audience, gauge interest in this open problem relevant to the responsible AI and data management community, and continue iterating on systems that bolster the trustworthiness and reproducibility of scientific discoveries.

Example 1. To illustrate how data management methods can boost the effectiveness of sensitivity analysis, we detail an example inspired by a real-world scenario [9, 33] and illustrated in Table 1:

Liora, an Economics PhD candidate, aims to analyze whether parole decisions are affected by duration relative to a food break, to demonstrate how legal processes can be influenced by seemingly irrelevant factors. Since parole decision sessions are scheduled throughout the workday already, Liora decided to analyze an existing dataset of parole decisions from multiple judges who send prisoners to several major prisons in the area, sorted by time of hearing (see the un-shaded

Keren's Additional Data, Attorney confounds relationship between time and parole decision

Judge	Prisoner	Location	Date	Time	Time after Session Start	Crime	Prisoner Age	Judge Age	Parole Decision	Scheduled Time	Attorney
A	MC	12	9/3	9:00AM	0 min	Theft	29	51	Accept	9:00AM	LW
A	JQ	12	9/3	9:39AM	39 min	Loitering	40	51	Accept	9:30AM	RT
...	...	...	...	...	...	...	...	...	...	...	...
B	PO	14	9/3	4:57PM	117 min	Tax Fraud	56	46	Deny	4:45PM	-
C	LU	11	9/3	5:08PM	128 min	Phishing	24	31	Deny	5:00PM	-
...	...	...	...	...	...	...	...	...	...	...	...

Liora's Original Data, Time after Session Start causally influences Parole Decisions

Table 1: Example Table with conclusions made with or without marshalled data

portion of Table 1). In her data analysis process, she derives a new 'Time after Session Start' metric as her treatment variable and observes a significant discrepancy (65% favorable decisions right after a break to almost 0% favorable right before the next break). She analyzes the robustness of her finding through: the introduction of hypothesized 'dummy' variables, which do not impact the estimated effect; and through incorporating all other possible explanatory factors found in her dataset (severity of crime, judge-specific decision inclinations, prisoner demographics, etc.) which also do not sway her conclusion. She publishes the paper and shares the news through widespread media coverage in popular science news outlets and magazines.

Keren, another Economics PhD candidate, reads Liora's paper and, through gathering of background knowledge and additional data tables from the same municipal data source (see the shaded region on Table 1), rebuts that when controlling for attorney representation, the magnitude of the effect goes away. Unfortunately, Keren's paper is not as widely-communicated to the public as Liora's, and the general public is unaware of this threat to validity.

In this example, if Liora had found the confounder Keren retrieved from the same data source, Liora may have better communicated her finding to the wider community. Liora already put in her best effort in her sensitivity analysis based on the data she had in her dataset; Keren had to put in the work in finding the relevant attribute(s) to compose a credible argument against validity.

Our key insight is this: **scientists show study robustness by describing necessary parameters of unobserved confounders; with data management techniques, we can now directly find and evaluate such threats to validity.** In this way, we can better support what scientists, like Liora and Keren, are already doing by using techniques from the data management research community.

2 PROBLEM DEFINITION & OPPORTUNITIES

Here, we explore how researchers currently find threats to validity, define our problem, and touch on technical challenges and opportunities for such an automatic data marshalling system.

2.1 Finding Threats to Validity

Causal analysts validate their estimations using *statistical refutation* as a first line of defense. These methods check that 1) the estimation remains stable after invariant transformations to the data (via data subsampling, introducing a random confounding variable) and 2) the estimation of related causal relationships drops to zero (via replacing the treatment with a random variable, randomly shuffling outcomes among study units). [30, 31] These are easily achievable methods which only use data the analyst already has, are non-parameterized, and do not require complex interpretation.

Beyond statistical refutation, *sensitivity analysis* methods [22] test the robustness of one's conclusions against possible unobserved threats to validity. One family of approaches [28] asks how much of a study population needed to have a different outcome result in order to change the causal conclusion of the study. For instance, if a study of 1000 patients only needed 2-3 people with an outcome demonstrating no observed treatment effect to change the study conclusion, the casual claims of the study would likely considered to be too weak to publish. Another family of approaches [8, 22] reasons that control and treatment groups can still differ on some unobserved confounding characteristic, so they determine how robust their conclusion is to hypothesized confounders of various strengths. Other tactics, like stratification, allow for one to inspect whether the causal estimation, in addition to meeting assumptions like independence, irrelevance, and positivity, remains stable [4, 23].

Prior art illustrates three primary approaches to confounding bias sensitivity analysis: simulation-based sensitivity analysis [25], linear partial-correlation-based sensitivity analysis [7, 32], and non-parametric sensitivity analysis [6], adapted to work for multiple robust estimations, partial linear models, and machine learning based estimations. All of these methods are parameterized by partial correlations and are compatible with our proposed system (see Problem 1). These techniques focus on characterizing a *single* unobserved confounder; related methods [34] consider scenarios with *multiple* unobserved confounders as seen in gene expression studies and other high-dimensional causal analysis scenarios.

As mentioned in Section 1, scientists conclude study robustness by demonstrating that confounders with high threat to validity are incredibly unlikely to exist [8, 15]. However, hidden explanatory factors occasionally do come to light [12], sometimes after multiple publications [9, 33] as seen in Example 1.

2.2 Problem Definition

We aim to automate sensitivity analysis via automatic data marshalling; given existing methodological literature detailed in Section 2.1, we propose the following problem:

Problem 1 (Confounder Discovery and Retrieval)**.** Given:

(1) **base table:** a single-relation database $\mathbf{D}$ with a schema $\mathbf{A} = (A_1, A_2, \cdots, A_s)$ where:
   (a) each A_i is categorical, discrete, or continuous,
   (b) its tuples (rows) contain each study unit's entry $d_i = (a_{i1}, a_{i2}, \cdots, a_{is})$ according to the schema $\mathbf{A}$, and
   (c) the user annotates a treatment variable $T \in \mathbf{A}$, an outcome variable $O \in \mathbf{A}$, and optionally some observed confounders $\mathbf{C} = \{C_1, C_2, \cdots\} \subset \mathbf{A}$;
(2) **correlation constraints:** a pair of constraints for a potential confounding threat to validity H:

(a) a user-defined minimum $R^2_{T\sim H|\mathbf{C}}$, or partial correlation with the treatment given observed confounders and
(b) a user-defined minimum $R^2_{O\sim H|\mathbf{C}\cup\{T\}}$, or partial correlation with the outcome given confounders and treatment,

(3) **additional data sources:** a set of target data sources $\mathbf{S} = \{S_1, S_2, \cdots\}$ from which the user wishes to find hypothesized confounding threats to validity,

construct an augmented database $\mathbf{D}'$ with a schema $\mathbf{A}' = (A_1, A_2, \cdots, A_s, H_1, H_2, \cdots, H_k)$ where each H_i is a potential confounder which meets the user-defined minimum partial correlation values $R^2_{T\sim H|\mathbf{C}}$ and $R^2_{O\sim H|\mathbf{C}\cup\{T\}}$, and each study unit d_i is populated with $(a_{i1}, a_{i2}, \cdots, a_{is}, h_{i1}, h_{i2}, \cdots, h_{ik})$.

Recalling the example in Section 1, Liora's base table $\mathbf{D}$ is the unshaded portion of Table 1 where each row represents a parole decision. She indicates the treatment T is "Time after Session Start" and the outcome O is "Parole Decision". Keren uses additional data sources $\mathbf{S}$ to find attributes "Scheduled Time" as H_1 and "Attorney" as H_2. Using Liora's base table $\mathbf{D}$ and the two attributes H_1 and H_2, he constructs the augmented data table $\mathbf{D}'$, illustrated in Table 1 including both the shaded and unshaded portions.

2.3 Technical Challenges & Opportunities

Our primary technical challenge is to enable fast, scalable potential confounder search based on bounds on *partial* correlations with conditioning sets. Case Study 3 illustrates a naive solution to such a problem which follows three predominant steps:

(1) *find* join paths linking external data to the base table,
(2) *augment* the base table using those join paths, and
(3) *filter* augmented attributes for sufficiently relevant confounders.

Such naive algorithms are inefficient due to the computational consequences of joins; if we proceed by joining all n tables available in a data lake, we can expect an $O(e^n)$ space and runtime complexity due to the curse of dimensionality [36] when constructing the denormalized table. Even if filtering the universal relational table is accomplished in linear time, that time complexity is linear *relative* to the size of that table. Furthermore, since empirically less than 0.5% of augmented attributes are applicable to the end goal [14], we may expect around 99.5% of the augmented variables to be not applicable. Consequently, an exponential-complexity amount of discovery and augmentation work would be completed and then pruned away in order for the system to succeed. Users working with such systems could execute joins to augment large numbers of causal variables that are mostly not consequential to their potential refutation. A time-honored approach to this performance bottleneck is to push down the filter past the join, which is critical from both a computational performance and human review perspective.

Using rigorous parameters to shrink the candidate confounder search space can also limit concerns around contradictory metadata (e.g., randomization or interference assumptions) that arise from joining large pools of discovered data gathered from disparate tables [36]. Rather than finding all possibly relevant confounders and summarizing the results, our data marshalling approach could focus on finding only attributes that are potential threats to validity.

Traditional correlation- or similarity-based methods in data discovery [5, 11, 14] fall short on the confounder discovery task due to their focus on direct pairwise correlations. For instance, knowledge graph based methods like [5, 19] excel in capturing pairs of highly related attributes; directly using these methods to capture partial correlation relationships, say by adding a new edge type per conditioning set, can be impractical given the space of conditioning sets is exponential (2^k for k possible confounders). Still, these methods in data discovery, table search, and integration [5, 10, 11, 14, 19, 20, 26] can play a critical component in the overall goal by finding join paths that link potential confounders to the base table.

While Youngmann et al. [35] tackle a virtually identical confounder search problem, their work relies on calculating conditional mutual information for *high-dimensional* conditioning sets on a *single* integrated table of analysis. Estimating conditional mutual information with high-dimensional conditioning sets is challenging to do accurately; fortunately, in our problem formulation, we do not need high-dimensional conditioning sets to determine confounder candidates. We focus our interest in calculating conditional mutual information, or otherwise evaluating potential confounders, across multi-table and multi-hop joins without executing the joins.

Greedy search approaches to finding augmentation candidates [14, 35] may not effectively address confounding bias when these factors are not *individually* strongly correlated with either the treatment or outcome, but multiple confounders in a set have strong aggregate explanatory power of the remaining unexplained residuals after adjusting for known confounders.

Galhotra et al. [14] provide a compelling framework for implementing a greedy search for potential confounder retrieval using the guidance of a monotonic utility function (say, distance of causal estimation from 0); after clustering candidate augmentations into distinct groups of similar augmentations, their system greedily searches for augmentation groups based on a user-provided utility function until the search space is exhausted or a stopping criterion is reached (i.e. utility threshold). The main challenge to realizing a possible solution for Problem 1 based on this framework is in designing appropriate data profiles or sketches [1] that will act as sieves for compelling confounding threats to validity. For example, [29] retrieves joinable tables with correlated columns by constructing a sketch of the hypothetically-joined table for calculating correlations between attributes. Similar to [14, 35], this work would need to be nontrivially adapted for partial correlation calculations with distinct conditioning sets, without losing the pre-join correlation calculation feature of this work.

Furthermore, for treatments defined by thresholds on a discrete or continuous distribution, future approaches employing scaled mean difference calculations or fixed-value filters based on robustness values [7, 32] can offer a performance improvement in the construction of any data sketches; such metrics only require subtracting two aggregate values based on partitions of the data, which is computationally cheaper than evaluating partial correlation or conditional mutual information. On the other hand, re-parameterizing our search problem into conditional mutual information can also provide robustness in non-linear scenarios. To avoid redundant work, if the database already has existing aggregation metrics, precomputed correlation statistics, or sketches with stratified data summaries, we can take advantage of them to accelerate the search.

To overcome limitations of statistical methods in automatic covariate selection [36], use of large language models (LLMs) or knowledge graphs (KGs) [18, 21, 36] can introduce semantic knowledge critical for evaluating confounder candidacy; for example, an artificial agent can employ happens-before causality and commonsense reasoning as examples of background knowledge.

3 CASE STUDY: CATHOLIC SCHOOL $\overset{?}{\rightarrow}$ MATH

In this case study, we aim to assess whether students attending Catholic high school (as opposed to public school) causes higher math grades in their senior year. To do this, we use a sample of 5671 students from the National Education Longitudinal Study of 1988 (NELS:88) [13] whose households had an annual income less than $75,000 in the first year of the survey; this is a commonly used example to discuss confounder bias adjustment [2, 24]. The dataset contains columns with demographics like race and gender, pre-high-school academic achievement metrics like 8th grade math test scores, familial circumstances like parental education, and student behavior attributes like frequency of class disruption or fighting.

We calculate the causal effect of attending Catholic high school on 12th grade math scores using the `dowhy` package [30] by specifying the treatment (whether the student attended Catholic high school), outcome (12th grade math score), and observed covariates [16, 24] including household income, 8th grade math scores, and disruptiveness. Without adjustment, we estimate that students attending Catholic high schools score 3.895 more points in math than those in public school; with adjustment, the score difference drops to an estimated 1.437 point boost for Catholic high school students.

To analyze the robustness of this textbook example finding, we follow the procedure detailed in [7], which bounds the necessary parameters that an unobserved confounder must have to either nullify the observed causal effect or render it no longer statistically significant. The analysis provides a *robustness value*, which asserts that any unobserved confounder that explains less than that proportion of the variance in the treatment and outcome would not be strong enough to explain away the observed effect.

Our robustness value is 8.05%. If an unobserved confounder explains more than that proportion of residual variation in the treatment and outcome, we may have found a threat to validity of our finding. When we hypothesize the existence of an independent unobserved confounder with partial correlation strength equal to, twice, or thrice that of all observed confounders, the hypothesized confounders are an order of magnitude too weak to lead to any noticeable changes in the causal effect. We have demonstrated the existence of a small but statistically robust causal effect.

While this is a small example, this narrative is the pattern of least resistance for scientists working with such data analyses. With large data resources like NELS:88, statisticians [2, 24] generally request the data needed for their specific task and proceed with analysis the way we did in this case study so far; the rest of the unretrieved data is unlikely to be purused or incorporated into the analysis, due to the scale of the dataset (NELS:88 has over 10,000 attributes in the full survey). Beyond the attributes already used in the causal analysis, we have a curated set of 16 additional columns from the NELS:88 dataset to consider in our analysis, including those around race, gender, parental education level, and non-math

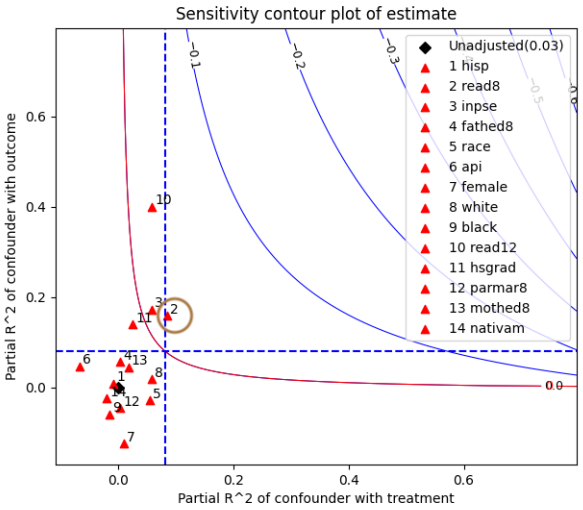

Figure 1: Contour plot for the causal estimation of *Catholic school attendance* on *math scores* in NELS:88, based on potential confounders from other parts of NELS:88 rather than hypothetical variables. `read8` is circled in brown (Section 3).

academic performance. With a brute force search through partial correlation pairs for each such attribute, we repeat the sensitivity analysis procedure [7] with marshalled potential confounders.

Looking at Figure 1, we see a black diamond at (0, 0) labeled with a Cohen's d (magnitude of causal effect adjusted to standard deviation) of 0.03 given the observed confounders; this represents the proverbial 'starting point' of our sensitivity analysis. Each red triangle represents the potential inclusion of that variable in the adjustment set, plotted according to the partial correlation values on the x- and y- axes. The red line indicates the frontier at which causal effect estimation is nullified, or turns to 0, by including additional confounders; if any triangles are found to lie to the upper right of the red line, the inclusion of that variable in the adjustment set may change the causal conclusion entirely. Potential confounding variables that lie above the red frontier line indicate the discovery of previously hidden confounders which pose a threat to validity.

We find `read8` (reading ability in 8th grade) meets the criteria put forth by our sensitivity analysis, with a partial correlation with the treatment $R^2_{T\sim H|\mathbf{C}}$ of 8.48% and partial correlation with the outcome $R^2_{O\sim H|\mathbf{C}\cup\{T\}}$ of 15.94%. When incorporated into the estimation, the 12th grade math score difference between Catholic school students and public school students drops 15% to 1.204. While it did not invalidate our causal claim, it did impact a considerable portion of the effect, which may have gone unnoticed without this search. Based on domain knowledge, we dismiss other attributes from being valid confounders, as they are factors not influential to Catholic high school enrollment: `inpse` (whether the student is in post-secondary education), `read12` (reading ability in 12th grade), and `hsgrad` (whether the student graduated from high school).

4 CONCLUSION

In this work, we discussed data marshalling for sensitivity analysis, defined the confounder discovery problem and its technical challenges and opportunities, and demonstrated it in a case study.

Acknowledgments. We gratefully appreciate the support of the DARPA ASKEM project (Award No. HR00112220042).

REFERENCES

[1] Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. 2015. Profiling relational data: a survey. *The VLDB Journal* 24, 4 (Aug. 2015), 557–581. https://doi.org/10.1007/s00778-015-0389-y

[2] Joseph G. Altonji, Todd E. Elder, and Christopher R. Taber. 2005. Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *Journal of Political Economy* 113, 1 (2005), 151–184. https://doi.org/10.1086/426036 arXiv:https://doi.org/10.1086/426036

[3] Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533, 7604 (May 2016), 452–454. https://doi.org/10.1038/533452a

[4] Hailey R Banack, Elizabeth Rose Mayeda, Ashley I Naimi, Matthew P Fox, and Brian W Whitcomb. 2024. Collider Stratification Bias I: Principles and Structure. *American Journal of Epidemiology* 193, 2 (Feb. 2024), 238–240. https://doi.org/10.1093/aje/kwad203

[5] 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

[6] Victor Chernozhukov, Carlos Cinelli, Whitney Newey, Amit Sharma, and Vasilis Syrgkanis. 2023. Long Story Short: Omitted Variable Bias in Causal Machine Learning. arXiv:2112.13398 [econ.EM]

[7] Carlos Cinelli and Chad Hazlett. 2019. Making Sense of Sensitivity: Extending Omitted Variable Bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82, 1 (12 2019), 39–67. https://doi.org/10.1111/rssb.12348 arXiv:https://academic.oup.com/jrsssb/article-pdf/82/1/39/49320681/jrsssb_82_1_39.pdf

[8] Jerome Cornfield, William Haenszel, E Cuyler Hammond, Abraham M Lilienfeld, Michael B Shimkin, and Ernst L Wynder. 1959. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute* 22, 1 (1959), 173–203.

[9] Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. 2011. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences* 108, 17 (April 2011), 6889–6892. https://doi.org/10.1073/pnas.1018033108

[10] Grace Fan, Jin Wang, Yuliang Li, and Renée J. Miller. 2023. Table Discovery in Data Lakes: State-of-the-art and Future Directions. In *Companion of the 2023 International Conference on Management of Data* (Seattle, WA, USA) *(SIGMOD '23)*. Association for Computing Machinery, New York, NY, USA, 69–75. https://doi.org/10.1145/3555041.3589409

[11] Grace Fan, Jin Wang, Yuliang Li, Dan Zhang, and Renée Miller. 2023. Semantics-aware Dataset Discovery from Data Lakes with Contextualized Column-based Representation Learning. arXiv:2210.01922 [cs.DB]

[12] Amy Finkelstein, Matthew J. Notowidigdo, Frank Schilbach, and Jonathan Zhang. 2024. *Lives vs. Livelihoods: The Impact of the Great Recession on Mortality and Welfare.* National Bureau of Economic Research, Cambridge, Mass.

[13] National Center for Education Statistics. 1988. *National Education Longitudinal Study of 1988.* Retrieved April 14, 2024 from https://nces.ed.gov/surveys/nels88/

[14] Sainyam Galhotra, Yue Gong, and Raul Castro Fernandez. 2023. Metam: Goal-Oriented Data Discovery. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. 2780–2793. https://doi.org/10.1109/ICDE55515.2023.00213

[15] Chad Hazlett. 2020. Angry or Weary? How Violence Impacts Attitudes toward Peace among Darfurian Refugees. *Journal of Conflict Resolution* 64, 5 (May 2020), 844–870. https://doi.org/10.1177/0022002719879217

[16] Andrew Heiss. 2021. *Unobserved Confounding.* Retrieved April 14, 2024 from https://evalf21.classes.andrewheiss.com/example/confounding-sensitivity/

[17] Jie Kate Hu, Eric J. Tchetgen Tchetgen, and Francesca Dominici. 2023. Using negative controls to adjust for unmeasured confounding bias in time series studies. *Nature Reviews Methods Primers* 3, 1 (Aug. 2023), 1–14. https://doi.org/10.1038/s43586-023-00249-4

[18] Moe Kayali, Anton Lykov, Ilias Fountalis, Nikolaos Vasiloglou, Dan Olteanu, and Dan Suciu. 2023. CHORUS: foundation models for unified data discovery and exploration. *arXiv preprint arXiv:2306.09610* (2023).

[19] Aamod Khatiwada, Grace Fan, Roee Shraga, Zixuan Chen, Wolfgang Gatterbauer, Renée J. Miller, and Mirek Riedewald. 2023. SANTOS: Relationship-based Semantic Table Union Search. *Proc. ACM Manag. Data* 1, 1, Article 9 (may 2023), 25 pages. https://doi.org/10.1145/3588689

[20] Aamod Khatiwada, Roee Shraga, Wolfgang Gatterbauer, and Renée J. Miller. 2022. Integrating Data Lake Tables. *Proc. VLDB Endow.* 16, 4 (dec 2022), 932–945. https://doi.org/10.14778/3574245.3574274

[21] Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. arXiv:2305.00050 (May 2023). https://doi.org/10.48550/arXiv.2305.00050 arXiv:2305.00050 [cs, stat].

[22] Weiwei Liu, S. Janet Kuramoto, and Elizabeth A. Stuart. 2013. An Introduction to Sensitivity Analysis for Unobserved Confounding in Non-Experimental Prevention Research. *Prevention science: the official journal of the Society for Prevention Research* 14, 6 (Dec. 2013), 570–580. https://doi.org/10.1007/s11121-012-0339-5

[23] Jessica L. Markham, Troy Richardson, John R. Stephens, James C. Gay, and Matt Hall. 2023. Essential Concepts for Reducing Bias in Observational Studies. *Hospital pediatrics* 13, 8 (Aug. 2023), e234–e239. https://doi.org/10.1542/hpeds.2023-007116

[24] Richard J Murnane and John B Willett. 2010. *Methods matter: Improving causal inference in educational and social science research.* Oxford University Press.

[25] Tommaso Nannicini. 2007. Simulation-Based Sensitivity Analysis for Matching Estimators. *The Stata Journal* 7, 3 (Sept. 2007), 334–350. https://doi.org/10.1177/1536867X0700700303

[26] Fatemeh Nargesian, Erkang Zhu, Ken Q Pu, and Renée J Miller. 2018. Table union search on open data. *Proceedings of the VLDB Endowment* 11, 7 (2018), 813–825.

[27] Graham E. Quinn, Chai H. Shin, Maureen G. Maguire, and Richard A. Stone. 1999. Myopia and ambient lighting at night. *Nature* 399, 6732 (May 1999), 113–114. https://doi.org/10.1038/20094

[28] Paul R Rosenbaum. 2005. Sensitivity analysis in observational studies. *Encyclopedia of statistics in behavioral science* (2005).

[29] Aécio Santos, Aline Bessa, Fernando Chirigati, Christopher Musco, and Juliana Freire. 2021. Correlation sketches for approximate join-correlation queries. In *Proceedings of the 2021 International Conference on Management of Data*. 1531–1544.

[30] Amit Sharma and Emre Kiciman. 2020. DoWhy: An End-to-End Library for Causal Inference. *arXiv preprint arXiv:2011.04216* (2020).

[31] Ilias Tsoumas, Georgios Giannarakis, Vasileios Sitokonstantinou, Alkiviadis Koukos, Dimitra Loka, Nikolaos Bartsotas, Charalampos Kontoes, and Ioannis Athanasiadis. 2023. Evaluating Digital Agriculture Recommendations with Causal Inference. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 1212 (June 2023), 14514–14522. https://doi.org/10.1609/aaai.v37i12.26697

[32] Tyler J. VanderWeele and Peng Ding. 2017. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Annals of Internal Medicine* 167, 4 (Aug. 2017), 268–274. https://doi.org/10.7326/M16-2607

[33] Keren Weinshall-Margel and John Shapard. 2011. Overlooked factors in the analysis of parole decisions. *Proceedings of the National Academy of Sciences* 108, 42 (Oct. 2011), E833–E833. https://doi.org/10.1073/pnas.1110910108

[34] Feng Xie, Zhengming Chen, Shanshan Luo, Wang Miao, Ruichu Cai, and Zhi Geng. 2024. Automating the Selection of Proxy Variables of Unmeasured Confounders. arXiv:2405.16130 [cs.LG]

[35] Brit Youngmann, Michael Cafarella, Yuval Moskovitch, and Babak Salimi. 2023. On Explaining Confounding Bias. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE, Anaheim, CA, USA, 1846–1859. https://doi.org/10.1109/ICDE55515.2023.00144

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

[37] Karla Zadnik, Lisa A. Jones, Brett C. Irvin, Robert N. Kleinstein, Ruth E. Manny, Julie A. Shin, and Donald O. Mutti. 2000. Myopia and ambient night-time lighting. *Nature* 404, 6774 (March 2000), 143–144. https://doi.org/10.1038/35004661