Outcome-Explorer: A Causality Guided Interactive Visual Interface for Interpretable Algorithmic Decision Making

ANONYMOUS AUTHOR(S)

The widespread adoption of algorithmic decision-making systems has brought about the necessity to interpret the reasoning behind these decisions. The majority of these systems are complex black box models, and auxiliary models are often used to approximate and then explain their behavior. However, recent research suggests that such explanations are not overly accessible to non-expert users and can lead to incorrect interpretation of the underlying model. In this paper, we show that a predictive and interactive model based on causality is inherently interpretable, does not require any auxiliary model, and allows both expert and non-expert users to understand the model comprehensively. To demonstrate our method we developed Outcome Explorer, a causality guided interactive interface, and evaluated it by conducting think-aloud sessions with three expert users and a user study with 18 non-expert users. All three expert users found our tool to be comprehensive in supporting their explanation needs while the non-expert users were able to understand the inner workings of the model easily.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Anonymous Author(s). 2018. Outcome-Explorer: A Causality Guided Interactive Visual Interface for Interpretable Algorithmic Decision Making. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In recent years, algorithmic and automated decision-making systems have been deployed in several critical application areas across society [6, 9, 35]. However, researchers and organizations have raised ethical concerns about the discriminatory effects that can be associated with these systems. For example, COMPAS [6], a recidivism risk assessment system used in several courtrooms in the United States, has been found to have discriminatory effects on the African-American population [3]. The opaqueness of algorithmic decision-making systems has subsequently led to the General Data Protection Regulation (GDPR), which mandates that the data controller should provide sufficient information to a data subject and do so in a *concise, transparent, intelligible, and easily accessible form* [43]. There thus exists a genuine need for human-interpretable machine-generated decisions.

To address this need, several algorithmic methods [30, 38] and visual analytics systems have been proposed [1, 2, 19–21, 23]. Such systems and methods are shown to be effective in explaining the decision-mechanism of a wide variety of models, but they are often limited to only support expert users such as machine learning researchers or practitioners [8].

They require a degree of mathematical, machine learning, and visualization literacy that is hard to expect from people who are not algorithmic experts [8].

Supporting both expert and non-expert users in a single explainable AI (XAI) platform is challenging since the two user groups have different goals, reasons, and skillsets for interpreting a machine learning model. An expert user such as a machine learning researcher or practitioner will want to interpret a model to ensure its accuracy and fairness. A non-expert user on the other hand will want to interpret the model to gain trust and understand the service the model facilitates and delivers. Supporting both of these user groups requires a model with a sound algorithmic foundation as well as an easy-to-understand visual representation.

To exhibit human-friendly interpretability, a predictive model also needs to produce answers to explanation queries such as "Why does this model make such decisions?", "What if I change a particular input feature?" or "How will my action change the decision?" [34]. Existing methods and systems often employ an auxiliary model (post-hoc) to answer such questions since black-box models do not readily provide these explanations [30, 33, 38]. However, recent research suggests that such approximations can lead to incorrect interpretations of the underlying model and further complicate model understanding [27, 39]. Thus, an XAI interface for non-expert users should be inherently interpretable and be able to answer causal questions without the need for an auxiliary model.

In this paper, we show that a predictive model based on causality (i.e., a causal model) meets the aforementioned criteria for a model that is understandable by non-expert users. A causal model is formulated as a Directed Acyclic Graph (DAG) which is an intuitive representation. Based on the causal DAG, a user can then gain a good understanding of how variables are related to each other and how they affect the outcome variable, without the need for any mathematical or algorithmic expertise. As such, causal models are inherently interpretable and can provide exact and truthful answers to causal explanation queries without the need for any auxiliary model.

To support our proposed method, we first developed a computational pipeline that would facilitate predictions in a causal model. This was needed since causal models do not support predictions by default. Informed by prior research and a formative study with ten non-expert users, we designed and developed Outcome-Explorer, our causality-guided interactive decision-making platform. Outcome-Explorer lets both expert and non-expert users interact with the causal DAG and supports common XAI functionalities such as answering What-If questions, exploring nearest neighbors, and comparing data instances. To evaluate the effectiveness of Outcome-Explorer, we first invited three expert users (machine learning researchers) to develop a causal model and then interpret the model using our tool. All three expert users found Outcome-Explorer to be comprehensive in terms of causal and explanation functionalities.

We then conducted a user study with 18 participants who did not possess any algorithmic expertise to understand how Outcome-Explorer would help non-expert users in interpreting a predictive model. Although Outcome-Explorer did not improve the users' overall performance in decision-making tasks, further investigation revealed that participants were able to reduce interactions with variables that did not affect the outcome by 47%, meaning that participants understood which variables to change while using our tool. A similar reduction (36%) was also found for the magnitude of changes made to non-impacting variables. These outcomes confirm the potential of Outcome-Explorer in both XAI research and application.

In summary, we make the following research contributions-

- A mechanism to facilitate prediction in a causal model based on an established statistical causality framework.
- The design and implementation of Outcome-Explorer, a causality-guided interactive visual tool to support the explanation needs of both expert and non-expert users.

- Results from think-aloud sessions with three expert users. We found that expert users were able to build and interpret the causal model correctly.
- A user study with 18 non-expert participants to evaluate the effectiveness of Outcome-Explorer. We found that our tool helped users understand a model without any algorithmic knowledge and that it helped them identify important input features for prediction.

2 RELATED WORK

Current tools and methods in XAI research can be broadly categorized into two distinct categories: (1) post-hoc explanation, and (2) explanation via interpretability. Post-hoc explanation methods explain the prediction of a model using local estimation techniques, without showing or explaining the workflow of the model. Instead, they give very detailed information on the impact of the model variables at a user-chosen data instance and seek to build trust into the model one instance at a time. Prominent examples of post-hoc explanation methods are SHAP [30] and LIME [38]. They are model agnostic and can explain the predictions of any machine learning model. While post-hoc methods are shown to be effective for explaining the decisions of complex black-box models, recent research has argued that these explanations can be misleading, and can complicate model understanding even more [39].

Given these shortcomings. there has been a growing interest in devising inherently interpretable models. The Generalized Additive Model (GAM) [18] and the Bayesian Case Model (BCM) [26] are examples of this kind of model where the workflow is directly observable by a human user. In this paper, we advocate for a causal model. Unlike other models, causal models are highly interpretable for two reasons: (1) they use cause and effect relations which are easy to mentally decode and understand [32], and (2) they can be readily visualized as a graph or network [45].

2.1 Causal Models

Causal models directly answer to Rudin's [39] core message to use models that are interpretable inherently (rather than post-hoc). They have found frequent use to support reasoning about different forms of bias and discrimination [14]. A first application was presented in [25] but mitigation strategies were reduced to simply removing edges and variables and with limited explanatory power, making it difficult to gauge human sensitivity to certain levels of bias. Other work on causality-based algorithmic bias understanding has emerged [24, 28, 29, 31, 47–49] but none uses interactive visualization within the model as an explainable AI paradigm.

2.2 Visual Analytics and Interactive XAI

People are more likely to understand a system if they can tinker and interact with it [10]. In that spirit, interactive visual systems have proven to be an effective way to augment a human's understanding of a given automated decision-making paradigm. These interactive systems broadly categorize into the aforementioned post-hoc explanation and interpretable interfaces. Examples of interactive post-hoc explanation interfaces include the What-if Tool [46], RuleMatrix [33], Vice [16], Model Tracker [2], Prospector [13], and others. These tools use scatter plots, line plots, and text interfaces to allow users to query and compare the outcomes of different decision models, but without showing the model itself. Though this helps to understand the model's behavior in a counterfactual sense (the 'if'), it does not explain, and allow a user to play with the reasoning flow within the system (the 'why').

On the other hand, interpretable interfaces such as GAMUT [19], and SUMMIT [21] allow a user to interact with the model itself and provide explanations that are faithful to the model. Our tool, Outcome-Explorer, falls into the latter category. It contributes to visual analytics research as an easy-to-understand interactive visual tool guided by statistical causality. It is a decision-making tool that combines the predictive and explanatory components of an XAI system into a single interface. Thus, there is no explanation paradigm in our tool; rather the predictive platform simultaneously works as a predictive and explanatory platform.

2.3 Users

Irrespective of whether an XAI system is post-hoc explainable or interpretable, there are recent critiques that XAI systems tend to support the explanation needs of the model-builders, not the actual users who are the recipients of the decisions [1, 8]. It is not clear from the empirical evidence whether these XAI systems are understandable and usable by non-expert users [1, 11]. One significant advantage of the causal model is that it is explainable through a self-explanatory path diagram, which has the potential to support the explanation needs of non-expert users. Our work advances research toward designing the XAI interface for non-experts and provides critical findings through a user study with 18 non-expert users.

3 BACKGROUND

We follow Pearl's Structural Causal Model (SCM) [36, 37] to define causal relationship between variables. According to SCM, causal relations between variables are expressed in a Directed Acyclic Graph (DAG), also known as a Path Diagram. In a path diagram, variables are categorized as either exogenous (U) or endogenous (V). Exogenous variables have no parents in a path diagram and are considered to be independent and unexplained by the model. On the other hand, endogenous variables are fully explained by the model and presented as the causal effects of the exogenous variables. Figure 1 presents two exogenous variables (A, and B) and three endogenous variables (C, D, and E). Formally, the Causal Model is a set of triples (U, V, F) such that

- U is the set of exogenous variables, and V is the set of endogenous variables.
- Structural equations [4] (F) is a set of functions $\{f_1, ..., f_n\}$, one for each $V_i \subseteq V$, such that $V_i = f_i(pa_i, U_{pa_i})$, $pa_i \subseteq V \setminus \{V_i\}$ and $U_{pa_i} \subseteq U$.

The notation "pa_i" refers to the "parents" of V_i.

3.1 Causal Structure Search

The causal structure between variables (F) can be obtained in three different ways: (1) causal structure defined from domain-expertise or prior knowledge; (2) causal structure learned from automated algorithms; and (3) causal structure learned from human-AI collaboration.

The first method is the prevalent way to operate causal analysis in domains such as social science, or medical science where expert-users or researchers are solely responsible for defining the causal structure [22]. In such scenarios, researchers utilize prior knowledge, domain expertise, and empirical evidence gathered from experiments such as randomized trials to hypothesized causal relations and then test the validity of the model through Structural Equation Modelling (SEM). Software such as "IBM SPSS AMOS", and "Lavaan" are build upon this principle.

On the other hand (in the second approach), automated causal search algorithms utilize conditional independence tests to find causal structure among data [15]. These algorithms help the user identify underlying causal structures between a large set of variables. The "Tetrad" software provides a comprehensive list of such algorithms.

The third approach is the combination of the first two approaches. One pitfall of the automated algorithms is that it may not find the true causal structure since multiple causal structures can meet the constraints set out by the algorithms.

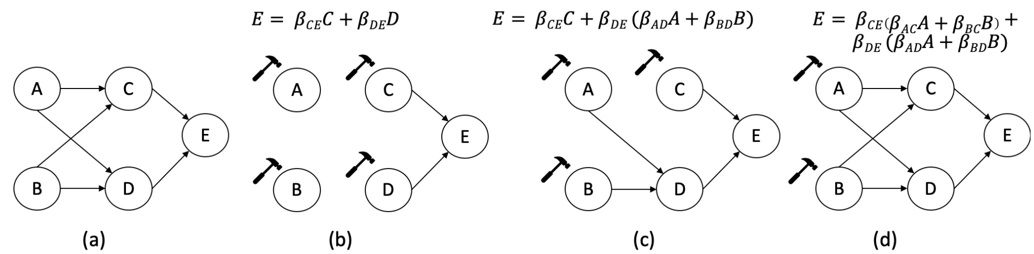


Fig. 1. Prediction in a causal model. The hammer icon represents intervention. (a) true causal model. (b) Interventions on all feature variables. The causal links leading to node C and D are removed since the values of C, and D are set externally. (c) Interventions on node A, B, and C. (d) interventions on node A, and B.

In such scenarios, human verification is necessary to validate the causal structure obtained from automated algorithms. Prior research and empirical evidence suggest this human-centered research approach can identify the causal structure better than automated algorithms alone [40].

Finally, once the causal structure (F) is learned via any of the above approaches, we can use Structural Equation Modelling (SEM) to parameterize the model, obtaining the path coefficients.

3.2 Prediction

We define the outcome variable Y as an endogenous variable, the variable we want to predict from a set of feature variables X. In a causal model (M), estimating Y from X is analogous to applying an intervention (fixing variables to specific values) on X [42]. This is achieved through the *do* operator which simulates a causal intervention by deleting certain edges from M while fixing X to specific values [36]. The resultant causal model M' is a subgraph of the original model M. Figure 1(b) shows M' when intervention is applied to all variables of X. Note that the edges leading to Node C, and D from the original model are removed in figure 1(b). This is because nodes A, and B can no longer causally effect C, and D, once we fix them to specific values. In this scenario, Y can be estimated using the following equation:

$$P_M(Y|do(X = x)) = P_{M'}(Y) = \beta_{CE}C + \beta_{DE}D \tag{1}$$

Figure 1(c) and (d) present different intervention scenarios on the original model M. The key idea is that Y is independent of its ancestors conditioned on all of its parents. Once we know the direct parents Y, other variables in the causal DAG can no longer influence Y.

4 FORMATIVE STUDY WITH NON-EXPERT USERS

Several research papers provide concrete sets of guidelines for creating an XAI interface for expert users or ML practitioners. To understand the expectations that non-expert users have for an automated decision-making interface, we conducted formative interviews with 10 users (5 female, 5 male) who did not have any algorithmic expertise. We refer to the participants as P1-P10 below.

Need for transparency. We noticed a general need for transparency among non-expert users. They appeared to prefer an automated system over human assistance, but feared the systems might not give them the optimal service. When prompted about their experience with automated systems such as car insurance and loan approval, the participants mentioned that automated platforms allowed them to obtain service quickly and efficiently, whereas to get human assistance they often had to wait on the phone for a long time. Yet, several participants mentioned that eventually, they needed to contact the human agents since they did not get answers to all of their questions from the automated interfaces. Specific rules and provisions were often buried under long textual descriptions which made it difficult for them to find useful information.

How can I improve the decision? When asked about the process of evaluating algorithmic decisions, our participants mentioned that they would repeatedly update the input features to change the decision in their favor. They would try to make sense of the underlying algorithm by changing the values of variables and then observe the effect this had on the outcome variable. We note that this process is the same as asking what-if questions, a popular method to interpret predictive models. The participants also mentioned that this process sometimes led to frustrations since they did not know which input feature to change.

How am I different than my friend? Users of automated decision-making systems often employ a mental process of comparing themselves with others and try to make sense of why different people received different decisions. Our participants shared several such cases where they wondered why they received one decision, while their friends received different decisions. We note that this is similar to neighborhood exploration, a common technique applied in XAI to compare a data point to its nearest neighbors.

5 DESIGN GUIDELINES

Based on prior research and the insights gathered from the formative study, we formulate the following design guidelines:

Creating the Causal Model. The ability to learn a causal model for prediction was a core challenge in the development of Outcome-Explorer since causal models do not provide prediction by default. Section 3.1 presented three different methods to obtain the causal structure. We opted for the third method as it combines the first two methods and matches well with our human-centered methodology. At the time of the development of this work, no open-source software or package was available for human-centered causal analysis. Additionally, none of the existing tools supported prediction in a causal model. Hence, we decided to support model creation in our tool such that an expert user would be able to obtain the best possible causal model from a dataset. Note that while this was not the main focus of this work, we emphasize that this step was necessary for facilitating accurate predictions in a causal model.

We also decided to create a different module for model creation since we anticipated that model creation will require multiple tools and techniques which would not be feasible to integrate into one module alongside the functionalities required for interpreting the model. Note that, by definition, a non-expert user is not supposed to create the model which has also motivated us to segregate the model creation module from the interpretation module. The above design rationale resulted in the following design guideline, DG1:

DG1. Our tool should have two modules: Model Creation Module; and Model Interpretation Module. Using the Model Creation module, an expert user should be able to create a causal model interactively with the help of state-of-the-art techniques and evaluate the prediction accuracy of the model.

Interpreting the Model: Supporting Expert and Non-expert Users. As discussed, the two intended user groups have different reasons and goals for interpreting a predictive model and they also have different algorithmic and visualization expertise. Supporting both of these explanations needs in a single predictive interface was a key challenge in the development of Outcome-Explorer. The key here is the causal DAG, which is inherently interpretable and easy to understand by any type of user. The causal DAG alone should allow a user to understand the model without any algorithmic and mathematical knowledge. This gives rise to the following design guideline, DG2:

DG2. Both expert and non-expert users should be able to understand and interact with the causal DAG. Users should be able to set values to the input features in the DAG to observe the changes in the outcome.

Explanation Queries. Interestingly, the formative study revealed that non-expert users ask explanation queries similar to those already well-studied in XAI research. In particular, we identified four out of the six capabilities listed by Hohman et al. [19] to be common among expert and non-expert users. We briefly discuss them below.

- **C1** *Local instance explanations*: Explain the prediction for a single instance through feature contribution.
- **C2** *Instance explanation comparisons*: Allow a comparison between instances in terms of the prediction mechanism.
- **C3** *Counterfactuals*: Allow the user to explore "What-If" questions as a probing method.
- **C4** *Nearest neighbors*: Allow the exploration of nearest neighbors of a data instance in terms of a prediction or feature.

Based on these capabilities, we articulate the following guideline, DG3:

DG3. Our tool should support C1-C4. The capabilities should be implemented keeping in mind the algorithmic and visualization literacy gap between the expert and non-expert users.

Input Feature Configuration. Our tool is completely transparent and a non-expert user can change the input features freely in the interface. However, when engaging in this activity, it is possible that to obtain a certain outcome a user might opt for a feature configuration that is unlikely to be realistic [39]. To address this problem, we derive the following guideline, DG4:

DG4. Our tool should allow non-expert users to evaluate not only the value of the outcome, but also how realistic the input configuration is when compared to existing data points.

6 A DEMO OF OUTCOME-EXPLORER

Based on the design guidelines outlined above, we iteratively developed Outcome-Explorer. Before outlining the implementation details, we present a usage scenario of Outcome-Explorer. Throughout the usage scenario, we demonstrate the functionalities of Outcome-Explorer and link the design guidelines with each functionality.

Suppose, Adam is a Research Engineer at a technological company and is responsible for creating a housing price prediction model. The model will eventually be used by people who do not have algorithmic expertise. Since non-experts are the target users, Adam also needs to create an interface where users can easily understand the model and interact with it. Based on these requirements, Adam decides to use an interpretable model for prediction. Adam recently came to know that causal models are interpretable and can be used for prediction. Adam researched software and tools that are available for causal prediction and determined that Outcome-Explorer matches the requirements perfectly.

6.1 Creating the Model (DG1)

Adam starts off Outcome-Explorer by uploading the housing dataset into the interface (Figure 2). Adam has prior knowledge about the dataset but is unsure about the underlying causal structure. Adam selects the PC algorithm for searching the causal structure (not depicted). Upon seeing the causal DAG obtained from the PC algorithm, Adam uses prior knowledge to refine the causal relations. For example, Adam notices that the initial model has an undirected edge between "INDUSTRIALIZATION" and "DISTANCE_FROM_CITY". From domain expertise, Adam knows that

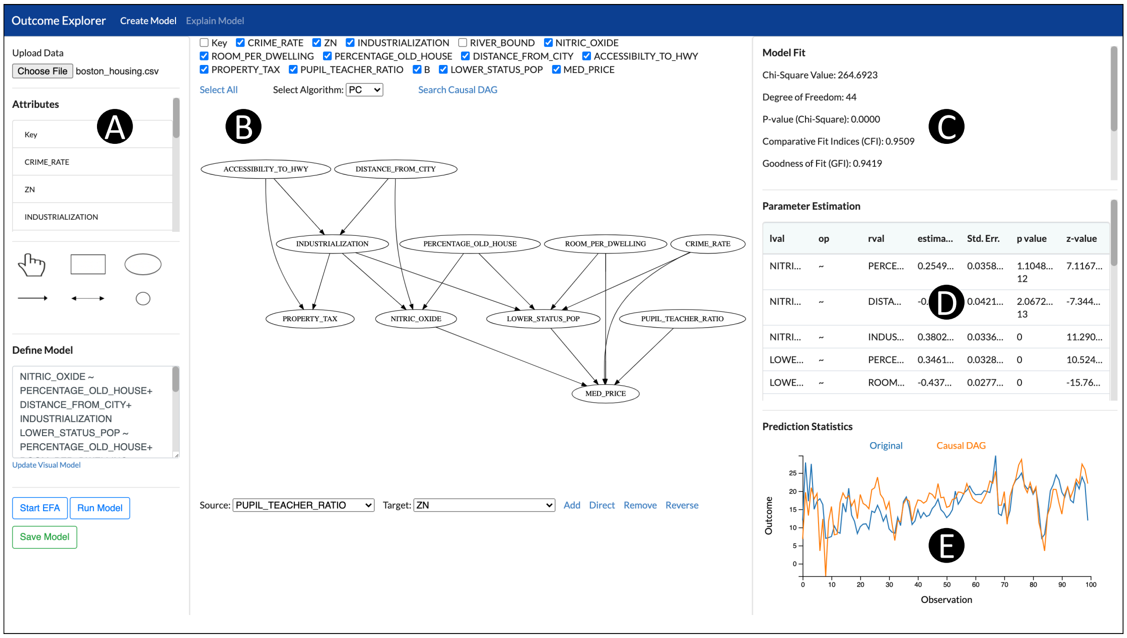


Fig. 2. Model Creation Module of Outcome-Explorer. A) Control Panel. B) Causal DAG obtained from causal search algorithms. Users can interactively add, remove, and direct edges in the causal DAG. C) Model fit measures obtained from Structural Equation Modelling (SEM). D) Parameter estimation for each relations (path coefficients). E) A line chart showing the prediction accuracy of the Causal DAG on the test set.

"DISTANCE_FROM_CITY" can be a cause of "INDUSTRIALIZATION", but the opposite relation is not plausible. Adam directs the edge from "DISTANCE_FROM_CITY" to "INDUSTRIALIZATION" and notices that the model fit and the prediction accuracy has also increased. Figure 2 presents the final causal DAG obtained in this iterative process (please see supplemental video for the intermediate steps). Adam hits the "Save Model" button and moves onto the "Model Interpretation" module.

6.2 Interacting with the Causal DAG and Exploring Nearest Neighbors (DG2, DG3)

After creating the model, Adam wants to explore and verify the model before making it public. Adam starts off the exploration process by selecting a sample data row from Figure 3(B). Adam observes that the selected row is immediately reflected in the feature values and the outcome of the model has changed to 22.61K. Adam also observes that the edges entering the endogenous variables became blurred (deactivated) since those variables were set to reflect the selected row and can no longer be estimated from the exogenous variables. From the profile map in Figure 3(C), Adam notices that the selected housing has a relatively small price. Adam decides to compare the selected housing with higher prices. To do so, Adam selects a data point with a higher housing price from the profile map. Immediately, an orange profile is created in the causal DAG. From the two profiles, Adam easily understands where the two housing differed and how that affected the outcome. At this point, Adam hits the "Save Profile" button to save both the profiles in the tracker (Figure 3(D)). In a similar manner, Adam explores several other data points to get a concrete idea of the model. At the

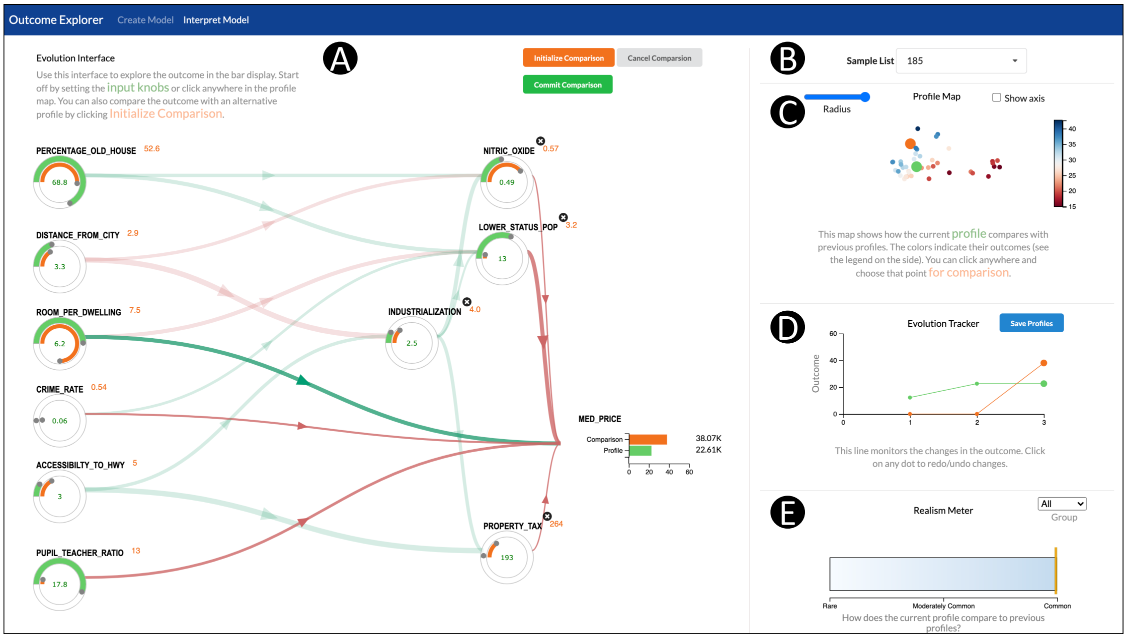


Fig. 3. Model Interpretation Module of Outcome-Explorer. A) Interactive causal DAG showing causal relations between variables. Each node includes two circular knobs (green and orange) to facilitate profile comparisons. The edge thickness and color depict the effect size and type of each edge. B) Sample selection panel. C) A biplot showing the position of green and orange profiles compared to nearest neighbors. D) A line chart to track the model outcome and to go back and forth between feature configuration. E) Realism meter allowing users to determine how common a profile is compared to other samples in the dataset.

end of the analysis, Adam is confident that the model is accurate, interpretable, and is ready for deployment. Adam then publishes the interface with the name *causalX*.

6.3 Understanding the Causal Relations (DG2)

Suppose, Emily is a middle-aged female who would like to purchase a new house. Emily came to know about a web service (*causalX*) which lets the user explore different neighborhoods and estimate the prices of houses in those neighborhoods. Emily decides to use the service to understand the quality of the desired neighborhood.

Emily visits the site and starts off by watching a small tutorial on how to use the interface. After that, Emily starts to make sense of the variables and how they are connected to each other. Emily notices that Property Tax is calculated from a region's accessibility to the highway and the industrialization index. Emily further notices that Industrialization depends on both the region's distance from the city and its accessibility to the highway. Based on that, Emily concludes that Property tax depends on three factors: Accessibility to Highway, Distance from City, and Industrialization. Interestingly, Emily observes that the Median Price has a red edge from Property Tax, meaning houses located in areas with higher property taxes are priced lower than houses from areas with lower property taxes.

6.4 Exploring Outcomes and What-if Analysis (DG2, DG3, DG4)

After understanding the causal relations between variables, Emily now starts putting values to the knobs in the interface. Emily observes that once the value of an internal node is set, edges leading to that node become blurred. For example,

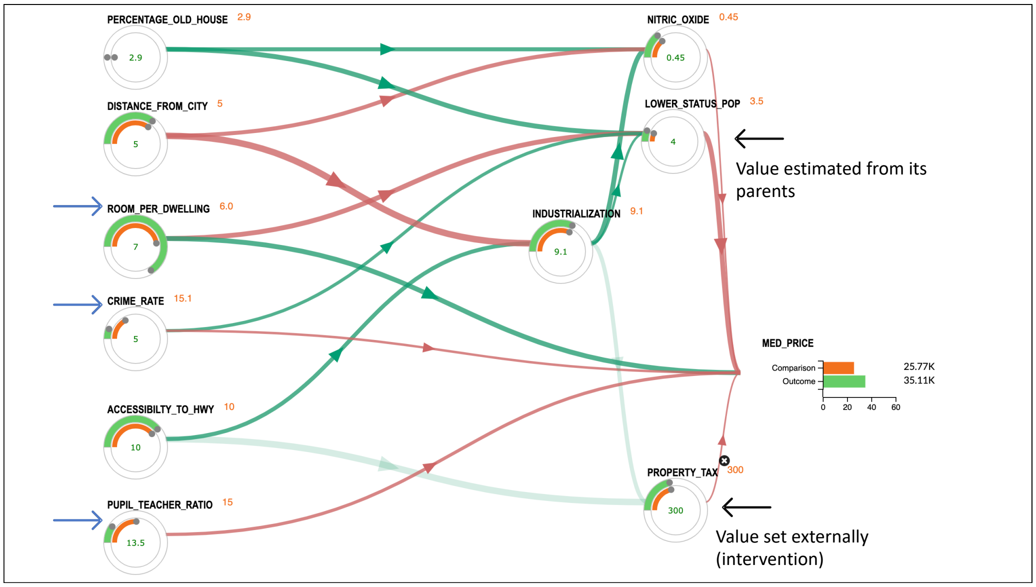


Fig. 4. Asking what-if questions in Outcome-Explorer. A user can keep one profile (green) fixed, and change the other profile (orange) to ask what-if questions. The blue arrows indicate the changes in the orange profile. Note that property tax is set to 300 by the user. As a result, changing its parents will not affect property tax. The other endogenous variables are estimated from their parents.

Emily sets the value of Property Tax to 300, which decreases the Median Price of the houses, but also blurs the edges entering Property Tax (figure 4).

After finding the ideal neighborhood, Emily notices that the Median Price of that neighborhood is around $35,000. But, Emily only has a budget of around $25,000. Based on that, Emily decides to change the variables so that the median price comes down to $25,000. Emily fires off the comparison mode by clicking the "Initialize Comparison" Button. Immediately, a new orange profile is created in the interface which is exactly the same as the current green profile. Keeping the green profile constant, Emily iteratively changes the variables to take down the median housing prices to $25,000. In this iterative process, Emily utilizes the tracker regularly to go back and forth between different profile configurations. Emily also consults the realism meter regularly to see how common the selected housing is compared to the existing neighborhoods.

7 IMPLEMENTATION

Outcome-Explorer is a web-based interface. We used Python as the back-end language and D3 [5] for interactive visualization. Several causal analysis packages such as Tetrad, factor_analyzer, and semopy were used for causal analysis.

7.1 Causal DAG Visualization

The model creation module and interpretation module use a different visual representation to present the graphs. The interpretation module accepts a DAG as input and employs topological sort to present that DAG in a left to right manner. In the model creation module, the graph returned by the automated algorithms are not necessarily a DAG; it can contain undirected edges. The graph can not be represented the same way we represent it in the interpretation module. Besides, in the model creation module an expert user can edit the graph, resulting in a change of the structure of the graph regularly. Hence, we decided to use GraphViz [12], a well-known library for graph visualization.

We follow the visual design of Wang et al. [44, 45] to encode the edge weights in the causal DAG presented in the interpretation module. We note that the primary purpose of Wang et al. and other general-purpose causal analysis tools is to find causal relations among variables. In contrast, the main focus of this paper is to design and evaluate an interpretable predictive interface based on causality where obtaining the causal structure is only one part of the overall computational pipeline.

7.2 Realism Meter

To determine how (dis)similar a profile is to the existing users, we opt for a multi-dimensional method similar to detecting an outlier in one dimension using the z-score. At first, we fit a Gaussian Mixture Model on the existing users. A mixture model with K Gaussians or components is defined as

$$P(X) = \prod_{n=1}^{N} \sum_{k=1}^{K} P(X_n|C_k)P(C_k) = \prod_{n=1}^{N} \sum_{k=1}^{K} \phi_k N(X_n|\mu_k, \Sigma_k) \tag{2}$$

where N is the number of datapoints, $\phi_k = P(C_k)$ is the mixture weight or prior for component k, and μ_k, Σ_k are the parameters for the k-th Gaussian. Once the parameters are learned through the Expectation-Maximization algorithm, we can calculate the probability of a datapoint x belonging to a component C_i using the following equation

$$P(C_i|x) = \frac{P(C_i)P(x|C_i)}{\sum_{k=1}^{K} P(C_k)P(x|C_k)} = \frac{\phi_i N(x|\mu_i, \Sigma_i)}{\sum_{i=1}^{K} \phi_k N(x|\mu_k, \Sigma_k)} \tag{3}$$

A high value of $P(C_i|x)$ implies that x is highly likely to belong to C_i, whereas a low value $P(C_i|x)$ implies that the features of x is not common among the members of C_i. Thus, $P(C_i|x)$ can be interpreted as a scale of how "real" a datapoint is to the other members of a component. We translate $P(C_i|x)$ to a human understandable meter with $P(C_i|x) = 0$ interpreted as "Rare", $P(C_i|x) = 0.5$ as "Moderately Common", and $P(C_i|x) = 1$ as "Common".

8 EVALUATION

We evaluated Outcome-Explorer in two phases: first, we conducted think-aloud sessions with three ML practitioners to gather expert feedback and real-world potential of Outcome-Explorer; second, we conducted a user study with 18 users to assess the effectiveness and usability of Outcome-Explorer in supporting the explanation needs of non-expert users.

8.1 Expert Evaluation

We invited three ML practitioners as expert users (1 female, 2 male) to examine Outcome-Explorer. Participation was voluntary with no compensation. All three participants had post-graduation degrees and had conducted research in the field of XAI, Fairness, and Data Ethics for at least five years. They were also familiar with statistical causal analysis.

The tool was deployed on a web server and the sessions were conducted via Skype. Participants shared their screen as they performed the tasks. One author communicated with the participants during the sessions while another author took notes. Each session started with the participant signing a consent form, and then observing a live demo of the tool. After that, participants were asked to choose one out of two datasets: Boston Housing [17]; and the PIMA Diabetes dataset [41]. We converted the outcome variable of the Diabetes dataset to probability as our interface does not support classification yet. Once a participant chose a dataset, we provided them with the textual descriptions of the features, and a task list. The task list was designed to guide the participant in exploring different components of Outcome-Explorer. Participants started off by creating a causal model, and then gradually moved into examining different explanation methods available in Outcome-Explorer. While performing the tasks, participants thought-aloud and conversed with the authors continuously. We sorted their feedback in the following four thematic categories:

8.1.1 Comprehensive and Generalizable. All three expert users found the "Create Model" module to be "comprehensive", and "generalizable". Participants found the accuracy statistics to be most helpful as that feature is not available on other comparable causal analysis tools. According to E1, Outcome-Explorer was "rigorous" in terms of causal functionality and should enable users to obtain the "best possible causal model".

The participants also found the explanation module to be comprehensive. E1 mentioned that the interactive path diagram alone should allow a non-expert user to understand the model. Additionally, they found the two profile comparison mechanism to be helpful and appreciated the fact that the user can ask the what-if questions directly to the model.

8.1.2 Engaging, Thought Provoking, and Fun. Participants continuously engaged themselves in making sense of the causal relations. Throughout the session, they enthusiastically initiated discussions with the authors to share their personal experiences related to causal relations.

Participants also found the visual design of the Explanation module to be ascetically pleasing and fun to interact with. They mentioned that the interface has a "certain gaming flavor" to it. According to E3, the thought-provoking and interactive nature of Outcome-Explorer might entice curious non-expert users to gather knowledge on a domain of interest.

8.1.3 Prior Knowledge and Position in the ML Pipeline. Participants suggested that Outcome-Explorer could be used once an expert user has preprocessed and explored the dataset. It would provide users the necessary background knowledge for creating and explaining the causal model. According to E1,

"I can see that the user might need to tweak the initial causal model iteratively to reach the final model, but that is also true for many ML models. The process of creating a predictive model is often messy, and require several iterations, each of which requires users to utilize prior knowledge to refine the model."

8.1.4 Disclaimers. Causal relations make stronger claims than associative (correlative) relations. Participants suggested that the implication of the causal relations should appropriately be communicated to the end-users. For example, a particular causal relation may hold true for a particular task or domain, but not in general. Expert users should be aware of such potential misleading relations in the causal model, and should provide disclaimers to the non-expert users whenever needed. This will ensure that non-experts are not misled into thinking that the causal relations in Outcome-Explorer are ubiquitously true.

8.2 User Study with Non-expert Users

In order to understand how Outcome-Explorer might help non-expert users understand the prediction mechanism of a machine learning model and affect their efficiency in obtaining automated decisions, we conducted a user study. Through this user study, we aimed at finding potential advantages and disadvantages of our tool compared to existing state-of-the-art explanation methods. We chose SHAP [30] as a comparison case for Outcome-Explorer as it is a widely used post-hoc explanation technique which can explain the prediction of any machine learning model. The interpretable nature of our tool has also motivated us to compare it to SHAP which approximates the prediction mechanism of a model without showing the model itself, an approach fundamentally different than ours. Additionally, SHAP is open-source and provides several visualizations to showcase explanation which could ensure a fair comparison with our visual interface. Hence, we conducted a repeated-measures within-subject experiment with the following two conditions.

C1. SHAP: This condition included input boxes which users could use to change variables. Users had access to two charts provided by SHAP: a bar chart showing global feature importance and a variant of stacked bars (force plot) showing feature contribution for a decision.

C2. Outcome-Explorer-Lite: This prototype included the interactive causal DAG with other components of Outcome-Explorer hidden.

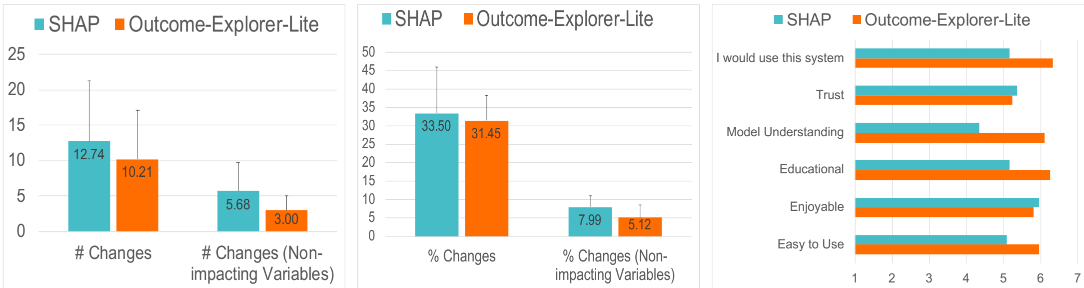
We chose to include only the causal DAG in the study as the other components provide auxiliary tools to understand the model, but are not necessary to interpret the model. The inclusion of these components could hinder a fair comparison between Outcome-Explorer and SHAP. To minimize the learning effect, we included two datasets and counterbalanced the ordering of study conditions and datasets. The two datasets are the Boston housing and PIMA Diabetes dataset, both appeared previously in the XAI literature [19, 33].

8.2.1 Participants. We recruited 18 participants (10 males, 8 females) through local mailing lists, university mailing lists, and social media posts. Participation was voluntary with no compensation. The participants varied in age from 19 to 35 ($M = 25$, $SD = 4.21$). None of the participants had machine learning expertise, except one who had a bachelor's degree in Computer Science. The participants were comfortable in using web technology and had a high-level idea of automated decision-making through exposure to credit-card approval and loan approval systems. Additionally, two participants had experience with interactive visualization through interactive online news.

8.2.2 Tasks. XAI interfaces are frequently evaluated on "proxy tasks" such as how well humans predict the AI's decision, and subjective measures of trust and understanding [7]. Recent research suggests that proxy tasks and subjective measures are not good predictors of how humans perform on actual decision-making tasks [7]. Based on that, we decided to evaluate our tool on actual decision-making tasks. Each participant was provided with a scenario and an input feature configuration. After that, the participants were asked to obtain a series of alternative decisions (goals) by interacting with the interface while minimizing the number of changes and magnitude of changes. We anticipated that the different approach of interpretability provided by our interface and SHAP might reveal interesting insights on how users utilize interpretability for performing actual decision-making tasks. Note that both conditions made predictions based on the same underlying causal model. Since the evaluation was based on actual tasks, not proxy tasks, we also collected self-reported subjective measures such as model understanding, trust, and usability which might reflect users' preference between post-hoc interpretability and inherently interpretable models.

8.2.3 Study Design. Similar to the sessions with the expert users, we conducted the study sessions via web and Skype. A study session began with the participant signing a consent form. Following this, the participants were introduced with the assigned first condition and received a brief description of the interface. The participants then interacted with the system (with a training dataset), during which they were encouraged to ask questions until they were comfortable. Each participant was then given a scenario and a task list for the first condition. After completing the tasks, the participants filled out a questionnaire with subjective measures. The same process was carried out for the second condition. Each session lasted around ~1 hour and ended with an exit-interview.

8.2.4 Results.



(a) Number of input feature changes (b) Magnitude of input feature changes (c) Subjective Measures (1: Strongly Disagree, 7:Strongly Agree).

Fig. 5. Study Results. (a) The average number of changes made to the input features to reach the goals. (b) The average amount of changes made to the input features to reach the goals. (c) Average self-reported subjective measures. Error bars show +1 SD.

Quantitative Measures. In a causal model, the prediction is facilitated through intervention. Exogenous variables may not affect the outcome if endogenous variables are set to specific values. While Outcome-Explorer visualizes the interplay between endogenous and exogenous variables, SHAP only estimates feature contributions to the decision, and does not explain why some variables are not affecting the outcome. We refer to such variables as *non-impacting variables*, input features that do not contribute to the outcome. We measured four quantitative metrics to account for the changes users made on the input features to reach the target outcomes. They are (1) number of changes, (2) number of changes (non-impacting) (3) magnitude of changes, and (4) magnitude of changes (non-impacting). Here, non-impacting refers to the changes made on non-impacting variables, while absence of the term mean changes on all variables. The magnitude of changes refers to the percentage of changes made to the input features.

As shown in Figure 5a, the average number of changes were 12.74 (SD = 8.59) for SHAP, and 10.21 (SD = 6.88) for Outcome-Explorer-Lite. The difference was not statistically significant. On average, the participants made 5.68 (SD = 4.01) changes to the variables that did not have any effect on the outcome when using SHAP (Figure 5a), while for Outcome-Explorer-Lite, the average was 3.00 (SD = 1.97). Participants were able to reduce changes made to the non-impacting variables by 47%, which was statistically significant ($p < 0.02$), and Cohen's effect size value ($d = 0.68$) suggested a medium significance.

We noticed a similar trend in the magnitude of changes users made to the input features to reach the goals. While we did not find any significant difference between the overall magnitude of changes users made in each study condition, we found a significant difference between the magnitude of changes users made on non-impacting variables (36% reduction with $p < 0.001$, Figure 5b). We also measured the time taken to complete the tasks, but no statistically significant difference was found.

Participants could see from the SHAP visualization that some features were not affecting the outcome, but this did not prevent them from interacting with the non-impacting variables. Participants constructed several hypotheses why the non-impacting variables did not have an effect on the outcome while using SHAP, including the possibility of a change of impact in the future. They also hypothesized that even though these variables did not have an impact on the outcome they might have indirect effects on other variables. This may be the reason for the increased interaction with the non-impacting variables. On the other hand, in Outcome-Explorer-Lite the lack of edges or blurred edges provided users with clear evidence for non-impacting variables and their relations with impacting variables.

In order to understand how study conditions and dataset related to each other with respect to quantitative measures, we constructed five mixed-effect linear models, one for each metric. We tested for interaction between study condition and dataset while predicting a specific metric. While there were no interaction effects and the dataset did not play any significant role in predicting the metrics, we found study condition to be the main effect in predicting the number of changes on non-impacting variables ($F(1, 26.033) = 7.723, p = 0.01$) and the magnitude of changes on non-impacting variables ($F(1, 26.992) = 13.140, p = 0.001$).

Self-reported Subjective Measures. We follow [8] to analyze the self-reported subjective measures using regression. Each participant rated the study conditions (interfaces) on a Likert scale ranging from 1 (Strongly Disagree) to 7 (Strongly Agree) based on six metrics. As shown in Figure 5, participants rated Outcome-Explorer-Lite higher on four metrics: Easy to use ($M : 5.96, SD : 0.96$), Educational ($M : 6.26, SD : 0.87$), Model Understanding ($M : 6.11, SD : 0.50$), and I would use this system ($M : 6.33, SD : 0.6$). In comparison, the scores for SHAP were: Easy to use ($M : 5.08, SD : 1.35$), Educational ($M : 5.15, SD : 0.964$), Model Understanding ($M : 4.34, SD : 0.911$), and I would use this system ($M : 5.16, SD : 1.42$).

Similar to the quantitative metrics, we constructed six mixed-effect linear models, one for each subjective measure. The results revealed no interaction effects and dataset did not play any significant role in predicting the measures. We found study condition to be a main effect in predicting Easy to Use ($F(1, 31.875) = 6.157, p = 0.019$), Educational ($F(1, 25.084) = 6.337, p = 0.018$), Model Understanding ($F(1, 25.123) = 9.997, p = 0.004$), and I would use this system ($F(1, 23.483) = 4.842, p = 0.038$).

The results matched our anticipation that Outcome-Explorer-Lite will improve users' model understanding and that they will learn more about the prediction mechanism using our tool. In general, participants appreciated the visual design of the causal DAG which might be the reason why they found Outcome-Explorer-Lite to be easy to use and want to see it in practice.

Qualitative Feedback. While using SHAP, most of the participants constructed hypotheses around non-impacting variables. In contrast, while using Outcome-Explorer-Lite they frequently constructed several hypotheses to interpret the causal relations between the input features. To that extent, two participants continued to use Outcome-Explorer even after completing the tasks. Participants found the explanation behind intervention natural, and understood why some features were not affecting the outcome. According to one participant, "It makes sense, if I know the Property Tax of an area, why would I want to estimate it from other things."

One of the participants compared Outcome-Explorer-Lite with a waterfall. According to the participant, "it seemed like streams of water coming from different branches and flowing towards the steep end of the waterfall, some branches are narrow and some branches are wide". Analogies related to gaming were also common: "It is just like a game, for each action, I can see the variables changing in the interface. I was really curious to know how I can obtain my desired score."

On average, the participants spent slightly more time when using Outcome-Explorer. While familiarizing with the interface was one factor for that, in the post-study interview, several participants mentioned that they felt curious, spent more time to learn the relations, and put some thought before taking an action. A participant who is a senior college student mentioned:

"The interface (Outcome-Explorer) is explanatory. I felt like I learned something. The interface is fun, attractive as well as educational. I did not know much about housing prices before this session. But, I think I now have a much better understanding of housing prices. If available in public when I buy a house in the future, it will help me make an informed decision."

9 DISCUSSION AND LIMITATION

Outcome-Explorer promotes graphical models as an interpretable set of machine learning models. Our finding suggests representing variables as nodes and showing their relations via edge connections is an effective way to convey the inner workings of the predictive models. Our design and visual encoding can be extended to other graphical models. For example, a Bayesian Network is also represented as a DAG, and the design of Outcome-Explorer can be transferred to interactive systems based on Bayesian networks.

9.1 Limitations and Future Work

As with every system there are certain limitations. One of these is inherent to the data available for constructing the causal model. As for any machine learning paradigm, a model is only as good as the data allows. As for causal models specifically, it is important to have a sufficient number of variables that cover all or at least most aspects determining the predicted outcome. Using computational causal inferencing under incomplete domain coverage can result in islands of variables or a causal skeleton where some links are reduced to correlations only.

We have experimented with evolutionary and confirmatory factor analysis to introduce additional variables that can complement the native set of variables. These variables are often not directly measurable and serve as latent variables. Others can be obtained by obtaining additional measurements. In that way this type of analysis can help drive the acquisition of further data.

Apart from wrongly directed edges which our visual interface allows users to correct, another source of error can be confounders which can lead to an overestimation or underestimation of the strength of certain causal edges. There are several algorithms available for the detection and elimination of counfonding effects and we are presently working on a visual interface where expert users can take an active role in this type of effort.

Another issue is scalabilty, especially when it comes to the visualization of the model and the interaction with it. Too many variables and their edges can clutter the display, and level of detail methods are needed to help users focus on certain aspects of the model. Level of detail methods can be driven by the in and out degree of variables weighted by the strength of the edges [50].

Finally, we acknowledge that causal models might not reach the prediction accuracy of complex machine learning models, but a linear model such as ours can compete with the deep neural networks if features are carefully engineered [39, 42]. Our computational pipeline also does not support classification yet, which we intend to integrate in the future.

10 CONCLUSION

We presented Outcome-Explorer— an interactive visual interface that exploits the explanatory power of the causal model and provides a visual design that can be extended to other graphical models. Outcome-Explorer advances research towards interpretable interfaces and provides critical findings through user study and expert evaluation.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–18.

[2] Saleema Amershi, Max Chickering, Steven M Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 337–346.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016), 2016.

[4] Peter M Bentler and David G Weeks. 1980. Linear structural equations with latent variables. *Psychometrika* 45, 3 (1980), 289–308.

[5] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics* 17, 12 (2011), 2301–2309.

[6] Tim Brennan, William Dieterich, and Beate Ehret. 2009. Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior* 36, 1 (2009), 21–40.

[7] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 454–464.

[8] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[9] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*. 134–148.

[10] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64, 3 (2018), 1155–1170.

[11] Finale Doshi-Velez and Been Kim. 2017. A roadmap for a rigorous science of interpretability. *arXiv preprint arXiv:1702.08608* 2 (2017).

[12] John Ellson, Emden Gansner, Lefteris Koutsofios, Stephen C North, and Gordon Woodhull. 2001. Graphviz—open source graph drawing tools. In *International Symposium on Graph Drawing*. Springer, 483–484.

[13] Ruth C Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3429–3437.

[14] Bruce Glymour and Jonathan Herington. 2019. Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms. In *Proceedings of the conference on fairness, accountability, and transparency*. 269–278.

[15] Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics* 10 (2019), 524.

[16] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. 2020. ViCE: visual counterfactual explanations for machine learning models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 531–535.

[17] David Harrison Jr and Daniel L Rubinfeld. 1978. Hedonic housing prices and the demand for clean air. (1978).

[18] Trevor J Hastie and Robert J Tibshirani. 1990. *Generalized additive models*. Vol. 43. CRC press.

[19] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[20] Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2018. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE transactions on visualization and computer graphics* 25, 8 (2018), 2674–2693.

[21] Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Polo Chau. 2019. Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 1096–1106.

[22] Rick H Hoyle. 1995. *Structural equation modeling: Concepts, issues, and applications*. Sage.

[23] Minsuk Kahng, Pierre Y Andrews, Aditya Kalro, and Duen Horng Polo Chau. 2017. A cti v is: Visual exploration of industry-scale deep neural network models. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 88–97.

[24] Aria Khademi, Sanghack Lee, David Foley, and Vasant Honavar. 2019. Fairness in Algorithmic Decision Making: An Excursion Through the Lens of Causality. In *The World Wide Web Conference*. ACM, 2907–2914.

[25] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*. 656–666.

[26] Been Kim, Cynthia Rudin, and Julie A Shah. 2014. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in neural information processing systems*. 1952–1960.

[27] I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. 2020. Problems with Shapley-value-based explanations as feature importance measures. *Proceedings of International Conference on Machine Learning* (2020).

[28] Matt J Kusner, Chris Russell, Joshua R Loftus, and Ricardo Silva. 2018. Causal Interventions for Fairness. *arXiv preprint arXiv:1806.02380* (2018).

[29] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. 2018. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859* (2018).

[30] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[31] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2019. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 349–358.

[32] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[33] Yao Ming, Huamin Qu, and Enrico Bertini. 2018. Rulematrix: Visualizing and understanding classifiers with rules. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 342–352.

[34] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. 2020. Causal Interpretability for Machine Learning-Problems, Methods and Evaluation. *ACM SIGKDD Explorations Newsletter* 22, 1 (2020), 18–33.

[35] Ziad Obermeyer and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm that guides health decisions for 70 million people. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 89–89.

[36] Judea Pearl. 2009. *Causality*. Cambridge university press.

[37] Judea Pearl. 2010. An introduction to causal inference. *The international journal of biostatistics* 6, 2 (2010).

[38] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[39] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[40] Xinpeng Shen, Sisi Ma, Prashanthi Vemuri, and Gyorgy Simon. 2020. challenges and opportunities with causal Discovery Algorithms: Application to Alzheimer's pathophysiology. *Scientific reports* 10, 1 (2020), 1–12.

[41] Jack W Smith, JE Everhart, WC Dickson, WC Knowler, and RS Johannes. 1988. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, 261.

[42] Shruti Tople, Amit Sharma, and Aditya Nori. 2019. Alleviating Privacy Attacks via Causal Learning. *arXiv preprint arXiv:1909.12732* (2019).

[43] Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* (2017).

[44] Jun Wang and Klaus Mueller. 2015. The visual causality analyst: An interactive interface for causal reasoning. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 230–239.

[45] Jun Wang and Klaus Mueller. 2017. Visual causality analysis made practical. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 151–161.

[46] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The What-If Tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.

[47] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. 2019. PC-Fairness: A Unified Framework for Measuring Causality-based Fairness. In *Advances in Neural Information Processing Systems*. 3399–3409.

[48] Jing Nathan Yan, Ziwei Gu, Hubert Lin, and Jeffrey M Rzeszotarski. 2020. Silva: Interactively Assessing Machine Learning Fairness Using Causality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[49] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making—the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[50] Zhiyuan Zhang, Kevin T McDonnell, and Klaus Mueller. 2012. A network-based interface for the exploration of high-dimensional data spaces. In *2012 IEEE Pacific Visualization Symposium*. IEEE, 17–24.