Mining Explanations from Knowledge Graphs

Brit Youngmann
CSAIL MIT
brity@mit.edu

Michael Cafarella
CSAIL MIT
michjc@csail.mit.edu

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

Yuval Moskovitch
University of Michigan
yuvalm@umich.edu

ABSTRACT

When analyzing large datasets, analysts are often interested in the explanations for surprising or unexpected results that were produced by their queries. Previous approaches to results explanations have focused on generating explanations from the data accessed by the query. However, in many real-life scenarios, the explanations are not solely contained in the input table(s). In this work we are interested in generating explanations in terms of a set of *confounding variables that explain away unexpected correlations* observed in query results. We propose to mine query result explanations from a Knowledge Graph (KG). We present an efficient algorithm that finds a bounded-size subset of attributes (mined from a KG and the dataset) that explain away unexpected correlations observed in user queries. This algorithm is embodied in a system called MESA. We demonstrate experimentally over multiple real-life datasets and through a user study that our approach generates insightful explanations, outperforming existing methods that search for explanations only in the input data. We further demonstrate the robustness of our approach to missing data and the ability of MESA to handle large input datasets and a KG containing millions of tuples.

PVLDB Reference Format:
Brit Youngmann, Michael Cafarella, Babak Salimi, and Yuval Moskovitch. Mining Explanations from Knowledge Graphs. PVLDB, 16(1): XXX-XXX, 2023.
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at URL_TO_YOUR_ARTIFACTS.

1 INTRODUCTION

When analyzing large datasets, users are often interested in explanations for surprising observations. Namely, they would like to find what might lie behind the surprising or unexpected results that were produced by their queries. Query results are often hard to interpret, especially for aggregate query answers [65, 74]. Further, good explanations might be found outside the narrow query results that the user observes and the database being used [39]. Thus, there is a need to develop automated solutions that can explain query results to data analysts in a meaningful way, which goes beyond just the data accessed by the user query.

In this work, we focus on aggregate SQL queries (select-from-where-groupby) that are aggregating an *outcome variable* based on some group of interest indicated by a grouping or *exposure variable*. A major challenge that hinders the interpretation of such queries is *confounding bias* [60] that can lead to spurious association between the exposure and outcome and hence perplexing conclusions. Confounding bias is a systematic type of an error that arises due to the uneven or unbalanced distribution of a third variable(s), known as the *confounding variable(s)* in the considered groups. Previous work detected uncontrolled confounding variables from the data [65]. However, in many cases, such variables cannot be found in the input data, but could be mined from external sources. We are interested in generating explanations in terms of a set of confounding variables (from the input data and from external sources) that **explain away unexpected correlations observed in query results.** To illustrate, consider the following example.

Example 1.1. *Ann is a data analyst in the WHO organization who aims to understand the coronavirus pandemic for improved policymaking. She examines Covid-Data [3], a dataset containing information describing Covid-19-related facts in multiple cities worldwide (as of May 2020). It consists of the number of deaths-/recovered-/active-/new-cases and the number of deaths-/recovered-/active-/new- per-100-cases in each city. Ann evaluates the following query over this dataset:*

```
Q = SELECT Country, avg(Deaths_per_100_cases)
FROM Covid-Data
GROUP BY Country
```

A visualization of the query results is given in Figure 1. Here, the exposure attribute is Country *and the outcome is* Deaths_per_100_cases*. Ann observes a big difference in the death rate among countries, i.e., she sees a spurious correlation between* Country *and* Deaths_per_100_cases*. She is interested in finding a set of confounding variables that explain away the relationship between* Country *and* Deaths_per_100_cases*. She uses statistical methods to see whether this correlation can be explained using attributes from Covid-data. She learns that the attribute* Confirmed_cases *is correlated with* Deaths_per_100_cases*. However, this attribute alone is not enough to explain away the correlation. For example, she sees that while Germany had the fifth-most confirmed cases in the world, it had only a fraction of the death toll seen in other countries.*[1] *She understands that other factors (that are not in this data) affect the relationship between death rate and country. For example, it was shown that as a country's success (defined*

Proceedings of the VLDB Endowment, Vol. 16, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

[1] As reported in https://tinyurl.com/2p9y8xz6.

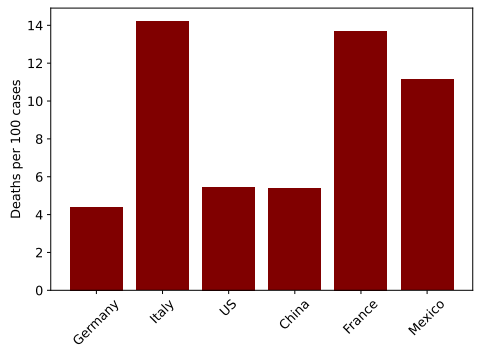


Figure 1: Visualization of the results of the query Q.

by multiple variables including GDP[2] and HDI[3]) grows, the death rate decreases [35, 72]. However, such economic features of countries are not available in the dataset and thus are not considered in the analysis. But such features could be mined from knowledge graphs.

We propose to mine query result explanations from a Knowledge Graph (KG). KGs are an emerging type of knowledge representation that has gained much attention in recent years [14, 22, 23, 77]. KGs typically contain a very large amount of data. The sheer breadth of coverage that makes knowledge graphs potentially valuable is also what creates the need to automate the process of mining relevant confounding variables. There are multiple general-purpose (e.g., Wikidata [12], DBpedia [8], Yago [62]) or domain-specific (e.g., for medical proteomics [66], or protein discovery [54]) KGs that act as central storage for data extracted from multiple sources (e.g., Wikipedia). We argue that such valuable data could be utilized for explaining away unexpected correlations observed in user queries in a wide range of different datasets.

To this end, we present an efficient algorithm that finds a bounded-size subset of attributes (mined from a KG and the dataset) that explain away spurious correlations observed in user queries. Those attributes can assist analysts in understating the results. This algorithm is embodied in a system called MESA (for **M**ining **E**xplanation**S** from knowledge gr**A**phs), that automatically mines candidate attributes form a KG. The KG may be provided by the user (for a specific domain) or could be any publicly available KG, as long as it can be integrated with the input table.

Example 1.2. *Ann uses MESA to search for an explanation for her query. MESA mines all available features about countries that appear in Covid-Data from DBpedia. She learns that* Confirmed_cases*, HDI, and GDP are uncontrolled confounding attributes. She sees that the death rate is similar in countries with a similar number of confirmed cases, HDI, and GDP. She is pleased because she found a plausible real-world explanation for her query results [35, 72].*

Previous work provides explanations for trends and anomalies observed in query results in terms of predicates on attributes that are shared by one (group of) tuple in the results but not by another (group of) tuple [29, 39, 63, 64, 73]. However, those methods do not account for correlations among attributes, and are thus inapplicable for explaining the correlation between the outcome and the exposure. [65] presented a system that provides explanations based on causal analysis, measured by correlation among attributes. However, this system only considers the input dataset, and its running times are exponential in the number of candidate confounding attributes. We share with CajaDE [39] the motivation of considering explanations that are not solely drawn from the input table. CajaDE is a system that generates insightful explanations based on contextual information mined from tables related to the table accessed by the query. Their explanations are a set of patterns that are unevenly distributed among groups in the query results, and are *independent of the outcome attribute*. Thus, CajaDE may generate explanations that are irrelevant for understanding the correlation between the exposure and outcome. (See discussion in Section 6).

There are several challenges that need to be overcome to facilitate an efficient system producing high-quality explanations. The first challenge is how to measure the explanatory power of an attribute subset, ensuring the explanations are interpretable and useful (challenge **C1**). The attributes are mined from a KG only *after the query arrives* (as the KG may be a part of the input). Since data in KGs is sparse, the extracted attributes may contain many missing values. Previous work showed that common approaches for handling missing data could cause substantial bias if many values are missing [67]. In contrast to prediction, the quality of explanations is more sensitive to missing data [55]. Thus, a second challenge is to ensure the explanations are robust to missing data (challenge **C2**). Last, since there are potentially hundreds of attributes that could be extracted from a KG, there is a need to develop an efficient algorithm to search for the optimal attribute subset in this extensive search space. Further, the search for the optimal attribute set involves estimating high-dimensional partial correlations (to measure the correlation between the outcome and the exposure, while controlling for the effect of confounding variables). However, estimating partial correlation (in this paper we use conditional mutual information) for high-dimensional conditioning set is notoriously difficult [40] (challenge **C3**).

Given a query Q, we denote by T and O the exposure and outcome attributes in Q, resp. To address **C1**, we present the Correlation-Explanation problem that seeks a subset of attributes (extracted from a KG or the input table), referred to as *the explanation for* Q, that explain away the correlation between T and O. Namely, when conditioning on these attributes, T and O become independent. We formalize Correlation-Explanation as the problem of finding an attribute set that minimizes the partial correlation between T and O. Previous work indicated the importance of minimal-size explanations [64]. Thus, besides the explanatory power (measured by partial correlation), we also consider the cardinality of attribute sets. Further, MESA enables analysts to learn the individual responsibility of selected attributes and to automatically identify unexplained data subgroups that correspond to refinements of the query in which a different explanation is required. (Section 2).

Given an input database $\mathcal{D}$ and a KG (which may be specified by the user), we extract attributes from the KG that can be joined to $\mathcal{D}$, utilizing off-the-shelf Named Entity Disambiguation tools [56]. To address **C2**, we present a principled way for handling missing values. Restricting the analysis only to tuples with no missing values

[2]Gross domestic product (GDP) is the monetary value of all goods and services made within a country during a specific period.
[3]The Human Development Index (HDI) is a statistic composite index of life expectancy, education, and per capita income indicators, which is used to rank countries.

may induce *selection bias* [67] . Inverse probability weighting (IPW) is a common method to correct this bias [67]. We provide sufficient conditions to detect selection bias and explain how IPW can be employed in our setting (Section 3).

To address **C3**, we propose the MCIMR algorithm. This algorithm does not require iterating over all possible attribute subsets, and avoids estimating high-dimensional conditional mutual information (a common measure for partial correlation). It selects attributes based on a minimal-conditional-mutual-information and a minimal-redundancy criteria, yielding a polynomial-time algorithm that finds the optimal k-size explanation where k is given. We then define a stopping criterion, allowing the algorithm to stop when no further improvement is found. We propose multiple heuristic pruning techniques to speed up computation time. (Section 4).

We conducted an experimental study based on four commonly used datasets that evaluate the quality and efficiency of the MCIMR algorithm. Our approach is effective whenever the explanation can be found in a KG. We show that this was the case in 72.5% of random queries evaluated on these datasets. To evaluate the quality of our explanations, we focus on 14 representative queries suffering from confounding bias. Our queries are inspired by real-life analysis reports, ranging from Stack Overflow annual reports [5] to academic research about Covid-19 [72]. We ran a user study consisting of 150 subjects to evaluate the quality of our explanations compared with six other approaches. We show that the explanations generated by MCIMR are almost as good as those of a computationally infeasible naive method that iterates over all attribute subsets, and are much better than those of feasible competitors. We also show that our substantive explanations are supported by previous findings in each domain. We then study the effect of different parameters on the performance. Our experimental results demonstrate the robustness of our solution to missing data and indicate the effectiveness of our algorithm in finding explanations in less than 10s for queries evaluated on datasets containing more than 5M tuples (Section 5).

Related work is presented in Section 6 and we conclude in Section 7. For space constraints, all proofs are deferred to [11].

2 MODEL AND PROBLEM FORMULATION

2.1 Data Model and Assumptions

We operate on a dataset consisting of a single relational table database $\mathcal{D}$. The table's attributes are denoted by $\mathcal{T}$. For an attribute T_i we denote its domain by $Dom(T_i)$. We use bold letters for sets of attributes $\boldsymbol{A}\subseteq\mathcal{A}$. We make the following assumption, standard in statistics [65]. The database $\mathcal{D}$ is a uniform sample from a large population (e.g., all developers, all countries in the worlds, etc.), obtained according to some unknown distribution $Pr(A)$. We expect the reader is familiar with basic information theory measures, such as entropy and conditional mutual information, associated to the probability distribution Pr.

In this work we focus on simple single-block SQL queries with a single aggregate function [39, 65], as shown in Listing 1. A query Q represents an empirical estimate $E[O|T = t_i, C]$, for each $t_i\in Dom(T)$. We restrict the queries to group-by-average queries. We do not consider more complex queries (e.g., nested sub-queries), and we also do not consider aggregate operators other than average. Since the data is assumed to be a uniform sample from a large population, the queries we examine compare among subpopulations. Thus, it does not make sense to consider other aggregations (e.g., estimating COUNT requires further assumptions).

Listing 1: A group-by-average query.

```
Q = SELECT T, avg(O)
FROM D
WHERE C
GROUP BY T
```

We denote by O the outcome attribute of the query Q and by T the exposure attribute used for grouping. To simplify the exposition, we assume a single grouping attribute. However, our results can be naturally generalized for cases with multiple grouping attributes. To handle a numerical exposure variable, one may bin this attribute. We call the condition C the context for the query Q. We interpret the query as follows: given the context, we aim to explain the difference among $avg(O)$ for each $T{=}t_i$, where $t_i\in Dom(T)$.

We use the following example based on the Stack Overflow (SO) dataset throughout this paper. In our experiments, we demonstrate the operation of MESA over four datasets, including Covid-Data.

Example 2.1. *SO dataset contains information about people who code around the world, such as their age, gender, income, and country. Consider the following query, denoted as Q_{so}:*

```
SELECT Country, avg(Salary)
FROM SO
WHERE Continent = Europe
GROUP BY Country
```

Here, O is Salary, T is Country and the context C is Continent = Europe. We aim to explain the difference in the average salary of developers from each country in Europe. While some attributes from the dataset may partially explain this difference (e.g., Gender, DevType), other important attributes that can cast light on this difference cannot be found in the SO dataset, but could be found in a KG.

External Knowledge. The system has access to a Knowledge Graph (KG) denoted as $\mathcal{G}$. The KG may be a provided by the user for a specific domain (e.g., [54, 66]), or any publicly available KG (e.g., [8, 12, 62]), as long as it can be integrated with the input table. The extracted attributes may be irrelevant for a given dataset, e.g., data collected for a specific period. We thus let the user decide which KG MESA should use. From this KG, we extract a set of attributes $\mathcal{E}$ representing additional properties of entities from $\mathcal{D}$. Continuing with our example, $\mathcal{E}$ could be a set of properties of the countries extracted from some KG, such as the density, population size, and HDI. See Section 3.1 for more details about the extraction process. We can potentially join $\mathcal{E}$ and $\mathcal{T}$ together, by linking values from $\mathcal{T}$ with their corresponding entities in $\mathcal{G}$ that were used for the attributes extraction. However, $\mathcal{E}$ may contain many attributes, most of them are most likely irrelevant for explaining the query results. Our goal is therefore to identify a subset of attributes $\boldsymbol{E}\subseteq\mathcal{E}\cup\mathcal{T}$ that *explains the correlation between T and O.*

2.2 Problem Formulation

Given a query Q, we assume the user observed an unexpected correlation between the exposure T and the outcome O attributes that she would like to investigate. In this work, we assume that there

is confounding bias that causes a spurious association between T and O. Confounding bias is a systematic error due to the uneven or unbalanced distribution of a third variable(s), known as the confounding variable(s) in the competing groups. Uncontrolled confounding variables may preclude finding a true effect; it may lead to an inaccurate estimate of the true association between T and O. Our goal is to discover the confounding variables, to control their impact on the observed correlation between T and O.

Let $\mathcal{A}$ denote $\mathcal{E} \cup \mathcal{T} \setminus \{O, T\}$, referred to as the candidate attributes. We denote by $\boldsymbol{E_k} \subseteq \mathcal{A}$ a k-size set of attributes, and by $E_i \in \mathcal{A}$ a single attribute. We assume that $\mathcal{A}$ contains confounding attributes that affect both T and O. We aim to find a set of attributes from $\mathcal{A}$ that control the correlation between O and T. Namely, when conditioning on this attribute set, the correlation between O and T is diminished. We call such a set *the explanation of the query results*.

Example 2.2. *It is very likely that countries' economic features (such as GDP, Gini Coefficient[4], HDI, HDI Rank) affect developers' salaries. To unearth the unexpected association between Country and Salary, one must measure the correlation while controlling such attributes. This will allow users to understand which factors affect the differences in developers' salaries in different countries. Intuitively, we expect the average developers' salaries to be similar in countries with similar economic characteristics.*

Ideally, we would like to find a minimal-size set of attributes $\boldsymbol{E} \subseteq \mathcal{A}$ s.t: $(O \perp\!\!\!\perp T | \boldsymbol{E}, C)$. However, in practice, we may not find these perfect explanations (that entirely explains away the correlation), hence we are looking for a minimal-size set of attributes that *minimize the partial correlation between T and O*.

Partial correlation measures the strength of a relationship between two variables, while controlling for the effect of other variables. A common measure of partial correlation is multiple linear regression, which is sensitive only to linear relationship. Here we use *Conditional Mutual Information* (CMI), a common measure of the mutual dependence between two variables, given the value of a third. CMI may suffer from underestimation, especially when quantifying dependencies among variables with high associations [76]. However, we avoid such cases since, as we explain in Section 4.2, we discard all attributes that are logically dependent on T or O. We chose CMI because of its estimation techniques and popularity [20]. It also allows us to develop information-theory-based optimizations. Note that $(O \perp\!\!\!\perp T | \boldsymbol{E}, C)$ holds iff $I(O; T | \boldsymbol{E}, C)$=0. Thus, we formalize the Correlation-Explanation problem as follows:

Definition 2.1 (**Correlation-Explanation**). *Given a set of candidate attributes $\mathcal{A}$ and a query Q, find a set of attributes $\boldsymbol{E}^*$ s.t.:*

$$\boldsymbol{E}^* = argmin_{\boldsymbol{E} \subseteq \mathcal{A}} I(O; T | \boldsymbol{E}, C) \cdot |\boldsymbol{E}|$$

Following previous work [44, 61, 64], besides the explanatory power, we also consider the cardinality of the attribute sets.

Example 2.3. *Among other attributes, we extracted from the KG the Gini coefficient (E_1), Density (E_2), and HDI (E_3) attributes. An attribute from SO is the developers Gender (E_4). According to our data, we have $I(O; T|C)$=2.6. When conditioning on E_1, we get: $I(O; T|C, E_1)$=1.3. Namely, conditioning on Gini coefficient, the correlation between Country and Salary decreases. That is, in countries with a similar Gini coefficient, there is less correlation between the country of developers and their salaries. When also considering Density, we get: $I(O; T|C, E_1, E_2)$=0.03. Thus, this set of attributes explains away the correlation in Q_{so}. When conditioning on HDI, on the other hand, we get: $I(O; T|C, E_3)$=2.5. Since the HDI of all countries in Europe is similar[5], this attribute does not explain the observed correlation. Similarly, when conditioning on Gender we get: $I(O; T|C, E_4)$=2.3, implying that the gender of the developers also cannot explain the correlation in Q_{so}.*

To assist analysts in interpreting the results, we enable users to learn the individual *responsibility* of each selected attribute. Given an explanation $\boldsymbol{E} \subseteq \mathcal{A}$, we rank the attributes in $\boldsymbol{E}$ in terms of their responsibilities as follows:

Definition 2.2 (Degree of responsibility). *Given a query Q and set of attributes $\boldsymbol{E}$, the degree of responsibility of the attribute $E_i \in \boldsymbol{E}$ is defined as follows:*

$$Resp(E_i) := \frac{I(O; T|\boldsymbol{E} \setminus \{E_i\}, C) - I(O; T|\boldsymbol{E}, C)}{\sum_{E_j \in \boldsymbol{E}} (I(O; T|\boldsymbol{E} \setminus \{E_j\}, C) - I(O; T|\boldsymbol{E}, C))}$$

The degree of responsibility of an attribute $E_i \in \boldsymbol{E}$ is the normalized value of its individual contribution. When all attributes in $\boldsymbol{E}$ contribute to the explanations (i.e., the numerator is positive), the denominator is non-negative. The degree of responsibility of E_i is positive if E_i contributes to the explanation. Thus, a negative responsibility score indicates that adding E_i only harms the explanation (it happens since E_i has a negative interaction information with O and T). The higher the responsibility score of an attribute, the greater is its individual explanation power.

Example 2.4. *Recall that E_1= Gini coefficient, E_2= Density, E_3= HDI, and E_4= Gender. Let $\boldsymbol{E}$=$\{E_1, E_2\}$. According to our data we have: $I(O; T|C, E_2)$=1.51. We get: $Resp(E_1)$=0.54, and $Resp(E_2)$=0.46. Now consider the attribute Hobby (E_5), indicating whether a developer is coding as an hobby. It has a negative interaction information with O and T. We have $I(O; T|C, E_5)$=2.7 >$I(O; T|C)$. Let $\boldsymbol{E}$=$\{E_1, E_5\}$. We get: $I(O; T|C, \boldsymbol{E})$=1.5, $Resp(E_1)$=1.2, and $Resp(E_5)$= − 0.2. Since E_5 did not contribute to the explanation, its responsibility is negative.*

Key Assumption. We generally believe that attributes with low responsibility are of little interest to a user and that XOR-like explanations (in which the explanation power of each individual attribute is low, but their combination makes a good explanation) are hard to understand; thus, they are less likely to be considered good explanations. Our view is motivated by [45]. A similar assumption is often made in feature selection [19, 71], where they assume the optimal feature set does not contain multivariate associations among features, which are individually irrelevant to a target class but become relevant in the presence of others. We further believe true xor style phenomena are likely to be uncommon in real datasets; the practical success of feature selection methods that make this assumption [20] is some evidence for this view. Further, generating XOR explanations would be a substantial additional technical challenge. It would eliminate our ability to prune low-relevance attributes and to define a stopping criterion for our algorithm (see Section

[4]The Gini index is a measure of statistical dispersion intended to represent the income inequality within a nation or a social group.

[5]As reflected in https://en.populationdata.net/rankings/hdi/europe/.

4). Also, extending our algorithm to consider XOR explanations would mean estimating CMI for a high-dimensional conditioning set, which is notoriously difficult [40].

3 ATTRIBUTES EXTRACTION

3.1 Extracting the Candidate Attributes

The first step is to map values that appear in the table $\mathcal{T}$ to their corresponding unique entities in the KG $\mathcal{G}$. This task is often referred to as the Named Entity Disambiguation (NED) problem [56]. We can use any off-the-shelf NED algorithm (e.g., [56, 78]) to match any non-numerical value in $\mathcal{T}$ to an entity in $\mathcal{G}$. At the next step, given an entity from $\mathcal{T}$, we extract all of its properties from the KG and store them in a dictionary. We then organize all the extracted dictionaries into a table, setting a null value to all properties whose values were missing. This process is equivalent to building the *universal relation* [30, 46] out of all of the entity specific relations that were derived from the KG. In cases where a property leads to multiple entities, in our implementation we aggregate the values by averaging them (or using mode for categorical attributes).

Note that one of the strengths of a KG is that most of the attributes are already reconciled. Namely, we will not have to match, e.g., different versions of HDI across different entities.

Multi-hops. To extract more attributes and potentially improve the explanations, one may "follow" links in $\mathcal{G}$. Namely, extract also all the properties of values which are entities in $\mathcal{G}$ as well. This process can be done up to any number of hops in $\mathcal{G}$. All properties are then flattened and stored as a single table.

Example 3.1. *A country's leader name is an attribute extracted for each country. We can extract all properties of the countries' leaders, such as their age and gender. In this case, we add to $\mathcal{E}$ additional properties such as Leader Age, and Leader Gender. Other properties may point to multiple entities. The US entity has the property Ethnic-Group, which points to different ethnic groups (which are entities in $\mathcal{G}$ as well). Each group has the property Population size. One may add the property of Avg Population size of Ethnic-Group to $\mathcal{E}$ by averaging the population size of the ethnic groups.*

3.2 Handling Missing Data

The extracted attributes may contain missing values. The simplest approach to dealing with missing values is to restrict the analysis to complete cases. Namely, discard cases that have missing values.However, this can induce *selection bias* if the excluded tuples are systematically different from those included. For example, if the HDI values of only the countries with a very high HDI are missing, restricting the analysis only to complete cases may lead to inaccurate and misleading explanations.

Handling missing data is an enduring problem for many systems [27]. We do not aim to make a novel intellectual contribution in this area, but rather to choose and adapt an existing method for our problem. A common approach is to impute missing values. Previous work showed that data imputation is unlikely to cause substantial bias if few data are missing, but bias may increase as the number of missing data increases [67]. Another common approach is Multiple Imputations (MI) [58]. While MI is useful in supervised learning as long as it leads to models with an acceptable level of accuracy, one must use it with care when generating explanations, as inaccurate imputation can lead to highly misleading explanations. Also, MI makes a missing-at-random assumption [27], which is often not the case in our setting. The approach that we followed is Inverse Probability Weighting (IPW), a commonly used method to correct selection bias [67]. In IPW, we restrict the attention only to complete cases, but more weight is given to some complete cases than others. We next explain how to employ this approach in our setting.

For simplicity of presentation, we assume that $\mathcal{T}$ and $\mathcal{E}$ have been joined into a single table. As we will explain in Section 4, for an attribute $E{\in}\mathcal{E}$ we estimate $I(O;T|E,C)$ and $I(E;E')$ for $E'{\in}\mathcal{E}$. Therefore, we need to recover the probabilities $P(O|C,E)$, $P(O|C,T,E)$, $P(E)$, and $P(E|E')$. But since E may contain missing values, we must ensure that those probabilities are *recoverable*. Given an attribute $E{\in}\mathcal{E}$, let R_E denote a selection attribute that indicates if the values of E for the i-th tuple in the results of Q is missing. I.e., $R_E[i]{=}1$ if the value of E for the i-th tuple was extracted, and $R_E[i]{=}0$ otherwise. A complete cases analysis means that we examine only cases in which $R_E[i]{=}1$. Let $R_E{=}1$ denote the selection of all tuples in which for them $R_E[i]{=}1$ holds. We say the probability of an event X which involves an attribute $E{\in}\mathcal{E}$ (e.g., $P(O|E)$, $P(E)$) is recoverable if: $P(X){=}P(X|R_E{=}1)$. If both E, E' contain missing values, then $P(E|E')$ is recoverable if: $P(E|E'){=}P(E|E', R_E{=}1, R_{E'}{=}1)$.

$P(O|E,C)$ is recoverable if the completeness of a case is independent of O given E and the context C. $P(O|T,E,C)$ is recoverable if the completeness of a case is independent of O given E, T, and C. Namely, the complete cases are a representative sample of the original sample, and each complete case is a random sample from the population of individuals with the same E and T values.

Proposition 3.1. *If $(O \perp\!\!\!\perp R_E = 1|E,C)$ and $(O \perp\!\!\!\perp R_E = 1|E,T,C)$, then $I(O;T|C,E) = I(O;T|C,R_E = 1,E)$.*

$P(E)$ and $P(E|E')$ are recoverable if the completeness of a case is independent of E, and remains independent of E given E'. In this case, we get that $I(E;E')$ is recoverable.

Proposition 3.2. *For $E_i, E_j{\in}\mathcal{E}$, if $(E_i \perp\!\!\!\perp R_{E_i}{=}1, R_{E_j}{=}1)$ and $(E_i \perp\!\!\!\perp R_{E_i}{=}1, R_{E_j}{=}1|E_j)$, then $I(E_i;E_j|R_{E_i}{=}1,R_{E_j}{=}1){=}I(E_i;E_j)$.*

In situations other than described above, the probabilities will generally not be recoverable. Following the IPW approach, we assign weights to complete cases, where the weight $W(X)$ of an event X is defined as: $P(X){=}P(X|R_E{=}1)W(X)$. We get: $W(X){=}P(R_E{=}1)/P(R_E{=}1|X)$. However, since E contains missing values, $P(X)$ is unknown, and thus we can not compute $W(X)$. We thus estimate $P(X)$. Commonly, a logistic regression model is fitted [33, 36]. Data available for this are the values of the attributes in $\mathcal{D}$. We therefore employ a logistic regression (at pre-processing) to predict missing values for E to estimate $P(X)$. We note that although, as in MI, we predict missing values, in the IPW approach, we only use those predicted values for weights computation and not for the entire analysis.

4 ALGORITHMS

4.1 The MCIMR Algorithm

Next, we present our algorithm, named the Minimal Conditional mutual Information Minimal Redundancy (MCIMR) algorithm for the Correlation-Explanation problem. This algorithm is inspired by a well-studied FS algorithm, called the MRMR algorithm

[59]. In the first stage of MRMR, features are selected according to Max Relevance Min Redundancy criteria. The main difference in MCIMR is that instead of the Max-Relevance criterion, we use a Min-Conditional-Mutual-Information criterion, adjusting the objective function to a minimum. We note that what makes this work different from ours is the details of the correctness proof, which is critical for obtaining the optimal output guarantee for our case.

We show that MCIMR is a PTIME algorithm that finds the optimal k-size explanation where k is given. We then define a stopping criterion, allowing our algorithm to stop when no further improvement is found.

When k equals 1, the optimal solution to Correlation-Explanation is the attribute $E \in \mathcal{A}$ that minimizes $I(O;T|C,E)$. When $k \geq 1$, a simple incremental solution is to add one attribute from $\mathcal{A}$ at a time: Given the explanation obtained at the $(k-1)$-th iteration $\boldsymbol{E_{k-1}}$, the k-th attribute to be added, denoted as E_k, is the one that contributes to the largest decrease of $I(O;T|C,\boldsymbol{E_{k-1}})$. Formally,

$$E_k = argmin_{E \in \mathcal{A} \setminus \boldsymbol{E_{k-1}}} I(O;T|C,\boldsymbol{E_k}) \quad (1)$$

where $\boldsymbol{E_k}=\boldsymbol{E_{k-1}} \cup \{E_k\}$.

Previous work has shown that it is difficult to get an accurate estimation for multivariate mutual information [59] (as in Equation (1)) Instead, our algorithm calculates only bivariate probabilities, which is much easier and more accurate. We thus incrementally select attributes based on Minimal-Conditional-mutual-Information (MCI) and Minimal-Redundancy (MR) criteria.

The idea behind MCI is to search a k-size set of attributes $\boldsymbol{E_k} \subseteq \mathcal{A}$ that satisfies Equation 2, which approximates Equation 1 with the mean value of all Conditional mutual Information (CI) values between the individual attributes in $\boldsymbol{E_k}$ and O and T:

$$\boldsymbol{E_k} = argmin_{\boldsymbol{E_k} \subseteq \mathcal{A}} CI(O,T,C,\boldsymbol{E_k}) \quad (2)$$

where $CI(O,T,C,\boldsymbol{E_k}) = \frac{1}{k} \sum_{E \in \boldsymbol{E_k}} I(O;T|C,E)$.

As in the MRMR algorithm, it is likely that attributes selected according to MCI are redundant. Therefore, the following Minimal Redundancy (MR) condition is added to select mutually exclusive attributes:

$$\boldsymbol{E_k} = argmin_{\boldsymbol{E_k} \subseteq \mathcal{A}} Rd(\boldsymbol{E_k}) \quad (3)$$

where $Rd(\boldsymbol{E_k}) = \frac{1}{k^2} \sum_{E_i,E_j \in \boldsymbol{E_k}} I(E_i;E_j)$.

Our goal is to minimize CI and Rd simultaneously. Namely, we look for a k-size set of attributes $\boldsymbol{E_k^*} \subseteq \mathcal{A}$ such that:

$$\boldsymbol{E_k^*} = argmin_{\boldsymbol{E_k} \subseteq \mathcal{A}} [CI(O,T,C,\boldsymbol{E_k}) + Rd(\boldsymbol{E_k})] \quad (4)$$

We present an incremental algorithm that finds the optimal k-size attribute set defined by Equation 4. It is defined as follows. In the k-th iteration we already have the set $\boldsymbol{E_{k-1}}$ with $k-1$ attributes. The goal is to select the k-th attribute to be added. This is done by selecting the attribute that minimizes the following condition:

$$E_k = argmin_{E \in \mathcal{A} \setminus \boldsymbol{E_{k-1}}} [I(O;T|C,E) + \frac{1}{k-1} \sum_{E_i \in \boldsymbol{E_{k-1}}} I(E;E_i)] \quad (5)$$

We can prove that the combination of the MCI and MRd criteria is equivalent to Equation 1.

Theorem 4.1. *The MCIMR incremental algorithm yields the optimal k-size solution to Equation 1.*

Stopping Criteria. The MCIMR algorithm assumes that the size of the explanation k is given. However, given two consecutive solutions of sizes k and k+1, we do not know which one provides a better explanation. Namely, we can not say if $I(O;T|C,\boldsymbol{E_k}) < I(O;T|C,\boldsymbol{E_{k+1}})$ or vice versa. As mentioned in Section 2.2, we assume that attributes in which their marginal explanation power is small are of no interest to the user. We thus stop the algorithm after the first iteration in which the responsibility score of the new attribute to be added is ≈ 0. Namely, we treat k as an upper bound on the explanation size. To this end, we propose the *responsibility test*. Given the set of attributes selected so far $\boldsymbol{E_k}$, this test verifies if the responsibility score of a candidate attribute E_{k+1} is ≈ 0.

Algorithm 1: The MCIMR Algorithm.

```
input : A number k, a set of attributes 𝒜, the outcome, treatment attributes O and T,
        and the context C
output: An explanation E.
1  MCIMR(k, 𝒜, O, T, C):
2  E ← ∅.
3  for i ∈ [1, k] do
4      E_i ← NextBestAtt (O, T, C, E, 𝒜)
5      if O ⊥⊥ E_i|E                          // The responsibility test for E_i
6      then
7          return E
8      E ← E ∪ {E_i}
9  return E
10 NextBestAtt (O, T, C, E, 𝒜):
11 E* ←None, v ← ∞
12 foreach E ∈ 𝒜 \ E do
13     /* Weights are added if selection bias was detected       */
14     v_1 ← I(O;T|C,E), v_2 ← 0                          // Min CI computation
15     foreach E' ∈ E do
16         /* Weights are added if selection bias was detected    */
17         v_2 ← v_2 + I(E;E')                  // Min redundancy computation
18     if v_1 + v_2/|E| < v then
19         E* ← E, v ← v_1 + v_2/|E|
20 return E*
```

Lemma 4.2 (Responsibility test). *If $O \perp\!\!\!\perp E_{k+1}|\boldsymbol{E_k}$ then $Resp(E_{k+1}) \leq 0$.*

We measure conditional independence using the highly efficient independence test proposed in [65].

The full MCIMR algorithm is depicted in Algorithm 1. First, the algorithm initializes the attribute set $\boldsymbol{E}$ to be returned with the empty set (line 2). Then, new attributes are iteratively added according to the NextBestAtt procedure (line 4). The algorithm then applies the responsibility test to the selected attribute. If the responsibility of this attribute is ≈ 0, the algorithm terminates and returns the solution obtained until this point (lines 5-7). Otherwise, the algorithm terminates after k iterations (line 9). Given the attribute set selected up until the i-th iteration, the NextBestAtt procedure finds the i-th attribute to be added. It implements Equation 5, by iterating over all candidate attributes and computing their individual explanation power (line 14), and their average pair-wise mutual information with all selected attributes (lines 16-18).

For simplicity, we omitted the parts dedicated to handling missing data from presentation. In our implementation, before executing lines 14 and 18, we check whether or not weights are needed to be added and adjust the computation accordingly.

The time complexity of this algorithm is $O(k|\mathcal{A}|)$. Nevertheless, note that the size of $\mathcal{A}$ is potentially very large because of the KG.

4.2 Optimizations

We propose several simple optimizations to reduce the size of $\mathcal{A}$ and thereby reduce execution times. These optimizations may cause the MCMIR algorithm to overlook some important attributes. However,

our experiments show that the solutions are almost unaffected in practice while running times significantly improve. We propose two types of optimizations: **Across-queries optimizations** that could be executed at pre-processing; and **Query-specific optimizations** that could be done only once O and T are known and are executed before running the MCIMR algorithm.

Preprocessing pruning. Attributes discarded at this phase either have a fixed value, a unique value for each tuple, or lots of missing values. Thus, such attributes are uninteresting as an explanation [39, 65]. **Simple Filtering**: We drop all attributes with a constant value (e.g., the attribute Type which has the value *Country* to all countries), and attributes in which the percentage of missing values is >90%. **High Entropy**: we discard attributes such as CountryCode, wikiID, that have high entropy and (almost) a unique value for each tuple (as was done in [65]).

Online pruning. **Logical Dependencies**: Logical dependencies can lead to a misleading conclusion that we found a confounding attribute, where we are, in fact, conditioning on an attribute that is functionally dependent on T or O. If for an attribute E we have: $FD: E \Rightarrow T$ then we get: $H(T|E) \approx 0$. We have: $I(O;T|E,C) = H(O|E,C) - H(O|T,E,C)$. But since T and E are dependent, we get: $H(O|T,E,C) \approx H(O|E,C)$ and thus $I(O;T|E,C)=0$. We thus discard all attributes E s.t. $H(T|E)=\epsilon$ and $H(E|T)=\epsilon$ for $\epsilon \approx 0$ (resp., for O) as was done in [65]. These tests correspond to approximate functional dependencies, such as CountryCode $\Rightarrow$ Country. **Low Relevance**: As mentioned, we assume that attributes with low responsibility are of little interest to a user, and thus we prune such attributes. Specifically, if $(O \perp\!\!\!\perp E|C)$ and $(O \perp\!\!\!\perp E|C,T)$ we get that: $H(O|E,C)=H(O|C)$ and $H(O|T,E,C)=H(O|T,C)$. Thus: $I(O;T|E,C) = H(O|E,C) - H(O|T,E,C) = H(O|C) - H(O|T,C) = I(O;T|C)$. That means that the individual explanation power of E is low.

Another possible optimization is to cluster attributes that are highly correlated, such as HDI and HDI Rank. This will reduce the redundancy among the attributes [39]. However, we found this optimization to be not useful because of the following reasons: (1) It could only be done after the query arrives, namely after we are done filtering, and the clustering process took longer than running our algorithm on all attributes. (2) We found that attributes clustered together were not necessarily semantically related.

4.3 Identifying Unexplained Subgroups

The MCIMR algorithm finds the explanations for the correlation between T and O. While the generated explanations are optimal considering the whole data, they may be insufficient for some parts in the data. To this end we propose an algorithm the user may use after getting the explanation, to identify unexplained data subgroups. The input of this algorithm is the original query Q and the explanation found by MESA. The output is a set of context refinements for the original query that may be of interest to the user. Each subgroup corresponds to a refinement of Q in which a different explanation is required and may require further exploration.

Example 4.1. *Consider a query compare the average salary of developers among countries. The explanation found by MESA is* $\boldsymbol{E}$={*HDI, Gini*}. *As mentioned, the HDI of all countries in Europe is similar. Thus, for of all countries in Europe, it is likely that* $\boldsymbol{E}$ *is not a satisfactory explanation. Here, the subgroup is all countries in Europe.*

Algorithm 2: Top-k unexplained data groups.

input : A number k, a set of attributes $\mathcal{A}$, the attributes O and T, the context C, an explanation $\boldsymbol{E}$, and a threshold τ.
output: Context refinements $\{C_1, \ldots, C_k\}$ s.t. the corresponding groups are the largest k groups and $I(O;T|C_i,\boldsymbol{E}) > \tau$

1 $\mathcal{R} \leftarrow \emptyset$
2 $MaxHeap \leftarrow$ GenChildren(C)
3 **while** $|\mathcal{R}| < k$ *or* $MaxHeap.isEmpty()$ **do**
4 $C' \leftarrow MaxHeap.exatractMax()$
5 **if** $I(O;T|C',\boldsymbol{E}) > \tau$ **then**
6 update($\mathcal{R}, C'$) // If none of the ancestors of C' are in $\mathcal{R}$, insert C' into $\mathcal{R}$.
7 **else**
8 **for** $C'' \in$ GenChildren(C') **do**
9 $MaxHeap.insert(C'')$
10 **return** $\mathcal{R}$

For simplicity, numerical attributes are assumed to be binned. Data groups are defined by a set of attribute-value assignments and correspond to refinement of the context C of Q. Treating the context C as a set of conditions, a refinement C' of C is a set s.t. $C' \subset C$. The goal is then to find the largest data groups s.t. $\boldsymbol{E}$ can not serve as their explanation. More formally, we are inserted in the top-k data groups (in terms of their size), each correspond to a context C' (a refinement of C), s.t. $I(O;T|C',\boldsymbol{E}) > \tau$ for some threshold τ (τ can be set based on the initial value $I(O;T|C,\boldsymbol{E})$).

Example 4.2. *Continuing with Example 4.1, we refine the query Q by adding a WHERE clause selecting only countries in Europe ($C' = \{$Continent = Europe$\}$). Let Q_{EU} denote this refinement query of Q. Indeed, we get: $I(O;T|C',\boldsymbol{E})$=2.13. As mentioned in Example 2.3, the optimal explanation for Q_{EU} is* {Gini, Density}.

A naive algorithm would traverse over all possible contexts C', check if $I(O;T|C',\boldsymbol{E}) > \tau$ for each refinement query, and choose the largest data groups for which $\boldsymbol{E}$ is not a satisfactory explanation. We propose a more efficient algorithm, exploiting the notion of pattern graph traversal [16]. Intuitively, the set of all context refinements can be represented as a graph where nodes correspond to refinements (set of conditions) and there is an edge between C and C' if C' can be obtained from C by adding a single value assignment. As shown in [16], the graph can be traversed in a top-down fashion, while generating each node at most once.

Algorithm 2 depicts the search for the largest k data groups that for which $\boldsymbol{E}$ is not a satisfactory explanation. It traverses the refinements graph in a top-down manner, starting for the children of the context C. It uses a max heap $MaxHeap$ to iterate over the refinements by their size. It first initialize the result set $\mathcal{R}$ (line 1) and $MaxHeap$ with the children of C (line 2). Then, while the $\mathcal{R}$ consists of less then k refinements (line 3), the algorithm extracts the largest (by data size) refinement C' (line 4) and computes $I(O;T|C',\boldsymbol{E})$. If it exceeds the threshold τ (line 5), C' is used to update $\mathcal{R}$ (line 6). The procedure `update` checks whether any ancestor of C' in the graph is already in $\mathcal{R}$ (this could happen because the way the algorithm traverses the graph). If not, C' is added to $\mathcal{R}$. If $I(O;T|C',\boldsymbol{E}) \leq \tau$ (line 5), the children of C' are added to the heap (lines 8– 9).

5 EXPERIMENTAL STUDY

We present experiments that evaluate the effectiveness and efficiency of our solution. We aim to address the following research questions. $Q1$: What is the quality of our explanations, and how

Table 1: Examined Datasets.

Dataset	n	$\|\mathcal{E}\|$	Columns used for extraction
SO	47623	461	Country, Continent
COVID-19	188	463	Country, WHO-Region
Flights	5819079	704	Airline, Origin/Destination city/state
Forbes	1647	708	Name

does it compare to that of existing methods? $Q2$: How robust are the explanations to missing data? $Q3$ What is the efficiency of the proposed algorithm and the optimization techniques? $Q4$: How useful are our proposed extensions?

5.1 Experimental Setup

Our code and datasets are available at [11]. We used DBPedia [8] as the KG, the SpaCy Entity Linker [10] for NED, and the Pyitlib library [9] for information-theoretic computations. The experiments were executed on a PC with a 4.8GHz CPU, and 16GB memory.

Datasets. We examine four commonly used datasets: **(1) SO**: Stack Overfow's (SO) annual developer survey is a survey of people who code around the world [7]. It has more than 47K records containing information about the developers' such as their age, gender, income, and country. **(2) Covid-19**: This dataset [3] includes information such as number of confirmed, death, and recovered cases in 2020 across the globe. **(3) Flights Delay**: This dataset [4] contains transportation statistics of over 5.8M domestic flights operated by large air carriers in the USA in 2015. **(4) Forbes**: This dataset [6] contains annual earning information of 1647 celebrities since 2005 extracted from the Forbes magazine. It contains the celebrities' annual pay, and category (e.g., Actors, Producers).

The attributes in which their values were linked to entities and the number of extracted attributes in each dataset are given in Table 1. By default, we follow 1-hop in the KG.

Baseline Algorithms. We compare MESA against the following baselines: **(1) Brute-Force**: The optimal solution according to Def. 2.1. This algorithm implements an exhaustive search over all subsets of attributes. To make it feasible, we run it after employing our pruning optimizations (see Section 4.2). **(2) Top-K**: This naive algorithm ranks the attributes according to their individual explanation power (i.e., their conditional mutual information with T and O). The lower the score, the higher the rank is. **(3) Linear Regression** (LR): This baseline employs the OLS method to estimate the coefficients of a linear regression that describes the relationship between the outcome and the candidate attributes. The explanations are defined as the top k attributes with the highest coefficients (and their p value is $< .05$). We note that Pearson's r is the standardized slope of LR and thus can be viewed as part of our competing baselines. **(4) HypDB** [65]: This system employs an algorithm for confounding variable detection based on causal analysis. The explanations are defined as the top-k attributes with the highest responsibility scores, as defined in [65]. *This baseline serves as a representative example for a causal-analysis-based approach.* **(5) MESA$^-$**: Last, to examine how pruning affects the explanation, we examine the explanation generated by MESA without the pruning optimizations.

We also examined the explanations generated by CajaDE [39], a system that generates query results explanations based on augmented provenance information. However, since in all cases, the explanations generated by CajaDE obtained the lowest scores, we omit its results from presentation. The reason for that is the explanations generated by CajaDE are a set of patterns that are unevenly distributed among groups in the query results. Such explanations are independent of the outcome variable. Thus, it cannot generate explanations that explain away the correlations between T and O.

Unless mentioned otherwise, we set the maximal explanation size, k, to 5 and extracted attributes for 1-hop in the knowledge graph. For a fair comparison, we run all baselines (except for MESA$^-$) after employing our pruning optimizations.

5.2 Quality Evaluation (Q1)

First, we validate our intuition that attributes extracted from KGs can explain correlations in common scenarios. To this end, we randomly generated 40 SQL queries (10 from each dataset) as follows. We set the exposure attribute to be one of the attributes used to extract additional attributes from the KG (as listed in Table 1). We set the outcome to be a numerical attribute that could be predicted from the data (e.g., Departure/Arrival Delay in Flights, Confirmed/New/Death Cases in Covid-19). We then added a WHERE clause to each query by randomly picking another attribute and one of its values, ensuring the selected subset contains more than 10% of the tuples in the original table. Full details can be found in [11]. We say our approach was useful for explaining the correlation between the exposure and outcome if (1) the conditional correlation (while conditioning on the explanation generated by MESA) is lower than the original correlation, and (2) the explanation contains at least one extracted attribute. We report this was the case in 72.5% percent of the queries. *This validates our assumption that, in many cases, data extracted from KGs is relevant to explaining correlations.*

Next, we aim to asses the quality of the generated explanations. To this end, we present a user study consisting of explanations produced by each algorithm. In this study we aim to validate our problem definition and evaluate the explanations quality.

Since a standard benchmark for results explanation does not exist, we consider 14 representative queries suffering from confounding bias, as shown in Table 2. Our queries are inspired by real-life analysis reports, ranging from SO annual reports [5] to news and media websites (e.g., Vanity Fair [1], and USA Today[2] for Forbes and Flights), and academic papers about Covid-19 [35, 72]. Similar experiments were conducted in [39, 42, 65].

We recruited 150 subjects on Amazon Mechanical Turk[6]. This sample size enables us to observe a 95% confidence level with a 10% margin of error. Subjects were asked to rank each explanation of each query (shown together with its corresponding query) on a scale of 1–5, where 1 indicates that the explanation does not make sense and 5 indicates that the explanation is highly convincing. The form we gave to the subjects is available at [11]. To compare the generated explanations with the "ground-truth" explanations we will show that our explanations are supported by previous findings. A similar approach was taken in [65].

The worst-case time complexity of HypDB is exponential in the size of $\mathcal{A}$ [65]. We run it over all attributes in $\mathcal{A}$ (after pruning) and report that it never terminates within 10 hours. Thus, we have no choice but to limit the number of attributes for HypDB, to allow it to generate explanations in a reasonable time. For HypDB,

[6] Amazon Mechanical Turk: https://www.mturk.com/

Table 2: User study: The best and second best explanations are marked in red and blue, resp.

Dataset		Query	Brute-Force	MESA-	MESA	Top-K	LR	HypDB
SO	Q_1	Average salary per country	-	**HDI Rank, Gini**	**HDI, Gini**	HDI, Established Date	Population Census, Language	GDP
	Q_2	Average salary per continent	-	**GDP Rank, Density**	**GDP,Density**	GDP,Area rank	GDP, Area Rank	GDP
	Q_3	Average salary per country in Europe	-	**Population Census, Gini Rank**	**Population Census, Gini**	Population Census, Population Estimate	Population Census, Language	Gini, Area Rank
Flights	Q_1	Average delay per origin city	-	**Precipitation Days, Year UV, Airline**	Population urban, Year Low F, Airline	Year Low F, Year Avg F, December Low F	Year Low F, December percent sun, Day	**Year Low F, May Precipitation Inch, Airline**
	Q_2	Average delay per origin state	-	**Density Sqmi, Year Snow Inch, Airline**	**Population estimation, Year Low F, Airline**	Population estimation, Population Urban, Population Rank	Population estimation, Median Household Income, Distance	Record Low F, Population estimation, Day
	Q_3	Average delay per origin cities in CA	-	**Density, Population Metropolitan, Security Delay**	Density, Population Total,Security Delay	Population Metropolitan,Security Delay	-	**Density, Population Ranking, Cancelled**
	Q_4	Average delay per origin state and airline	-	**Population Total, Fleet size**	**Population Ranking, Fleet size**	Density, Population Total	-	Revenue, Dec Record Low F
	Q_5	Average delay per airline	-	**Equity, Fleet Size**	**Equity, Fleet Size**	Equity, Net Income	**Equity, Fleet Size**	**Num of Employees, Revenue**
Covid-19	Q_1	Deaths per country	**HDI, GDP, Confirmed cases**	**HDI, GDP Rank, Confirmed cases**	**HDI, GDP, Confirmed cases**	GDP Rank, GDP Nominal, HDI	Area Rank, Currency, Recovered cases	Density, Time Zone, Confirmed cases
	Q_2	Deaths per country in Europe	**Gini Coefficient, Population Census, Confirmed cases**	**Gini Rank, Density, Confirmed cases**	**Gini Coefficient, Population Census, Confirmed cases**	Gini Rank, Gini Coefficient, GDP	Area Rank, Currency, Population Total	Currency, GDP, New cases
	Q_3	Average deaths per WHO-Region	**Density, Confirmed Cases**	**Density,Confirmed Cases**	**Density,Confirmed Cases**	**Density,Confirmed Cases**	-	**Area Km,Confirmed Cases**
Forbes	Q_1	Salary of Actors	**Net Worth, Gender, Age**	**Net Worth, ActiveSince, Gender**	Net Worth, Gender	Net worth, Awards	Citizenship, Honors	Gender, Honors
	Q_2	Salary of Directors/Producers	**Net Worth, Awards**	**Years Active, Net Worth**	**Net Worth, Awards**	Net Worth, Age	-	Years Active
	Q_3	Salary of Athletes	**Cups, Draft Pick, Active Years**	National Cups, Draft Pick	Cups, Draft Pick	Total Cups, National Cups	-	**Cups, Active Years**

Table 3: Average explanation scores according to the subjects (the higher the better).

Baseline	Average Score	Average Variance
Brute-Force	3.8	0.8
MESA-	3.7	1.1
MESA	3.5	0.9
HypDB	2.8	1.1
Top-K	2.1	0.8
LR	1.8	0.6

besides our pruning optimizations, we omitted candidate attributes uniformly at random, ensuring that $|\mathcal{A}| \leq 50$. We only report the results of Brute-Force for the small Covid-19 and Forbes datasets, as it was infeasible to compute the explanations for the larger datasets. We do not randomly drop attributes for computational efficiency here because Brute-Force is intended to be an optimal benchmark against which the others are judged. The explanations generated by different methods are given in Table 2, and the average scores given by the subjects for each algorithm are depicted in Table 3.

Result Summary: We summarize our main finding as follows:

- The subjects found the explanations generated by Brute-Force, MESA$^-$, and MESA to be the most convincing. This supports our mathematical definition of what constitutes a good explanation.
- MESA explanations are supported by previous in-domain findings, which serve as "domain-expert" explanations.
- Our pruning has little effect on explanation quality.
- The next best competitor is HypDB, which has the goal of discovering confounding attributes. However, it is unable to scale to a large number of candidate attributes.
- As expected, Top-k yields redundancy in selected attributes.

First, subjects found the explanations generated by Brute-Force, MESA$^-$, and MESA to be the most convincing. *This positive result supports our mathematical definition of what constitutes a good explanation (Def 2.1).* The pairwise differences between the average scores of these 3 methods are not statistically significant. Previous in-domain findings also support our generated explanations. For example, in SO Q_1, it was shown in [5] that there is a correlation between the developers salary and countries' economies (which is reflected in the HDI and Gini values). For Flights Q_1, it was stated in [2] that weather is one of the top reasons for flights delay in the US. For Covid-19 Q_1, it was shown that there is a correlation between countries' economies and Covid-19 death rate [35, 72]. More details can be found in the appendix. In all cases where the results of Brute-Force and MESA are different, it happens because MESA drops attributes with insignificant responsibility (according to the responsibility test). For example, in Forbes Q_1, MESA dropped Age, as its responsibility is ≈ 0. The low difference between the results of MESA$^-$ and MESA indicates that the pruning has little effect on the explanations quality. Namely, *MESA is able to execute efficiently without compromising on explanation quality.*

The next best competitor is HypDB (the average score is worse than that of MESA. This difference is statistically significant, $p < .05$). This is not surprising as HypDB finds confounding attributes using causal analysis. However, the main disadvantage of HypDB is its ability to scale for large number of confounding attributes. In cases where HypDB generated explanations that were considered not convincing, it was mainly because the most important attributes were dropped (since we sampled only a fraction of the attributes to

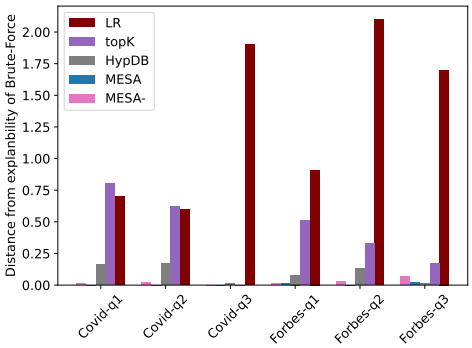
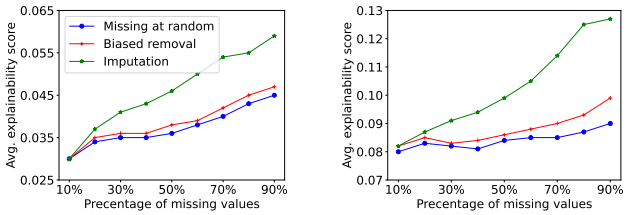


Figure 2: Distance from explainability scores of Brute-Force.

enable feasible execution times). *This demonstrates the limitation of causal-analysis-based solutions in handling large search spaces.* Not surprisingly, the explanations generated by Top-K and LR were considered to be less convincing (their average scores are statistically significant from all other methods, $p<.05$). For Top-K, this is substantially because it ignores redundancy among selected attributes. For example, in Flights Q_1, it chose the attributes Year Low F and Year Average F, which are highly correlated. For LR, in many cases, it failed to generate explanations, as there were no attributes in which their p-value was lower than .05. Even when it succeeded, the subjects found them to be less convincing than MESA. The reason is that LR focuses only on finding linear correlations.

Explainability scores. Let $\boldsymbol{E}$ denote the explanation found by an algorithm. We call $I(O;T|\boldsymbol{E})$ the explainability score. An Explainability score equal to 0 means that $\boldsymbol{E}$ perfectly explains away the correlation between O and T. The explainability scores of Brute-Force serve as ground truth. In some cases, the explanations generated by all algorithms cannot fully explain away the correlations. For example, in Flights Q_2, the explainability score of MESA is 0.25. This means that other factors that affect flight delays may not exist in the KG (e.g., labor problems). Nevertheless, most of the explanations of all methods consist of attributes extracted from the KG. This validates our intuition that *in many real-life scenarios, explanations could be mined from KGs.* The results are depicted in Figure 2. The y-axis is the distance between the explainability scores of each method and Brute-Force. The lower the distance the better is the explanation. Observe that the explainability scores of MESA are almost as good as the ones of Brute-Force and MESA$^-$, and are much better than those of the competitors.

Impact of pruning. We next examine how useful were our pruning techniques. **Offline Pruning**. At the offline phase, we filter the extracted attributes using the simple filtering and the high-entropy techniques. We found those two simple optimizations to be highly useful: On average, we dropped 41%, 59%, 45%, and 73% of the extracted attributes, in the SO, Flights, Covid-19, and Forbes dataset, resp. **Online Pruning**. At query time, we filter the extracted attributes using the logical dependency and the low relevance techniques. Not surprisingly, as most irrelevant attributes were already dropped, we dropped many fewer attributes at this phase. On average, we dropped 14%, 6%, 11% and 3% of the remaining attributes, in SO, Flights, Covid-19, and Forbes, resp.



SO **Covid-19**

Figure 3: Explainability as a function of missing data.

5.3 Robustness to Missing Data (Q2)

On average, the percentage of missing values in extracted attributes is 37%, 42%, 45% and 73% in Covid-19, SO, Flights and Forbes, resp. The high prevalence of missing values in Forbes is because DBpedia uses different attributes to describe a person from each category (e.g., actors, authors). In Covid-19, SO, Flights, and Forbes, the percentage of attributes with selection bias (according to the conditions defined in Section 3.2) is 13.3%, 14.1%, 24.2%, and 29.4%, resp. *This verifies that selection bias exists in attributes extracted from KGs, and thus should be appropriately handled.*

We examine the robustness of our explanations to missing data. To this end, we vary the percentage of missing values from the top 10 most relevant (w.r.t. the outcome) attributes. We examine two ways to omit values: missing-at-random and biased removal. To enable biased removal, the top-x highest values from the selected columns were omitted (when varying x). We consider the explanations generated by MESA and examined the effect on their average explainability score. Explainability should not be affected if an explanation is robust to missing data. We also examine the effect on the explainability scores while imputing missing values. Here we used the common mean imputation technique [75]. The results for the SO and Covid-19 datasets are depicted in Figure 3. As expected, data imputation has a great effect on explainability. On the other hand, our approach is much less sensitive to missing data. As can be seen, even with 50% missing values (at random or not), the explainability scores have hardly changed. When the percentage of missing values is above 50%, a lot of the information is missing, and thus it is harder to estimate partial correlation correctly. This demonstrates the robustness of our explanations to missing data, even when values are not missing at random.

5.4 Efficiency Evaluation (Q3)

To examine the contribution of our optimizations, we report the running times of the following methods: **No Pruning** —the MCIMR algorithm without pruning; **Offline Pruning** —MCIMR only with offline pruning; **MCIMR** —MCIMR with all pruning optimizations. We study the effect of multiple parameters on running times. For each dataset, we report the average execution time of the queries presented in Section 5.2. In all cases, the execution time of MCIMR was less than 10 seconds, a reasonable response time for an interactive system. In what follows, we omit the results obtained on the (smallest) Covid-19 dataset from presentation, as the results demonstrated similar trends to those of Forbes.

Candidate Attributes. In this experiment, we omitted from consideration attributes from $\mathcal{A}$ uniformly at random. The results are depicted in Figure 4. In all dataset, we exhibit a (near) linear growth

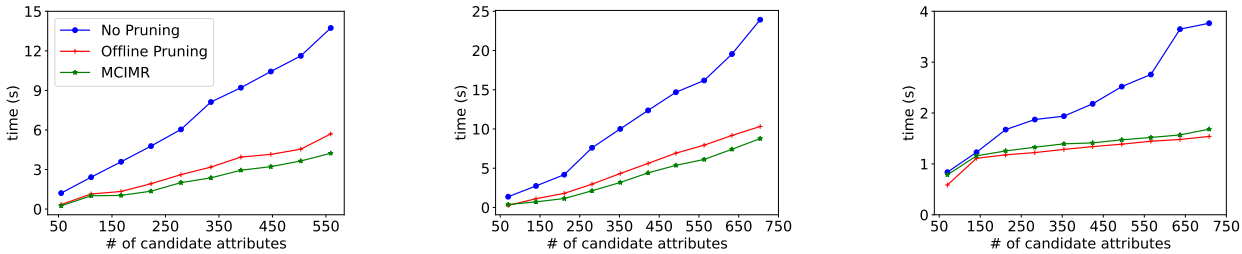


Figure 4: Running times as a function of the number of candidate attributes.

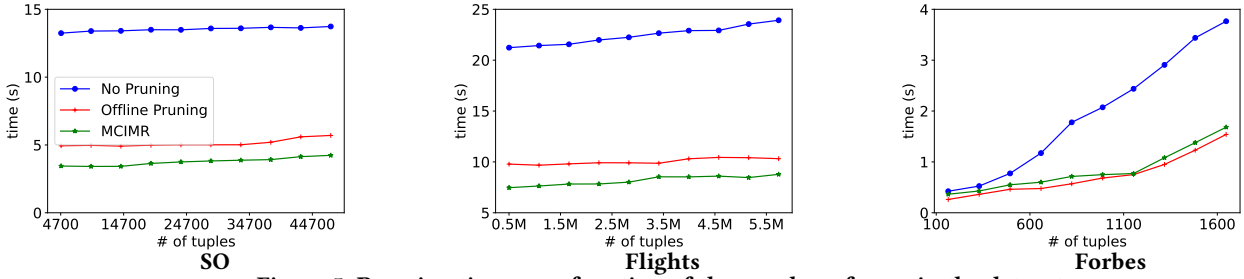


Figure 5: Running times as a function of the number of rows in the dataset.

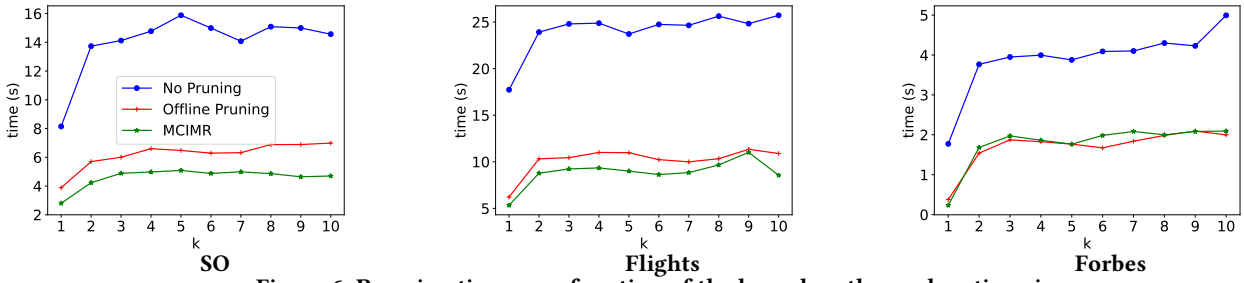


Figure 6: Running times as a function of the bound on the explanation size.

in running times as a function of the size of $\mathcal{A}$. *Observe that the execution times of No-Pruning are significantly higher than those of Offline Pruning and MCIMR, indicating the usefulness of the offline pruning techniques.* The difference in times across the datasets is due to the size of the dataset. Estimating mutual information on large datasets (e.g., Flights, SO) takes longer than on small datasets (e.g., Forbes). In Forbes, Offline Pruning is faster than MCIMR. This implies that in small datasets online pruning is not necessary, as it takes longer that running MCIMR.

Data Size. We vary the number of tuples in the datasets, by removing tuples uniformly at random. The results are depicted in Figure 5. In SO and Flights, observe that the dataset size has a little effect on running times. This is because the queries considered over those datasets are group-by queries. Thus, when randomly omitting tuples from the datasets, the number of considered groups is almost unchanged. On the other hand, since in Forbes the examined queries compared among tuples from the dataset (without a GROUP-BY statement), we exhibit a (near) linear growth in running times.

Explanation size. We vary the bound on the explanation size. Recall that given a bound k, MCIMR returns an explanation of size $\leq k$. It may return an explanation of size $l<k$ if the responsibility of the l+1 attribute is $\approx$0. The results are shown in Figure 6. In all cases, the size of the explanations was no bigger than 3. Thus, as can be seen, for all methods, k has almost no effect on running times, as the algorithms terminate after no more than 4 iterations.

5.5 Extensions (Q4)

Multi-Hops. We examine the effect of extracting attributes following more than one hop in the KG. We report that in the vast majority of cases, MESA's explanations were unaffected. Namely, almost all attributes extracted from 2 or more hops were found to be irrelevant (and were pruned). In some cases, we found at most one more attribute that was included in the explanations. For example, in Forbes Q_1, an attribute representing the average budget of the films played by actors (attribute extracted from 2-hops) was included in the explanation. In all cases, no attributes from 3 or more hops was considered to be relevant. Further, since the number of candidate attributes was increased (in 145%, on average), running times were increased (by up to 15 seconds). This indicates that most of the relevant information can be found in the first hop. Future research will predict which paths in the KG may lead to relevant attributes.

Unexplained Subgroups. We demonstrate the effectiveness of the Top-K unexplained groups algorithm by focusing on SO Q_1, setting

Table 4: Top-5 unexplained groups for SO Q1.

Rank	Size	Data group
1	18342	Continent = Europe
2	17899	Continent = Asia
3	15466	Continent = North America
4	14788	Currency = Euro
5	12754	Continent = Africa

$\tau > 0.2$. The top-5 largest unexplained groups for this query are given in Table 4. Observe that economy-related attributes (such as GDP, HDI, Gini Coefficient) of selected data groups are internally consistent (e.g., as mentioned, the HDI of countries in Europe is similar). Thus, it makes sense that the explanation generated for SO Q_1 ({HDI, Gini}) will not be a satisfactory explanation for these data groups. Indeed, as can be seen in Table 2, the explanation found by MESA for the top-1 unexplained data group (SO Q_3) is different from the explanation found for all countries. We ran this algorithm over the 14 queries depicted in Table 4. The average execution time is 4.4s. This demonstrates the ability of our algorithm to efficiently identify subgroups (corresponding to query refinements) that are likely to be of interest to the user.

6 RELATED WORK

We discuss multiple lines of work that are relevant to ours.

Results Explanations. Methods explaining why data is missing or mistakenly included in query results have been studied in [18, 21, 37, 70]. Explanations for unexpected tuples in the results have been presented in [17, 52]. Those works are orthogonal to our work, as we aim to explain unexpected correlations. Another line of work provides explanations on how a query result was derived by analyzing the query provenance and pointing out database facts that significantly affect the results [49, 50, 53]. Those methods are designed to generate tuple-level explanations and not attribute-level explanations that are required for unearthing correlations between attributes. Another type of explanation for query results is a set of patterns that are shared by one (group of) tuple but not by another (group of) tuple [29, 39, 63, 64, 73]. However, those works as well do not account for correlations among attributes. [65] presented HypDB, a system that provides explanations based on causal analysis, measured by the correlations among attributes. However, as mentioned in our experiments, HypDB only considers attributes from the input table, and it cannot efficiently handle a large amount of candidate attributes.

We share with [39] the motivation for considering explanations that are not solely drawn from the input table. [39] presented CajaDE, a system that generates explanations of query results based on information mined from tables related to the table accessed by the query. However, related tables often do not exist. Moreover, as mentioned in Section 5, their explanations are *independent of the outcome*. Thus, even if CajaDE is given the attributes mined from a KG, it may generate explanations that are irrelevant to the correlation between the exposure and outcome.

Feature Selection. The Correlation-Explanation problem is closely related to the well-studied Feature Selection (FS) problem [20, 32, 40], the problem of selecting a subset of the most relevant attributes for use in model construction. FS methods aim to select a concise and diverse set of attributes relevant to a target attribute [20]. The goal of FS is twofold: (1) *Simplification of models*: select as few features as possible. This makes models easier to interpret, reduces training times, and helps to avoid the curse of dimensionality [43]. (2) *Model accuracy*: minimize the classification error. If the underlying model is not given, The goal is to maximize the dependency between selected features and the target class [59]. We note the model simplification requirement corresponds to the conciseness and no-redundancy requirements of a "good" explanation, and model accuracy corresponds to explainability.

Closest to our project, there is a line of work using information-theoretical-based methods for FS [40]. Their main goal is to maximize the relevance of the selected features and minimize redundancy. Such methods are typically independent of the underlying learning algorithm. Algorithms in this family [28, 31, 38, 41, 51] define different criteria to measure the importance of features. Namely, to maximize feature relevance and minimize their redundancy. The relevance of a feature is typically measured by its correlation with the target attribute. The MCIMR algorithm is inspired by the first stage of a commonly used FS algorithm in this family, called MRMR [59]. The MRMR algorithm has been widely used in different domains, such as gene classification [24] and emotion recognition from electrodermal activity [68].

Confounding Bias. Confounding bias, referred to as a "mixing of effects", occurs when an analyst tries to determine the effect of an exposure on an outcome, but unintentionally measures the effect of another factor (i.e., the confounding variable) on the outcome. This results in a distortion of the actual association between the outcome and exposure [34, 69]. We share with [65] the motivation of identifying confounding bias in SQL queries for improved decision making. However, we consider attributes that are not solely drawn from the input table and are mined from KGs. Also, as demonstrated in our experiments, our approach is more scalable.

Explainable AI. A related line of work is Explainable AI (XAI), an emerging field in machine learning that aims to address how black box decisions of AI systems are made [13, 26]. Similar to our approach, XAI can be used to learn new facts, to gather information and thus to gain knowledge [13]. We share the motivation with posthoc XAI methods [15, 25, 47], which extract explanations from already learned models. The advantage of this approach is that it does not impact the performance of the model, which is treated as a black box. In MESA as well, we generate explanations after the SQL query was executed, independently from the database engine.

7 CONCLUSION AND LIMITATIONS

This paper presented the Correlation-Explanation problem, whose goal is to identify uncontrolled confounding attributes that explain away unexpected correlations observed in query results. We developed an efficient algorithm that finds the optimal subset of attributes. This algorithm is embodied in a system called MESA, which adapts the IPW technique for handling missing data. MESA is applicable for cases where the explanations can be found in a KG.

While we discussed simple SQL aggregate queries, an extension of our work can be made for more general queries (e.g., nested subqueries, other aggregations besides average). In the future, we plan to extend our work by mining explanations from text documents. Another interesting future work is to automatically identify which links in the KG are relevant to the explanation and worthy to follow.

REFERENCES

[1] 2018. The Vanity Fair. https://www.vanityfair.com/hollywood/2018/02/hollywood-movie-salaries-wage-gap-equality.

[2] 2019. The USA Today. https://www.usatoday.com/story/travel/airline-news/2022/06/19/why-us-flights-canceled-delayed-sunday/7677552001/.

[3] 2020. COVID-19 Dataset. https://www.kaggle.com/imdevskp/corona-virus-report?select=usa_county_wise.csv.

[4] 2020. Flights Delay Dataset. https://www.kaggle.com/usdot/flight-delays?select=flights.csv.

[5] 2021. 2021 Stackoverflow Developer Survey. https://insights.stackoverflow.com/survey/2021.

[6] 2021. Forbes Dataset. https://www.kaggle.com/datasets/slayomer/forbes-celebrity-100-since-2005.

[7] 2021. Stack Overfow developer survey. https://insights.stackoverflow.com/survey.

[8] 2022. DBPedia. https://www.dbpedia.org/.

[9] 2022. Pyitlib library. https://pypi.org/project/pyitlib/.

[10] 2022. spaCy Entity Linker. https://spacy.io/api/entitylinker.

[11] 2022. Technical Report. https://github.com/ResultsExplanations/ExplanationsFromKG.

[12] 2022. Wikidata. https://www.wikidata.org/wiki/Wikidata:Main_Page.

[13] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[14] Farahnaz Akrami, Mohammed Samiul Saeef, Qingheng Zhang, Wei Hu, and Chengkai Li. 2020. Realistic re-evaluation of knowledge graph completion methods: An experimental study. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 1995–2010.

[15] Anneleen Van Assche and Hendrik Blockeel. 2007. Seeing the forest through the trees: Learning a comprehensible model from an ensemble. In *European Conference on machine learning*. Springer, 418–429.

[16] Abolfazl Asudeh, Zhongjun Jin, and H. V. Jagadish. 2019. Assessing and Remedying Coverage for a Given Dataset. In *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*. IEEE, 554–565. https://doi.org/10.1109/ICDE.2019.00056

[17] Aline Bessa, Juliana Freire, Tamraparni Dasu, and Divesh Srivastava. 2020. Effective Discovery of Meaningful Outlier Relationships. *ACM Transactions on Data Science* 1, 2 (2020), 1–33.

[18] Nicole Bidoit, Melanie Herschel, and Katerina Tzompanaki. 2014. Query-based why-not provenance with nedexplain. In *Extending database technology (EDBT)*.

[19] Laura E Brown and Ioannis Tsamardinos. 2008. Markov blanket-based variable selection in feature space.

[20] Girish Chandrashekar and Ferat Sahin. 2014. A survey on feature selection methods. *Computers & Electrical Engineering* 40, 1 (2014), 16–28.

[21] Adriane Chapman and HV Jagadish. 2009. Why not?. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. 523–534.

[22] Xiaojun Chen, Shengbin Jia, and Yang Xiang. 2020. A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications* 141 (2020), 112948.

[23] Christopher De Sa, Alex Ratner, Christopher Ré, Jaeho Shin, Feiran Wang, Sen Wu, and Ce Zhang. 2016. Deepdive: Declarative knowledge base construction. *ACM SIGMOD Record* 45, 1 (2016), 60–67.

[24] Chris Ding and Hanchuan Peng. 2005. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology* 3, 02 (2005), 185–205.

[25] Yinpeng Dong, Hang Su, Jun Zhu, and Fan Bao. 2017. Towards interpretable deep neural networks by leveraging adversarial examples. *arXiv preprint arXiv:1708.05493* (2017).

[26] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 0210–0215.

[27] Bradley Efron. 1994. Missing data, imputation, and the bootstrap. *J. Amer. Statist. Assoc.* 89, 426 (1994), 463–475.

[28] Ali El Akadi, Abdeljalil El Ouardighi, and Driss Aboutajdine. 2008. A powerful feature selection approach based on mutual information. *International Journal of Computer Science and Network Security* 8, 4 (2008), 116.

[29] Kareem El Gebaly, Parag Agrawal, Lukasz Golab, Flip Korn, and Divesh Srivastava. 2014. Interpretable and informative explanations of outcomes. *Proceedings of the VLDB Endowment* 8, 1 (2014), 61–72.

[30] Ronald Fagin, Alberto O Mendelzon, and Jeffrey D Ullman. 1982. A simplied universal relation assumption and its properties. *ACM Transactions on Database Systems (TODS)* 7, 3 (1982), 343–360.

[31] François Fleuret. 2004. Fast binary feature selection with conditional mutual information. *Journal of Machine learning research* 5, 9 (2004).

[32] Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of machine learning research* 3, Mar (2003), 1157–1182.

[33] David Hinkley. 1985. Transformation diagnostics for linear models. *Biometrika* 72, 3 (1985), 487–496.

[34] KJ Jager, C Zoccali, A Macleod, and FW Dekker. 2008. Confounding: what it is and how to deal with it. *Kidney international* 73, 3 (2008), 256–260.

[35] A Kaklauskas, V Milevicius, and L Kaklauskiene. 2022. Effects of country success on COVID-19 cumulative cases and excess deaths in 169 countries. *Ecological indicators* (2022), 108703.

[36] Joseph DY Kang and Joseph L Schafer. 2007. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science* 22, 4 (2007), 523–539.

[37] Seokki Lee, Bertram Ludäscher, and Boris Glavic. 2020. Approximate summaries for why and why-not provenance (extended version). *arXiv preprint arXiv:2002.00084* (2020).

[38] David D Lewis. 1992. Feature selection and feature extraction for text categorization. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

[39] Chenjie Li, Zhengjie Miao, Qitian Zeng, Boris Glavic, and Sudeepa Roy. 2021. Putting Things into Context: Rich Explanations for Query Answers using Join Graphs. In *Proceedings of the 2021 International Conference on Management of Data*. 1051–1063.

[40] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. 2017. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)* 50, 6 (2017), 1–45.

[41] Dahua Lin and Xiaoou Tang. 2006. Conditional infomax learning: An integrated framework for feature extraction and fusion. In *European conference on computer vision*. Springer, 68–82.

[42] Yin Lin, Brit Youngmann, Yuval Moskovitch, HV Jagadish, and Tova Milo. 2021. On detecting cherry-picked generalizations. *Proceedings of the VLDB Endowment* 15, 1 (2021), 59–71.

[43] Huan Liu. 2010. *Feature Selection*. Springer US, 402–406.

[44] Brandon Lockhart, Jinglin Peng, Weiyuan Wu, Jiannan Wang, and Eugene Wu. 2021. Explaining Inference Queries with Bayesian Optimization. *Proc. VLDB Endow.* 14, 11 (2021), 2576–2585.

[45] Tania Lombrozo. 2007. Simplicity and probability in causal explanation. *Cognitive psychology* 55, 3 (2007), 232–257.

[46] David Maier, Jeffrey D Ullman, and Moshe Y Vardi. 1984. On the foundations of the universal relation model. *ACM Transactions on Database Systems (TODS)* 9, 2 (1984), 283–308.

[47] David Martens, Bart Baesens, Tony Van Gestel, and Jan Vanthienen. 2007. Comprehensible credit scoring models using rule extraction from support vector machines. *European journal of operational research* 183, 3 (2007), 1466–1476.

[48] Paulo R Martins-Filho. 2021. Relationship between population density and COVID-19 incidence and mortality estimates: A county-level analysis. *Journal of infection and public health* 14, 8 (2021), 1087.

[49] Alexandra Meliou, Wolfgang Gatterbauer, Katherine F Moore, and Dan Suciu. 2009. Why so? or why no? functional causality for explaining query answers. *arXiv preprint arXiv:0912.5340* (2009).

[50] Alexandra Meliou, Wolfgang Gatterbauer, Katherine F Moore, and Dan Suciu. 2010. The complexity of causality and responsibility for query answers and non-answers. *arXiv preprint arXiv:1009.2021* (2010).

[51] Patrick Emmanuel Meyer, Colas Schretter, and Gianluca Bontempi. 2008. Information-theoretic feature selection in microarray data using variable complementarity. *IEEE Journal of Selected Topics in Signal Processing* 2, 3 (2008), 261–274.

[52] Zhengjie Miao, Qitian Zeng, Boris Glavic, and Sudeepa Roy. 2019. Going beyond provenance: Explaining query answers with pattern-based counterbalances. In *Proceedings of the 2019 International Conference on Management of Data*. 485–502.

[53] Tova Milo, Yuval Moskovitch, and Brit Youngmann. 2020. Contribution Maximization in Probabilistic Datalog. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 817–828.

[54] Sameh K Mohamed, Vít Nováček, and Aayah Nounu. 2020. Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics* 36, 2 (2020), 603–610.

[55] Karthika Mohan, Felix Thoemmes, and Judea Pearl. 2018. Estimation with incomplete data: The linear case. In *Proceedings of the International Joint Conferences on Artificial Intelligence Organization*.

[56] Alberto Parravicini, Rhicheek Patra, Davide B Bartolini, and Marco D Santambrogio. 2019. Fast and accurate entity linking via graph embedding. In *Proceedings of the 2nd Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*. 1–9.

[57] R Pascoal and H Rocha. 2022. Population density impact on COVID-19 mortality rate: A multifractal analysis using French data. *Physica A: Statistical Mechanics and its Applications* 593 (2022), 126979.

[58] Patricia A Patrician. 2002. Multiple imputation for missing data. *Research in nursing & health* 25, 1 (2002), 76–84.

[59] Hanchuan Peng, Fuhui Long, and Chris Ding. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence* 27, 8

(2005), 1226–1238.
[60] Mohamad Amin Pourhoseingholi, Ahmad Reza Baghestani, and Mohsen Vahedi. 2012. How to control confounding effects by statistical analysis. *Gastroenterology and hepatology from bed to bench* 5, 2 (2012), 79.
[61] Romila Pradhan, Jiongli Zhu, Boris Glavic, and Babak Salimi. 2021. Interpretable Data-Based Explanations for Fairness Debugging. *arXiv preprint arXiv:2112.09745* (2021).
[62] Thomas Rebele, Fabian Suchanek, Johannes Hoffart, Joanna Biega, Erdal Kuzey, and Gerhard Weikum. 2016. YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames. In *International semantic web conference*. Springer, 177–185.
[63] Sudeepa Roy, Laurel Orr, and Dan Suciu. 2015. Explaining query answers with explanation-ready databases. *Proceedings of the VLDB Endowment* 9, 4 (2015), 348–359.
[64] Sudeepa Roy and Dan Suciu. 2014. A formal approach to finding explanations for database queries. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. 1579–1590.
[65] Babak Salimi, Johannes Gehrke, and Dan Suciu. 2018. Bias in olap queries: Detection, explanation, and removal. In *Proceedings of the 2018 International Conference on Management of Data*. 1021–1035.
[66] Alberto Santos, Ana R Colaço, Annelaura B Nielsen, Lili Niu, Maximilian Strauss, Philipp E Geyer, Fabian Coscia, Nicolai J Wewer Albrechtsen, Filip Mundt, Lars Juhl Jensen, et al. 2022. A knowledge graph to interpret clinical proteomics data. *Nature Biotechnology* (2022), 1–11.
[67] Shaun R Seaman and Ian R White. 2013. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research* 22, 3 (2013), 278–295.
[68] Jainendra Shukla, Miguel Barreda-Angeles, Joan Oliver, GC Nandi, and Domenec Puig. 2019. Feature extraction and selection for emotion recognition from electrodermal activity. *IEEE Transactions on Affective Computing* 12, 4 (2019), 857–869.
[69] Andrea C Skelly, Joseph R Dettori, and Erika D Brodt. 2012. Assessing bias: the importance of considering confounding. *Evidence-based spine-care journal* 3, 01 (2012), 9–12.
[70] Balder ten Cate, Cristina Civili, Evgeny Sherkhonov, and Wang-Chiew Tan. 2015. High-level why-not explanations using ontologies. In *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. 31–43.
[71] Ioannis Tsamardinos, Constantin F Aliferis, Alexander R Statnikov, and Er Statnikov. 2003. Algorithms for large scale Markov blanket discovery.. In *FLAIRS conference*, Vol. 2. St. Augustine, FL, 376–380.
[72] Ashwini Kumar Upadhyay and Shreyanshi Shukla. 2021. Correlation study to identify the factors affecting COVID-19 case fatality rates in India. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 15, 3 (2021), 993–999.
[73] Eugene Wu and Samuel Madden. 2013. Scorpion: Explaining away outliers in aggregate queries. (2013).
[74] Lin Yin, Youngmann Brit, Moskovitch Yuval, Jagadish H. V., and Milo Tova. 2022. On Detecting Cherry-picked Generalizations. *Proceedings of the VLDB Endowment* (2022).
[75] Zhongheng Zhang. 2016. Missing data imputation: focusing on single imputation. *Annals of translational medicine* 4, 1 (2016).
[76] Juan Zhao, Yiwei Zhou, Xiujun Zhang, and Luonan Chen. 2016. Part mutual information for quantifying direct associations in networks. *Proceedings of the National Academy of Sciences* 113, 18 (2016), 5130–5135.
[77] Weiguo Zheng, Jeffrey Xu Yu, Lei Zou, and Hong Cheng. 2018. Question answering over knowledge graphs: question understanding via template decomposition. *Proceedings of the VLDB Endowment* 11, 11 (2018), 1373–1386.
[78] Ganggao Zhu and Carlos A Iglesias. 2018. Exploiting semantic similarity for named entity disambiguation in knowledge graphs. *Expert Systems with Applications* 101 (2018), 8–24.

A MISSING PROOFS

In this part we provide missing proofs.

Proof of Corollary ??. For any two random variables, if $X \perp\!\!\!\perp Y$ we have: $H(Y|X) = H(Y)$. This can be generalized to conditional independence as well. We get:

$$I(O;T|E,R_E=1,C) = H(O|E,R_E=1,C) - H(O|T,E,R_E=1,C) = H(O|E,C) - H(O|T,E,C) = I(O;T|E,C)$$

□

Proof of Corollary ??. We have:

$$I(E_i;E_j|R_{E_i}=1,R_{E_j}=1) = H(E_i|R_{E_i}=1,R_{E_j}=1) - H(E_i|E_j,R_{E_i}=1,R_{E_j}=1) = H(E_i) - H(E_i|E_j) = I(E_i;E_j)$$

□

In what follows, to ease the exposition, we assume that there is no WHERE clause in the query, i.e., $C = \emptyset$. Our results also hold for cases where C is not empty.

Proof of Theorem 4.1. Recall that by definition of the algorithm, we assume that $\boldsymbol{E_{k-1}}$, i.e., the set of $k-1$ attributes, has already been obtained, and thus $\boldsymbol{E_{k-1}}, O$, and T are fixed when selecting the k-th attribute. The goal is to select the optimal k-th attribute to be added, E_k, from $\mathcal{D} \setminus \boldsymbol{E_{k-1}}$.

By the definition of conditional mutual information, we have:

$$I(O;T|\boldsymbol{E_{k-1}},E_k) = I(O;T|\boldsymbol{E_k}) = H(O;\boldsymbol{E_k}) + H(T;\boldsymbol{E_k}) - H(O;T;\boldsymbol{E_k}) - H(\boldsymbol{E_k})$$

We use the following definition of [59] for the attributes $E_1,\ldots,E_k$: $J(\boldsymbol{E_k}) = J(E_1,\ldots,E_k)$ where:

$$J(\boldsymbol{E_k}) = \sum\ldots\sum Pr(E_1,\ldots,E_k)\frac{Pr(E_1,\ldots,E_k)}{Pr(E_1)\cdot\ldots\cdot Pr(E_k)}$$

Similarly, we have:

$$J(O,T,\boldsymbol{E_k}) = \sum\ldots\sum Pr(E_1,\ldots,E_k,O,T)\frac{Pr(E_1,\ldots E_k,O,T)}{Pr(E_1)\ldots Pr(E_k)Pr(O)Pr(T)}$$

$$J(X,\boldsymbol{E_k}) = \sum\ldots\sum Pr(E_1,\ldots E_k,X)\frac{Pr(E_1,\ldots E_k,O)}{Pr(E_1)\cdot\ldots Pr(E_k)Pr(X)}$$

We can derive:

$$H(O;\boldsymbol{E_k}) + H(T;\boldsymbol{E_k}) - H(O;T;\boldsymbol{E_k}) - H(\boldsymbol{E_k}) =$$

$$H(O) + \sum_{i=1}^{k} H(E_i) - J(O,\boldsymbol{E_k}) + H(T) + \sum_{i=1}^{k} H(E_i) - J(T,\boldsymbol{E_k})$$

$$- H(O) - H(T) - \sum_{i=1}^{k} H(E_i) + J(O,T,\boldsymbol{E_k}) - \sum_{i=1}^{k} H(E_i) + J(\boldsymbol{E_k}) = J(O,T,\boldsymbol{E_k}) + J(\boldsymbol{E_k}) - J(O,\boldsymbol{E_k}) - J(T,\boldsymbol{E_k})$$

Thus we consider the following expression:

$$J(O,T,\boldsymbol{E_k}) + J(\boldsymbol{E_k}) - J(O,\boldsymbol{E_k}) - J(T,\boldsymbol{E_k}) \tag{6}$$

We argue that (6) is minimized when the k-th attribute minimizes the Min-CIM and Min-Redundancy criteria.

As stated in [59], the maximum of $J(O,\boldsymbol{E_k})$ is attained when all variables are maximally dependent. When $O, \boldsymbol{E_{k-1}}$ are fixed, this indicates that the attribute E_k should have the maximal dependency to O. In this case, we get that $J(O,T,\boldsymbol{E_k}) = J(T,\boldsymbol{E_k})$. Note that when the dependency of O or T in E_k increases, the conditional mutual information $I(O;T|\boldsymbol{E_k})$ decreases. This is the Min-CIM criterion.

Moreover, as noted in [59], the minimum of $J(\boldsymbol{E_k})$ is attained when the attributes $E_1,\ldots,E_k$ are independent of each other. As all the attributes $E_1,\ldots,E_{k-1}$ are fixed at this point, this pair-wise independence condition means that the mutual information between the attribute E_k and any other attribute E_i is minimized. This is the Min-Redundancy criterion.

Thus, we get that the overall expression in (6) is minimized (i.e., $J(O,\boldsymbol{E_k})$ is maximized, $J(O,\boldsymbol{E_k},T) = J(T,\boldsymbol{E_k})$, and $J(\boldsymbol{E_k})$

is minimized) when we are minimizing the Min-CIM and Min-Redundancy criteria.

□

Proof of Lemma **??**. First, since $(O \perp\!\!\!\perp E_{k+1}|\boldsymbol{E_k})$ we have: $I(O, E_{k+1}|\boldsymbol{E_k}) = 0$. We get:

$I(O;T|\boldsymbol{E_k}) - I(O;T|\boldsymbol{E_k}, E_{k+1}) =$

$H(O|\boldsymbol{E_k}) - H(O|T, \boldsymbol{E_k}) - H(O|\boldsymbol{E_k}, E_{k+1}) + H(O|T, \boldsymbol{E_k}, E_{k+1})$

Since $H(O|\boldsymbol{E_k}) - H(O|E_{k+1}, \boldsymbol{E_k}) = H(O|\boldsymbol{E_k}) - H(O|\boldsymbol{E_k}) = 0$, we get:

$$I(O;T|\boldsymbol{E_k}) - I(O;T|\boldsymbol{E_k}, E_{k+1}) = H(O|T, \boldsymbol{E_k}, E_{k+1}) - H(O|T, \boldsymbol{E_k}) \leq 0$$

For the last inequality we used the fact that for every three ransom variables X, Y, X: $H(X|Y) \leq H(X|Y, Z)$, since adding more conditions can only reduce the uncertainty of X.

We get that the numerator of the responsibility score of E_{k_1} is ≤ 0, and thus $Resp(E_{k+1}) \leq 0$ □

B EXPERIMENTS

Explanation quality. Next, we provide references supporting the explanations generated by MESA. These in-domain findings serve as "domain-expert" explanations.

SO Q1: It was shown in [5] that there is a correlation between the developers salary and countries' economies. In https://www.daxx.com/blog/development-trends/it-salaries-software-developer-trends, it was also shown that the countries with the highest salary for developers are countries with a relatively high HDI (e.g., the US, Switzerland, Denmark).

SO Q2 + Q3: It was mentioned in https://content.techgig.com/career-advice/what-is-the-average-salary-of-software-engineers-in-different-countries/articleshow/91121900.cms that countries that have a scarcity of software graduates tend to offer higher salaries than countries like India which produce hundred of thousands developers every year. This suggests that besides the economy of a country (resp., continent), the population size is also a factor that affects the average salary of developers.

Flights Q1: It was stated in [2] that weather is one of the top reasons for flights delay in the US.

Flights Q2-Q4: It was mentioned in https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations that besides weather conditions, main causes for flights delay in the US are heavy traffic volume, and air traffic control. Those two factors are highly correlated with population size. In bigger and more dense area, the air traffic increases.

Flights Q5: It was mentioned in https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations that a main cause of the delay of flights in the US is the airline's control (e.g., maintenance or crew problems).

Covid Q1: It was shown that there is a correlation between countries' economies and Covid-19 death rate [35, 72].

Covid Q2-Q3: It was stated in [48, 57] that population density impact on COVID-19 mortality rate.

Forbes Q1: It was shown in https://www.theguardian.com/world/2019/sep/15/hollywoods-gender-pay-gap-revealed-male-stars-earn-1m-more-per-film-than-women that there is a gender pay gap for actors in Hollywood. Thus, it make sense that gender is a factor affecting the average salary of actors. It was also stated in https://www.gobankingrates.com/money/jobs/how-much-do-actors-make/ that actors get paid according to their experience, which is reflected in their net worth.

Forbes Q2: It was mentioned in https://climbtheladder.com/producer-salary/ that what affects directors and producers salary is their level of experience (which is reflected in the awards and net worth attributes).

Forbes Q3: It was stated in that very often professional athletes salaries are performance-based. The performance quality is reflected in the Cups and Draft Pick attributes (for tennis, basketball and football athletes, which are the majority of athletes in the Forbes dataset).

We next discuss the effect of the entity linker.

Entity linker. Many of the missing values were caused by an unsuccessful matching of values from the table to their entities in the KG. For example, in SO, for some developers, their origin country is `Russian Federation`. However, the corresponding entity in DBpedia is called `Russia`. We thus failed to extract the properties of this country. In other cases, the values that appear in the tables were ambiguous, and thus we failed to match them to DBpedia entities. For example, in Forbes, one of the athletes is called `Ronaldo`. SpaCy entity linker could not decide whether to link this value to the entity `Ronaldo Luís Nazário de Lima` (Brazilian footballer) or to `Cristiano Ronaldo` (Portuguese footballer).