Explainable AI: Foundations, Applications, Opportunities for Data Management Research

Romila Pradhan
Purdue University
rpradhan@purdue.edu

Aditya Lahiri
University of California, San Diego
adlahiri@ucsd.edu

Sainyam Galhotra
University of Chicago
sainyam@uchicago.edu

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

ABSTRACT

Algorithmic decision-making systems are successfully being adopted in a wide range of domains for diverse tasks. While the potential benefits of algorithmic decision-making are many, the importance of trusting these systems has only recently attracted attention. There is growing concern that these systems are complex, opaque and non-intuitive, and hence are difficult to trust. There has been a recent resurgence of interest in explainable artificial intelligence (XAI) that aims to reduce the opacity of a model by explaining its behavior, its predictions or both, thus allowing humans to scrutinize and trust the model. A host of technical advances have been made and several explanation methods have been proposed in recent years that address the problem of model explainability and transparency. In this tutorial, we will present these novel explanation approaches, characterize their strengths and limitations, position existing work with respect to the database (DB) community, and enumerate opportunities for data management research in the context of XAI.

CCS CONCEPTS

• **Information systems**; • **Computing methodologies** → **Artificial intelligence**;

KEYWORDS

Explainable AI; Data Management

ACM Reference Format:
Romila Pradhan, Aditya Lahiri, Sainyam Galhotra, and Babak Salimi. 2022. Explainable AI: Foundations, Applications, Opportunities for Data Management Research . In *Proceedings of the 2022 International Conference on Management of Data (SIGMOD '22), June 12–17, 2022, Philadelphia, PA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3514221.3522564

1 INTRODUCTION

Artificial intelligence (AI) systems are increasingly deployed for decision-making in critical domains, such as healthcare, criminal justice, and finance. However, the opacity and complexity of these systems pose new threats. There is growing concern that the opacity of these systems can inflict harm to stakeholders distributed across different segments of society by perpetuating systemic biases and discrimination reflected in training data [37]. These calls for transparency created a resurgence of interest in *eXplainable Artificial Intelligence* (*XAI*– see [50] for a recent survey), which aims to provide human-understandable explanations of outcomes or processes of algorithmic decision-making systems.

The development of XAI methods is motivated by technical, social and ethical objectives [9, 14, 36, 38, 44]: (1) increasing societal acceptance of machine learning (ML)-based decision-making algorithms by establishing trust in decision results, (2) providing users with actionable insights to change the results of algorithms in the future, (3) facilitating the identification of sources of harms such as bias and discrimination, and (4) providing the ability to debug ML algorithms and models by identifying errors or biases in training data that result in adverse and unexpected behavior. The urgency of this matter has been furthered by governmental regulations that mandate businesses using automated decision-making systems to explain their decisions to end-users [1, 16].

Recently, several methods have been proposed to explain the behavior or predictions of an ML model. These approaches can be broadly categorized based on: (a) whether explainability is achieved by design (*intrinsic*) or by post factum system analysis (*extrinsic*), (b) whether the methods assume access to system internals (*model dependent*) or can be applied to any black-box algorithmic system (*model agnostic*), and (c) whether the explanations generated by the method cater to the prediction for an individual instance (*local*), explains the overall behavior of the model (*global*) or lies somewhere in between these two extremes.

In this tutorial, we will provide a detailed coverage of contemporary XAI techniques and highlight their strengths and limitations. In contrast to existing tutorials on XAI, we will discuss the scope of XAI in the context of the database community and outline a set of challenges and opportunities for data management research leveraging advances in XAI and contributing to challenges in XAI research. The learning outcomes of the tutorial are as follows.

(1) Understand the landscape of XAI techniques.
(2) Appreciate the connection between XAI techniques and existing techniques in the data management community.
(3) Exposure to key vulnerabilities of prior XAI proposals and how data management techniques can help in numerous instances.
(4) Exposure to some new opportunities to leverage data provenance and causal inference based techniques to explain model behavior and debugging AI pipelines.

Overall, our tutorial proposal is aimed to summarize key advances in the ML and AI communities with a data management lens and present numerous directions for future research.

SIGMOD '22, June 12–17, 2022, Philadelphia, PA, USA.

ACM ISBN 978-1-4503-9249-5/22/06..
https://doi.org/10.1145/3514221.3522564

2 COVERED TOPICS

Models and their predictions can be interpreted and explained according to a number of dimensions depending on the results generated by existing XAI techniques [50]. A variety of techniques are currently available that address these different dimensions of explainability. For example, some methods provide a comprehensive summary of features representing the data used to train a model, some return data points to make the model interpretable, some approximate it with an inherently interpretable model and so on. The tutorial is organized in five broad topics covering representative techniques along these various dimensions. The contents of each topic are summarized below.

2.1 Feature-based Explanations

One common way to explain black-box models is to attribute responsibility of the model outputs to its inputs. This approach is analogous to providing input feature importance. For instance, in case of Linear Regression, the coefficients of the features in the learned linear equation can be an indicator for the importance of the features. This end goal of assigning a real number to all features in the training data can be achieved in multiple ways. Furthermore, the number can indicate both the magnitude and direction of the influence exerted by the feature. We will cover the following feature attributions methods in the tutorial.

2.1.1 Surrogate Explainability. The basic idea behind explainability methods based on surrogate models is that we can approximate a (usually) local region of a complex black-box model using a simple surrogate model as a proxy. This surrogate model is an inherently interpretable model such as linear regression model [53] or decision rules [43]. The interpretation of the surrogate model is taken to be the explanation of the complex model's decision making. This method is also a subset of post hoc explainability where complex models are explained after they have been trained, making it convenient for practitioners to separate training and explanation processes. However, this method has various assumptions. For instance, LIME [53] assumes that the weighted linear regression surrogate model is able to model the complex model's decision making well enough. It also involves sampling of points near the local neighborhood which can be unreliable [73]. These components can be exploited to perform adversarial attacks that render the explanations futile [66]. There has been a lot of incremental work on creating more powerful methods based on the general LIME framework [42, 68]. However, these works still suffer from issues related to unreliable sampling.

2.1.2 Methods Based on Shapley Value. Shapley Values [63] is a concept from Game Theory that provides a method to fairly distribute payoffs obtained from a game among its players. It is the only method that follows a set of desired properties of such a payoff allocation. Recently, there has been an increased adoption of this concept in XAI. The idea is transferred by replacing the game with the prediction of the learning model and the players with the features involved in obtaining the prediction. By making this connection, we are able to convey how much each feature contributed to the prediction. A neat property by virtue of Shapley value allocation is that the attributions of the features add up to the difference of prediction given out for a particular point and the average prediction. Quantitative Input Influence [14] uses Shapley values to find average marginal influence of a feature across multiple sets. SHAP [47] unifies multiple additive feature importance measures and introduces model-agnostic Kernel SHAP to approximate Shapley values and produce attributions as explanations. TreeSHAP [46] introduces a polynomial-time algorithm to approximate Shapley values for tree-based complex models. It exploits properties of the tree structure for faster and efficient computation. It also suggests ways to combine local explanations to get a global understanding of the model. Computing Shapley values takes exponential time, since all possible feature orderings are considered. Existing methods, therefore, compute some approximation of these values, leading to certain issues with the attributions provided [40]. However, Broeck et al. suggest that even these approximations can have intractable computations for some common ML models such as logistic regression [70]. Other criticisms of using methods based on Shapley values include defining "games" in the context of a learning task and the inability of these methods to capture the indirect influences of features on the target label [40].

2.1.3 Causal Approaches. A natural human-centric way to explain predictions is to provide end-users with causal relationships between the input features and the target label. This notion goes beyond capturing mere correlations and actually tries to understand the underlying relationships in data which can help capture feature dependencies and interactions. In contrast to vanilla surrogate-based explainability methods that use standard learning algorithms to train proxy models, recent research has attempted to learn surrogates with causal objective functions instead [61]. There also have been some efforts on integrating causality with the concept of Shapley values. Asymmetric Shapely values [18] incorporate causality by discarding coalitions that do not follow causal ordering. In this process, they sacrifice the symmetry axiom of vanilla Shapley values. Causal Shapley values [30] use causal interventions and allow for explanations that are able to decompose a feature's influence into direct and indirect effects without violating any of the original Shapley value axioms. Shapley flow [74] interprets model based on assigning credit to the edges in a graph and in doing so, it extends the set-based view of Shapley values to a graph-based approach. Furthermore, nuanced causal notions such as those of sufficiency and necessity can also be used in these methods to provide more intuitive explanations [20, 75].

2.1.4 Counterfactuals and Algorithmic Recourse. Counterfactual explanations surface hypothetical examples that serve as comparison points [45]. Concretely, given that an input instance gets an output y, counterfactual explanations provide as output another input instance which is minimally different from the original instance and has an opposite output y'. Counterfactual-based explanation methods adhere to human reasoning in that they help us "compare" with other instances. They can also be extended to provide actionable recourse to the user, that is, providing users with steps to change their obtained decision [35, 69]. DiCE [51] generates a candidate set of diverse and feasible counterfactuals as explanations to understand ML classifiers. LEWIS [20, 21] uses contrastive probabilistic counterfactuals to both explain model output and provide counterfactual recourse to users who obtained a negative decision.

There are, however, challenges with counterfactuals. They can be gamed [67], and they also sometimes provide unrealistic and impossible counterfactual instances [5]. Combining counterfactual explanations with causality [48] or constraining the generated counterfactuals to obey the data manifold [5] can help overcome some of these issues.

2.2 Rule-based Explanations

Feature-attribution-based methods assign a real-valued importance score to each feature value. In contrast, rule-based explanations generate a set of rules as explanations to model behavior. The set of output rules satisfy a common property that whenever these rules are obeyed, the model provides a particular outcome. These rules should ideally be concise and be applicable to numerous data points. Longer rules (more than 5 clauses) are incomprehensible, while very specific rules are not generalizable. Anchors [54] is a method that attempts to generate short and widely applicable rules. It uses a multi-armed bandit-based algorithm to search for these rules. Lakkaraju et al. use interpretable decision sets to obtain a set of if-then rules which can be used to explain black-box models [43]. Their objective function is designed to balance and optimize both the accuracy and interpretability of these decision sets.

2.2.1 Rule-mining and rule-based data mining. Pattern recognition and rule mining is one of the fundamental topics of research in the data management community [3, 4, 26, 27]. Prior research on association rule and frequent itemset mining focussed towards identifying patterns across database. The research in this domain has evolved from pattern mining towards designing rule-based data mining techniques that leverage recent advances of weak-supervision for labelling datasets [7, 19, 71]. We will discuss how different paradigms from prior studies in data management could be leveraged to study different aspects of rule-based explanations. recent work proposed the use of abductive reasoning to computing provably correct explanations for ML predictions.

2.2.2 Logic-based methods. Recent work proposed the use of abductive reasoning and logic-based diagnosis to computing provably correct explanations for ML predictions. These methods work with a logical representation of ML algorithms [12, 32, 32, 65]. In this context, the fundamental concepts of prime implicate/implicant are closely related to sufficiency and necessary causation when the underlying causal model is a *logical circuit* [13, 15, 25, 25, 31]. It can be shown that the notion of sufficient/necessary explanations proposed in [65] translates to explanations in terms of a set of attributes that have a sufficiency/necessary score of 1. However, these methods can generate explanations only in terms of a set of attributes, are intractable in model-agnostic settings, fail to account for the causal interaction between attributes, and cannot go beyond deterministic algorithms.

2.3 Training-data-based Explanations

In contrast to feature-attribution methods, training-data-based methods attribute the output of ML algorithms to particular instances of the training dataset [10]. The central idea of data-based explanations is that training data affects the model and thus, indirectly affects the outcomes predicted by the model. To understand the predictions of a model, data-based explanations trace the model parameters and predictions back to the training data used to train the model. These methods explain the behavior of the model not in terms of the features of the data (e.g., age, gender etc.) but with respect to specific data points (e.g., enumerate 20 data points responsible for a particular model output). Data-based explanations help in debugging ML models, and understanding and explaining model behavior and model predictions. We will cover the following training-data-based methods in this tutorial.

2.3.1 Data Valuation Explanations. Shapley values [63] have also been applied to quantify the *value* of data [24, 34, 41]. Data Shapley [24] assigns values to individual training data points based on their contribution to the performance of the model (in terms of some performance metric e.g., accuracy, precision) over a test dataset. Computing exact Shapley values requires the model to be retrained for each data point, and is intractable for real-world datasets that comprise of tens of thousands of data points. Ghorbani and Zou propose Monte-Carlo based and gradient-based approaches to efficiently approximate data Shapley values of data points [24]. Jia et al. introduce practical Shapley value estimation algorithms by making assumptions on the stability of the model in terms of its performance metric and loss function [34]. Note that the Data Shapley value of a data point is specific to the learning algorithm, the performance metric and how the data point is related to other training data points. Several researchers have posited that measuring the Data Shapley value of training data points with respect to a fixed data set ignores the fact that the training data is in fact sampled from an unknown underlying distribution [23, 41]. Moreover, the assigned values may not be meaningful for the data points in the context of a new dataset. Distributional Shapley [23, 41] addresses these concerns by extending Data Shapley to quantify the value of data points in the context of an underlying data distribution. Ghorbani et al. estimate the distributional Shapley value of a data point by considering the expected value of its data Shapley value with respect to the underlying distribution [23]. Kwon et al. introduce analytical expressions for distributional Shapley for commonly used supervised and unsupervised learning algorithms such as linear regression, binary classification and non-parametric density estimation [41].

2.3.2 Influence-based Explanations. We next consider methods that identify training data points that are the most *influential* for estimating the model parameters, and in turn, for the model predictions. The naïve way of computing the influence of a data point is by removing it, retraining the ML model on the reduced dataset and computing the difference in model parameters or model predictions. Retraining the model is computationally prohibitive when there are numerous data points to consider as is the case with most of the real-world datasets.

Recently, influence functions [11], a classic technique of robust statistics that measures how optimal model parameters depend on training data points, have been used to rank individual training data points based on their influence on model predictions. For parametric models with twice-differentiable loss functions, Koh and Liang compute the first-order approximate change in model parameters by upweighting the data point by a small amount [39]. This approach avoids retraining the model by estimating the change

in model parameters effected by a slight change in the weight of a data point. First-order approximations imply that the effect of removing a *group* of data points can be obtained by simply adding the influences of individual data points. However, applying first-order approximations to a group of data points can be inaccurate because they do not capture the correlations among data points in the group. Basu et al. estimate the influence of a coherent group of data points using second-order approximations [8]: the intuition is that in the presence of correlations between data points, the effect of higher-order approximations is not negligible. These approaches are not applicable to non-parametric models such as decision trees. Sharchilev et al. extend influence functions to the non-parametric gradient boosted decision trees by proposing methods for estimating influences based on proxy metrics [64]. Their approach further develops efficient approximations to compute influence by fixing the tree ensemble structure and analyzing changes in leaf values with respect to the weights of the training data points.

2.4 Explanations for Unstructured Data

Deep learning has been very successful, especially in tasks such as image classification and language translation that involve images and texts. Although existing XAI approaches primarily focus on structured data, there have been significant advances on explaining ML model predictions over unstructured data. For example, explanations for image classification models can be found under various names such as sensitivity map, saliency map, pixel attribution maps, gradient-based attribution methods, feature relevance, feature attribution, and feature contribution [50]. These explanations typically highlight and rank the input pixels in terms of their importance toward the classification outcome. However, individual pixels may not have a large direct influence on the outcome of a classifier, but can indirectly influence its outcome by contributing to the abstract features and concepts learned by neural networks from the raw pixels. It has been shown that these methods are computationally expensive and could be highly misleading, fragile and unreliable [2, 22, 52]. Similarly, LIME [53] can be applied to textual data to identify specific words that explain the outcome of a text classification model. Another popular type of explanations in computer vision are counterfactual explanations that are generated by changing minimal regions of an image that lead to a change in the outcome of a classification [72]. In this tutorial, we will focus on structured data since it is more relevant to the DB community.

3 OPPORTUNITIES FOR DATA MANAGEMENT RESEARCH

In the final stage of the tutorial, we will outline data management research directions in the context of XAI including:

Efficiency of Feature-based Explanations. The intractability of feature-based explanations [6, 70] is a major challenge going forward. Furthermore, counterfactual explanations must be *plausible*, *feasible*, and given the huge search space of perturbations, generated in real time. Recent efforts in this direction includes GeCo [60], but several aspects, such as explanation robustness to small changes in data distribution, automatically inferring plausibility and feasibility constraints, and addressing data privacy during explanation generation, are yet to be covered.

Data-Based Explanations. The central idea in data-based explanations is to estimate updated model parameters when a subset of training data is removed. An interesting new direction is to adopt database techniques such as incremental view maintenance to estimate the parameters of the updated model by incrementally retraining the model [59, 77]. Recently, Wu et al. proposed a system that uses influence functions to explain SQL queries by identifying data points that are responsible for an error in a query result (where the query includes predictions from an ML model trained over that data) [76]. Wu et al. develop a provenance-based approach for incremental computation of model parameters and the influence of removing subsets of training data points [77]. An important future challenge is to design algorithms that generate compact, diverse explanations that describe homogeneous subsets of training data.

Provenance-Based Explanations. Existing data-based XAI techniques focus on identifying training data points responsible for error in model predictions. However, training data errors may get introduced or exacerbated during different data preparation stages. To hold particular stages accountable for ML decisions, the flow of training data points must be monitored through different stages using provenance techniques [29]. Provenance information can be harnessed to generate explanations for an ML model outcome in terms of the actions taken and decisions made throughout the ML pipeline that led to the model outcome.

Explanations in Databases. Explaining database query results has been an active area of research where the focus is on providing justification and evidence that establish the validity of or assist with the interpretation of a query answer [49, 55]. We believe that the large body of work on explanations for database query results can benefit from advances in XAI research and vice versa. As an example, recent developments in XAI have inspired novel explainability approaches such as Shapley value-based methods to generate explanations for SQL query answers [62] and database repairs [17].

User study and evaluation. Explainable Artificial Intelligence (XAI) and explanations in general are aimed towards helping end-users understand the internal of complex data science pipelines. However, evaluation of different explanation techniques requires carefully designed experiments that understand user's understanding. User studies are generally performed to evaluate faithfulness of explanations [28, 53]. Recent work has exposed the vulnerabilities of many prior proposals [33]. We will discuss the connections between these vulnerabilities and commonly considered evaluation strategies in data management. We will raise open questions to improve the guidelines to run user studies to streamline the design of XAI techniques.

4 TUTORIAL ORGANIZATION

Target Audience. Explainable AI has been a fast-growing and recent area of interest for the database and ML communities. We expect this tutorial to have widespread appeal among the SIGMOD 2022 attendees. The tutorial aims at researchers, developers and students with an interest in XAI. Database researchers can expect the tutorial to provide interesting research opportunities at the junction of data management and ML. From the perspective of system developers and practitioners, the concepts and techniques presented in the tutorial can serve as potent mechanisms for understanding,

explaining and debugging production ML models. Finally, the tutorial will help students comprehend and appreciate the complexities of XAI, going far beyond the toy examples typically covered in a classroom setting. We will use running examples to demonstrate the advantages and limitations of different approaches in a hands-on fashion. All the materials that will be used for the tutorial will be publicly available.

Prerequisites. The background expected is that of an introductory ML course covering supervised ML and optimization techniques. The tutorial has been carefully structured to accommodate both attendees unfamiliar with the topic and experienced participants by providing required background knowledge, shared terminology and common understanding of the basic concepts in XAI.

Intended Duration. We are aiming for a 1.5-hour tutorial.

5 TUTORIAL PRESENTERS

Romila Pradhan is an Assistant Professor at Purdue University. Her research interests lie in data management with an emphasis on building trustworthy and responsible data management systems. She received her undergraduate degree in Mathematics and Computing from the Indian Institute of Technology, Kharagpur, and the PhD degree in Computer Science from Purdue University.

Aditya Lahiri is a Masters student at University of California, San Diego, majoring in Computer Science with a specialisation in Machine Learning and Artificial Intelligence. His interests lie in explainable ML and causal inference. He received his undergraduate degree from BITS Pilani, Goa. Post that, he worked at American Express, AI Labs for two years on problems related to ensemble algorithms, explainable AI and NLP.

Sainyam Galhotra is a Computing Innovation Postdoctoral researcher at University of Chicago. He received his Phd from University of Massachusetts Amherst in 2021. Previously, he was a researcher at Xerox Research and received his Bachelor's degree in computer science from Indian Institute of Technology, Delhi in 2014. His research interests span the area of Data Management with a focus towards building equitable systems. He is a recipient of the Best Paper Award in FSE 2017 and Most Reproducible Paper Award in SIGMOD 2017 and 2018. He is a DAAD AInet Fellow and the first recipient of the Krithi Ramamritham Award at UMass for contribution to database research.

Babak Salimi is an Assistant Professor at the Halıcıoğlu Data Science Institute at the University of California at San Diego (UCSD). He is also affiliated with the Database and AI Labs of the UCSD Computer Science and Engineering department. His research spans causal inference and responsible data management, and fairness and transparency. Salimi has made several contributions to the understanding of various aspect of trustworthy data analysis, including explainability, fairness (won the SIGMOD'19 best paper award), reliability (won the VLDB'18 best Demo award), robustness [20, 56–58].

REFERENCES

[1] 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). (2016).

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 9525–9536. https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html

[3] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*. 207–216.

[4] Rakesh Agrawal, Ramakrishnan Srikant, et al. 1994. Fast algorithms for mining association rules. PVLDB.

[5] Emanuele Albini, Jason Long, Danial Dervovic, and Daniele Magazzeni. 2021. Counterfactual Shapley Additive Explanations. *arXiv preprint arXiv:2110.14270* (2021).

[6] Marcelo Arenas, Pablo Barceló, Leopoldo E. Bertossi, and Mikaël Monet. 2021. The Tractability of SHAP-Score-Based Explanations for Classification over Deterministic and Decomposable Boolean Circuits. In *AAAI*. 6670–6678. https://ojs.aaai.org/index.php/AAAI/article/view/16825

[7] Alexander Ratner Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid Training Data Creation with Weak Supervision. *PVLDB* 11, 3 (2017).

[8] Samyadeep Basu, Xuchen You, and Soheil Feizi. 2020. On Second-Order Group Influence Functions for Black-Box Predictions. In *International Conference on Machine Learning*. PMLR, 715–724.

[9] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*. 1–14.

[10] Jonathan Brophy. 2020. *Exit Through the Training Data: A Look into Instance-Attribution Explanations and Efficient Data Deletion in Machine Learning*. Technical Report. Available at https://www.cs.uoregon.edu/Reports/AREA-202009-Brophy.pdf.

[11] Dennis R. Cook and Sanford Weisberg. 1980. Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression. *Technometrics* 22 (1980).

[12] Adnan Darwiche and Auguste Hirth. 2020. On The Reasons Behind Decisions. *arXiv preprint arXiv:2002.09284* (2020).

[13] Adnan Darwiche and Judea Pearl. 1994. Symbolic causal networks. In *AAAI*. 238–244.

[14] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE symposium on security and privacy*. 598–617.

[15] Johan De Kleer, Alan K Mackworth, and Raymond Reiter. 1992. Characterizing diagnoses and systems. *Artificial intelligence* 56, 2-3 (1992), 197–222.

[16] Lydia de la Torre. 2018. A guide to the california consumer privacy act of 2018. *Available at SSRN 3275571* (2018).

[17] Daniel Deutch, Nave Frost, Amir Gilad, and Oren Sheffer. 2021. Explanations for Data Repair Through Shapley Values. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 362–371.

[18] Christopher Frye, Colin Rowat, and Ilya Feige. 2019. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *arXiv preprint arXiv:1910.06358* (2019).

[19] Sainyam Galhotra, Behzad Golshan, and Wang-Chiew Tan. 2021. Adaptive rule discovery for labeling text data. In *Proceedings of the 2021 International Conference on Management of Data*. 2217–2225.

[20] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. 2021. Explaining black-box algorithms using probabilistic contrastive counterfactuals. In *Proceedings of the International Conference on Management of Data*. 577–590.

[21] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. 2021. Feature Attribution and Recourse via Probabilistic Contrastive Counterfactuals.

[22] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3681–3688.

[23] Amirata Ghorbani, Michael Kim, and James Zou. 2020. A Distributional Framework For Data Valuation. In *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119. 3535–3544. http://proceedings.mlr.press/v119/ghorbani20a.html

[24] Amirata Ghorbani and James Y. Zou. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97. 2242–2251.

[25] Joseph Y Halpern and Judea Pearl. 2005. Causes and explanations: A structural-model approach. Part II: Explanations. *The British journal for the philosophy of science* 56, 4 (2005), 889–911.

[26] Jiawei Han, Jian Pei, and Micheline Kamber. 2011. *Data mining: concepts and techniques*. Elsevier.

[27] Jiawei Han, Jian Pei, and Yiwen Yin. 2000. Mining frequent patterns without candidate generation. *ACM sigmod record* 29, 2 (2000), 1–12.

[28] Leo A Harrington, Michael D Morley, A Ščedrov, and Stephen G Simpson. 1985. *Harvey Friedman's research on the foundations of mathematics*. Elsevier.

[29] Melanie Herschel, Ralf Diestelkämper, and Houssem Ben Lahmar. 2017. A Survey on Provenance: What for? What Form? What From? *The VLDB Journal* 26, 6 (Dec. 2017), 881–906. https://doi.org/10.1007/s00778-017-0486-1

[30] Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. 2020. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *arXiv preprint arXiv:2011.01625* (2020).

[31] Mark Hopkins and Judea Pearl. 2003. Clarifying the usage of structural models for commonsense causal reasoning. In *Proceedings of the AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*. AAAI Press Menlo Park, CA, 83–89.

[32] Alexey Ignatiev. 2020. Towards Trustable Explainable AI.. In *IJCAI*. 5154–5158.

[33] Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685* (2020).

[34] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J. Spanos. 2019. Towards Efficient Data Valuation Based on the Shapley Value. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 89)*. PMLR, 1167–1176. https://proceedings.mlr.press/v89/jia19a.html

[35] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 353–362.

[36] Atoosa Kasirzadeh. 2021. Reasons, Values, Stakeholders: A Philosophical Framework for Explainable Artificial Intelligence. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 14–14.

[37] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea). 3819–3828. https://doi.org/10.1145/2702123.2702520

[38] Jakko Kemper and Daan Kolkman. 2019. Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society* 22, 14 (2019), 2081–2096.

[39] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*. PMLR, 1885–1894.

[40] I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. 2020. Problems with Shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*. PMLR, 5491–5500.

[41] Yongchan Kwon, Manuel A. Rivas, and James Zou. 2021. Efficient Computation and Analysis of Distributional Shapley Values. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*. 793–801. http://proceedings.mlr.press/v130/kwon21a.html

[42] Aditya Lahiri and Narayanan Unny Edakunni. 2020. Accurate and Intuitive Contextual Explanations using Linear Model Trees. *arXiv preprint arXiv:2009.05322* (2020).

[43] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.

[44] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology* 31, 4 (2018), 611–627.

[45] David Lewis. 1973. Counterfactuals and comparative possibility. In *Ifs*. Springer, 57–85.

[46] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* 2, 1 (2020), 56–67.

[47] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*. 4768–4777.

[48] Divyat Mahajan, Chenhao Tan, and Amit Sharma. 2019. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277* (2019).

[49] Alexandra Meliou, Wolfgang Gatterbauer, Katherine F. Moore, and Dan Suciu. 2010. WHY SO? or WHY NO? Functional Causality for Explaining Query Answers. In *Proceedings of the Fourth International VLDB workshop on Management of Uncertain Data (MUD 2010) in conjunction with VLDB 2010, Singapore, September 13, 2010 (CTIT Workshop Proceedings Series, Vol. WP10-04)*, Ander de Keijzer and Maurice van Keulen (Eds.). Centre for Telematics and Information Technology (CTIT), University of Twente, The Netherlands, 3–17. http://ewi1276.ewi.utwente.nl:3000/papers/MUD2010_whyso.pdf

[50] Christoph Molnar. 2020. *Interpretable Machine Learning*. Lulu. com.

[51] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 607–617.

[52] Supun Nakandala, Kabir Nagrecha, Arun Kumar, and Yannis Papakonstantinou. 2020. Incremental and Approximate Computations for Accelerating Deep CNN Inference. *ACM Trans. Database Syst.* 45, 4 (2020), 16:1–16:42. https://doi.org/10.1145/3397461

[53] 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.

[54] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[55] 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* (Snowbird, Utah, USA) *(SIGMOD '14)*. Association for Computing Machinery, New York, NY, USA, 1579–1590. https://doi.org/10.1145/2588555.2588578

[56] Babak Salimi, Corey Cole, Peter Li, Johannes Gehrke, and Dan Suciu. 2018. HypDB: a demonstration of detecting, explaining and resolving bias in OLAP queries. *Proceedings of the VLDB Endowment* 11, 12 (2018), 2062–2065.

[57] Babak Salimi, Harsh Parikh, Moe Kayali, Lise Getoor, Sudeepa Roy, and Dan Suciu. 2020. Causal Relational Learning. In *Proceedings of the International Conference on Management of Data*. 241–256.

[58] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*. ACM, 793–810.

[59] Sebastian Schelter, Stefan Grafberger, and Ted Dunning. 2021. HedgeCut: Maintaining Randomised Trees for Low-Latency Machine Unlearning. In *Proceedings of the 2021 International Conference on Management of Data*. 1545–1557.

[60] Maximilian Schleich, Zixuan Geng, Yihong Zhang, and Dan Suciu. 2021. GeCo: Quality Counterfactual Explanations in Real Time. *Proc. VLDB Endow.* 14, 9 (2021), 1681–1693. http://www.vldb.org/pvldb/vol14/p1681-schleich.pdf

[61] Patrick Schwab and Walter Karlen. 2019. Cxplain: Causal explanations for model interpretation under uncertainty. *arXiv preprint arXiv:1910.12336* (2019).

[62] Moshe Sebag, Benny Kimelfeld, Leopoldo Bertossi, and Ester Livshits. 2021. The Shapley Value of Tuples in Query Answering. *Logical Methods in Computer Science* 17 (2021).

[63] Lloyd S Shapley. 1953. *A value for n-person games*. Princeton University Press.

[64] Boris Sharchilev, Yury Ustinovsky, Pavel Serdyukov, and M. de Rijke. 2018. Finding Influential Training Samples for Gradient Boosted Decision Trees. In *ICML*.

[65] Andy Shih, Arthur Choi, and Adnan Darwiche. 2018. A symbolic approach to explaining bayesian network classifiers. *arXiv preprint arXiv:1805.03364* (2018).

[66] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 180–186.

[67] Dylan Slack, Sophie Hilgard, Himabindu Lakkaraju, and Sameer Singh. 2021. Counterfactual Explanations Can Be Manipulated. *arXiv preprint arXiv:2106.02666* (2021).

[68] Kacper Sokol, Alexander Hepburn, Raul Santos-Rodriguez, and Peter Flach. 2019. bLIMEy: surrogate prediction explanations beyond LIME. *arXiv preprint arXiv:1910.13016* (2019).

[69] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 10–19.

[70] Guy Van den Broeck, Anton Lykov, Maximilian Schleich, and Dan Suciu. 2021. On the tractability of SHAP explanations. In *AAAI*.

[71] Paroma Varma and Christopher Ré. 2018. Snuba: Automating weak supervision to label training data. In *PVLDB*, Vol. 12. NIH Public Access, 223.

[72] Tom Vermeire and David Martens. 2020. Explainable Image Classification with Evidence Counterfactual. *arXiv preprint arXiv:2004.07511* (2020).

[73] Giorgio Visani, Enrico Bagli, Federico Chesani, Alessandro Poluzzi, and Davide Capuzzo. 2020. Statistical stability indices for LIME: obtaining reliable explanations for machine learning models. *Journal of the Operational Research Society* (2020), 1–11.

[74] Jiaxuan Wang, Jenna Wiens, and Scott Lundberg. 2021. Shapley flow: A graph-based approach to interpreting model predictions. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 721–729.

[75] David Watson, Limor Gultchin, Ankur Taly, and Luciano Floridi. 2021. Local explanations via necessity and sufficiency: unifying theory and practice. *arXiv preprint arXiv:2103.14651* (2021).

[76] Weiyuan Wu, Lampros Flokas, Eugene Wu, and Jiannan Wang. 2020. Complaint-driven Training Data Debugging for Query 2.0. In *Proceedings of the International Conference on Managment of Data*.

[77] Yinjun Wu, V. Tannen, and S. Davidson. 2020. PrIU: A Provenance-Based Approach for Incrementally Updating Regression Models. *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* (2020).