Demonstrating Nexus for Correlation Discovery over Collections of Spatio-Temporal Tabular Data

Yue Gong
yuegong@uchicago.edu
The University of Chicago
Chicago, USA

Raul Castro Fernandez
raulcf@uchicago.edu
The University of Chicago
Chicago, USA

ABSTRACT

Causal analysis is crucial for understanding cause-and-effect relationships in observed data to inform better decisions. However, conducting precise causal analysis on observational data is usually impractical, so domain experts often begin their exploration by identifying correlations. In this paper, we demonstrate Nexus [10], a system that aligns tabular datasets across space and time, handles missing data, and identifies correlations deemed "interesting", facilitating the exploration of causal relationships.

CCS CONCEPTS

• **Information systems** → **Information integration**; *Specialized information retrieval*.

KEYWORDS

Data Discovery, Correlation Analysis, Spatio-Temporal Data

ACM Reference Format:
Yue Gong and Raul Castro Fernandez. 2024. Demonstrating Nexus for Correlation Discovery over Collections of Spatio-Temporal Tabular Data. In *Companion of the 2024 International Conference on Management of Data (SIGMOD-Companion '24), June 9–15, 2024, Santiago, AA, Chile.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3626246.3654747

1 INTRODUCTION

Correlation analysis is a vital initial step for investigating causation, essential for understanding complex phenomena and making informed choices. While it is hard to establish causality from vast observational data without assumptions and expert knowledge [12], identifying correlations remains a key strategy to "cast a wide net" and detect potential causal links. This paper demonstrates Nexus [10], a system that identifies correlations over collections of spatio-temporal tabular data, aiming to identify interesting hypotheses and provide a good starting point for further causal analysis. Nexus focuses on two personas:

Persona 1: Exploring an Established Hypothesis. *A researcher at a medical school, Bob, has a dataset with asthma attack incidences in hospitals across various zip codes in Chicago. He has noticed a link between air quality and asthma attacks in his prior analysis. His goal is to establish causality between these two variables, instead of a mere correlation. Thus, Bob wants to consider other variables that could influence asthma attacks.* **Persona 1 is someone who has an initial dataset and an established hypothesis, and seeks to enrich such a dataset with additional variables relevant to the analysis**. *In the example, Bob uses our new system Nexus on Chicago Open Data [11] and discovers additional variables correlated with asthma attacks such as "household income" and "violence crime rate", inspiring him to include socio-economic factors in the analysis.*

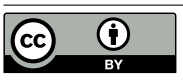
Persona 2: Data-Driven Hypothesis Generation. *Amy, a social scientist in Chicago, is seeking to discover intriguing phenomena within the city for her research. To avoid limiting her analysis to existing knowledge, she employs a data-driven strategy. Recognizing that Chicago Open Data has a wealth of datasets on diverse societal aspects such as education, business, and crime, Amy wants to identify interesting correlations automatically to generate new hypotheses.* **Persona 2 has a large repository of tabular data and wants to automatically identify interesting correlations to formulate new hypotheses for further causal analysis.** *In the example, Amy uses Nexus to analyze Chicago open data, and 40k variable correlations are identified. Nexus then assists her in navigating these correlations and finding interesting ones such as between the variables 'number of bike-sharing stations' and 'household income' across different neighborhoods. This leads her to question if bike-sharing locations are biased toward richer neighborhoods in Chicago.*

Nexus addresses the following challenges to identify correlations interesting to Persona 1 and 2.

Challenge 1. Spatio-Temporal Alignment. Datasets need to be aligned first before computing correlations. However, many datasets lack a common key for joining. Nexus circumvents this by leveraging the abundant spatial and temporal attributes in datasets, aligning them across space and time. This requires efficient indexing techniques to handle millions of records as well as applying transformation and aggregations that resolve granularity inconsistencies, e.g., "household income" is aggregated to zip code level while "violence crime rate" is at the neighborhood level.

Challenge 2. Identifying Correlations in the Presence of Missing Data. Missing data, present either in the original dataset or resulting from merging two datasets, can hinder correlation discovery. Simply ignoring missing data can omit potential correlations, while using outer joins to include all observed data is expensive.

Challenge 3. Identifying "Interesting" Correlations. Addressing Challenges 1 and 2 leads to a vast array of correlations, even in modest datasets. For example, more than 40k correlations are identified in Chicago Open Data. This overwhelming number of correlations hampers Persona 1 and 2 from identifying correlations useful for their analysis.



SIGMOD-Companion '24, June 9–15, 2024, Santiago, AA, Chile

ACM ISBN 979-8-4007-0422-2/24/06
https://doi.org/10.1145/3626246.3654747

In this paper, we demonstrate how Nexus addresses Challenges 1-3 and aids Persona 1 and 2 in analyzing real-world datasets and discovering interesting correlations that generate intriguing hypotheses, which have been validated in the existing literature.

Related Work. Data Polygamy [7] explores the spatio-temporal relationships in open data. However, it uses a specialized metric different from correlation. Another demonstration [15] identifies potential confounding variables for a target correlation, assuming users have a specific correlation in mind. In contrast, our work assists users in identifying correlations worth further exploration before they have a specific target in mind. Data discovery systems [3, 5] focus on finding or enriching a dataset based on some criteria. Auctus [5] is a discovery engine that can filter datasets based on a spatio-temporal range but lacks the functionality of identifying correlated attributes. There are systems [3, 13] that enrich a dataset with correlated variables. However, those are not tailored to spatio-temporal datasets (Challenge 1), do not handle missing values (Challenge 2), and do not organize the resulting vast array of correlations to facilitate the identification of interesting ones (Challenge 3). Nexus is the first end-to-end system that addresses Challenges 1-3 collectively to satisfy the needs of Persona 1 and 2.

2 NEXUS OVERVIEW

This section gives an overview of Nexus.

Problem Definition. Given an input dataset $)_{in}$, Nexus first identifies all spatio-temporal datasets $\mathcal{L}$ that are joinable with $)_{in}$ according to a specified overlap threshold and a spatio-temporal granularity; then calculate all significant correlations between attributes from $)_{in}$ and $\mathcal{L}$ according to a specified correlation coefficient threshold. This problem definition focuses on a single input dataset, reflecting the scenario of Persona 1 (Explore an existing hypothesis). More broadly, the problem can be extended to identify all correlations within a collection of tables, which is the scenario of the second persona (Data-Driven Hypothesis Generation).

System Architecture. Fig. 1 shows the architecture of Nexus. Nexus is designed based on the reference architecture originally proposed in Ver [9]. Inputs for the system include i) data collection, ii) spatio-temporal granularities, iii) and aggregate functions. The system returns a collection of correlations as the intermediate output. Since the output is often too large for a human to examine one by one, Nexus further distills the structure of those correlations.

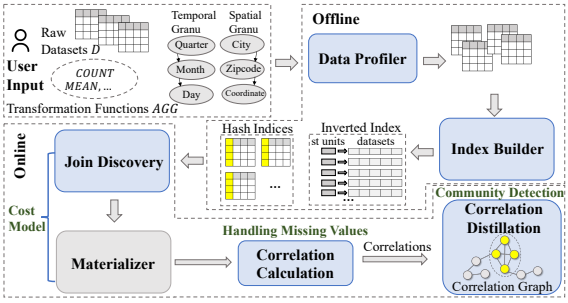


Figure 1: System Overview of Nexus.

In an offline stage, Data Profiler and Index Builder accept the input datasets, label spatial and temporal attributes, and their granularity, and convert them to the user-provided granularity. It also collects statistics about each attribute that is subsequently used to address missing values. The created indices facilitate efficient dataset alignment, with all resulting data products stored in an RDBMS (PostgreSQL in our implementation) (Challenge 1).

In the online stage, Nexus uses a cost model to identify and materialize spatio-temporal datasets efficiently (Challenge 1). Then, Correlation Calculation decomposes the correlation calculation process and takes into account the impact of missing values (Challenge 2). The output is a list of correlations processed by Correlation Distillation, which groups them according to the structure manifested in a correlation graph (Challenge 3).

2.1 Spatio-Temporal Alignment

We next explain how Nexus identifies and materializes spatio-temporal joinable datasets efficiently.

2.1.1 Transform and Index Datasets (Offline). Nexus processes raw spatio-temporal datasets into the desired granularity by evaluating all possible combinations of spatio-temporal granularities for each dataset and applying transformation functions to the spatio-temporal attributes. Nexus establishes two types of indices for efficiency: a hash index on the spatio-temporal attributes to accelerate materialization through hash join, and an inverted index where each spatio-temporal value is a key linked to a list of datasets containing that value.

2.1.2 Spatio-Temporal Datasets Alignment. Nexus implements two execution strategies for aligning spatio-temporal datasets:

Index Search. The first mode of finding spatio-temporal joinable datasets is via the inverted index. Given a transformed dataset, Nexus first queries this index to identify all datasets with intersecting spatio-temporal values. It then counts how frequently each dataset appears, reflecting the degree of overlap with the input dataset. Datasets with an overlap exceeding a predefined threshold are added to a candidate list. Nexus then merges the input dataset with these joinable datasets to produce materialized views.

Exhaustive Join. This method consolidates finding and materializing joinable datasets into a single stage. It skips the phase of querying an index. Instead, it directly joins the input dataset with every other dataset. If the result yields the number of rows larger than the overlap threshold, it means two datasets are joinable and the materialized view is added to the candidate list.

2.1.3 A Cost-based Model for Choosing between Index Search and Exhaustive Join. Index-Search is beneficial when the number of potentially joinable datasets is small. But when the number of joinable datasets is sufficiently large, the additional overhead introduced by the inverted index makes Exhaustive-Join a better choice. To manage this trade-off, Nexus introduces a cost-based model. Specifically, Nexus conducts a formal cost analysis for each execution strategy, estimates costs using sampling methods, and then chooses the most efficient strategy for a given dataset.

2.2 Correlation with Missing Values

A full outer join is needed to include all observed samples from two datasets, which introduces missing values if a row in one dataset does not have a match in the other. These missing values need to be handled before subsequent correlation calculation. Managing missing values is essential to prevent biases in data processing [2], and the approach depends on the nature of the data and the source of the missing values, which Nexus does not have full knowledge of. Thus, the goal of Nexus is not to devise a missing value method that avoids selection bias in all scenarios. Instead, it aims to provide transparency regarding the methods employed and let users decide whether the results are useful downstream. Nexus considers three strategies to handle missing values in the full outer join result.

(1) Drop all missing values: Drop all rows containing missing values, which is equivalent to the inner join result.
(2) Fill with zero: Fill missing values with zero.
(3) Fill with mean: Fill missing values with the mean of an attribute.

A naive method for applying these strategies would involve first computing the outer join result of two spatio-temporal joinable datasets, then applying each strategy to this result to calculate correlations. However, full outer joins are computationally expensive. Instead, Nexus efficiently calculates correlations without resorting to outer join operations. The key insight is that the full outer join is not required for the last two strategies (fill-zero and fill-avg). Instead, these correlations can be accurately calculated using the results of inner joins, combined with attribute-level statistics such as the sum and square sum, which are collected offline. This method allows Nexus to acquire correlation coefficients for all three strategies efficiently by materializing only the inner join, bypassing the high computational demands of full outer joins.

2.3 Correlation Distillation

We now discuss how Nexus organizes result correlations. On large repositories of data, it is common to find many correlations e.g., in our demonstration, more than 40K correlations are found on a data collection with 338 tables.

Signal-based Exploration. Nexus offers various *signals* for users to filter correlations, such as the ratio of missing values in an attribute, and the number of samples used to compute the correlation (these signals are collected offline by the Data Profiler component). A user with specific interests in certain aspects of a correlation can employ these signals to navigate correlations.

Exposing the structure of correlations. To further assist users in processing the correlations, Nexus exposes the structure of correlations based on the following intuition: The correlation graph, which depicts the correlations between attributes, tends to reflect aspects of the underlying causal graph. For instance, in a ground truth causal graph that is disconnected, variables in separate components do not influence each other, resulting in few correlations across these components, mostly due to noise (green pair in Fig. 2). This means that sparse areas in the causal graph typically correspond to sparse sections in the correlation graph (green region). In contrast, dense area in the causal graph (e.g. the presence of a common cause) tends to be represented as densely connected nodes in the correlation graph (yellow region). Consequently, presenting each correlation separately to users would be both impractical and overwhelming when searching for meaningful correlations. A more effective way to organize these correlations is to extract their clustering structure from the graph, thereby helping the user to identify causal links and confounders.

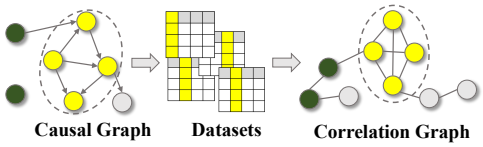


Figure 2: Causal graph encoded within a correlation graph.

Identify variable clusters on the correlation graph. Correlation Distillation extracts variable clusters from the correlations. First, it selects a subset of correlations using signals. Then it builds the correlation graph by iterating over correlations: each variable becomes a node in the graph, and edges indicate correlation between the respective variables. Since dense regions of the causal graph show up as densely connected components in the correlation graph, it employs a community detection algorithm [4] to identify communities (*variable clusters*) in the correlation graph.

3 DEMONSTRATION

To demonstrate Nexus, we illustrate how Nexus assists Persona 1 and 2 with the analysis of real-world datasets, enabling them to discover interesting correlations that lead to compelling hypotheses, validated by established literature.

3.1 Persona 1: Exploring a Hypothesis

We first demonstrate how Nexus helps Persona 1, Bob, to enrich his asthma analysis with additional variables from Chicago open data as shown in Fig. 3a.

Index Creation. In the offline stage, Nexus creates discovery indices over Chicago open data. This includes converting and aggregating spatio-temporal datasets to various granularities and creating inverted and hash indices for efficient join discovery and materialization. MEAN and Count are used as aggregate functions.

Identify Correlations. To find variables correlated with asthma visits, Bob uses the find-correlations API, inputting the asthma dataset and specifying the desired spatio-temporal granularity. Bob chooses the spatial granularity as zipcode because the asthma dataset's spatial attribute is at the zipcode level. Nexus then identifies 109 correlations with asthma visits. In the correlation list shown in Fig. 3a, C1; denotes the table name, 9>8=_:4~ is the spatio-temporal attribute used for joining two tables. Their original granularities may differ. For instance, while the asthma dataset is at the zipcode level, the crime dataset is at the coordinate level originally. Nexus automatically resolves such granularity inconsistencies. OCCA represents the attributes involved in the correlation calculation, and 2>AA4;OC8>= indicates the Pearson's correlation coefficient.

Filter and Rank Correlations. Nexus provides various signals for Bob to filter and rank the identified correlations. He can filter

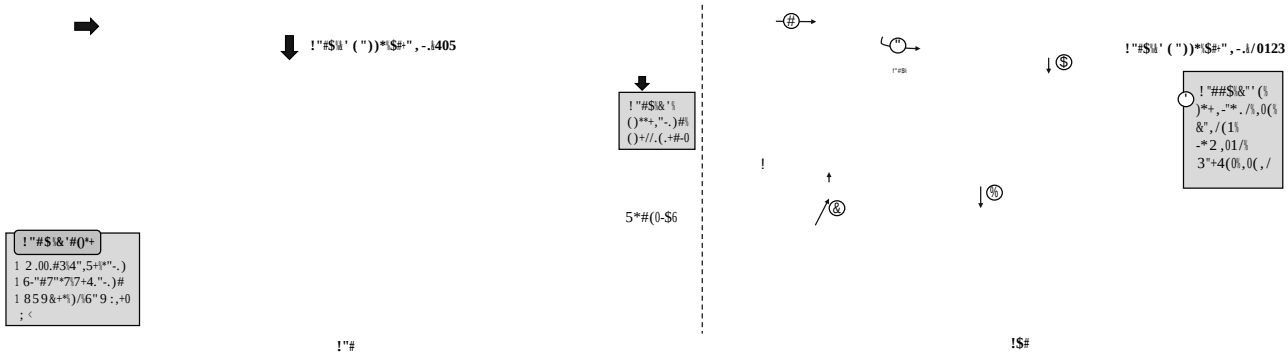


Figure 3: (a) Bob using Nexus to identify variables correlated with the asthma dataset (b) Amy using Nexus to explore and identify interesting hypotheses within Chicago Open Data

correlations based on table names, and attribute names. Nexus also provides detailed data profiles for each correlation, including the correlation coefficient and data quality indicators like missing value ratio, standard deviation, and the number of samples used in the correlation calculation. Bob can click the "Show Profile" button to see detailed profiles of a correlation. He can rank or filter these correlations based on any combination of these signals.

Insight from Correlations. The correlations identified by Nexus reflect that the number of asthma visits is correlated with many poverty indicators, such as *OAA4BCB_E8>;4=C_>5 5 4=B4B*, which is the number of times the individual has been arrested for a violent offense, and the number of crimes happened within a zipcode. This prompts Bob to model the factor *?>E4AC~* in his model for asthma attacks. Notably, the relationship between asthma attacks and the poverty level is verified in a report from CDC [6].

3.2 Persona 2: Hypothesis Generation

We next showcase how Nexus assists Persona 2, Amy, in discovering interesting correlations within the city to form new hypotheses for her research as depicted in Fig. 3b.

The deluge of correlations. Amy identifies all correlations within Chicago open data using the FIND-ALL-CORRELATIONS API, specifying the data source, along with temporal and spatial granularities. In this case, Amy chooses temporal granularity as "Month" and spatial granularity as "Census Tract". Nexus identifies more than 40k correlations, an overwhelming number that makes it impractical for Amy to manually find interesting correlations (Step 1).

Correlation Distillation. Nexus helps Amy reduce the burden of examining correlations by extracting a few variable clusters from the vast correlations. Nexus identifies 22 clusters after the EXACT-CLUSTERS API is executed (Step 2). As shown in Fig. 3b, there is a cluster with tables related to divvy bike stations[1], taxi trips, and Chicago covid-19 community vulnerability index (CCVI) (Step 3). Amy can use the SHOW VARIABLES button to examine all variables and the SHOW CORRELATIONS button to list correlations within the cluster (Step 4). After a list of correlations is displayed, Amy can use the GET-ALIGNED-RESULT API to examine the merged dataset where the correlation is calculated, specifying the index of the correlation and the desired spatio-temporal granularity (Step 5).

Insight from variable clusters and Interesting Hypotheses. As shown in Fig. 3b, there are six correlations between divvy bike docks and CCVI score. CCVI score [1] measures a community's susceptibility to the negative impacts from COVID-19 based on various social and economic factors. A lower CCVI score means less vulnerability, indicating an area has a more advanced socio-economic status. These significant negative correlations between ccvi score and divvy bike docks inspire Bob to form a hypothesis that *Divvy bike locations are biased towards richer areas* (Step 6). Notably, this hypothesis has been verified in existing studies [8].

Demonstration Engagement. During the demonstration, we will guide participants through the use cases of the two personas. Participants can use Nexus's API to write correlation queries, browse the identified correlations, and examine variable clusters. We will prepare two datasets for participants to explore - Chicago Open Data [11], and United Nations Data [14].

[1] divvy is a bike-sharing system in Chicago

REFERENCES

[1] 2021. Chicago CCVI. https://chicago.gov/content/dam/city/sites/covid/reports/012521/Community_Vulnerability_Index_012521.pdf.
[2] Elias Bareinboim and Judea Pearl. 2012. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*. PMLR, 100–108.
[3] Jannis Becktepe and et al. 2023. Demonstrating MATE and COCOA for Data Discovery. In *SIGMOD*. 119–122.
[4] V. D Blondel and et al. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.
[5] Sonia Castelo and et al. 2021. Auctus: A dataset search engine for data discovery and augmentation. *Proceedings of the VLDB Endowment* 14, 12 (2021), 2791–2794.
[6] Centers for Disease Control and Prevention. 2023. Asthma - Health, United States. https://www.cdc.gov/nchs/hus/topics/asthma.htm.
[7] Fernando Chirigati and et al. 2016. Data polygamy: The many-many relationships among urban spatio-temporal data sets. In *SIGMOD*. 1011–1025.
[8] Elizabeth Flanagan and et al. 2016. Riding tandem: Does cycling infrastructure investment mirror gentrification and privilege in Portland, OR and Chicago, IL? *Research in Transportation Economics* 60 (2016), 14–24.
[9] Yue Gong and et al. 2023. Ver: View discovery in the wild. In *ICDE*. IEEE, 503–516.
[10] Yue Gong and et al. 2024. Nexus: Correlation Discovery over Collections of Spatio-Temporal Tabular Data. *Proc. ACM Manag. Data* 2, 3, Article 154 (2024).
[11] City of Chicago. 2023. Chicago Open Data. https://data.cityofchicago.org/
[12] Judea Pearl. 2009. Causal inference in statistics: An overview. (2009).
[13] Aécio Santos and et al. 2021. Correlation sketches for approximate join-correlation queries. In *SIGMOD*. 1531–1544.
[14] United Nations. 2023. UNdata. https://data.un.org/
[15] Brit Youngmann and et al. 2023. NEXUS: On Explaining Confounding Bias. In *SIGMOD*. 171–174.