# A Data Management Approach to Explainable AI

Marcelo Arenas
marenas@ing.puc.cl
Pontificia Universidad Católica de Chile and IMFD Chile
RelationalAI, USA

## ABSTRACT

In recent years, there has been a growing interest in developing methods to explain individual predictions made by machine learning models. This has led to the development of various notions of explanation and scores to justify a model's classification. However, instead of struggling with the increasing number of such notions, one can turn to an old tradition in databases and develop a declarative query language for interpretability tasks, which would allow users to specify and test their own explainability queries. Not surprisingly, logic is a suitable declarative language for this task, as it has a well-understood syntax and semantics, and there are many tools available to study its expressiveness and the complexity of the query evaluation problem. In this talk, we will discuss some recent work on developing such a logic for model interpretability.

## CCS CONCEPTS

• **Information systems** → **Query languages for non-relational engines**; • **Computing methodologies** → **Artificial intelligence**.

## KEYWORDS

Explainable artificial intelligence, query language, explainability language

**ACM Reference Format:**
Marcelo Arenas. 2024. A Data Management Approach to Explainable AI. In *Companion of the 43rd Symposium on Principles of Database Systems (PODS Companion '24), June 9–15, 2024, Santiago, AA, Chile.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3635138.3654762

## 1 MOTIVATION

The growing necessity to understand the reasoning behind decisions made by machine learning (ML) models has catalyzed significant research in explainable AI (XAI) methods [27]. This research has led to the development of various queries and metrics designed to understand the individual predictions of these models. For instance, a number of techniques have been developed to assess the impact of one or more features on the output of an ML model. These methods help users identify the key features that predominantly influence the model's decision regarding a specific input [14, 25, 28].

Nevertheless, it is frequently not a single query or metric, but rather a combination of them, that yields the most comprehensive explanation [12, 26]. Additionally, research has demonstrated that some explainability metrics, despite being considered theoretically sound and robust, can exhibit unexpected behavior under certain circumstances [7, 16, 17, 19, 24, 31]. These findings have prompted the proposal for the development of "explainability languages". These general-purpose languages would provide users with the flexibility to interact with an ML model by posing various queries in pursuit of the optimal explanation.

Building on the momentum about the potential of explainability languages to enhance interactions with ML models, the data management community is well-positioned to make substantial contributions. Their extensive background in query language development offers significant insights into structuring complex systems that are both efficient and user-friendly. As we explore the introduction of an explainability language designed to facilitate understanding of ML decisions, defining a clear set of criteria—or desiderata—becomes crucial. These criteria should draw from proven strategies in database query development to ensure the language is robust, accessible, and effective. Below are key points of these desiderata, outlining the fundamental features and capabilities an ideal explainability language should possess:

- **Declarative:** The language should allow users to articulate what explanation they need without detailing the computational method to achieve it. This characteristic enables users to focus on interpreting results rather than navigating complex computational processes, making the system easier to use.
- **Simple syntax and semantics:** The explainability language should be built with simple syntax and semantics, leveraging well-known database query languages. This simplicity will make the language easier to learn and use, broadening its appeal to a diverse range of users, including those without specialized knowledge in machine learning.
- **Specific query capability for explainability:** The language must have the capability to consistently define explanation concepts across different models, irrespective of their size or the type of classification model employed. This crucial feature should guarantee that a given explanation concept can be effectively represented by a single, fixed query, independent of any specific characteristics of the model. This ensures the language's adaptability and effectiveness across a broad spectrum of applications.
- **Expressiveness:** To effectively serve its purpose, the language must be able to represent a wide array of common explanation concepts [2, 4, 11, 15, 18, 21, 30].
- **Exploratory operators:** Including operators that enable exploration within a model is crucial. These operators should


*PODS Companion '24, June 9–15, 2024, Santiago, AA, Chile*

ACM ISBN 979-8-4007-0483-3/24/06
https://doi.org/10.1145/3635138.3654762

facilitate the investigation of various aspects of an explanation, allowing users to refine their understanding incrementally. For instance, a classification model may yield a large number of possible explanations for a given result. The language should allow users to retrieve these explanations, filtering them according to various criteria.

- **Combination of explanations:** The language should support the combination of different explanation approaches through specific operators. This capability allows for richer, more nuanced insights by integrating multiple explanatory perspectives.
- **Efficient data complexity:** As an explanation notion should be representable in the language by a fixed query, the appropriate way to measure the evaluation complexity of the language is through the concept of data complexity [32]. Although we expect the data complexity of certain fragments of the language to be polynomial, we must be more permissive given the inherently high complexity of certain explanation tasks [3, 4]. In particular, a data complexity such as $P^{NP}$ would still be desirable, as this level would enable the use of SAT solvers for query evaluation. SAT solvers are a mature technology that has proven effective in computing explanations for various ML models [20, 23, 33].
- **Verification versus computation:** Beyond the ability to efficiently verify whether a possible explanation indeed meets certain criteria, it should also be feasible to compute these explanations efficiently. In this regard, we expect the language to demonstrate efficient data complexity for both the verification and computation problems.

Several questions arise from the previous desiderata. Should the language be model agnostic, so that it does not depend on the specific type of ML model being employed? If this is the case, the language can be used to provide explanations for any ML model, treating it merely as a black box. However, this approach will inevitably lead to higher complexity, as it limits the ability to develop more efficient evaluation algorithms for the language that are tailored to specific features of the ML model being used.

If we move away from the model-agnostic approach, which ML models should we consider? A natural starting point would be to focus on decision trees, as they have been extensively studied for explainability in the literature [2–4, 21–23]. Following this, various forms of decision diagrams and circuits could be considered, particularly ordered binary decision diagrams (OBDDs), which, along with decision trees, are regarded as easily interpretable [8, 10, 13, 27, 29]. Other more expressive forms of decision diagrams and circuits with advantageous properties should also be explored, especially those that are decomposable and deterministic [1], if we aim to achieve efficient data complexity. Clearly, many more alternatives should be explored, particularly probabilistic ML models [9].

How should an explanation be presented to the user, and how can it be proved that such an explanation is trustworthy? The database community has much to contribute here, particularly since many concepts developed in this area, such as data provenance [5, 6], could be helpful in addressing these questions. It is also important to note that different levels of detail may be required by different users; this too needs to be considered when responding to a query in the explainability language.

We are convinced that incorporating the aforementioned desiderata and addressing the aforementioned questions could be instrumental in the development of powerful explainability languages that enhance the transparency and accessibility of ML models. This would help bridge the gap between complex algorithmic decisions and actionable, understandable insights. In this talk, we will discuss some recent work on developing such an explainability language that tries to meet the criteria discussed in this article.

## ACKNOWLEDGMENTS

The work presented in this talk was partially funded by the ANID - Millennium Science Initiative Program under Code ICN17002 and was developed in part while the author was visiting the Simons Institute for the Theory of Computing in Berkeley, CA, USA.

## REFERENCES

[1] Antoine Amarilli, Marcelo Arenas, YooJung Choi, Mikaël Monet, Guy Van den Broeck, and Benjie Wang. 2024. A Circus of Circuits: Connections Between Decision Diagrams, Circuits, and Automata. *arXiv preprint arXiv:2404.09674* (2024).

[2] Marcelo Arenas, Daniel Baez, Pablo Barceló, Jorge Pérez, and Bernardo Subercaseaux. 2021. Foundations of Symbolic Languages for Model Interpretability. In *NeurIPS 2021*. 11690–11701.

[3] Marcelo Arenas, Pablo Barceló, Miguel Romero Orth, and Bernardo Subercaseaux. 2022. On Computing Probabilistic Explanations for Decision Trees. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 28695–28707. https://proceedings.neurips.cc/paper_files/paper/2022/file/b8963f6a0a72e686dfa98ac3e7260f73-Paper-Conference.pdf

[4] Pablo Barceló, Mikaël Monet, Jorge Pérez, and Bernardo Subercaseaux. 2020. Model Interpretability through the lens of Computational Complexity. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 15487–15498. https://proceedings.neurips.cc/paper/2020/hash/b1adda14824f50ef24ff1c05bb66faf3-Abstract.html

[5] Peter Buneman, Sanjeev Khanna, and Wang Chiew Tan. 2001. Why and Where: A Characterization of Data Provenance. In *Database Theory - ICDT 2001, 8th International Conference (Lecture Notes in Computer Science, Vol. 1973)*. Springer, 316–330.

[6] Peter Buneman and Wang-Chiew Tan. 2018. Data Provenance: What next? *SIGMOD Rec.* 47, 3 (2018), 5–16.

[7] Oana-Maria Camburu, Eleonora Giunchiglia, Jakob N. Foerster, Thomas Lukasiewicz, and Phil Blunsom. 2019. Can I Trust the Explainer? Verifying Post-hoc Explanatory Methods. *CoRR* abs/1910.02065 (2019).

[8] Hei Chan and Adnan Darwiche. 2003. Reasoning about Bayesian Network Classifiers. In *UAI '03, Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 107–115.

[9] Y Choi, Antonio Vergari, and Guy Van den Broeck. 2020. Probabilistic circuits: A unifying framework for tractable probabilistic models. *UCLA. URL: https://starai.cs.ucla.edu/papers/ProbCirc20.pdf* (2020), 6.

[10] Karine Chubarian and György Turán. 2020. Interpretability of Bayesian Network Classifiers: OBDD Approximation and Polynomial Threshold Functions. In *International Symposium on Artificial Intelligence and Mathematics*.

[11] Adnan Darwiche and Auguste Hirth. 2020. On the Reasons Behind Decisions.. In *ECAI*. 712–720. https://doi.org/10.3233/FAIA200158

[12] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 [stat.ML]

[13] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael A. Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *5th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2018*. IEEE, 80–89.

[14] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5 (2019), 93:1–93:42.

[15] Xuanxiang Huang, Martin C. Cooper, António Morgado, Jordi Planes, and João Marques-Silva. 2023. Feature Necessity & Relevancy in ML Classifier Explanations.. In *ETAPS*. 167–186. https://doi.org/10.1007/978-3-031-30823-9_9

[16] Xuanxiang Huang and João Marques-Silva. 2023. The Inadequacy of Shapley Values for Explainability. *CoRR* abs/2302.08160 (2023).

[17] Alexey Ignatiev. 2020. Towards Trustable Explainable AI. In *IJCAI*, Christian Bessiere (Ed.). ijcai.org, 5154–5158.
[18] Alexey Ignatiev, Nina Narodytska, and João Marques-Silva. 2019. Abduction-Based Explanations for Machine Learning Models. In *AAAI*. AAAI Press, 1511–1519.
[19] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. 2019. On Validating, Repairing and Refining Heuristic ML Explanations. *CoRR* abs/1907.02509 (2019).
[20] Alexey Ignatiev and João P. Marques Silva. 2021. SAT-Based Rigorous Explanations for Decision Lists. In *SAT (LNCS, Vol. 12831)*, Chu-Min Li and Felip Manyà (Eds.). Springer, 251–269.
[21] Yacine Izza, Alexey Ignatiev, and João Marques-Silva. 2020. On Explaining Decision Trees. *CoRR* abs/2010.11034 (2020).
[22] Yacine Izza, Alexey Ignatiev, and João Marques-Silva. 2022. On Tackling Explanation Redundancy in Decision Trees. *J. Artif. Intell. Res.* 75 (2022), 261–321.
[23] Yacine Izza and João Marques-Silva. 2021. On Explaining Random Forests with SAT. In *IJCAI*, Zhi-Hua Zhou (Ed.). ijcai.org, 2584–2591.
[24] I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle A. Friedler. 2020. Problems with Shapley-value-based explanations as feature importance measures. In *ICML*. 5491–5500.
[25] João Marques-Silva. 2022. Logic-Based Explainability in Machine Learning. *CoRR* abs/2211.00541 (2022).
[26] Joao Marques-Silva and Alexey Ignatiev. 2023. No silver bullet: interpretable ML models must be explained. *Frontiers in Artificial Intelligence* 6 (Apr 2023), 1128212. https://doi.org/10.3389/frai.2023.1128212
[27] Christoph Molnar. 2022. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book
[28] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *AAAI*. 1527–1535.
[29] 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.
[30] Andy Shih, Arthur Choi, and Adnan Darwiche. 2018. A symbolic approach to explaining Bayesian network classifiers. *arXiv preprint arXiv:1805.03364* (2018).
[31] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *AIES*. 180–186.
[32] Moshe Y. Vardi. 1982. The Complexity of Relational Query Languages (Extended Abstract). In *Proceedings of the 14th Annual ACM Symposium on Theory of Computing*. ACM, 137–146.
[33] Jinqiang Yu, Alexey Ignatiev, Peter J. Stuckey, and Pierre Le Bodic. 2020. Computing Optimal Decision Sets with SAT. In *CP (LNCS, Vol. 12333)*, Helmut Simonis (Ed.). Springer, 952–970.