Data Exploration: A Roll Call of All User-Data Interaction Functionality

Anna Gogolou[1,2] Marialena Kyriakidi[1] Yannis Ioannidis[1,2]
[1] Dept. of Informatics & Telecom, National & Kapodistrian University of Athens, Greece
[2] Institute for the Management of Information Systems, "Athena" Research Center, Greece
{agogolou, marilou, yannis}@di.uoa.gr

ABSTRACT
Data exploration encompasses a variety of interaction types and data functionality, such as search, data analysis, curation, constraint satisfaction, data mining, and visualization. Data exploration naturally begins when a user is given a set of data and ends when the user extracts all information and knowledge hidden in the data. Although a plethora of systems have been developed to tackle different data exploration aspects, there is no framework devoted to it as a whole. In this paper, we claim that "any" user-data interaction is essential for data exploration and sketch a prototype with both automated and user-induced functionality.

CCS Concepts
• Information systems → Database management system engines • Human-centered computing → Visual analytics.

Keywords
Data exploration, data curation, recommendations.

1. INTRODUCTION
There is no manual for data exploration. In fact, there is no proper definition of it either. Data exploration as a technical term came to be gradually, as a realization that when a user comes in contact with data, there is no predefined manner of interaction with it. She spends much time investigating and analyzing it, attempting to understand, clean and prepare it for the next steps in her current task. Depending on the context, user-data interaction can take different forms. Data profiling, identification, analysis, validation, transformation, and visualization are only a few of the most common functions identified in a data exploration process. In a database-related environment, interaction with the data occurs through some form of queries. However, if the user is not so familiar with the language of the current engine or if she is uncertain of what she wants to do with the data, the usual system responses may not only be irrelevant but may also be considered as distractions to her goal. In this context, it is imperative for modern systems to facilitate users in exploring, interpreting, and discovering the information they need.

In this paper, we outline a general framework for data exploration and its multiple manifestations. It captures an iterative workflow over diverse user-data interactions that depend on the task at hand and the nature of the underlying data. In addition to a loop of user-initiated interactions, there is also a loop of system-initiated data profiling actions that assist users and enrich their experience.

Interestingly, the data exploration concept presented has evolved from the initial functionality of the Data Curation and Validation tool (DCV), which we have developed for data cleaning. During the tool's development, increasingly more functionality was evidently needed to improve data cleaning, to the point that data cleaning became but one aspect of general data exploration.

In Section 2, we present the architecture and overall functionality of our data exploration framework. In Section 3, we describe how DCV has evolved to an interactive data exploration system enhanced with recommendation services. In Section 4, we proceed with the related work where we discuss how DCV differentiates from several related and well-established systems. Finally, in Section 5, we present our conclusions and future work.

ExploreDB'16, June 26–July 01, 2016, San Francisco, CA, USA.

DOI: http://dx.doi.org/10.1145/2948674.2955105

2. SYSTEM ARCHITECTURE
Data exploration is an iterative, continuous, user-initiated activity that ends when the user has reached her goals. To serve this, a data exploration system integrates three major components, i.e., Activity, Profiling, and Presentation (Fig. 1). The user triggers the Activity component by submitting a specific action request (e.g., overview, search, analysis, curation). This activates the relevant processing engines in the component, which execute the action requested. The results are sent back and/or saved to the database.

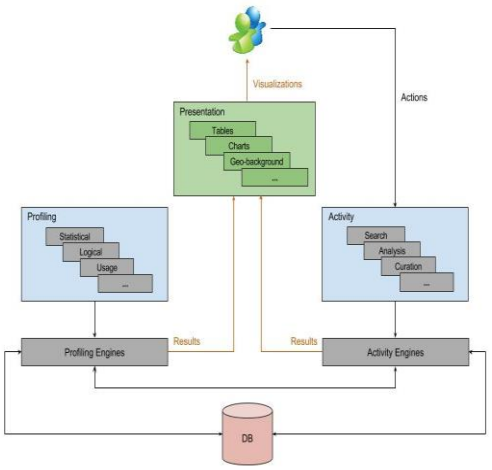

Figure 1. Data Exploration Architecture and Flow.

On the other hand, it is the system that triggers the Profiling component, when new data comes in or the user proceeds with an action. It extracts diverse data profiles (e.g., logical, structural, statistical, usage) by having the corresponding profiling engines operate on the relevant data. The data and user profiles thus

created are sent to the Activity component for further use and/or are combined to send recommendations of actions (e.g., queries, transformations, repairs), data products, or visualizations.
Finally, the Presentation component offers a wide range of visualizations for the Activity and Profiling output, from simple tables and charts to interactive visualizations.

3. DCV

DCV is a web-based tool that has been initially designed and developed exclusively for data cleaning, able to automatically detect inconsistencies and outliers in tabular datasets (Fig. 2). Its initial target audience included doctors and other biomedical personnel, as their data are often manually collected and contain many errors. A typical example is data coming from questionnaires filled out by patients with a pen during a clinical research. Not only is this manual task error-prone, but it is also followed by another manual task, of a lab person entering the data into a computer, leading to more errors.
Profiling is at the heart of DCV. To tackle different cleaning (and curation) problems, DCV calculates relevant data profiles and uses them as a basis to offer the user (the curator) decision support functionality on cleaning/curation actions. For example, statistical profiles are used to detect potentially erroneous outliers, while the logical profile is used as a set of constraints and detects violations. Currently, for each column of a tabular dataset, the statistical profiling engine calculates a histogram on the value distribution; for numeric columns, it also calculates quartiles and the min, max, avg, and standard deviation values. The logical profiling engine handles constraints in the form of association rules for two or more columns, which are automatically detected and confirmed or manually declared and inserted into the system by the user. Both types of potential errors, i.e., outliers and inconsistencies (rule violations, nulls) are highlighted by the Presentation component in an information rich fashion. With this information on display, the user interacts with the cleaning engine in the Activity component for her next action. For example, she may select a string distance metric, e.g., Hamming or Damerau-Levenshtein distance, to cluster similar text values together. The value distribution within a cluster may indicate possible typographical errors in the values of a text column, as they may actually all be the same with some character insertions/deletions/substitutions in the true value.
As DCV started being used, however, it progressively became evident that increasingly more functionality was necessary in the Activity and Profiling components to better support data cleaning, but also that data cleaning would be beneficial to other activities. This led to the realization that essentially all forms of user-data interaction are part of data exploration and may be interleaved in arbitrary ways depending on the user needs.
Hence, DCV has been evolving in the direction of becoming a comprehensive data exploration tool. In addition to data cleaning, its Activity component now provides data search, analysis, and curation functionality. Especially regarding curation, in principle, it includes extremely diverse functionality, ranging from cleaning and transformation to integration and deduplication. Currently, DCV offers data transformation possibilities through a string and math expression language, with more functionality under development.
Its Presentation component offers various interactive and interdependent visualizations (e.g., histograms, scatterplots, action histories, various charts) that can help the user gain better insights to the data, detect more subtle errors, redo or undo past actions, or formulate her next action by isolating specific data subsets. Incidentally, horizontal and vertical virtual scrolling scales to rather large datasets by loading in memory only their visible part.
Finally, work is under way on adding usage profiling to the Profiling component of DCV, so that recommendations may be given to further facilitate the user in her goals. The profiler monitors the users' actions and choices as they interact with the data and builds their profiles. The recommender engine can then provide recommendations on any relevant aspect of possible future actions, e.g., the actions themselves or whole sequences of them (workflows), their results, source datasets, visualizations, so as to further improve the exploration. For example, it may recommend a search or analysis on a particular area of a dataset, on the basis that its results may reveal similar errors to those the user corrected earlier in a different area of the dataset. Or it may recommend a more specialized search than those already posed by the user, based on how other similar users proceeded with their

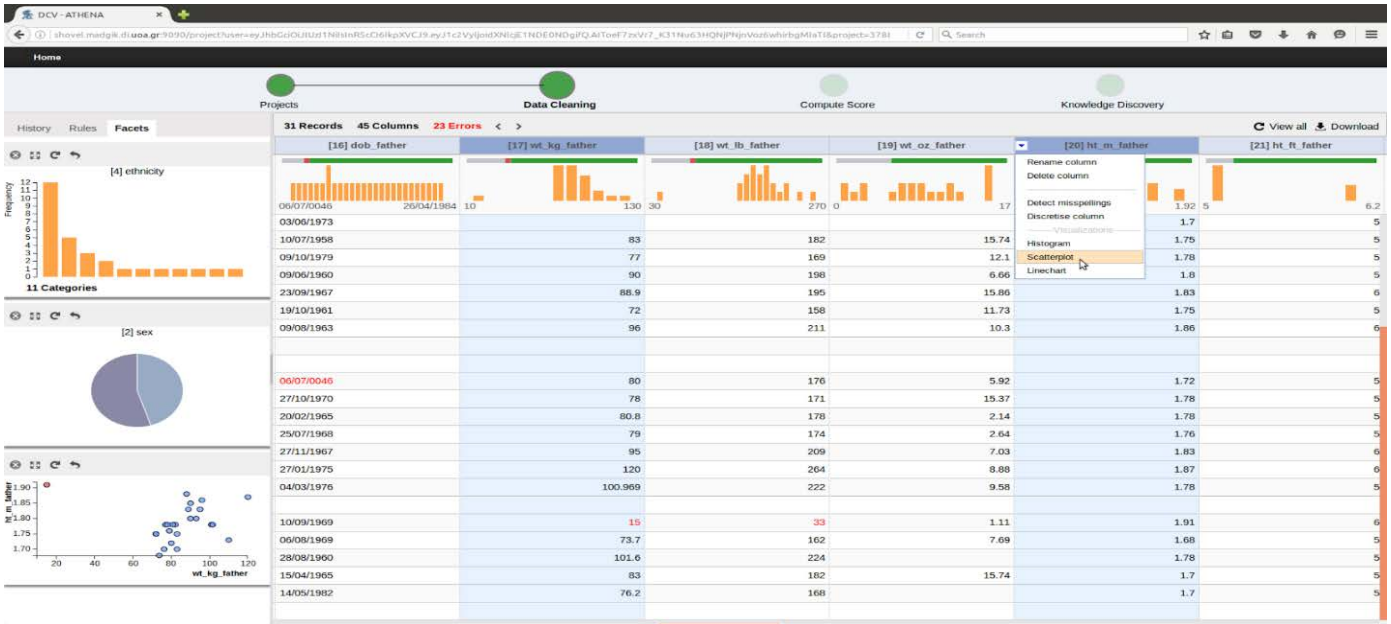

Figure 2. DCV Interface – Sample Screenshot.

exploration and eventually succeeded, having followed similar earlier searches as well. It may also suggest specific data transformations, based on user-defined mathematical formulas that have been used on similar data; additions to the logical profile of data (e.g., functional dependencies), when elements of the corresponding statistical profile appear (almost) universal; visualization approaches that have proved useful to other users on similar data; or corrections on erroneous data, by generalizing the pattern of errors and their corrections performed earlier. Beyond actions, it may also directly recommend results of searches similar to the ones posed (generalizations or alternative specializations usually), actual visualizations, and possible repairs produced by profiling, detection, and cleaning algorithms.

4. RELATED WORK

Given the claim that data exploration may include "any" functionality that engages the user in an interaction with the data, it should be clear that the work related to this effort is enormous and cannot be described in this short paper. As DCV is presently primarily for data curation, we highlight existing systems in that area alone. Data curation is a broad concept and affects every data-based scientific, commercial, or other human activity. It is typically the first step after data collection, before any data analysis and/or knowledge discovery tasks may produce meaningful valid results. A significant number of data curation tools have already been developed: *Data Wrangler* [1] (now *Trifacta* [2]), *Data Tamer* [3] (now *Tamr Inc.* [4]), *NADEEF* [5, 6], *KATARA* [7, 8], *OpenRefine* [9], and *Google Sheets* [10] are among them. Interestingly, all of them offer some elements of data exploration functionality, thereby in some sense supporting the main claim of this paper.

Each tool focuses on different aspects of curation (i.e., cleaning, transformation, integration, identification). Data Wrangler and its successor Trifacta focus mainly on data transformation tasks, offering a direct manipulation interface on grid data and automatic inference of relevant transforms [1]. On the other hand, Data Tamer, now Tamr Inc., focuses mainly on data integration, taking as input a sequence of data sources and performing attribute identification and record deduplication using machine learning algorithms [3]. *NADEEF* and *KATARA* mostly serve as cleaning tools, while *OpenRefine* and *Google Sheets* provide the basic functionality of a spreadsheet application for cleaning and transforming unclean data.

All these tools offer interactive data visualizations, while some actually offer recommendations as well to help the user. *Data Wrangler*'s recommendations are not based on usage history and user profiles, as is the case in the design of DCV, but solely on the current user-data interaction session. *NADEEF* recommends specific types of cleaning rules (mainly functional dependencies), while *KATARA* recommends possible repairs of detected errors. Finally, *Google Sheets* offers extensive recommendations on appropriate data visualizations, depending on the type and actual distribution of the data.

5. CONCLUSIONS & FUTURE WORK

In this short paper, we have outlined our ongoing work on DCV, a data curation tool that has been naturally evolving into a full-blown data exploration tool. Driven by our DCV experience, we have presented a general framework for data exploration, which captures the iterative and interactive nature of the process, is able to incorporate "any" user-data interaction functionality, and integrates both user-initiated and system-initiated activities in a symmetric fashion for more effective data exploration.

As part of our future work, we plan to gradually add a variety of functions in DCV regarding all three of the basic components of the framework, raising it to a true data exploration tool. In the Profiling component, we will focus on usage profiling and the recommender engine, transferring ideas from our earlier effort on PAROS [11]. Additionally, we plan to enrich the Presentation component with more available visualizations that match different types of data. Finally, in the Activity component we intend to create a richer data analytics package, by including specialized machine learning and data mining algorithms that are appropriate for data from different fields.

6. ACKNOWLEDGEMENTS

This work has been partially supported by the Seventh Framework Programme (FP7) of the European Commission under contracts #600932 (MD-Paedigree) and #604102 (Human Brain Project).

7. REFERENCES

[1] Kandel, S., Paepcke, A., Hellerstein, J., and Heer, J. Wrangler: Interactive Visual Specification of Data Transformation Scripts. *CHI*, May 2011. DOI=http://doi.acm.org/10.1145/1979444.

[2] *Trifacta*, http://www.trifacta.com/

[3] Stonebraker, M., et al. Data Curation at Scale: The Data Tamer System. *Proc. of the 6th Biennial Conf. on Innovative Data Systems Research*, Jan 2013.

[4] *Tamr Inc.*, http://www.tamr.com/

[5] Dallachiesa, M., et al. NADEEF: A Commodity Data Cleaning System. *SIGMOD Conference*, (Jun. 2013). DOI=http://doi.acm.org/10.1145/2465327.

[6] Ebaid, A., et al. NADEEF: A Generalized Data Cleaning System. *Proc. of the VLDB Conf,*, Aug. 2013. DOI=http://doi.acm.org/10.1145/2536280.

[7] Chu, X., et al. KATARA: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing. *Proc. SIGMOD Conf.*, June 2015. DOI=http://doi.acm.org/10.1145/2749431.

[8] Chu, X., et al. KATARA: Reliable Data Cleaning with Knowledge Bases and Crowdsourcing. *Proc. VLDB Conf.*, Aug. 2015. DOI=http://doi.acm.org/10.1145/2824109.

[9] *OpenRefine*, http://openrefine.org/

[10] *Google Sheets*, https://www.google.com/sheets/about/

[11] Ioannidis, Y., et al. Profiling Attitudes for Personalized Information Provision. IEEE Data Eng. Bulletin, 34(2), 2011. DOI=http://sites.computer.org/debull/A11june/Yannis.pdf