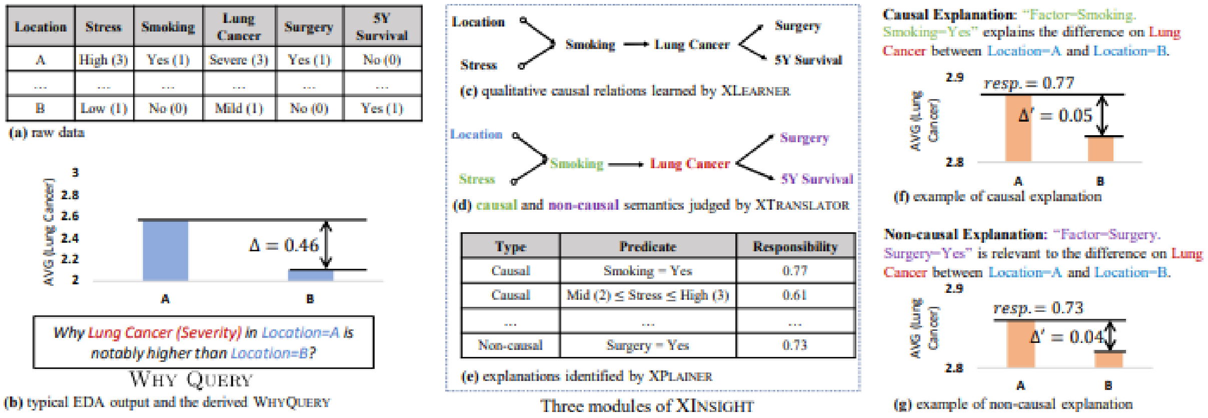
XInsight: eXplainable Data Analysis Through The Lens of Causality

Group 25

Yash Khandelwal and Victor Samsonov

Introduction:

EDA has clearly become one of the most prevalent aspects in tasks heavily focused on analytics. The main goal of EDA is to extract insights and leverage the uncovered knowledge to potentially make an informed decision, however, understanding certain patterns and behaviors within EDA can be challenging. XDA (Explainable Data Analysis) is a new movement within the field which aims to provide explanations of a particular observation that a user may see which can be causal/non-causal and qualitatively/quantitatively explain a target variable at hand, which at a high level is what XInsight aims to do. Xinsight, the topic of the report, is a framework consisting of an XLearner, XTranslator and Xplainer to provide qualitative and quantitative explanations at the predicate level based on a given WHYQUERY (Δ) which as user provides the “pipeline” as one of the first steps. In the following sections we will cover in detail how each step works within XInsight, go over results based on both synthetic data and real-world data, and finally provide both our and the paper’s conclusions.


Location | Stress | Smoking | Lung Cancer | Surgery | 5Y Survival
A | High (3) | Yes (1) | Severe (3) | Yes (1) | No (0)
... | ... | ... | ... | ... | ...
B | Low (1) | No (0) | Mild (1) | No (0) | Yes (1)
3
2.8
2.6
2.4
2.2
2
AVG (Lung Cancer)
Δ = 0.46
A
B
Why Lung Cancer (Severity) in Location=A is notably higher than Location=B?
Why Query
Location
Stress
Smoking
Lung Cancer
Surgery
5Y Survival
Type | Predicate | Responsibility
Causal | Smoking = Yes | 0.77
Causal | Mid (2) ≤ Stress ≤ High (3) | 0.61
... | ... | ...
Non-causal | Surgery = Yes | 0.73
Causal Explanation: "Factor=Smoking. Smoking=Yes" explains the difference on Lung Cancer between Location=A and Location=B.
2.9
2.8
AVG (Lung Cancer)
resp. = 0.77
Δ' = 0.05
A
B
Non-causal Explanation: "Factor=Surgery. Surgery=Yes" is relevant to the difference on Lung Cancer between Location=A and Location=B.
2.9
2.8
AVG (Lung Cancer)
resp. = 0.73
Δ' = 0.04
A
B

(a) raw data

(b) typical EDA output and the derived WHYQUERY

(c) qualitative causal relations learned by XLEARNER

(d) causal and non-causal semantics judged by XTRANSLATOR

(e) explanations identified by XPLAINER

Three modules of XINSIGHT

(f) example of causal explanation

(g) example of non-causal explanation

Fig. 1. Illustrative example of XINSIGHT.

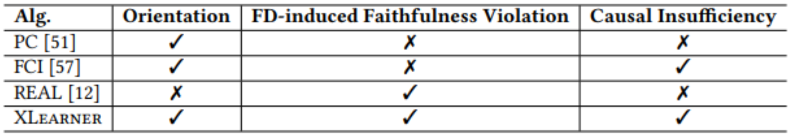
How XInsight Works:

XInsight as previously mentioned can be divided into three main steps, XLearner (leverages a causal graph, causal sufficiency, and FCI to determine a Partial ancestral Graph (PAG), XTranslator (uses the PAG generated by XLearner to define causal and non-causal explanations while excluding the context and target), and XPlainer (taking into account similarities of DB Causality mapping the problem to a predicate level instead of a tuple/cause level, is able to generate explanations in the format of: <type, predicate, responsibility>).

XLearner

Xlearner handles the first step of the framework and even has it's own offline phase, which consists in extracting/learning the relevant PAG (compute intensive) alleviating cost, which allows the user to simply specify a WhyQuery and almost immediately move to the XTranslator step (compute intensive tasks are done first in the offline stage and the online stage uses "lighter" computations). In past work, one of the key concepts used in a constraint-based Causal Discovery such as the FCI algorithm (XLearner builds upon/enhances the capabilities of this algorithm which will be discussed) while leveraging the Global Markov Property and the Faithfulness assumption. Both definitions allow us to interpret a data distribution to a causal graph representation and vice versa, however, the problem of earlier work done within causal discovery include the issue of Faithfulness violations because of Functional Dependencies (FD), which is a problem that XInsight tackles. Additionally, XInsight (XLearner in this case) aims to address the problem of causal insufficiency which takes place when a latent variable isn't present in the data and as a result can introduce an incorrect graph under certain conditions, previous literature either addresses FD-induced Faithfulness Violations or Causal Insufficiency, but never addresses both, something that the model at hand aims to do.

Table 2. Comparing different causal discovery algorithms. ✓ denotes "support" whereas ✗ denotes "no support".



Alg.	Orientation	FD-induced Faithfulness Violation	Causal Insufficiency
PC [51]	✓	✗	✗
FCI [57]	✓	✗	✓
REAL [12]	✗	✓	✗
XLearner	✓	✓	✓

XLearner can be decomposed into 3 main steps, which include obtaining the FD-induced Graph, obtaining a harmonious skeleton while leveraging ***Theorem 3.1*** and finally output a FD-augmented PAG.

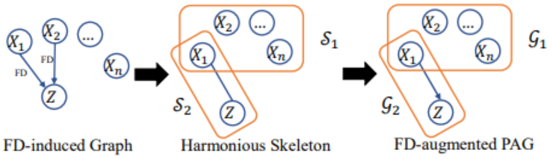
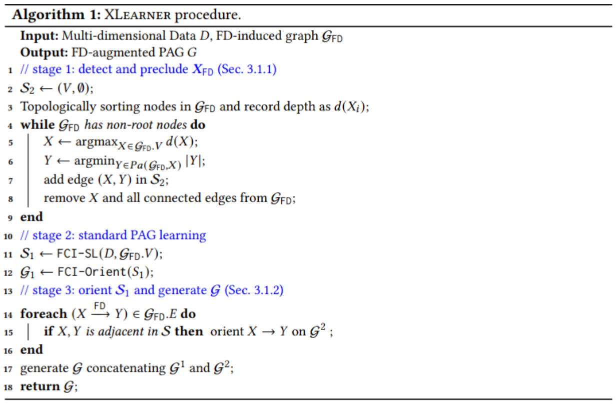


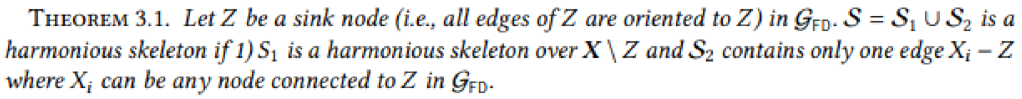
Fig. 5. Running example of XLearner.

Below we will include the algorithm and discuss at a high level what it is able to achieve and how it differentiates from a Vanilla FCI.



```
Algorithm 1: XLearner procedure.
Input: Multi-dimensional Data $D$, FD-induced graph $\mathcal{G}_{FD}$
Output: FD-augmented PAG $G$
// stage 1: detect and preclude $X_{FD}$ (Sec. 3.1.1)
$\mathcal{S}_2 \leftarrow (V, \emptyset)$;
Topologically sorting nodes in $\mathcal{G}_{FD}$ and record depth as $d(X_i)$;
while $\mathcal{G}_{FD}$ has non-root nodes do
    $X \leftarrow \text{argmax}_{X \in \mathcal{G}_{FD}.V}\, d(X)$;
    $Y \leftarrow \text{argmin}_{Y \in Pa(\mathcal{G}_{FD}, X)}\, |Y|$;
    add edge $(X, Y)$ in $\mathcal{S}_2$;
    remove $X$ and all connected edges from $\mathcal{G}_{FD}$;
end
// stage 2: standard PAG learning
$\mathcal{S}_1 \leftarrow \texttt{FCI-SL}(D, \mathcal{G}_{FD}.V)$;
$\mathcal{G}_1 \leftarrow \texttt{FCI-Orient}(\mathcal{S}_1)$;
// stage 3: orient $\mathcal{S}_1$ and generate $\mathcal{G}$ (Sec. 3.1.2)
foreach $(X \xrightarrow{FD} Y) \in \mathcal{G}_{FD}.E$ do
    if $X, Y$ is adjacent in $\mathcal{S}$ then orient $X \rightarrow Y$ on $\mathcal{G}^2$ ;
end
generate $\mathcal{G}$ concatenating $\mathcal{G}^1$ and $\mathcal{G}^2$;
return $\mathcal{G}$;
```

When comparing to FCI, XLearner is able to addresses issues related to FD and the faithfulness assumption for causally insufficient data with a harmonious skeleton framework, learning real world data better. Additionally, this approach provides us a causally sufficient environment, which includes the hypothesis that a FD provide a more complete orientation for an underlying causal graph with less undetermined edges, i.e. XInsight enhances the FCI algorithm. Lines 1-9 in the algorithm leverage Theorem 3.1 which essentially states that if Z is a sink node, and an edge X_i-Z forms a skeleton (S2), if there exists a harmonious skeleton over **X,** then the union of both skeletons is also considered a harmonious skeleton. Lines 1-9 essentially construct a harmonious skeleton recursively applying theorem 3.1 to connect while applying a prior topological sord over the FD-induced Graph.



THEOREM 3.1. *Let Z be a sink node (i.e., all edges of Z are oriented to Z) in $\mathcal{G}_{\text{FD}}$. $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ is a harmonious skeleton if 1) $\mathcal{S}_1$ is a harmonious skeleton over $\mathbf{X} \setminus Z$ and $\mathcal{S}_2$ contains only one edge $X_i - Z$ where X_i can be any node connected to Z in $\mathcal{G}_{\text{FD}}$.*

Subsequent lines in the algorithm (excluding 10-12 which simply infer a PAG on our initial graph) leverage FD as a tool to reflect a causal relationship instead of orienting edges by simply using conditional independence and graphical structural relationships as the main criteria. The method proposed within the literature is ANM (Additive Noise Model) which is able to orient FD-related edges. The main idea encountered is that if an asymmetric AN $Y = f(x) + N_y$ exists from X to Y and the noise factor is independent of X, then X causes Y. The researchers of our paper use this idea to hypothesize that if $N_y\ X \rightarrow_{FD} Y\ in\ G_{FD}\ implies\ causation\ of\ X \rightarrow Y$. While the FCI approach is still good, it isn't suitable for FD-related edges due to faithfulness violations. A functional dependency is a lot more reliable given that it tends to describe a deterministic relation and the results from a FD is compatible with the result of the FCI, implying the GMP isn't violated. The researchers conclude that the implementation of ANM results in an augmented graph that is more informative and represents an overcomplete graph w.r.t. the ground-truth MAG's Markov equivalence class, i.e. better precision than previous approaches.

XTranslator

A causal graph extracted from the previous step doesn't reveal if a given variable explains a WHYQUERY and it also doesn't explain if an explanation is causal or non-causal. Causal primitives need to be translated into XDA semantics. XTranslator in order to make the process more efficient leverages the principle of Explainability if our given WHYQUERY has an aggregator of type AVG to determine if a variable X has explainability, which it does when a measure M isn't m-separated by a foreground variable F in a causal graph. If X and M are independent given a foreground variable then providing an explanation over our target measure simply isn't feasible. Essentially if a variable has explainability (determined by m-separation), it can contribute to answering the WHYQUERY, otherwise it can't do so. XTranslator determines if X provides a causal explanation based on if it has explainability and the structure of the possible path (if X is a parent/ancestor then it is a causal explanation, otherwise it is non-causal), and leveraging the concept allows us the chance to prune any variables that cannot provide us with a relevant explanation.

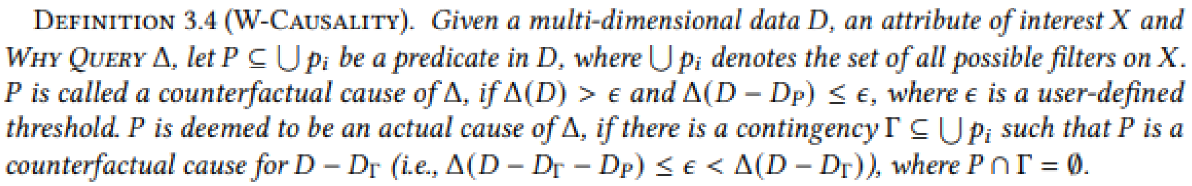
This formulation can be extended to a SUM aggregator. If X has no explainability when X = x, $\Delta(D_{X=x})$ is enforced it can merely be affected by the number of rows, implying that a count-based explanation is provided which could result in inconsistencies in terms of how data analysis is interpreted.

Overall, for the XTranslator step, we can conclude that a given variable can have a higher/lower impact in terms of how strongly connected it is with respect to our target measure M, the goal of this step is to focus on the variables that are the most likely to provide more desirable explanations, which along with the principle of explainability we are able to prune the uninformative variables form aggregates.

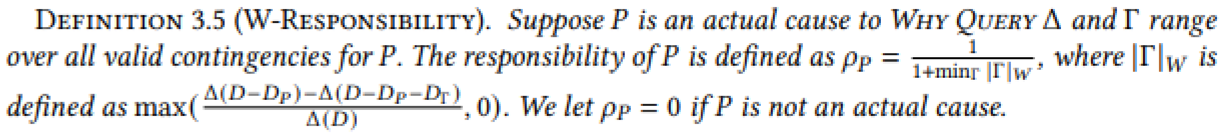
XPlainer

XLearner and XTranslator provide an ideal foundation for variable-level explanations to a WHYQUERY. This current step closely follows the main ideas of DB causality, which leverages a cause t in over a set of contingencies in order to determine if t is an actual cause to our WHYQUERY, however, we are not interested in tuples which is how the Data Provenance DB causality approach works, we are interested in generating predicate level explanations, but concepts from this framework are leveraged since it has useful normalized metrics such as responsibility -> (0, 1] and uses minimal contingency to reflect which are the additional influential factors when it comes to determining the responsibility of the outcome.

Mapping this problem to provide predicate level explanations, we leverage W-Causality and W-Responsibility which are defined below:



Definition 3.4 (W-Causality). *Given a multi-dimensional data D, an attribute of interest X and* Why Query Δ, *let* $P \subseteq \bigcup p_i$ *be a predicate in D, where* $\bigcup p_i$ *denotes the set of all possible filters on* X. P *is called a counterfactual cause of* Δ, *if* $\Delta(D) > \epsilon$ *and* $\Delta(D - D_P) \leq \epsilon$, *where* ϵ *is a user-defined threshold. P is deemed to be an actual cause of* Δ, *if there is a contingency* $\Gamma \subseteq \bigcup p_i$ *such that P is a counterfactual cause for* $D - D_\Gamma$ *(i.e.,* $\Delta(D - D_\Gamma - D_P) \leq \epsilon < \Delta(D - D_\Gamma)$*), where* $P \cap \Gamma = \emptyset$.



Definition 3.5 (W-Responsibility). *Suppose P is an actual cause to* Why Query Δ *and* Γ *range over all valid contingencies for P. The responsibility of P is defined as* $\rho_P = \frac{1}{1+\min_\Gamma |\Gamma|_W}$, *where* $|\Gamma|_W$ *is defined as* $\max(\frac{\Delta(D-D_P)-\Delta(D-D_P-D_\Gamma)}{\Delta(D)}, 0)$. *We let* $\rho_P = 0$ *if P is not an actual cause.*

As previously implied, the ideas are similar to what is encountered within DB Causality however, we generalize to a predicate level instead of using tuples. Furthermore, using responsibility as a unique criteria is not sufficient in data analysis, therefore the optimal explanation is formulated by the following:

$$\underset{P \subseteq \bigcup p_i}{\operatorname{argmax}} \rho_P - \sigma|P| \qquad (4)$$

Where sigma is > 0 and an ideal value would be 1/m where m is the number o filters selected within our predicate.

Multiple research papers have pointed out that the computation of responsibility is intractable and solving the optimization problem in the equation above is complex. The brute force approach while being optimal is quite expensive, XInsight proposes to approximate based on different aggregates with different compromises at hand:

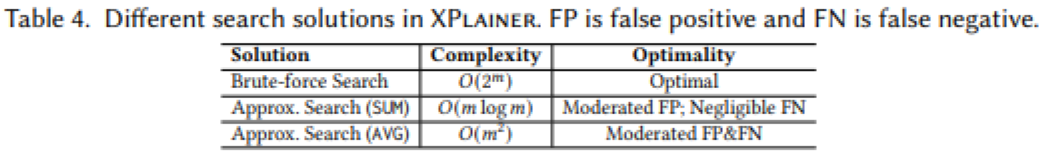
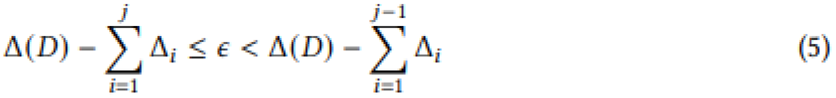


Table 4. Different search solutions in XPlainer. FP is false positive and FN is false negative.

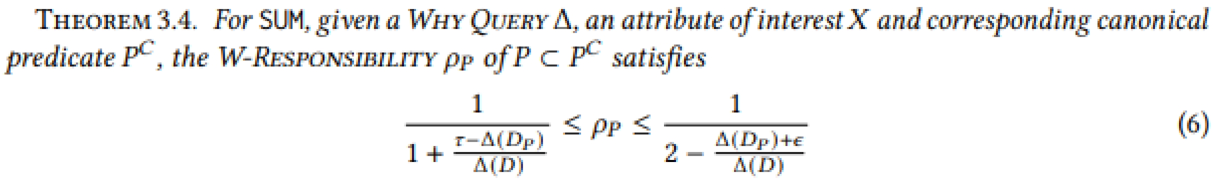
Solution	Complexity	Optimality
Brute-force Search	$O(2^m)$	Optimal
Approx. Search (SUM)	$O(m \log m)$	Moderated FP; Negligible FN
Approx. Search (AVG)	$O(m^2)$	Moderated FP&FN

The reason why the different aggregations result in lowered complexity, is because of certain assumptions. When using a **SUM** aggregate we can leverage the additive property to do the following: $\Delta(D_{P_1} + D_{P_2}) = \Delta(D_{P_1}) + \Delta(D_{P_2})$ which allows us to prune the search space. The researchers are able to extrapolate this information and determine that search algorithms can omit filters with a non-positive Δ_i. Since equation 4 requires an optimal explanation we only focus on the queries with the highest deltas without using optimality. Canonical fitlers are introduced in order to guarantee completeness through a minimal counterfactual cause entailed by the following equation:



$$\Delta(D) - \sum_{i=1}^{j} \Delta_i \leq \epsilon < \Delta(D) - \sum_{i=1}^{j-1} \Delta_i \qquad (5)$$

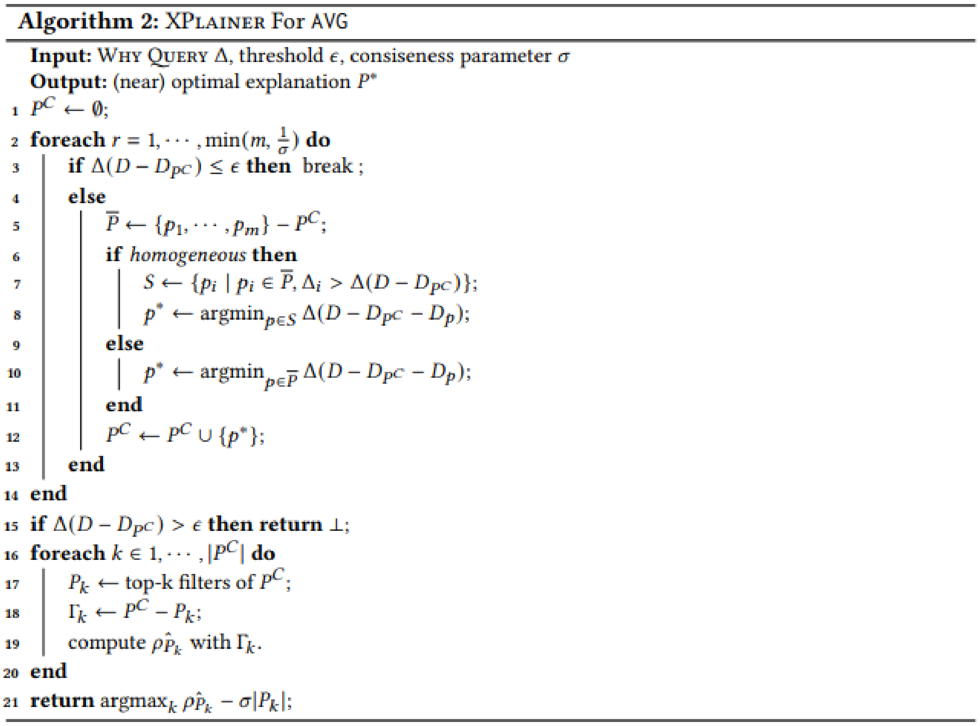
Completeness states that when leveraging a SUM aggregate, given a WHYQUERY Q, an attribute of interest X and it's corresponding canonical predicate, implies the existence of an optimal explanation which is a subset of the canonical predicate, which is a proposition that allows us to focus on canonical filters only when searching for the optimal explanation. The researchers conclude that confirming valid explanations can be done without exhaustive enumerations, and the why-responsibility can be rewritten as the following theorem where tau is a hyperparameter, implying that responsibility can be efficiently approximated using theoretical guarantees:



Theorem 3.4. *For* SUM, *given a* Why Query Δ, *an attribute of interest* X *and corresponding canonical predicate* P^C, *the* W-Responsibility ρ_P *of* $P \subset P^C$ *satisfies*

$$\frac{1}{1 + \frac{\tau - \Delta(D_P)}{\Delta(D)}} \leq \rho_P \leq \frac{1}{2 - \frac{\Delta(D_P) + \epsilon}{\Delta(D)}} \qquad (6)$$

When using AVG aggregates, optimizations are more challenging due to the absence of the additive property, however, there still exists pruning to be done. Leveraging the homogenous sibling subspace property, we can rely on a greedy-based heuristics with a pruning strategy from one of the propositions which implies that for a homogenous AVG with a WHY QUERY Q, an attribute of interest and a subset of the union of all the filters along with a filter within said subset if a WHYQUERT over a single filter is greater than a WHYQUERY over the subset of filters, then we know that the difference between both is less than the WHYQUERY over the subset of filters. This is leveraged in the following algorithm:



```
Algorithm 2: XPlainer For AVG
Input: Why Query $\Delta$, threshold $\epsilon$, consiseness parameter $\sigma$
Output: (near) optimal explanation $P^*$
$P^C \leftarrow \emptyset$;
foreach $r = 1, \cdots, \min(m, \frac{1}{\sigma})$ do
    if $\Delta(D - D_{P^C}) \leq \epsilon$ then break ;
    else
        $\overline{P} \leftarrow \{p_1, \cdots, p_m\} - P^C$;
        if homogeneous then
            $S \leftarrow \{p_i \mid p_i \in \overline{P}, \Delta_i > \Delta(D - D_{P^C})\}$;
            $p^* \leftarrow \mathrm{argmin}_{p \in S}\ \Delta(D - D_{P^C} - D_p)$;
        else
            $p^* \leftarrow \mathrm{argmin}_{p \in \overline{P}}\ \Delta(D - D_{P^C} - D_p)$;
        end
        $P^C \leftarrow P^C \cup \{p^*\}$;
    end
end
if $\Delta(D - D_{P^C}) > \epsilon$ then return $\perp$;
foreach $k \in 1, \cdots, |P^C|$ do
    $P_k \leftarrow$ top-k filters of $P^C$;
    $\Gamma_k \leftarrow P^C - P_k$;
    compute $\rho\hat{P}_k$ with $\Gamma_k$.
end
return $\mathrm{argmax}_k\ \rho\hat{P}_k - \sigma|P_k|$;
```

The idea behind the algorithm is similar to SUM since a connonical predicate is being constructed, such that said predicate forms a counterfactual clause. One of the differences to highlight is that SUM does ensure the optimality of the resulting explanation due to incompleteness of the canonical predicate. The paper also points out that the strategy is greedy-based for constructing the canonical predicate progressively and prunes if proposition 3.4 is met while iteratively adding the necessary filters.

NOTE:

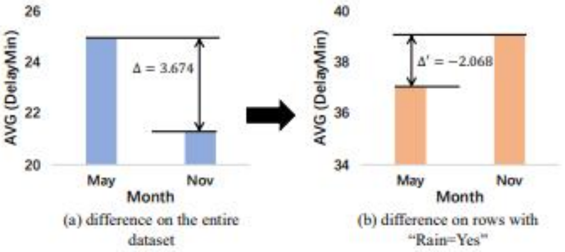
XInsight considers only deterministic FDs, stochasticity is not considered in the framework.

XINSIGHT EVALUATION

The researchers formulated 3 research questions to answer: how can XInsight facilitate end users in explainable data analysis (End-To-End Performance), does XLearner effectively recover causal relations from observational data (XLearner evaluation), does XPlainer accurately and efficiently yield explanations (Xplainer Evaluation)? Additionally some of the datasets used by the researchers are the following: Hotel Booking dataset, Web service behavior data set, synthetic data A, and synthetic data B.

End-To-End Performance

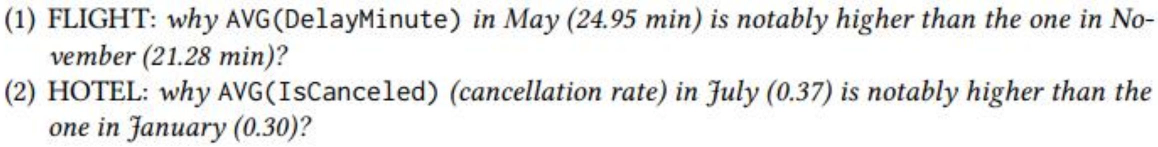
The researchers asked a WHYQUERY on the Flight and Hotel datasets and allowed users to rate how well it was answered.



(a) difference on the entire dataset

(b) difference on rows with "Rain=Yes"

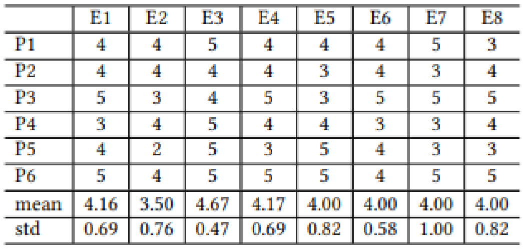
Fig. 6. Explanation of WHY QUERY on the FLIGHT dataset.



(1) FLIGHT: *why* AVG(DelayMinute) *in May (24.95 min) is notably higher than the one in November (21.28 min)?*

(2) HOTEL: *why* AVG(IsCanceled) *(cancellation rate) in July (0.37) is notably higher than the one in January (0.30)?*

Table 5. Results of explanation assessment. E*i* and P*i* stand for the *i*th explanation and the *i*th participant, respectively.



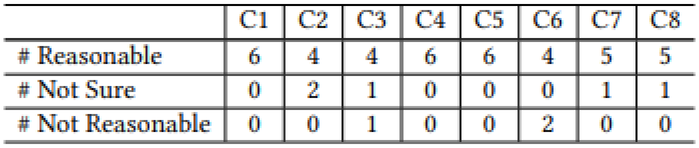
	E1	E2	E3	E4	E5	E6	E7	E8
P1	4	4	5	4	4	4	5	3
P2	4	4	4	4	3	4	3	4
P3	5	3	4	5	3	5	5	5
P4	3	4	5	4	4	3	3	4
P5	4	2	5	3	5	4	3	3
P6	5	4	5	5	5	4	5	5
mean	4.16	3.50	4.67	4.17	4.00	4.00	4.00	4.00
std	0.69	0.76	0.47	0.69	0.82	0.58	1.00	0.82

It is noted in the paper that some participants later on changed their opinion for how well the explanation was done after further discussing with participants with lower ratings, they concluded that the participants misunderstood some of the outcomes.

XLEARNER EVALUATION

The researchers when adopting a FD-based approach contrary to FCI, had considerable interested to evaluate how the performance would differ. The results indicate an increase for F1 and recall and that as the number FD's grows, the performance of XInsight gradually increases compared to FCI. XLearner experiments are analyzed based on the web dataset and they presented 40 participants causal claims that were extracted said dataset, resulting in 6.3% of participants determining the claims not to be reasonable and 83.3% of the participants determined the claims to be reasonable.

Table 7. User study. Ci stands for the ith causal claim.



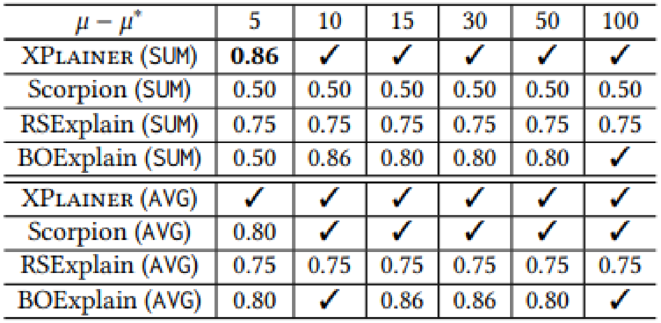
	C1	C2	C3	C4	C5	C6	C7	C8
# Reasonable	6	4	4	6	6	4	5	5
# Not Sure	0	2	1	0	0	0	1	1
# Not Reasonable	0	0	1	0	0	2	0	0

The researchers after discussing with some of the participants allowed one of them to change their opinion to "Reasonable" after understanding that the initial assumption they made wasn't necessarily correct. The conclusion is that XLearner generates plausible graphs that are consistent with expert knowledge.

XPLAINER EVALUATION

The paper after performing experiments is able to conclude that XPlainer is highly efficient particularly for high cardinality regimes where both Scorpion and RSExplain run out of time when the cardinality exceeds 30. Another conclusion that is extracted is that XPlainer is more robust than previous solutions given $\mu^* - \mu = 5$ (subtle data differences) while other methods have difficulties, the contrast is especially clear while using an AVG aggregate.

Table 9. XPLAINER and baselines with different $\mu^* - \mu$. ✓ denotes that the result is identical to the ground truth (F1=1.0).



$\mu - \mu^*$	5	10	15	30	50	100
XPLAINER (SUM)	**0.86**	✓	✓	✓	✓	✓
Scorpion (SUM)	0.50	0.50	0.50	0.50	0.50	0.50
RSExplain (SUM)	0.75	0.75	0.75	0.75	0.75	0.75
BOExplain (SUM)	0.50	0.86	0.80	0.80	0.80	✓
XPLAINER (AVG)	✓	✓	✓	✓	✓	✓
Scorpion (AVG)	0.80	✓	✓	✓	✓	✓
RSExplain (AVG)	0.75	0.75	0.75	0.75	0.75	0.75
BOExplain (AVG)	0.80	✓	0.86	0.86	0.80	✓

Overall after evaluation XPlainer, the researchers concluded that this step shows scalability to very large datasets and also accurately generates explanations in difficult settings. The tradeoff compared to other solutions results in slightly less precision while having a large gain in Recall.

Conclusion:

XInsight is a SOTA XDA framework which allows the user to obtain quantitative and qualitative explanations while determining if it's causal or non-causal. From an XLearner standpoint, It enhances the capabilities of the FCI algorithm with assumptions made leveraging the harmonious skeleton to output a FD-augmented PAG, instead of either focusing on FD-induced faithfulness violations or Causal insufficiency, both problem are addressed at the same time. When it comes to comparing XPlainer the performance is not only superior but much faster allowing to deal with large datasets, which is something that existing tools in the literature do not possess. Future research remains when it comes to designing translation rules which seem to have been one of the leading factors for a considerable jump in performance compared to previous existing methods. Overall the results made involving human feedback and different metrics conclude that this XDA approach has a lot of potential.

We think that it will be interesting to follow the field of XDA along with XAI, they are both related and possible advancements in one of these topics might result in improvements being made in both. It is especially important that advancements are made to XAI in order to increase the transparency of DL architectures, especially in the context of transformers which at a high level are easier to understand but diving into lower level explanations seems to become exponentially taxing and complicated.