Support the Data Enthusiast: Challenges for Next-Generation Data-Analysis Systems

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 of next-generation visual analytics services. We argue that these services should have three related capabilities: support visual and interactive data exploration as they do today, but also suggest relevant data to enrich visualizations, and 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 next-generation visual analytics services.

1. INTRODUCTION

The need for effective analysis of data is widely recognized today and many tools aim to support professional data scientists from industry and science with this task. There is, however, another growing group of users who need the ability to analyze data. These users are without formal training in data science. They are called *Data Enthusiasts* [8, 24]. A common example is journalists who increasingly use data and visualizations to illustrate their stories. This paper presents a vision for the next generation of data analysis tools aimed at supporting data enthusiasts.

Recent *self-service visual analytics services* already strive to support these users. Tableau Public [1], Fusion Tables [5], and Many Eyes [22] are among the most popular examples. These tools enable the *sensemaking model* [3]: the typical analytical process starts with a question that a data enthusiast seeks to answer. The data enthusiast then forages 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 or adding dimensions from other datasets.

These existing systems provide several desirable features to support data enthusiasts. They enable users to *visually* explore their data as illustrated in Figure 1, which removes the need for learning any programming or query languages. They facilitate the integration and study of *multiple* datasets at the same time. Finally, they support *collaborations* through sharing visualizations and data online for both viewing and editing by others.

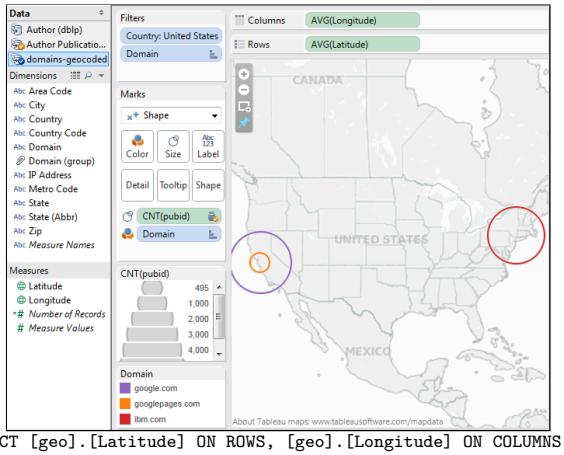
In recent work [16], we found that today's visual analytics services are attracting hundreds or thousands of new accounts each month, but most users author only one visualization and never return. A recent interview study of Open Government Data consumers [6] corroborates that current visualization tools are underserving their users. We see three key limitations that make these systems unsuitable for many users. First, these tools assume the data is clean and in a well-structured relational format, which is typically not the case. As a result, the data cleaning capabilities are minimal and not well integrated with the visualization aspects of the tools. Second, the tools provide little or no help in discovering relevant datasets to enrich a visualization. Users must perform that discovery offline. Finally, the data integration capabilities remain primitive, often limited to equi-joins on fields with the same names.

We present a vision for how visual analytics services should be redesigned to meet users' needs and the associated research challenges. Based on interactions with Tableau customers as well as a study of how Tableau Public and Many Eyes are used [16], we argue that visual analytics services need to improve in four dimensions:

(1) Combined Data Visualisation and Cleaning: (Section 2). 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 [4]). The data is useless until that labor is accomplished up front. Currently, they must use unrelated tools for cleaning and visualization. Since visualizations help identify anomalies and trends, there is an opportunity for combining data cleaning and error detection into the visual data analytics cycle. There is also an opportunity to leverage the collaborative nature of today's tools to propagate cleaning actions across users.

(2) Data Enrichment for Visual Analytics: (Section 3) While many data sources are available on the Web or (more conveniently) shared by other users of the visual analytics service, identifying interesting data to enrich a visualization is challenging. Different datasets have different schemas, different levels of granularity (*e.g.,* we may have state-level unemployment data but zip code-level income data), or different levels of cleanliness. They may also contain different subsets of relevant data. Next-generation visual analytics services should help users *identify* datasets that they can potentially leverage for their current data

Proceedings of the VLDB Endowment, Vol. 7, No. 6




```
SELECT [geo].[Latitude] ON ROWS, [geo].[Longitude] ON COLUMNS,
   [geo].[Domain] ON COLOR, COUNT([pub].[pubid]) ON SIZE
FROM [pub] LEFT JOIN [geo] ON [pub].[Domain] = [geo].[Domain]
WHERE [pub].[Domain] IN {"google.com","googlepages.com","ibm.com"}
AND [geo].[Country] = "United States"
```

Figure 1: The Tableau visual interface and visualization (top) and corresponding VizQL (bottom) shows duplicate entries for Google publications and missing entries for IBM Almaden in the Bay Area.

analysis task. The recommendation needs to take into account the visualizations that the user is creating and could create and not just the underlying data.

(3) Seamless Data Integration in the Context of Visual Analytics: (Section 4) 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. Visual analytics services should support this integration with a focus on producing useful visualizations: It should help the user identify and fix data integration issues in current visualizations and in other visualizations that the user may subsequently derive. It should minimally interrupt the user's analysis task.

(4) A Common Formalism: (Section 5) *Most importantly, the three capabilities above should be seamlessly combined 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 require a mental context switch 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.

While finding, cleaning, and integrating data are each a well-established research area, their exploration in the context of visual analytics services is different for several reasons: First, users are non-technical. Second, all interactions should aim to improve the output of the user's visualizations rather than the base data. Third, the speed should be interactive to support sensemaking. The accepted limit for keeping a user's attention focused on a given task is 10 seconds [15]. All secondary tasks such as cleaning should be minimally disruptive to the user's primary task.

We next present the research challenges of our vision in *supporting data enthusiasts with structured datasets.*

2. DATA VISUALIZATION & CLEANING

Visual interfaces like Figure 1(top) enable data enthusiasts to author sophisticated queries using drag-and-drop actions in a GUI and view the answer(s) through single (or multiple linked) visualizations. This type of visual programming environment supports the sensemaking model well because it provides interactive response times both in authoring the question and in producing visual results; such interactive Q&A is possible because the user only has to focus on the semantics of the query and not its syntax. The first goal of our envisioned system is to enable users to perform data exploration and cleaning through not just any GUI but an *analysis-oriented interface* and as part of their data analysis activity. Visualizations can easily help users identify problems with their data. The research challenge lies in managing the complexity of the transformations required to clean that data while minimizing disruptions to the analysis. The following example illustrates the challenge.

The Tableau visualization in Figure 1(top) demonstrates how integrating two public data sets, DBLP and Freegeoip (http://freegeoip.net), results in dirty data which affects the visualization. In this example, the user specified through GUI actions that she was interested in comparing the academic productivity of the research branches of several top companies. Since DBLP does not contain affiliation information for each publication (only homepages with IP/domain information), she used the homepage as a proxy and combined it with the IP-to-geolocation mapping database, Freegeoip, to produce the map in Figure 1. Two data-related problems clearly manifest. First, we see that certain research branches are missing: there is one red circle for IBM's publications from TJ Watson in the NY area, but other actively-publishing IBM branches such as Almaden in the Bay Area are strangely absent. Second, we see evidence of duplicate entries. Google, for example, has two representations in the Bay Area on this map: `google.com` (purple circle) and `googlepages.com` (orange circle).

Recent work [25] also uses visualizations to reveal outliers that correspond to errors in the input data set. Another system [13] generates annotations to help explain outliers and trends in point-based visualizations. However, many data quality issues are much more difficult to identify and resolve than what existing tools support; entity resolution is one example. In the above scenario, the user could easily indicate to the system that some data appears to be missing (*e.g.,* IBM Almaden) and other data seems to be duplicated in the visualization (*e.g.,* the two Google records). However, determining what needs to be done to fix the problem is non-trivial: should the user find additional datasets to integrate, or perform entity-resolution on the existing data, or manually fix some entries? Even if entity resolution is a clear approach, existing methods are computationally expensive, yet visual analytics impose a real-time computation constraint: data cleaning is a disruption. Current tools [12] apply cleaning transformations over the entire data set, regardless of how it is being explored by the visual analytics system. Computation time is wasted on transforming data that is not used in the visualization. This problem worsens as data sizes increase. One promising approach is to prioritize cleaning the records that impact the visualization first. This view-at-a-time approach will need new incremental approaches to cleaning, where the objective is to resolve the records that affect the view, not the entire input dataset.

Additionally, the system could suggest relevant cleaning actions performed by others (*e.g.,* captured as scripts by the formalism). The second challenge then is how to manage and share such cleaning actions among different users. Suppose the user zooms into the Google-specific publications and notices some duplicate entries due to inconsistent spellings for conference names. These fixes must be reflected in the user's current visualization but should they be applied to the base data, which may be publicly available in the system and shared with other users? Should they be applied automatically to data beyond what is being visualized? Should other users have the option to access the cleaned data? What if some of the cleaning actions are wrong? Should the system manage different versions of datasets with different scripts of cleaning actions applied to them? Some systems have been developed for managing conflicting updates [10], but focus on experts, and require the specification of trust relationships, which is inconsistent with global sharing by non-technical users.

3. DATA ENRICHMENT

An important step in the analysis process is to add context to a dataset by combining it with other relevant data sources. 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 add useful context to the analysis, however, is 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. 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 those 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 [7] that data recommendation is challenging due to schema and semantic mismatches between datasets (among other challenges). Recent prior work [19] has tackled the problem of finding tables from a large Web corpus that are related to an input table. This tool, however, only considers extending each entity, as identified by a set of key attributes, with additional fields. In recent work [16], we observed 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.

Given the large volumes of public data that are searchable and the real-time constraints of an interactive visual analysis system, data recommendation is especially challenging. To limit the search space and to improve recommendation quality, the data recommender should leverage the context of the visualization of the current data source. Our idea is to 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, or aggregation/filters/projections used. One key challenge is to identify the pertinent information in a user's current visualization that is most relevant for 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. Generalizing this intuition to other types of views and other types of information, however, is not trivial.

Another question is whether the recommender could better support data enthusiasts by leveraging historical actions by expert users (*e.g.,* how others have integrated and used similar datasets) when recommending new data? 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.

4. SEAMLESS DATA INTEGRATION

Data integration is a well-known difficult problem. To assist data enthusiasts, Tableau offers a pay-as-you-go data integration feature called *data blending* [17]. This feature automatically creates mediated schemas and wrappers as the user interactively builds a visualization on-the-fly. It also joins in only the necessary information from a data source (*e.g.,* as specified by the user through the GUI) to create the view with minimal data movement, as queries are federated to the data sources. A key aspect of the Tableau data blending feature is its ability to integrate data without causing any significant disruption to the analysis cycle: Tableau automatically infers how to combine the two datasets. Its inference abilities, however, are limited to joining datasets on the columns that share the same name and aggregating the new dataset if necessary and possible. The next-generation of such tools should expand these automatic data integration capabilities by taking the visualization context into account. While the general problem statement is not specific to visual analytics, its solution in that context raises new opportunities: For example, if a user is viewing data displayed on a map, the system should try to determine if any columns in the new dataset can possibly correspond to geographic locations. Other types of integration are not relevant in the current context. In contrast, if a user examines a bar chart (e.g., average salary per category of worker), the x-axis labels are good candidates for the join key. Furthermore, successful integration of only the rows corresponding to the visualized data is all the user needs at that point.

Beyond automated inference, a second promising approach is to leverage the work done for prior visualizations by the same user or by other users. Imagine, for example, two tables with salary information for two different companies and a user who 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 research opportunity is to identify such relevant actions by past users, determine how and when they generalize, 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. Extending further, if a user is looking at weekly network traffic data, the system could automatically suggest to break down the information at the granularity of hours to join with intrusion detection statistics at that granularity.

Of course, it remains to be shown how far such reuse will go toward solving the general problem.

Once schemas have been mapped, the next problem is to ensure that the join produces meaningful results. Inconsistent representations of the same entity across data sets (as in the entity resolution problem) can compromise the visual analysis and affect decision making as we discussed in Section 2. Current integration systems [20, 5, 11] provide users no assistance with detecting or correcting data quality issues. Other integration systems [21] offer basic cleaning operations, but lack support for more complex scenarios such as resolving duplicate entities.

A final core challenge is how to handle the case when the user continues to interact with the integrated visualization. What if the user alters the granularity of the visualized data by drilling-down to the details or rolling-up to summarize. As an 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. To accomplish this task, the data integration 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 data integration 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.

5. A COMMON FORMALISM

To realize this vision, we advocate developing a single formalism for a fully integrated visual analysis sensemaking cycle. In the existing Tableau system, VizQL and its underlying data model serve the purpose of capturing a formal specification of a user's actions and their mappings onto underlying database queries. For example, the visualization shown in Figure 1 is driven by the VizQL query shown below it. These SQL-like statements perform the task of querying the underlying data source (the VizQL queries are compiled into SQL or MDX queries) and rendering the results visually. The formalism represents the clear semantics underlying the tool. The user does not author the language directly, but rather her interactions with the data through the GUI automatically result in the generated code.

VizQL, however, supports only data exploration. Similarly, there has been work on supporting either data cleaning [12], data integration [18], or collaboration [9, 23] from a GUI, however, no tool today supports the complete analysis cycle. A simple union of state-of-the-art tools is insufficient for supporting data enthusiasts in two ways. First, we need to integrate activities that are both typically performed using different interfaces and that yield very different actions: visual analysis creates views over underlying base data, data cleaning edits the data, and data integration creates schema mappings, a mediated schema, and wrappers for data sources. Cleaning integrated data may require changes to these mappings as well as changes to the underlying data sources. Second, these different tools typically operate at different time-scales. For example, current entity-resolution systems [2] require that users label tens to hundreds of examples and then take minutes to hours to run, which is at odds with keeping the user focused on their data analysis tasks. Larger data sets put additional pressure on maintaining the interactivity required by the sensemaking 10-second window. Current approaches [14] include pre-computing data cubes and parallelizing the workload. However a unified system with a formal language requires extending such optimizations across all actions in the analytical workload.

6. AND BEYOND

Extending our vision for unstructured or semi-structured data presents even greater challenges. For unstructured data, users typically apply data extractors (from n-grams to sentiment analysis). Since most of these extractors are approximate, the challenge is how to assist the data enthusiast in analyzing, cleaning, or integrating such approximate, probabilistic, and possibly conflicting data. For semi-structured data, visualizations may contain errors due to the heterogeneity in the structure of the data (*e.g.,* address represented as one string vs. a set of tokens), further complicating the identification and cleaning of the data.

7. ACKNOWLEDGEMENTS

This work is partially supported by NSF CDI grant IIA-1028195. We thank Dan Halperin, Marianne Shaw, and the anonymous reviewers for their helpful feedback.

8. REFERENCES

[1] Tableau Public. http://www.tableaupublic.com/, 2012.
[2] K. Bellare et al. Active sampling for entity matching. In *SIGKDD*, 2012.
[3] S. K. Card et al. Using Vision to Think. In *Readings in Information Visualization*. Morgan Kaufmann, 1999.
[4] T. Dasu et al. *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons, New York, NY, 2003.
[5] H. Gonzalez et al. Google Fusion Tables: Data Management, Integration and Collaboration in the Cloud. In *SOCC*, 2010.
[6] A. Graves et al. Visualization tools for open government data. In *Proc. of the 14th International Conf. on Digital Government Research*, 2013.
[7] A. Halevy et al. *Principles of Data Integration*. Morgan Kaufmann, 2012.
[8] P. Hanrahan. Analytic database technologies for a new kind of user: the data enthusiast. In *SIGMOD*, 2012.
[9] D. Huynh et al. Piggy bank: Experience the semantic web inside your web browser. In *Proc. of ISWC*, 2005.
[10] Z. G. Ives et al. The orchestra collaborative data sharing system. *ACM SIGMOD Record*, 37(3):26–32, 2008.
[11] Z. G. Ives et al. Interactive data integration through smart copy & paste. In *CIDR*, 2009.
[12] S. Kandel et al. Wrangler: Interactive Visual Specification of Data Transformation Scripts. In *CHI*, 2011.
[13] E. Kandogan. Just-in-time annotation of clusters, outliers, and trends in point-based data visualizations. In *VAST*, 2012.
[14] Z. Liu et al. immens: Real-time visual querying of big data. In *EuroVis*, 2013.
[15] R. Miller. Response Time in Man-Computer Conversational Transactions. In *AFIPS Fall Joint Computer Conf.*, 1968.
[16] K. Morton et al. A Measurement Study of Two Web-based Collaborative Visual Analytics Systems. Technical Report UW-CSE-12-08-01, U. of Washington, Aug 2012.
[17] K. Morton et al. Dynamic Workload Driven Data Integration in Tableau. In *SIGMOD*, 2012.
[18] A. Raffio et al. Clip: a Visual Language for Explicit Schema Mappings. In *ICDE*, 2008.
[19] A. D. Sarma et al. Finding Related Tables. In *SIGMOD*, 2012.
[20] M. Stonebraker et al. Data curation at scale: The data tamer system. In *CIDR*, 2013.
[21] R. Tuchinda et al. Building mashups by example. In *IUI*, 2008.
[22] F. B. Viegas et al. Many eyes: A site for visualization at internet scale. *IEEE TVCG*, 13(6), 2007.
[23] W. Willett et al. Commentspace: Structured support for collaborative visual analytics. In *CHI*, 2011.
[24] G. Wolf et al. The Quantified Self. *TED*, 2010.
[25] E. Wu et al. Scorpion: Explaining Away Outliers in Aggregate Queries. In *VLDB*, 2013.