Zorro: Quantifying Uncertainty in Models & Predictions Arising from Dirty Data

Kaiyuan Hu
University of California, San Diego
USA
kah032@ucsd.edu

Jiongli Zhu
University of California, San Diego
USA
jiz143@ucsd.edu

Boris Glavic
University of Illinois Chicago
USA
bglavic@uic.edu

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

Abstract

Machine learning models are increasingly employed in high-stakes decision making in domains such as personalized medicine, policing, and many others. As data quality issues are prevalent and recovering a ground truth clean dataset is often impossible or prohibitively expense, heuristic cleaning techniques are employed in practice to clean training data. The net result are models whose predictions can fundamentally not be trusted as we do not know how much the model's predictions differ from a model trained on the unknown ground truth clean data. We present Zorro, a principled framework for modeling the uncertainty in model parameters and predictions arising from the multiplicity of datasets that could feasibly be the ground truth clean version of a dirty training / test dataset. Under the hood, Zorro employs a novel framework for training and prediction with linear models over uncertain data. Given training and test datasets that are subject to data quality issues, we compute a sound over-approximation of all possible models, the set of models generated by training a model on each possible clean version of the dataset, and then over-approximate all possible predictions based on these models. Using Zorro, we can certify the robustness of models, i.e., to what degree are the model parameters impacted by data quality issues, and of individual and aggregated predictions. The demonstration video is available at https://drive.google.com/file/d/1gxkvRY3pLM0ATco2qvBnM1EcR8-U9F5h/view?usp=sharing.

CCS Concepts

• **Information systems** → **Data cleaning**; **Data analytics**.

Keywords

Data debugging, Explanations, Interpretability, Fairness

ACM Reference Format:
Kaiyuan Hu, Jiongli Zhu, Boris Glavic, and Babak Salimi. 2025. Zorro: Quantifying Uncertainty in Models & Predictions Arising from Dirty Data. In *Companion of the 2025 International Conference on Management of Data (SIGMOD-Companion '25), June 22–27, 2025, Berlin, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3722212.3725143

SIGMOD-Companion'25, Berlin, Germany

ACM ISBN 979-8-4007-1564-8/2025/06
https://doi.org/10.1145/3722212.3725143

1 Introduction

Most real-world datasets contain data quality issues such as missing values, outliers, and constraint violations, which can significantly impact downstream analyses and machine learning models [1, 8, 13]. Since recovering the ground truth dataset is often infeasible, practitioners rely on heuristic data cleaning techniques that correct errors, e.g., choosing a repair that best preserves statistical properties or requires minimal modifications to the input data [3, 10, 11]. However, different cleaning strategies can lead to substantially different versions of the data, resulting in models whose predictions may diverge significantly from those trained on the unknown ground truth [6]. This uncertainty raises fundamental concerns about the reliability and robustness of models trained on cleaned data, particularly in high-stakes applications where prediction errors can have serious consequences.

Example 1. *Consider the training dataset with missing values and noisy labels shown in Figure 1, collected to train a linear model for predicting savings based on age and income. Since training requires a complete dataset, missing values must be handled using techniques such as mean imputation or predictive modeling. However, different imputation strategies yield different cleaned datasets, leading to the training of distinct models, as shown in Figure 1 (b). Consequently, as illustrated in Figure 1 (d), for an individual with age 70 and income 50K, the predictions from two plausible models differ by 2.5K due to the uncertainty introduced by these data quality issues.*

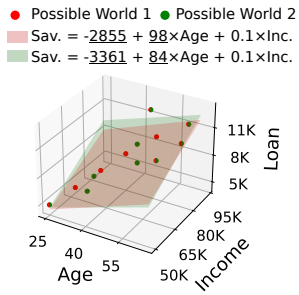
In the example above, the choice of cleaning method significantly affects the predictions made by a model trained on the repaired data. Because it is generally impossible to identify which repair recovers the true clean dataset, models trained on a single cleaned version cannot be fully trusted. What is needed is an approach that computes all possible predictions for any datapoint across all feasible repairs, enabling analysts to assess prediction reliability, determine whether additional data collection is necessary, or escalate uncertain cases to a human expert.

In this work, we demonstrate Zorro [12], a framework that systematically addresses this uncertainty by treating each valid repair of a dirty dataset as a distinct *possible world* resulting in an incomplete database as in consistent query answering [2]. The naive approach for computing all plausible predictions for a test data point is to train a model in each possible world and then compute a prediction with this model. Rather than enumerating every repair —- which is often infeasible when the space of repairs is large or unbounded -— we employ *abstract interpretation* [4, 9], a formal method that compactly represents a set of possible worlds using a single element from an *abstract domain*. Specifically, Zorro uses

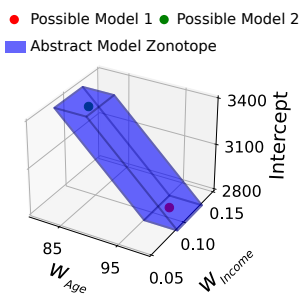
Training Data

Age	Income	Savings
25	50K	5K
NULL	60K	6K
35	NULL	7K
NULL	NULL	[8K,9K]
45	90K	12K
50	NULL	[10K,12K]
55	75K	9K
60	85K	10K
65	80K	13K

(a)



(b)



(c)

Possible Model 1 Predictions

Age	Income	Savings
20	90K	8.1K
70	50K	9K

Possible Model 2 Predictions

Age	Income	Savings
20	90K	7.3K
70	50K	11.5K

Zorro Prediction Ranges

Age	Income	Savings
20	90K	[7K,8.2K]
70	50K	[8.5K,12K]

(d)

Figure 1: A training dataset with missing values (a) used to train a linear regressor predicting `Savings`; (b) shows two example worlds and the corresponding linear models; (c) shows a zonotope that over-approximates the set of models based on the possible worlds of the input; and (d) shows possible model predictions (red and green), and prediction intervals computed by Zorro that bound all possible prediction outcomes (blue).

zonotopes, a special class of convex polytopes that provide a compact symbolic representation of uncertainty, to over-approximate the sets of repairs, model parameters, and predictions. Instead of training separate models for each possible repair, Zorro performs *symbolic gradient descent* within the zonotope domain, simultaneously learning an over-approximation of all possible models. This allows Zorro to efficiently construct a zonotope enclosing all feasible model parameters and, for each test datapoint, return a prediction interval that covers all outcomes consistent with any valid repair. By providing a principled way to reason about uncertainty in model predictions, Zorro enables analysts to make more informed and robust decisions. Zorro is available open-source at https://github.com/lodino/Zorro.

Example 2 (Possible Models and Predictions). *Continuing with Example 1, we run Zorro on the dirty training dataset from Figure 1 (a) and obtain the blue zonotope in Figure 1 (c) that is guaranteed to enclose all possible models, including the two models shown in Figure 1 (b). Given the test dataset from Figure 1 (d), Zorro computes a prediction interval for each data point as shown on the bottom of Figure 1 (d), over-approximating the set of outcomes for the data point over all possible repairs.*

Zorro not only captures uncertainty in model training but also provides a principled way to quantify its impact on predictions (Example 2). Unlike prior work by Karlas et al. [7], which focuses on identifying *certain models* — models that remain unchanged across all possible data repairs — Zorro provides a more general framework by explicitly modeling the full range of plausible model parameters and predictions. This allows analysts to assess reliability of predictions in the presence of data quality issues, both at the granularity of individual data points as well as at the granularity of the whole model, rather than relying on a single, potentially misleading cleaned dataset.
We make the following contributions in this demonstration:

- We present Zorro, a system that models data uncertainty using possible worlds semantics and provides sound over-approximations of model parameters and predictions. Zorro enables users to analyze the impact of data quality issues in a practical, interactive manner.
- Our demonstration allows users to upload datasets with data quality issues, train possible models, and explore prediction ranges to evaluate robustness guarantees. By comparing Zorro's output to models trained on imputed data, we demonstrate the risks of relying solely on heuristic data cleaning techniques and emphasize the necessity of quantifying model and prediction uncertainty in the presence of data quality issues.

2 System Overview

Given a training dataset with quality issues, Zorro systematically explores all *possible repairs* for the data and propagates the resulting uncertainty through the model training process, yielding an over-approximation of the set of *possible models*. Based on these models, Zorro produces a range of *possible predictions*, which quantify the impact of data quality issues on model outputs. These possible predictions serve as robustness certificates, providing a quantified measure of prediction reliability in the face of data uncertainty.

Possible World Semantics. The notion of *possible world semantics*, which has been studied extensively by the database and AI communities, is a foundational tool for modeling uncertainty arising from data issues such as missing values and outliers: a dataset with quality issues can be represented by a set of *possible clean datasets* called *repairs*, e.g., for missing values these are all datasets generated by replacing each missing value with some domain value. Applying a learning algorithm to each possible dataset yields a set of *possible models*. Predictions for a given input are derived from these models, resulting in a range of outcomes reflecting the uncertainty in predictions. The most straightforward approach to this is training one model for each possible world, which is computationally infeasible, especially as the number of possible worlds is typically exponential in the number of uncertain data points. We address this challenge through *abstract interpretation*, a technique widely used in the formal verification of neural networks and control systems [4, 9].

Abstract Interpretation. Abstract interpretation provides a framework for compactly over-approximating a set of possible worlds using a single element from an abstract domain and for evaluating computations in the abstract domain that preserve this over-approximation. The abstract domain that Zorro employs is *zonotopes*, a type of convex polytope well-suited for representing

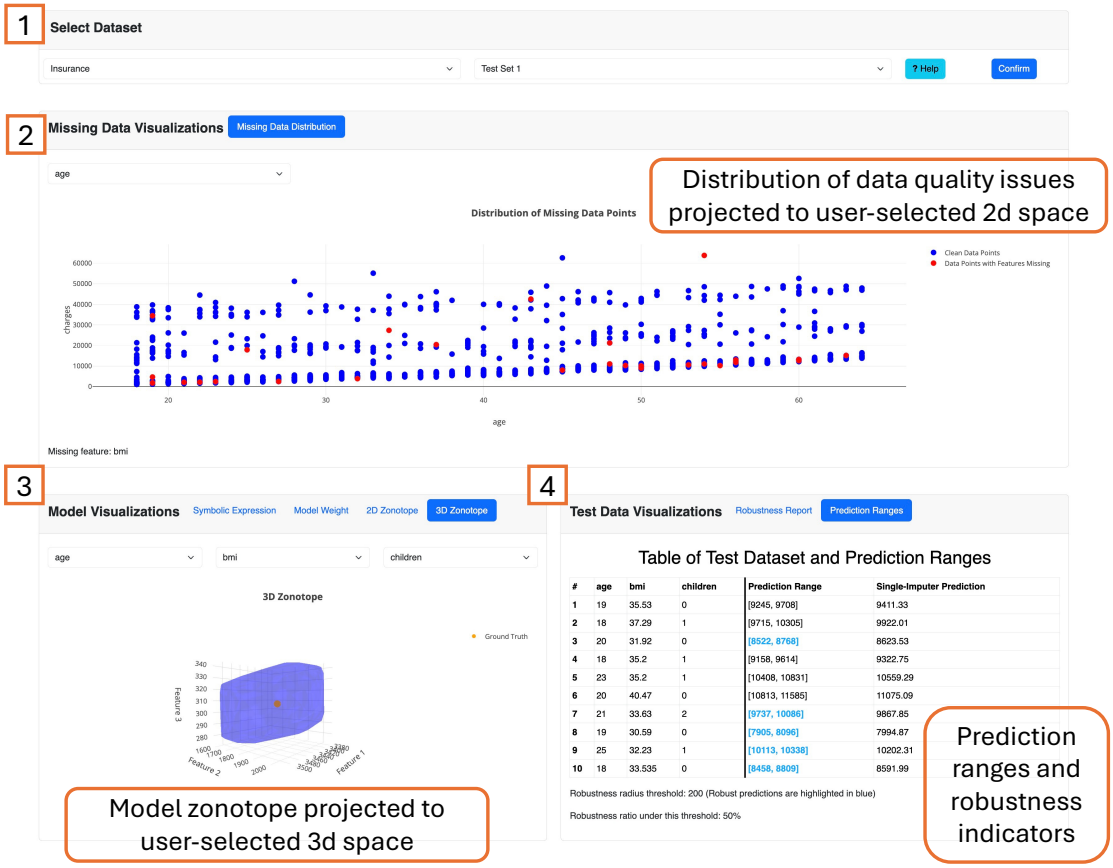


Figure 2: Zorro's UI. The user selects training and test datasets with errors, and one of the supported classifiers. Zorro then computes and visualizes the set of possible models as a zonotope and computes prediction intervals for each test data point.

high-dimensional uncertainty in a compact, symbolic form. For example, an uncertain training dataset with m features and n data points is represented as a $n \cdot m$ zonotope. Each dimension of a such zonotope is a linear combination of variables called *error terms*. A zonotope represents a set of possible worlds, each derived by assigning each error term a value in $[-1, 1]$ and evaluating these linear expressions. We demonstrate in [12] how to encode the possible repairs for a wide variety of data quality issues using zonotopes. For instance, assume that income ranges from $0K$ to $200K$ and age ranges from 18 to 120. Consider the 3rd and 4th row from Figure 1 (a) and a zonotope encoding their possible repairs:

Zonotope Encoding Possible Repairs

Age	Income	Savings
$69 + 51 \cdot \epsilon_1$	60K	6K
$69 + 51 \cdot \epsilon_2$	$100K + 100K \cdot \epsilon_3$	$8.5K + 0.5K \cdot e_4$

Rather than training a separate model for each possible dataset, Zorro performs symbolic execution of gradient descent using zonotopes to encode the training data and model weights, which is guaranteed to *over-approximate* the set of possible model parameters in each step. Symbolic execution of gradient descent, however, introduces new challenges. During the iterative process, the multiplication between symbolic expressions leads to an exponential growth of polynomial terms in the symbolic expressions representing the model parameters, which makes the computation intractable. To address this, Zorro employs two over-approximation techniques: *linearization*, which approximates higher-order symbolic terms with linear ones, and *order reduction*, which simplifies the representation by reducing the number of error terms without losing much precision. These techniques ensure computational feasibility, enabling Zorro to maintain scalability without compromising theoretical soundness or robustness guarantees.

A Closed-form Solution for Symbolic Fixed-points. While linearization and order reduction effectively accelerate computation, they also introduce over-approximation errors that can accumulate during the iterative gradient descent process. In some cases, the accumulated errors cause the model weight zonotope to diverge. Determining whether a fixed point exists -— where the zonotope representation of model weights stabilizes and stops evolving -— is challenging. To address this, [12] establishes a sufficient condition under which the fixed point is guaranteed to exist. Building on this sufficient condition, Zorro uses a novel closed-form solution for the fixed point of abstract gradient descent, which translates to solving a system of linear equations, and can be performed efficiently. In addition to improved computational efficiency, using a closed form solution has the additional benefit of circumventing the accumulation of over-approximation errors inherent to iterative gradient descent in the abstract domain.

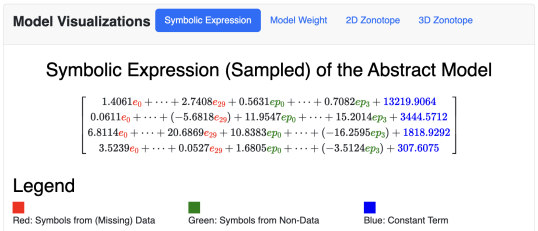


Figure 3: Symbolic model weights returned by Zorro.

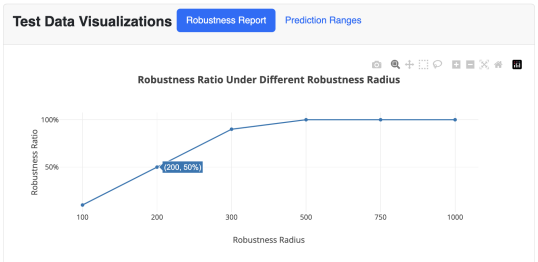


Figure 4: Robustness ratio when varying robustness radius.

3 Demonstration Details

Dataset Selection. We will demonstrate Zorro using several datasets including a US Health Insurance dataset [5] that records health insurance charges for 1,338 people based on their health and family information. We will start by guiding the user through training and inference with Zorro using this dataset and use an example test dataset to demonstrate how to use Zorro for evaluating the robustness of predictions on individual data points. Users may also upload new training and test datasets (Figure 2 ①). To help the user understand how data errors are distributed in the dataset, Zorro allows the user to select two columns and visualizes how missing values are distributed in these two columns. (Figure 2 ②).

Model Uncertainty. The core idea of Zorro is to over-approximate the set of possible models using a zonotope. We provide several visualizations for the uncertain model parameters. Projections of the zonotope onto 1D, 2D, and 3D subspaces empower the user to understand which model parameters are more or less affected by uncertainty in the training data and how individual uncertain model parameters are correlated (Figure 2 ③).

Users can also explore the symbolic form of a zonotope, represented by a vector of linear symbolic expressions, as shown in Figure 3. Zorro uses colors to distinguish error terms that exist in the training data (red) and error terms due to over-approximation in the training process (green). This enables the user to understand how uncertainty in the training data affects model parameters and how model parameters are correlated.

Prediction Uncertainty and Evaluating Robustness. Given the model zonotope, inference with Zorro amounts to computing prediction intervals covering all possible predictions for a set of test data points based on uncertainty in the model (Figure 2 ④). Based on the width of the predictions interval for a data point, the user can decide whether to trust the model's prediction for this data point. For the example dataset the task is to predict medical charges. The domain of the predicted feature is [1K, 60K]. Assume for this use case we can tolerate if the predicted charges are off by up to $250. Then as shown in Figure 2 ④, all the predictions highlighted in blue will be considered to be robust while the ones shown in black should not be trusted as they may differ more than $250 from the unknown prediction based on the ground truth training dataset. Intuitively, features for which a test data point has a large value and whose model weights, as shown in Figure 5 using the 1D visualization, have a larger range of possible values have more impact on the size of prediction intervals. In Figure 5 BMI's weight has a higher uncertainty compared to the weight for age. Therefore, tuple # 2, which has a low age = 18 and high BMI = 37.29, has a relatively wide prediction interval of $590. In contrast, tuple # 9 with high age = 25 and low BMI = 32.23, only has a prediction interval of size $225, which is less than 50% of tuple # 2. We also show the predictions of a model trained on a cleaned version of the dataset generating using standard data repair techniques, to demonstrate how much such a prediction may differ from the ground truth prediction contained in the set of possible predictions. As shown in Figure 4, Zorro also computes the aggregated robustness of a model on a test dataset as the fraction of data points with a robust prediction (*robustness ratio*) for a given maximum prediction intervals size that should be considered as robust (the *robustness radius*).

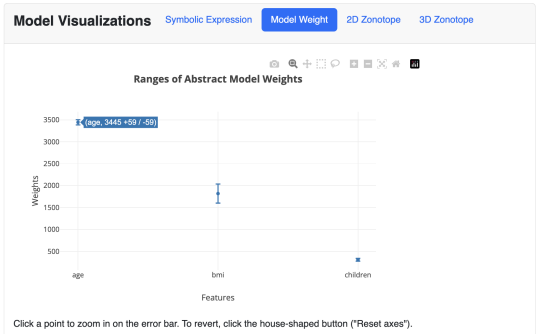


Figure 5: 1D projection of the model weight zonotope encoding the range of possible weights for each feature.

References

[1] Ziawasch Abedjan, Xu Chu, Dong Deng, Raul Castro Fernandez, Ihab F Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker, and Nan Tang. Detecting data errors: Where are we and what needs to be done? *PVLDB*, 9(12):993–1004, 2016.
[2] Leopoldo E. Bertossi. *Database Repairing and Consistent Query Answering*. Morgan & Claypool Publishers, 2011.
[3] Fei Chiang and Renée J Miller. A unified model for data and constraint repair. In *ICDE*, pages 446–457, 2011.
[4] Patrick Cousot and Radhia Cousot. Abstract interpretation: a unified lattice model for static analysis of programs by construction or approximation of fixpoints. In *POPL*, pages 238–252, 1977.
[5] Dheeru Dua and Casey Graff. Uci machine learning repository, 2017.
[6] Tlamelo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago, and Oteng Tabona. A survey on missing data in machine learning. *Journal of Big data*, 8:1–37, 2021.
[7] Bojan Karlas, Peng Li, Renzhi Wu, Nezihe Merve Gürel, Xu Chu, Wentao Wu, and Ce Zhang. Nearest neighbor classifiers over incomplete information: From certain answers to certain predictions. *PVLDB*, 14(3):255–267, 2020.
[8] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.
[9] Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *ICML*, pages 3578–3586, 2018.
[10] Felix Neutatz, Binger Chen, Ziawasch Abedjan, and Eugene Wu. From cleaning before ML to cleaning for ML. *IEEE Data Eng. Bull.*, 44(1):24–41, 2021.
[11] Jef Wijsen. Database repairing using updates. *TODS*, 30(3):722–768, 2005.
[12] Jiongli Zhu, Su Feng, Boris Glavic, and Babak Salimi. Learning from uncertain data: From possible worlds to possible models. In *NeurIPS*, 2024.
[13] Jiongli Zhu and Babak Salimi. Overcoming data biases: Towards enhanced accuracy and reliability in machine learning. *IEEE Data Eng. Bull.*, 47(1):18–35, 2024.