Toward Supporting the Data Enthusiast: Unlocking the Potential of Data for Analysis

Kristi Morton, Magdalena Balazinska, and Dan Grossman
University of Washington, Seattle, WA, USA
{kmorton,magda,djg}@cs.washington.edu

Jock Mackinlay
Tableau Software
Seattle, WA, USA
jmackinlay@tableausoftware.com

ABSTRACT

We present a vision for the next-generation of visual analytics services. We argue that these services should provide three related capabilities: (1) support visual and interactive data exploration as they do today, but also (2) suggest relevant data to enrich visualizations, and (3) facilitate the integration and cleaning of that data. Most importantly, they should provide all these capabilities seamlessly in the context of an uninterrupted data analysis cycle. We present the challenges and opportunities in building such next-generation visual analytics services.

1. INTRODUCTION

The increasing availability of public datasets on the Web is encouraging individuals called *Data Enthusiasts* [6] to analyze data even though they have little background in databases, statistics, or programming. A typical example is an online news reporter who wants to use data and visualizations to illustrate a story on her blog.

To accommodate the needs of these users, a new type of tool, *visual analytics services*, has recently emerged Tableau Public [1], Fusion Tables [4], and Many Eyes [14] are among the most popular examples. These tools support what is called the *sensemaking model* [2]: The typical analytical process starts with a question that a data enthusiast seeks to answer. The data enthusiast then starts to forage for relevant data unless she already has a dataset to explore. Once the appropriate dataset is acquired, the data is explored through an appropriate visualization. The user continues to interact with the visualization by, for example, drilling down to the details or pivoting out some dimensions. To support such sensemaking, the core functionality of visual analytics systems is threefold: (1) They enable users to *visually* explore their data as illustrated in Figure 1. (2) These systems facilitate the integration and study of *multiple* datasets at the same time. (3) Finally, they support *collaborations* through sharing visualizations and data *online* for both viewing and editing by others.

...

As the world is increasingly becoming data centered, we argue that it is critical for data-management tools to support the data enthusiasts well. Current usage suggests that much room for improvement remains: In recent prior work [9], we found that today's visual analytics services are attracting hundreds or thousands of new accounts each month, but most users author only one visualization and never return. Only a small fraction of users make more than 2–3 visualizations.

This paper presents a vision for how visual analytics services should evolve to meet users' needs and the research challenges that this evolution raises. Based on interactions with Tableau customers as well as a recent study of Tableau Public and Many Eyes utilizations [9], we argue that visual analytics services need to improve in three dimensions:

- **Visual Data Cleaning:** Data enthusiasts have reported that cleaning and transforming their datasets is one of the most time-consuming and tedious steps in their analytical workflows (often comprising 80% of the work [3]). The data is useless until that labor is accomplished up front. Currently, they must use unrelated tools for cleaning and visualization. We argue that data cleaning should instead be an integral part of the visual data analytics cycle.
- **Data Enrichment:** While significant amounts of data are available on the Web or even already available in visual analytics services through contributions by other users, identifying interesting data to enrich a visualization is a challenging task. Different datasets have different schemas, different levels of granularity (*e.g.,* we may have state-level unemployment data but zip code-level income data), different levels of cleanliness, and they contain different subsets of relevant data. Next-generation visual analytics services should provide better support to help users identify datasets that they can potentially leverage for their current data analysis task.
- **Seamless Data Blending:** Once a user identifies a dataset of potential interest, integrating that dataset with a current visualization is also challenging due to all the well-known data integration barriers including schema matching, schema mapping, and entity resolution. The visual analytics service should help smooth this integration with a focus on producing useful visualizations. This means that the tools must be robust to: (1) semantic mismatches, (2) data fragility and domain mismatches, (4) combining data at different granularities, and (4) further exploration (roll-up,

drill-down, slice, etc.) of the integrated datasets.

Moreover, the three capabilities above should be **seamlessly integrated into a unified framework**. To the user, it should be a visualization system that enables jumping among the tasks of exploring data, finding new data, integrating data, and cleaning data in a consistent, integrated fashion. Current systems impose a mental context switch penalty every time a user needs to integrate another data source by forcing the user to deal with the details of cleaning and transforming it. We want to avoid these expensive context- and tool-switches.

We believe the key to achieving a unified *tool* is an underlying unified *formal description language* for describing visual data analytics, data integration, and data cleaning (Section 3.1). Here, we take inspiration from three advances from the graphics, programming languages, and human computer interaction communities on formal descriptive languages for transforming data: (1) Tableau's Visual Query Language (VizQL) [13] which is a language that turns data into a visualization, (2) Recent work that applies program synthesis to automatically generate data transformation scripts on tables given an example input and output table [7], and (3) the Data Wrangler [8] project which recommends relevant data cleaning operations and stores them as scripts. We advocate combining these technologies into one that enables interactive data exploration over multiple datasets.

The final key component is **public collaboration**. We envision that the tool will capture and make available the data and visualizations analyzed by its various users, just like visual analytics services do today. However, we propose to go further and capture the fundamental data transformations that were applied to the original data sources to produce their corresponding visualizations (including any cleaning, exploration, and integration actions). We propose to leverage these actions in addition to the raw data and visualization to support data discovery (Section 3.2), data cleaning (Section 3.3) and data integration (Section 3.4).

The rest of this paper presents our vision by first discussing what makes the visual analytics setting both more and less challenging in different ways. We then discuss the main components of the system we envision and the associated research challenges for the database community.

2. THE VISUAL ANALYTICS SETTING

While finding, cleaning, and integrating data are each a well-established research area, their exploration in the context of visual analytics services is very different from their standard problem statement.

In this context, the problems become *more* challenging for several reasons: First, users are completely non-technical; Second, all interactions must be through a visual interface; Third, the speed must be interactive to support sensemaking. Finally, interactive data exploration is a nonlinear process involving many workflow steps: At any stage of analysis a user may need to switch to a related task such as cleaning or integrating and build upon the progress made so far. Indeed, data exploration helps identify problems and clean the data. It also helps determine if the effort required by data cleaning is worth it.

At the same time, the problems are also *less* challenging in three ways. First, datasets are small. In our recent study

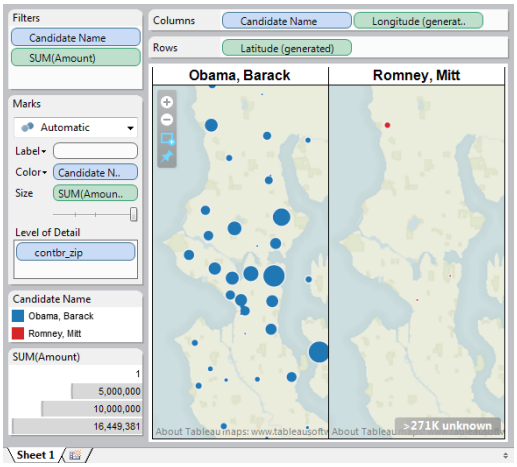

Figure 1: Tableau Visualization With Dirty Data (indicated via the ">271K unknown," bottom right)

```
// Standard VizQL
SELECT Latitude ON ROWS,
    ([Candidate Name] * Longitude ) ON COLUMNS,
    [Candidate Name] ON COLOR,
    SUM(Amount) ON SIZE,
    [Zip Code] on LEVEL_OF_DETAIL
FROM [Contributions View]
WHERE [Candidate Name] = {"Barack Obama", "Mitt Romney"}
// Possible extension for data cleaning
AND  [Zip Code] VALIDATE
```

Figure 2: Example of Extended VizQL Combining Visualization and Cleaning

of Tableau Public we found that datasets analyzed by users had fewer than 100K rows. Second, all operations are interactive with back-and-forth between system actions and user actions (*e.g.,* the system can recommend a mapping between attributes in two datasets, the user can then edit these mappings). Third, the systems are collaborative. It is thus possible to leverage the prior work of the community of users to support users in each new visualization task (*e.g.,* a user may have partially cleaned a dataset before another user needs it).

3. VISUAL DATA MANAGEMENT AND ANALYTICS SERVICE

We present the challenges and research opportunities behind our envisioned next-generation visual analytics service.

3.1 A Common Formalism

The most important goal of our envisioned visual analytics system is to enable users to perform data exploration, integration, and cleaning activities through a single visual interface and as part of a single data analysis activity. Users should never have to think that they are switching away from exploration when they clean and integrate their data.

The key idea to realize this vision is to develop a single formalism for a fully integrated and visual analysis cycle. In the existing Tableau system, VizQL and its underlying data model serve this purpose: User actions are translated into VizQL queries in a principled fashion. For example, the visualization shown in Figure 1 is driven by the VizQL query

shown in Figure 2 (except for the last line, which we discuss shortly). These queries are similar to SQL and perform both the task of querying the underlying data source (the VizQL queries are compiled into SQL or MDX queries) and rendering the results. The formalism represents the clear semantics underlying the tool. The user does not author the language directly, but rather her interactions with the GUI automatically result in the generated code.

VizQL, however, supports only data exploration. Similarly, there has been work on supporting either data cleaning [8] or data integration [11] from a graphical interface. But no tool today supports the complete analysis cycle. Developing a single interface and a single language, however, is not as simple as unioning existing interfaces and languages. The challenge is that we *need to integrate activities that are both typically performed using different graphical interfaces and that yield very different actions*: visual analysis creates views over underlying base data, data cleaning actually edits the data, and data integration creates schema mappings, a mediated schema, and wrappers for data sources. Combining such different activities behind a single language tied to a single visual interface is a challenge.

For example, Figure 2 shows how VizQL could be used together with a data cleaning language extension called, `VALIDATE`. In this example, the user specified through a GUI action that she was interested in seeing results only for valid zip codes on her map and this action translated into the added VALIDATE clause. Imagine that `VALIDATE` is a special User Defined Function that detects invalid values based on some stored domain information about zip codes, and uses the associated name/address to infer the correct zip codes. While we can use integrity constraints to detect dirty data, it raises some challenges such as how are these integrity constraints visually specified and how will they interact with user-specified edits to dirty values.

In addition to the challenge of developing the appropriate data model and language, a single formalism and tool also raise important system design questions. For example, consider combining data exploration and cleaning. Typically, user interactions with a visualization change the view definition underlying the displayed data. As the user drills-down, slices, and otherwise explores the data, her actions translate into adjustments to the view definition. One approach to incorporating data cleaning would be to translate cleaning actions also into changes to the view definition. The problem, however, is that data cleaning will often be significantly more detailed, affecting individual records. For example, the user may adjust the value of a couple of zip codes that the user knows. Capturing all such edits as view-definition changes can quickly render the view definition untenably intricate. The tool may need to actually apply the user edits on a materialized view of the data to speed-up the exploration. This approach, however, raises many questions. For example, what actions should be considered cleaning and what actions should be considered exploration? Does the distinction matter? Would the combination of a materialized and virtual view support both activities efficiently? If so, how should the system determine which changes to materialize? The design space becomes even more complicated once the user integrates multiple datasets.

Overall, the single tool and formalism thus raise the challenge of (1) developing a single data model and language similar to VizQL but able to also capture cleaning and data integration tasks and (2) building the underlying integrated system that efficiently supports all three types of activities.

3.2 Recommendation Component

A common important step in the analysis process includes adding context to a dataset by combining it with some other relevant data source. For example, a reporter may be interested in extending a dataset about obesity rates in different cities with another dataset showing the availability of bike paths in these same cities. Identifying datasets that could extend an analysis in a useful way, however, is in general far from trivial. The user must first find datasets (either on the Web or contributed to the visual analytics service by other users) that contain useful information: *i.e.,* datasets that have information about bike paths. He must then assess if the data can actually be integrated: *e.g.,* Is the new information at the granularity of cities or counties? Does the set of cities in the new dataset overlap with the set of cities in the original dataset?

To help users with this task, the next generation of visual analytics services should include powerful data recommenders that would help *identify datasets that both contain relevant information and can be successfully integrated.*

It is well-known in the data integration literature [5] that data recommendation is challenging due to schema and semantic mismatches between datasets (among other challenges). Recent prior work [12] has tackled the problem of finding tables from a large Web corpus that are related to an input table by leveraging metadata from the input table (*e.g.,* labels, schemas, data values). This tool only considers extending each entity, as identified by a set of key attributes, with additional fields. However, in recent work [9], we observed that that such a tool would not apply to more than half of all integration scenarios that occur in practice in a visual analytics service, as 50% of users joined multiple entities at either the same location or the same point in time.

We posit that a data recommender should leverage the context of the visualization of the current data source to perform a more focused search for related data. The idea is to both improve semantic matches and relevance of the recommended data by using clues from the current data and visualization including the schemas, axis labels, annotations, domain of values being visualized, primary keys, aggregation/filters/projections used, or functional dependencies among columns.

The key challenge lies in determining what information in a user's current visualization is most relevant for the purpose of recommending additional data. For example, a map view suggests latitude/longitude coordinates as a possible join key since the data currently displayed on the map could be enhanced with information about co-located objects. What about other types of views? Or axes labels? Are they also useful? Similarly, should the recommender use historical actions by a user (*e.g.,* the cleaning transformations that were applied to a common dataset) when recommending new data to him or her? How? When could this be helpful?

Beyond the user's current visualization, many options are open regarding the data to be recommended. For example, the engine could recommend either entire datasets or only individual columns. It could choose to recommend raw data or data cleaned by other users. It could leverage past history on how datasets were integrated and visualized by other users to improve the recommendation quality. The

key challenge lies in identifying what features of existing datasets and their past use in other visualizations to take into account in the recommendation process.

3.3 Cleaning Component

In our proposed system, the visualizations are part of the data cleaning and shaping effort and can help the user both to determine if a dataset has enough potential to be worth the investment of further cleanup and to actually perform the cleaning actions. Section 3.1 describes the overall integration of data cleaning into the visual analysis cycle. In this section, we discuss how a visual analytics service can leverage the collaborative nature of the engine to better support cleaning.

For example, consider the visualization in Figure 1, which shows the data from the Federal Election Commission on the 2012 presidential campaign contribution statistics, zoomed in to Seattle, Washington. This dataset is quite dirty, and Tableau detects that it contains over 271,000 unknown contributor zip codes. We cannot expect the user to manually clean each incorrect zip code. However, other users have probably worked with this dataset in the past or with similar datasets. The system can thus help the user by suggesting semantically similar datasets from its datastore that may have better overall data quality by considering the field value domains. This dataset is a common one and chances are good that it has been downloaded and cleaned up by other users, so our system should recommend these cleaned versions to our users.

Similarly, because the visual analytics service will capture all transformations performed by all users on their data, it will be able to learn various cleaning actions such as corrections to common typos, different spellings of the same entity names, the best datasets to use to clean specific types of attributes, such as street addresses, etc.

3.4 Data Integration Component

To make the integration as simple as possible for data enthusiasts, we envision building on Tableau's existing pay-as-you-go data blending feature [10], which (1) automatically creates the mediated schemas and wrappers as the user interactively builds a visualization on-the-fly and (2) only joins in the necessary information from a different data source (*e.g.,* as specified by the user through a filter or projection in the visualization's query expression) to create the view with minimal data movement, as queries are federated to the data sources. Tableau's data blending feature, however, is primitive and only correctly handles simple data integration scenarios.

To push the data integration further, we propose to leverage the context of prior visualizations from other users to help. One main challenge in combining data is ensuring that the join produces meaningful results, as visualizing `NULL` values provides little value (*i.e.,* due to the domain mismatch problem). For example, imagine two tables with salary information for two different companies and a user wants to join them. Perhaps, in the past, another user has transformed one salary from a weekly salary to an hourly salary. The system, when recommending the second table, would suggest applying the same transformation in order for the two tables to join in a meaningful way (*i.e.,* with consistent field value domains). The first research opportunity is to find ways to identify such relevant actions by past users and apply them in the context of a new data integration task. What if the original user was looking at salaries in euros when performing the transformation while the second user has a dataset with dollar-value salaries? Even though the domains do not match, the transformation remains applicable in both contexts. The system should identify and leverage such transformations.

A second key challenge is how to handle the case when the user continues to interact with the blended visualization by altering the granularity of the data powering the view by drilling-down to the details or rolling-up to summarize. For example, imagine that a user has created a map view that combines per-capita coffee production with coffee consumption at the granularity of countries. However, the user wants to continue exploring this view by drilling-down to the city-level. In order to accomplish this task, the blending system would attempt to pull the city-level data from each dataset. If one of the datasets does not have any information about cities, then the blending operation will not succeed. One solution could be to leverage the recommender tool to suggest a relevant dataset with the necessary city-level coffee consumption statistics.

4. CONCLUSION

We are proposing a next-generation visual analytics system with the goal of providing complete support for data enthusiasts. We believe a unified framework of interactive cleaning, blending, and recommendation is key to helping this growing group of data consumers unlock the potential of their data for gaining insight and knowledge.

5. REFERENCES

[1] Tableau Public. `http://www.tableaupublic.com/`, 2012.

[2] S. K. Card, J. D. Mackinlay, and B. Shneiderman. Using Vision to Think. In *Readings in Information Visualization*, pages 579–581. Morgan Kaufmann Publishers Inc., 1999.

[3] T. Dasu and T. Johnson. *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons, New York, NY, 2003.

[4] H. Gonzalez, A. Halevy, C. Jensen, A. Langen, J. Madhavan, R. Shapley, and W. Shen. Google Fusion Tables: Data Management, Integration and Collaboration in the Cloud. In *Proc. of SOCC*, pages 175–180, 2010.

[5] A. Halevy, A. Doan, and Z. Ives. *Principles of Data Integration*. Morgan Kaufmann, 2012.

[6] P. Hanrahan. Analytic database technologies for a new kind of user: the data enthusiast. In *Proceedings of the ACM SIGMOD Conf.*, pages 577–578, 2012.

[7] W. R. Harris and S. Gulwani. Spreadsheet Table Transformations From Examples. In *PLDI 2011*, pages 317–328, 2011.

[8] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Wrangler: Interactive Visual Specification of Data Transformation Scripts. In *Computer Human Interaction*, pages 3363–3372, 2011.

[9] K. Morton, M. Balazinska, D. Grossman, R. Kosara, J. Mackinlay, and A. Halevy. A Measurement Study of Two Web-based Collaborative Visual Analytics Systems. Technical Report UW-CSE-12-08-01, Univ. of Washington, `ftp://ftp.cs.washington.edu/tr/2012/08/UW-CSE-12-08-01.PDF`, Aug 2012.

[10] K. Morton, R. Bunker, J. Mackinlay, R. Morton, and C. Stolte. Dynamic Workload Driven Data Integration in Tableau. In *Proc. of SIGMOD*, pages 807–816, 2012.

[11] A. Raffio, D. Braga, S. Ceri, P. Papotti, and M. Hernandez. Clip: a Visual Language for Explicit Schema Mappings. In *Proceedings of ICDE 2008*, pages 30–39, 2008.

[12] A. D. Sarma, L. Fang, N. Gupta, A. Halevy, H. Lee, F. Wu, R. Xin, and C. Yu. Finding Related Tables. In *Proceedings*

of the ACM SIGMOD Conf., pages 817–828, 2012.

[13] C. Stolte, D. Tang, and P. Hanrahan. Polaris: a system for query, analysis, and visualization of multidimensional databases. *IEEE TVCG*, 8(1):52–65, 2002.

[14] F. B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. Many eyes: A site for visualization at internet scale. *IEEE TVCG*, 13(6):1121–1128, 2007.