Error Diagnosis and Data Profiling with Data X-Ray

Xiaolan Wang[♣] Mary Feng[♣♭] Yue Wang[♣] Xin Luna Dong[♯] Alexandra Meliou[♣]
[♣]University of Massachusetts
Amherst, MA, USA
{xlwang, yuewang, ameli}@cs.umass.edu
[♭]University of Iowa
Iowa City, IA, USA
mary-feng@uiowa.edu
[♯]Google Inc.
Mountain View, CA, USA
lunadong@google.com

ABSTRACT

The problem of identifying and repairing data errors has been an area of persistent focus in data management research. However, while traditional data cleaning techniques can be effective at identifying several data discrepancies, they disregard the fact that many errors are *systematic*, inherent to the process that produces the data, and thus will keep occurring unless the root cause is identified and corrected.

In this demonstration, we will present a large-scale diagnostic framework called DATAXRAY. Like a medical X-ray that aids the diagnosis of medical conditions by revealing problems underneath the surface, DATAXRAY reveals hidden connections and common properties among data errors. Thus, in contrast to traditional cleaning methods, which treat the symptoms, our system investigates the underlying conditions that cause the errors.

The core of DATAXRAY combines an intuitive and principled cost model derived by Bayesian analysis, and an efficient, highly-parallelizable diagnostic algorithm that discovers common properties among erroneous data elements in a top-down fashion. Our system has a simple interface that allows users to load different datasets, to interactively adjust key diagnostic parameters, to explore the derived diagnoses, and to compare with solutions produced by alternative algorithms. Through this demonstration, participants will understand (1) the characteristics of good diagnoses, (2) how and why errors occur in real-world datasets, and (3) the distinctions with other related problems and approaches.

1. INTRODUCTION

Current trends have seen data grow larger, more intertwined, and more diverse. Retrieving high quality datasets from voluminous and diverse sources is crucial for many data-intensive applications, such as data analysis, search, and strategy planning. However, data often contains errors, different sources may provide conflicting information, and extraction systems have inherent imperfections. As a result, the retrieved datasets often contain noise and other discrepancies, which in turn result in misguided actions [9, 16] and large financial cost [4, 6].

We will demonstrate DATAXRAY [19], a general-purpose, highly-scalable tool that explains why and how errors happen in a data generative process. Data management research has built an extensive arsenal of data cleaning tools [1, 6, 10, 14, 15] that aim to answer questions such as *"Which data is incorrect?"* However, errors in the data are often a symptom of a deeper problem, but traditional techniques offer little to no help with questions such as *"Why are there errors in the data?"* or *"How can I prevent further errors?"* DATAXRAY is fundamentally distinct from these techniques, and supports precisely these questions; its focus is to provide a *diagnosis* of the underlying condition, based on the symptoms (i.e., the errors). This is a crucial distinction, as many errors are *systematic* and will keep occurring until the problems are corrected at their source.

As part of the demonstration, we will present diagnoses that DATAXRAY derives in real-world datasets. For example, we will show how DATAXRAY automatically diagnoses errors in real-world knowledge extraction datasets, pointing to specific mistakes in the extraction systems themselves. In another example, we will show how DATAXRAY can analyze traffic and weather data to diagnose the leading causes of highway accidents. By highlighting that the error rate is high for data items with common properties, we can help users diagnose the possible causes. Using datasets from diverse domains, our demonstration will show that DATAXRAY is a generic framework that goes beyond diagnosis: It is a *data profiling* tool that derives common properties among collections of data items.

Key techniques. Since finding particular causes is often domain-specific, we instead aim to provide a generic approach that finds groupings of errors that may be due to the same cause; such groupings give clues for discerning the underlying problems. DATAXRAY identifies these groups based on their common characteristics, called *features*; e.g., a large group of highway accidents is associated with surface water level of more than 2cm. At the core of DATAXRAY lie three key techniques. (1) DATAXRAY organizes features in a hierarchical structure based on simple containment relationships; this hierarchy is a key to solving the problem efficiently. (2) DATAXRAY uses a top-down, iterative, and highly-parallelizable algorithm to explore the feature hierarchy and identify the set of features that best summarize all erroneous data elements. (3) The traversal algorithm makes local decisions based on a simple, additive cost function that approximates the Bayesian estimate of the probability that a set of features is the cause of the data errors.

Demonstration goals. Through the demonstration, participants will explore the feature hierarchy and the derived diagnoses interactively, starting with small toy examples, and continuing with real-world data from different domains. Our system will highlight the properties of good diagnoses by showing comparisons with results produced by other techniques. Our demonstration will offer insight for the following questions.

- *"What are the differences between diagnosis and cleaning? What do diagnoses look like?"*

Proceedings of the VLDB Endowment, Vol. 8, No. 12

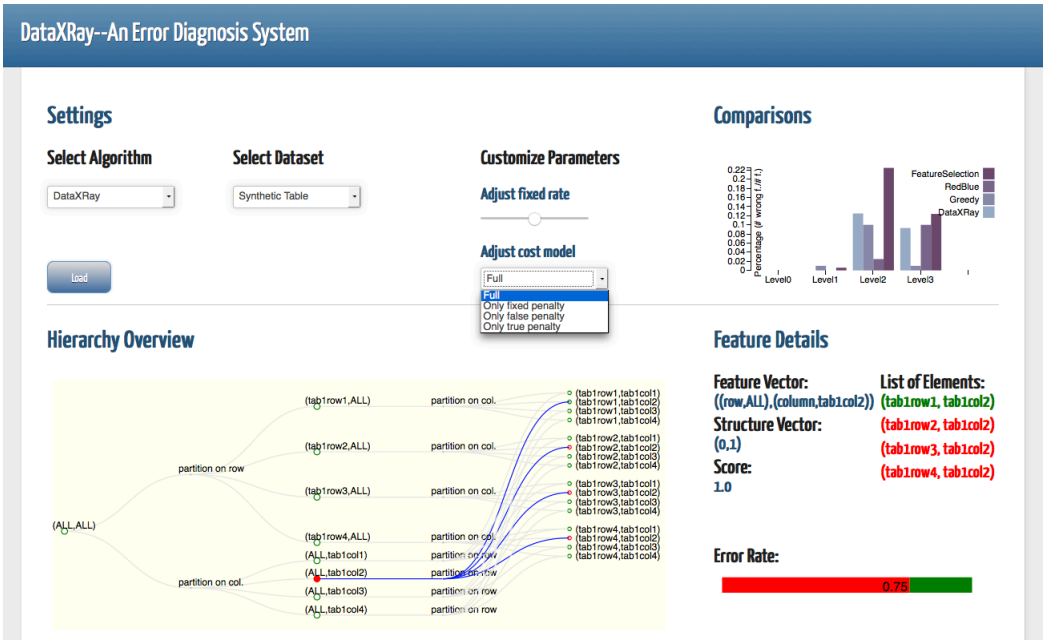

Figure 1: DATAXRAY provides a simple interface that allows users to load different datasets, to interactively adjust key diagnostic parameters, to explore the derived diagnoses, and to compare with solutions produced by alternative algorithms.

- *"How does the feature hierarchy support diagnosis? How is it derived?"*
- *"Which features comprise a diagnosis? Why and how do they explain the data errors?"*
- *"How do changes in the cost model affect diagnosis quality?"*
- *"Why use* DATAXRAY*? What are the differences between* DATAXRAY *and other classification or summarization tools? "*

In the rest of the proposal, we first give an overview of the inner-workings of DATAXRAY, focusing on the core model and algorithmic components (Section 2). We then proceed to describe the demonstration scenario; we introduce the datasets that we will use for the demonstration, and discuss the functionality of DATAXRAY that the demonstration will highlight (Section 3). The participants will have a chance to compare and understand the differences between DATAXRAY and other classification or summarization tools.

2. OVERVIEW OF DATAXRAY

In this section, we provide an overview of the key model and algorithmic components of DATAXRAY [19]: the *feature hierarchy*, the *diagnostic cost model*, and the *top-down iterative traversal*.

Feature hierarchy

DATAXRAY provides clues for discerning the underlying causes of error, by identifying groups of erroneous data with common characteristics. These characteristics, called *features*, can be derived from data using schema, values, and provenance metadata. Features form a natural hierarchy based on containment relationships; for example, the feature (season = "Spring") contains, and thus is an ancestor of, the feature (water level>2cm)∧(season = "Spring") and the feature (month = "April"). Figure 1 shows a screenshot of the DATAXRAY interface that displays the hierarchy of features for a simple web table. The root of the hierarchy (ALL,ALL) corresponds to the entire table, whereas the feature (tab1row1,ALL) corresponds to the first row of the table. Features at the leaf-level of the hierarchy are called *elements*, and correspond to a specific data item. For example, element (tab1row2,tab1col2) represents the data in the corresponding cell of the web table tab1.

DATAXRAY transforms the problem of error diagnosis to the problem of finding the features that best represent erroneous elements. This transformation enforces minimal assumptions, can model a large range of application scenarios, and allows for efficient exploration of possible diagnoses.

Diagnostic cost model

DATAXRAY uses a cost function based on Bayesian analysis to derive the set of features with the highest probability of being associated with the causes for the mistakes in a dataset. The set of features with the lowest cost corresponds to the highest *a posteriori* probability that these features are the real causes for the errors. The cost function is additive, and thus easy to compute. It contains three types of penalties, which capture the following three intuitions:
Conciseness: Simpler diagnoses with fewer features are preferable.
Specificity: Selected features should have a high error rate.
Consistency: Diagnoses should not include many correct elements.

Top-down iterative traversal

DATAXRAY derives diagnoses using a top-down, iterative algorithm with linear-time complexity that identifies the set of features with the lowest cost. This algorithm traverses the feature hierarchy from coarser to finer granularity features. It uses local stopping conditions to decide whether to accept the current feature or explore deeper. This top-down traversal is amenable to parallelization in the Map-Reduce framework, making DATAXRAY effective at large-scale diagnostic tasks [19].

2.1 Diagnostic interface and support

DATAXRAY provides intuitive interaction and algorithmic support to perform diagnosis on a variety of datasets. Our system comprises two main components, as shown in Figure 2:

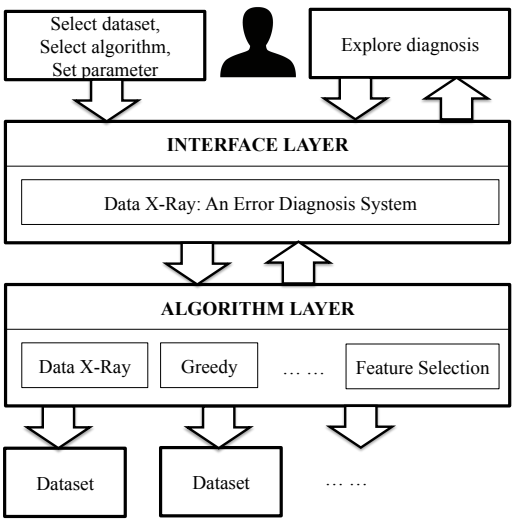

Figure 2: The DATAXRAY system provides visual abstractions to support user interactions with different datasets and derived diagnoses. It also supports multiple alternative algorithms to highlight the differences between DATAXRAY and other approaches.

- The interface layer provides visual abstractions to help users interact with the system. The demonstration participants will be able to load different datasets, control a choice of algorithms and parameters, and select different diagnostic and profiling tasks. The interface supports multiple interactions, allowing users to explore the feature hierarchies in different data domains, and displays the derived diagnoses by highlighting parts of the hierarchy. Users can see details by clicking on specific features, and can see different statistics of diagnoses derived by DATAXRAY, vs other classification and summarization techniques.
- The algorithms layer provides support for executing DATAXRAY and other techniques on the selected datasets. This will allow the demonstration participants to understand the differences in the derived diagnoses between DATAXRAY and other methods.

3. SYSTEM DEMONSTRATION

DATAXRAY provides a generic framework for data diagnosis and profiling that does not assume any underlying model for the data or provenance. Through our demonstration, participants will have the chance to interact and understand the causes of errors in real-world datasets from diverse domains. We describe here the datasets that our demonstration will support.

Synthetic web table. Our demonstration will start with a small toy example to help users familiarize themselves with the key concepts of DATAXRAY. The example dataset contains 16 knowledge triples extracted from a 4-by-4 web table. Three out of the four triples associated with the second column are incorrect, due to a reconciliation error of the extraction system.

Traffic incidents and weather. The demonstration participants will be able to analyze traffic incident and weather data collected by the US Department of Transportation, Federal Highway Administration [18]. The data contains detailed information about more than 1,500 traffic incidents on freeways around Portland, OR, over a two-month period, as well as weather data (e.g., temperature, humidity, windspeed) collected by multiple stations in the area during the same time period. Participants can use DATAXRAY to diagnose traffic incidents of different levels of severity.

Knowledge extraction. We will demonstrate diagnostic results for five real-world knowledge extraction systems provided by the ReVerb ClueWeb Extraction dataset [5]. The dataset applies different extraction systems on 500 sentences sampled from the Web, using Yahoo!'s random link service. The dataset provides the truth label for each extracted knowledge triple (indicating whether it is correct or erroneous). The five extractors have high error rates, ranging from 49% to 71%, resulting in more than 2,000 incorrect triples. The demonstration participants will use DATAXRAY to diagnose these errors and understand why and how they occur in the knowledge extraction process.

3.1 Demonstration highlights

Our demonstration is designed to showcase several features of DATAXRAY, and to help participants understand different aspects of the diagnosis problem. We describe here three main aspects of our demonstration.

Diagnostic features and key components of DATAXRAY

In their first experience with DATAXRAY, demonstration participants will load and explore the synthetic web table example (Figure 1). This dataset has a simple diagnosis and simple structure, which will allow participants to get familiar with the key components of the system. The feature hierarchy in this dataset is based on two dimensions: row and column. For example, the feature (tab1row1,ALL) describes all the triples that are associated with the first row of the web table. This example will help users understand the hierarchical structure of features. For example, the root of the hierarchy is feature (ALL,ALL), which corresponds to all rows and all columns. Thus, it contains (is an ancestor of) feature (tab1row1,ALL).

The visualization of the hierarchy (Figure 1) indicates the erroneous elements in red, and shows the links between parent and child features in light gray. *Clicking* on a feature highlights links to all its descendants, and displays information about the feature. In Figure 1, the feature (ALL, tab1col2) is highlighted. This feature is also the diagnosis that DATAXRAY derives for this dataset (highlighted in red in the hierarchy). This diagnosis identifies the connection among the errors in the data (they all come from the same column), indicating that there is a problem with the extraction process for that category.

Diagnosis of real-world errors

After getting familiar with the key components of the system through the toy example, our demonstration participants will access real-world datasets and use DATAXRAY to diagnose the errors in the different settings. DATAXRAY highlights the features in the hierarchy that it determines to be contributing factors to the errors. Users can zoom in and out, and scroll through the hierarchy to explore and understand a diagnosis. We include here two example diagnoses that our system derives.

EXAMPLE 1 (KNOWLEDGE EXTRACTION). *Diagnosing the errors in data generated by the* reverb *extractor, returns a grouping of triples with object structure ending in coordinating conjunction (e.g.,* and*,* but*,* for*). This indicates a clear problem with the extraction process, as it does not make sense to have a coordinating conjunction as an object in a knowledge triple (e.g., "newspapers and").*

EXAMPLE 2 (TRAFFIC INCIDENTS). *We label traffic incidents of all severities as "errors" and apply* DATAXRAY *to diagnose them.* DATAXRAY *returns several features that identify surface water level of more than 2cm is a contributing factor.*

Comparisons with other models and approaches

Our participants will have the opportunity to investigate the impact of the different penalties in the DATAXRAY cost model, by adjusting parameters of the cost function and eliminating some of its penalties. Through this analysis, participants will understand the impact of each factor in the cost model. Further, users of our system will compare the diagnoses derived by DATAXRAY with those derived by other related techniques. The DATAXRAY interface will display different statistics that compare diagnoses, including granularity of selected features, accuracy, and runtime. This analysis will highlight the differences between diagnosis, classification, and other related problems. Specifically, the demonstration will show that DATAXRAY is more efficient than techniques developed for other related problems, and derives diagnoses of higher quality.

Our demonstration participants will compare and understand the differences between DATAXRAY and three classes of techniques:

Set cover methods. The problem of deriving optimal diagnoses is related to the set cover problem: one needs to find a set of features that best cover the errors in the data, based on a cost function. Our interface will allow participants to select two different alternative algorithms for set cover. First, we apply a greedy approximation for weighted set cover [3] to select the set of features of minimum weight that cover all of the erroneous elements. We adjust the greedy algorithm to use the DATAXRAY cost model to allow set cover to penalize features that cover correct elements, which it does not do in its default objective. Second, we will provide an approximation algorithm that solves *red-blue* set cover [2,12]: Given a collection of sets with "blue" and "red" elements, the red-blue set cover problem looks for a sub-collection of sets that covers all "blue" elements and minimum number of "red" elements. In contrast to regular set-cover, red-blue set cover can model both correct and incorrect element coverage. Demonstration participants will observe that red-blue set cover favors features with high error rate, but does not consider the number of returned features, resulting in high recall but low precision.

Classification methods. Demonstration participants will compare DATAXRAY to two different classification methods, logistic regression [11, 17] and decision trees [13]. For each feature, logistic regression learns a weight between -1 and 1: a positive weight indicates that the feature is positive proportional to the class (in our context the feature is a cause), and a negative weight indicates the opposite. We use the labeled data as the training dataset, excluding features with only correct elements to speed up learning, and return features with positive weights. We also apply L_1-regularization, which favors fewer features for the purpose of avoiding over-fitting. While these techniques build good predication models, the chosen features do not make good diagnoses, as they often contain redundancy and have low error rates.

Summarization methods. Users will be able to compare DATAXRAY with diagnoses produced by Data Auditor [7, 8], a data quality exploration tool that uses rules and integrity constraints to construct *pattern tableaux*. We annotate erroneous data as a consequent (dependent) value in a FD, and use Data Auditor to learn a pattern for this rule. We treat the generated tableau as a diagnosis. Data Auditor produces diagnoses of low precision and recall, as it focuses on reducing the number of returned attributes (features), constrained on the coverage of satisfying and non-satisfying elements.

4. DEMONSTRATION SUMMARY

DATAXRAY introduces a novel extension to the problem of data quality, by providing support for error diagnosis and data profiling. Our demonstration will highlight the distinction between data cleaning and error diagnosis. Participants will gain a better understanding of the criteria that impact the quality of diagnoses through comparisons with related techniques, and will have the chance to explore error causes in real-world datasets.

Acknowledgements: This work was partially supported by the National Science Foundation under grants CCF-1349784 and IIS-1421322, and by Google Inc. via a Faculty Research Award.

5. REFERENCES

[1] S. Abiteboul, S. Cluet, T. Milo, P. Mogilevsky, J. Siméon, and S. Zohar. Tools for data translation and integration. *IEEE Data Engineering Bulletin*, 22(1):3–8, 1999.

[2] R. D. Carr, S. Doddi, G. Konjevod, and M. Marathe. On the red-blue set cover problem. In *Proceedings of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 345–353, 2000.

[3] V. Chvatal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):233–235, 1979.

[4] W. W. Eckerson. Data warehousing special report: Data quality and the bottom line. http://www.adtmag.com/article.asp?id=6321, 2002.

[5] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *EMNLP*, 2011.

[6] W. Fan, F. Geerts, and X. Jia. A revival of integrity constraints for data cleaning. *PVLDB*, 1(2):1522–1523, 2008.

[7] L. Golab, H. Karloff, F. Korn, D. Srivastava, and B. Yu. On generating near-optimal tableaux for conditional functional dependencies. *PVLDB*, 1(1):376–390, 2008.

[8] L. Golab, H. J. Karloff, F. Korn, and D. Srivastava. Data auditor: Exploring data quality and semantics using pattern tableaux. *PVLDB*, 3(2):1641–1644, 2010.

[9] D. Herring and M. King. Space-based observation of the Earth. *Encyclopedia of Astronomy and Astrophysics*, 2001.

[10] D. V. Kalashnikov and S. Mehrotra. Domain-independent data cleaning via analysis of entity-relationship graph. *ACM Transactions on Database Systems*, 31(2):716–767, 2006.

[11] A. Y. Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *ICML*, 2004.

[12] D. Peleg. Approximation algorithms for the label-cover MAX and red-blue set cover problems. *Journal of Discrete Algorithms*, 5(1):55–64, 2007.

[13] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[14] E. Rahm and H. H. Do. Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23(4):3–13, 2000.

[15] V. Raman and J. M. Hellerstein. Potter's wheel: An interactive data cleaning system. In *VLDB*, pages 381–390, 2001.

[16] A. Robeznieks. Data entry is a top cause of medication errors. *American Medical News*, 2005.

[17] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[18] US Department of Transportation, Federal Highway Administration. Freeway incident and weather data. http://portal.its.pdx.edu/Portal/index.php/fhwa.

[19] X. Wang, X. L. Dong, and A. Meliou. Data X-Ray: A diagnostic tool for data errors. In *SIGMOD Conference*, pages 1231–1245, 2015.