A Measurement Study of Two Web-based Collaborative Visual Analytics Systems

Kristi Morton[1], Magdalena Balazinska[1], Dan Grossman[1], Robert Kosara[2,3], Jock Mackinlay[3], and Alon Halevy[4]

[1]University of Washington
{kmorton,magda,djg}@cs.washington.edu

[2]UNC Charlotte
rkosara@uncc.edu

[3]Tableau Software
jmackinlay@tableausoftware.com

[4]Google
halevy@google.com

ABSTRACT

In this paper, we present a longitudinal study of the use of two popular Web-based, collaborative visual analytics systems, Tableau Public and Many Eyes. As data has become more widely accessible through the Web, online visual analytics systems have emerged as a popular tool for data analysis and sharing. In spite of their growing popularity, however, little is known about how these systems are being utilized. The study presented in this paper addresses this shortcoming and shows details about the workloads of these systems, their users, the types of analysis they perform over single or integrated datasets, and their degree of collaboration. To the best of our knowledge, this is the first study of its kind, and presents important details about the user of online, visual analytics systems.

1. INTRODUCTION

As data has become more publicly available on the Web, for example, through local and national government initiatives such as the Open Data movement [10], a broader audience has emerged as data consumers and knowledge-seekers (referred to as *data enthusiasts*) [14]. These people are not mathematicians or programmers, and believe that data can be used to answer a question or solve a problem. A typical example is a news reporter who wants to use data and visualizations to illustrate a story and make it available online (*e.g.* on her blog). As a result, online data visualization systems have increased in popularity to meet the demands of such users for data analysis and sharing [2, 1, 6, 22, 9, 12, 13].

The core functionality of these systems is threefold: (1) They enable users to *visually* explore their data: users have access to a graphical user interface through which they can easily create various charts and graphs as illustrated in Figures 1 and 2. (2) These systems also facilitate the integration and study of *multiple* datasets at the same time. (3) Finally, they support *collaboration* between users through sharing visualizations and data *online* for both viewing and editing by others [15, 24].

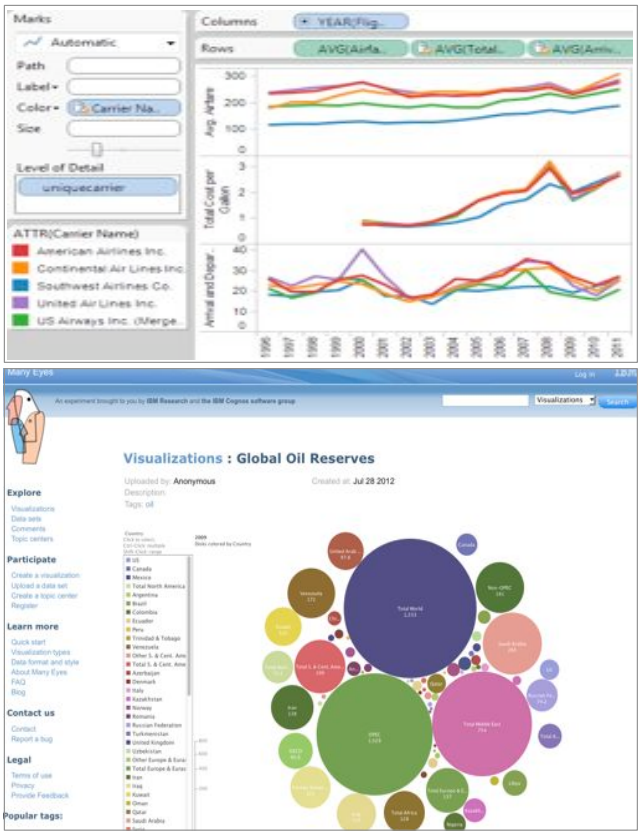

Figure 1: Screen Captures of Tableau (top) and Many Eyes (bottom)

While different systems have different architectures, several are based on the integration of a visualization front-end with a database management system (DBMS) back-end [23, 12]. For example, Tableau's architecture (see Figure 3) supports analysis across a variety of heterogeneous data sources (*e.g.* files, cubes, data marts, and databases), and issues live queries to these sources to obtain the necessary data to render each visualization.

In spite of their growing popularity, little is known about how these systems are being used. In most cases, even basic statistics such as the number of users are not published (*e.g.* Fusion Tables), let alone any details of user activity. As our

Proceedings of the VLDB Endowment, Vol. X, No. Y

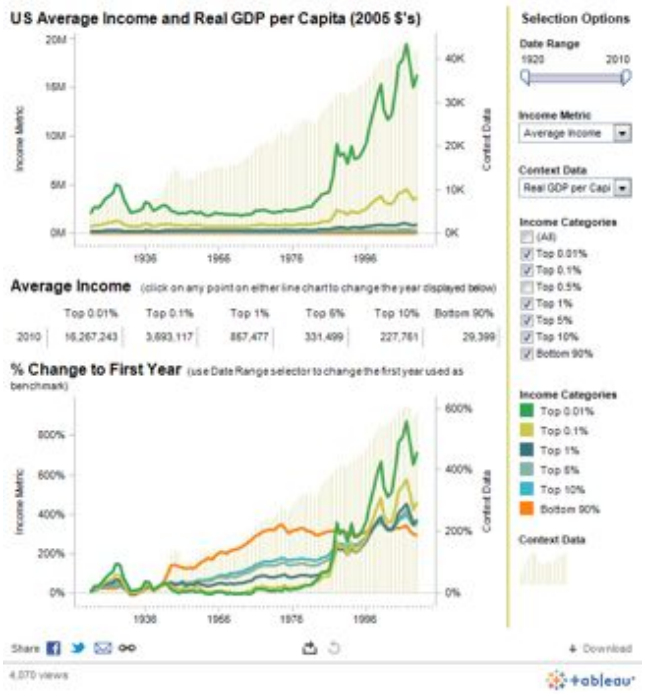

Figure 2: Screen Capture of Tableau Public

society continues to become "data-enabled", it is important that we continue to improve data management and analysis tools. If we are to build better online, data visualization and sharing systems, the first step is to understand how they are being used today. The key contribution of this paper is to shed light on this exact question: *How are online data visualization and sharing systems being used?*

We take a first step toward answering this questions through a longitudinal measurement study of two popular online data visualization and analysis systems: Tableau Public [6] and Many Eyes [22, 3]. Both systems allow users to create visualizations online, and both are free to use. Tableau Public requires the download of a Windows-only client, while Many Eyes is used entirely in the browser. Both systems provide a variety of different visualization techniques, which not only generate static images, but which the viewer can interact with in the browser. The data used in visualizations can be downloaded in both systems.

We tackle the question of how both of these systems are being used from the perspective of the database community. Through our study, we thus focus on the following core set of more detailed questions: (1) How heavy is the workload of these systems? And how is it evolving over time? (i.e., size of user-base, its growth, amount of data that users want to analyze); (2) What types of users are leveraging these systems? Are users mostly novices or do they accumulate expertise and become increasingly power-users? (3) What do users actually do with the data? How do they analyze it? How much data (in terms of relation cardinality and degree) do users choose to visualize at any given time? (4) Do users collaborate with each other? What is the extent of these collaborations? And finally (5) Do users integrate multiple data sources in their visualizations? And how do

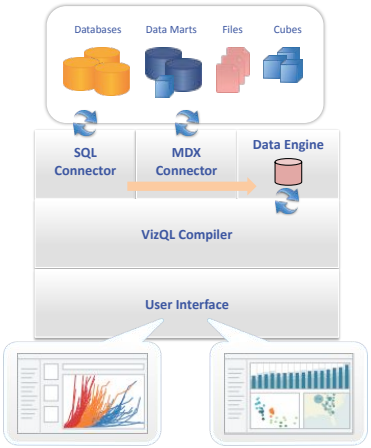

Figure 3: Architecture of Tableau

they perform these integrations?

Our study is based on a trace of both systems, and a high-level summary is provided in Table 1. Each trace includes detailed information (spanning multiple years) about the data and visualizations that are published to each system. Additionally, the Tableau Public traces include detailed traffic and impression data for each visualization. To the best of our knowledge, this is the first formal such study of these types of systems. We present a more detailed overview of these systems and the measurement data in Section 2.

Our key findings are focused along the following four dimensions:

- *Overall Workload and Author Demographics:* (Section 3.1) We first study the characteristics of the overall workload in these systems and the high-level behavior of users. Through our study, we find that these systems are growing in popularity, acquiring hundreds to thousands of new users every month. Retention, however, is a problem. Only a small fraction of users are active in these systems long-term. Today, these systems are also designed to analyze *small* data. Tableau Public, for example, limits users to only 50MB of data per account and the vast majority of users does not reach this limit. Furthermore, we find that both Tableau Public and Many Eyes authors consist primarily of one-time and light users who publish only a handful of datasets and visualizations. This is not to say that these systems are not useful. Indeed, several thousand new visualiations are created each month in either Tableau Public or Many Eyes and, together, these visualizations are viewed by millions of users.

- *User Interactions and Collaborations:* (Section 3.2) Next, we assess how often users view or interact with the online published visualizations and how often they collaborate over a common dataset or visualization (and subsequently publish new insights). We find that collaborations take primarily the form of amplifying a user's impact by attracting large numbers of viewers. For instance, we measured 20 million unique visitors to Tableau Public. However, a very small fraction of viewers take further action with the visualizations:

only 2.4% are downloaded, modified, and republished by other authors. We also see a similar trend on Many Eyes where 6% of the datasets are shared among different authors.

- *Single-Dataset Analytics* (Section 3.3) We next study the details of how users perform their visual analysis on a single data source. We find that most visualizations are powered by small data sizes (*e.g.* 90% of datasets on Tableau Public contain fewer than 22,000 rows and for Many Eyes 90% contain fewer than 1,000 rows) and use few columns from the data (*e.g.* 50% of Tableau Public visualizations use up to three columns) relative to what is available (*e.g.* 50% of the data sources contain up to 28 columns). Furthermore, users of Tableau Public tend to visualize their data primarily with simpler visualizations which require fewer columns, such as bar views (37%), and Many Eyes visualizations tend to be more text-centric (40%).

  Overall, there is thus an obvious gulf between the data that is available and the data that is actually used in the visualizations on both systems. While visualizations are often used to summarize the data at an aggregate level, these systems support complex visualizations of multi-dimensional data, yet they remain underutilized.

- *Multi-Dataset Analytics* (Section 3.4) Finally, we asses how users perform their visual analytics tasks on multiple (joined) data sources. Compared to their single dataset counterparts, multi-dataset visualizations tend to use more columns: 45% contain more than five columns while only 14% of single source visualizations contain more than five. We also find different trends in visualization techniques for those containing multiple data sources versus single ones. For example, multi-dataset views tend to be more prevalent in maps and scatter views; these visualization types tend to be more complex (*i.e.* use more columns). While places and time collectively account for almost half of the join keys, the most common type of join key is one that represents object entities such as alphanumeric identifiers. Overall, these results indicate that multi-dataset visualizations tend to leverage more columns from their data sources and thus tell a richer story.

2. BACKGROUND AND METHOD

This section first presents an overview of the Tableau Public and Many Eyes visual analysis systems, and later discusses in detail the collected statistics on the systems.

2.1 Tableau Public Overview

Tableau Public is an extension to the desktop-based Tableau visual analysis system. It is also a cloud service. We first present Tableau and then the Tableau Public extension.

2.1.1 Tableau

Tableau [7] is a data visualization tool that sits between the end-user and the database and allows the user to create visualizations by dragging-and-dropping fields from their datasets onto a visual canvas as illustrated in Figure 1(top). In response to these actions, the system issues formal VizQL (Visual Query Language) [21] statements to build the requested visualization (see Figure 3). VizQL is a structured query language with support for rendering graphics. Each VizQL statement is compiled into the SQL or MDX queries necessary to generate the data for the visualization.

The platform currently supports connections to several data sources, including Microsoft Access, Microsoft Excel, delimited text files, OData [5], and Microsoft Azure Data Marketplace [4]. Additionally, the following visualization types are supported: bar view, text tables, line view, pie view, scatter view, map view, gantt view, area view, and circle view.

2.1.2 Tableau Public

Tableau Public is a Web-based, collaborative platform for visual analysis that launched in February of 2010. It is free-to-use, but has the following restrictions: visualizations are limited to 100K rows of data, accounts are limited to 50 MB of storage, and all content is published to the Web-facing Tableau Public servers. All Tableau Public visualizations and data are public; any user with a Web browser can view, interact with, and download the visualizations and data. Also, any published visualization can be embedded on other websites or shared through social media or email.

In order to publish any content to the Tableau Public server, a user must first create an account and also have installed the Tableau Desktop client. The user first operates on his or her data locally using Tableau: The user opens the Tableau Desktop Client, connects to any appropriate data sources, and creates desired visualizations. Once the visualizations are ready, the user can publish them. When the user publishes a collection of visualizations from the client to the Tableau cloud, the subset of the back-end data used in the visualizations is copied over to Tableau Public. The details of the visualizations on top of the data are also copied. Once the visualizations are published, any user in the world can view them through their Web browser. These users can also interact with these visualizations (*e.g.* by filtering content, drilling down to details, and pivoting on data sources), but cannot persist their changes to the server without the Tableau Desktop Client.

2.1.3 Terms

In this section, we clarify the common nomenclature found in Tableau's visual analytics platform. First, authors can publish two primary types of visual content: visualizations (called *worksheets*) and and collections of visualizations (called *workbooks*). Additionally, the data supporting each visualization (*i.e.* the *view*) is published along with each workbook. Visualizations can be created with a single data source or by joining multiple data sources together, referred to as *blending* [18]. The data blending feature allows users to join data on-the-fly from multiple heterogeneous sources without having to specify any of the schema mappings or mediators. A user authors a visualization by dragging the columns from an initial data source (referred to as a *primary* data source) which establishes the context for subsequent blending operations in that visualization. Data blending happens when the user drags in fields from a different data source, known as a *secondary* data source. Additionally, the visualization can be further modified by, for example, adding more secondary data sources or drilling down to finer-grained details. Finally, *authors* are users who have published their data and

System	Start Date	End Date	# Visualizations/Worksheets	# Workbooks	# Datasets	Users
Tableau Public	February 10, 2010	March 31, 2012	165,144	45,598	63,812	15,729
Many Eyes	January 1, 2007	March 31, 2012	125,249	n/a	286,440	37,904

Table 1: Tableau Public and Many Eyes Trace Information

visualizations to the online repository.

2.2 Many Eyes Overview

Many Eyes [3, 22] is a Web-based visualization service that allows users to upload datasets and create visualizations. Unlike Tableau Public, all visualizations are created and published directly through a Web browser. The site was launched in early 2007 by IBM's *Visual Communication Lab* as the first online service that provided ways to not only create static charts, but interactive visualizations that could easily be embedded in blogs and other websites. While both systems share many of the same view types (*i.e.* bar, line, text, pie, area, scatter, and maps), Many Eyes includes a number of unique techniques that are not available in any other software. In particular, Many Eyes' text views – including word clouds, phrase nets, and word trees – allow users to experiment with text data in ways that are still unmatched in most other visualization tools and services.

In addition to the visualization tools themselves, Many Eyes also pioneered the notion of social visualization. The typical Web 2.0 feature of leaving a comment has the added twist that it also contains a live thumbnail reflecting the configuration of the visualization the user was looking at when writing the comment. Users can also browse existing data sets and visualizations and create new ones from what others have uploaded. Many users would thus benefit from the work of somebody scraping or otherwise collecting data.

2.3 Statistics Collection and Approach

All statistics and measurements shown for Tableau Public come from a snapshot of the Tableau Public online repository taken from the end of day March 31, 2012. This snapshot contains information about all operations performed by users since the service launched in February 10, 2010. The data collected thus spans 9 three-month quarters: from Q1 2010 through Q1 2012. Additionally, the trace for Many Eyes was collected from end of day January 1, 2007 until end of day March 31, 2012 (*i.e.* spanning 21 three-month quarters). The content of the statistics repository for both systems is summarized in Table 1. The repository contains 15,729 Tableau Public user accounts and 37,904 Many Eyes accounts. Each account contains all the data sources, workbooks, and worksheets (and their metadata) published by the user since registering. For Tableau Public, each workbook specifies the data sources analyzed (including all of the schema metadata), the types of visualizations produced, and all of the specific VizQL queries that produce each visualization. In addition to these user-level statistics, our repository includes detailed traffic and impressions information for each visualization on Tableau Public. For Many Eyes, however we have a limited subset of the information available in Tableau Public that includes only user-level information (no traffic/impressions data is available). Each user account contains aggregated statistics about the number of data sources (no detailed schema-level info), number visualizations produced, and the visualization types used.

Since the two collected traces are not identical, in the rest of the paper, we show graphs for both datasets when possible and show only data for one system when the corresponding data was not available for the other system.

3. MEASUREMENT RESULTS

In this section, we present the key measurement results organized around our four core questions of overall workload and author demographics (Section 3.1), user interactions and collaborations (Section 3.2), single-dataset analytics (Section 3.3), and multi-dataset analytics (Section 3.4).

3.1 Workload and Author Demographics

We first explore the primary characteristics of authors and their usage of data for visual analysis in both Tableau and Many Eyes.

What is the growth of new authors?

Since its inception in February 2010, Tableau Public has grown to a user-base of around 16,000 authors who have contributed over 45,000 workbooks, 63,000 datasets, and 165,000 visualizations (worksheets). For Many Eyes, the user-base has grown to 38,000 authors who published over 286,000 data sources and 125,249 visualizations (there is no equivalent version of a workbook on Many Eyes).

Figure 4 shows the detailed cumulative quarterly growth of new authors on Tableau Public and Many Eyes. We see that the growth has steadily increased, reaching hundreds of new users per month for Tableau Public and thousands of new users per month for Many Eyes. Since Many Eyes has been in existence longer than Tableau Public, it has grown to 37,904 accounts over the course of 21 quarters. By the end of its 9th quarter, Tableau Public has grown to 15,729 unique user accounts. After 9 quarters, Many Eyes had 9,004 users. These systems thus have moderate numbers of users today, but their popularity is continuing to grow.

How productive are authors in publishing content?

There are two types of content that can be published to Tableau Public and Many Eyes: data sources and visualizations (for Tableau Public authors publish at the level of whole workbooks). Figure 5 shows the Cumulative Distribution Functions (CDFs) of the number of published data sources and visualizations per author on Tableau Public and Many Eyes. The publication trends for both systems are quite similar. For Tableau Public, 47% of the authors uploaded one data source, 83% contributed at most five, and 90% contributed at most eight. Similarly for Many Eyes, 51% published one dataset, 90% published at most five, and 95% published at most eight. Overall these statistics reveal that most authors on Tableau Public and Many Eyes publish only a few data sources. Finally, the long tail starting around 14 data sources shows that while these most prolific authors are the minority (with 2–5% representation), their contributions are quite varied (ranging up to 623 data sources for Tableau Public and 12,119 for Many Eyes).

We see similar trends in workbooks published per author

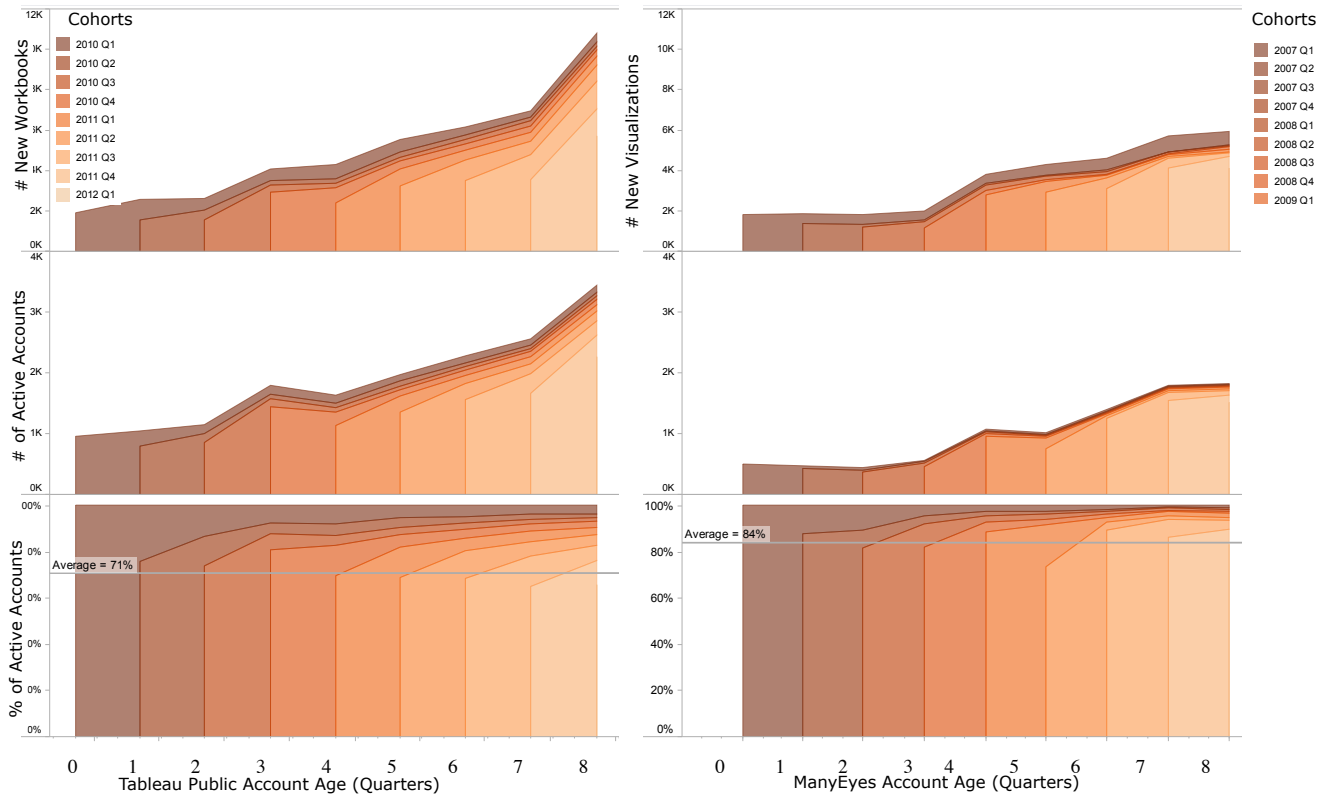

Figure 6: Cohort Analysis of Tableau Public (left) and Many Eyes (right) Authors (includes avg. of new author participation in each quarter)

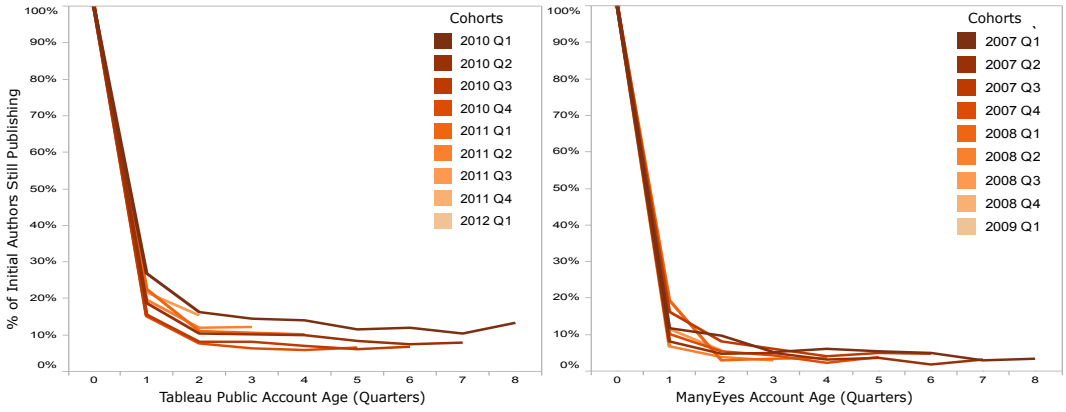

Figure 7: Publication Activity of Tableau Public (left) and Many Eyes (right)

in Figure 5 for both systems. On Tableau Public, 55% of authors publish one workbook, and 90% publish at most five. Furthermore, for Many Eyes 44% publish one visualization and 90% publish at most four. There is also a long tail of the remaining top 10% contributing authors. The number of workbooks (visualizations for Many Eyes) published for these authors range from six to 515 for Tableau Public and five to 8,820 for Many Eyes.

Given these results, we categorize Tableau Public and Many Eyes authors into three main groups based on their publication activity:

1. One-time users:
   (a) Tableau Public: 47% publish one dataset and 55% publish one workbook of visualizations
   (b) Many Eyes: 51% publish one dataset and 44% publish one visualization
2. Light users:

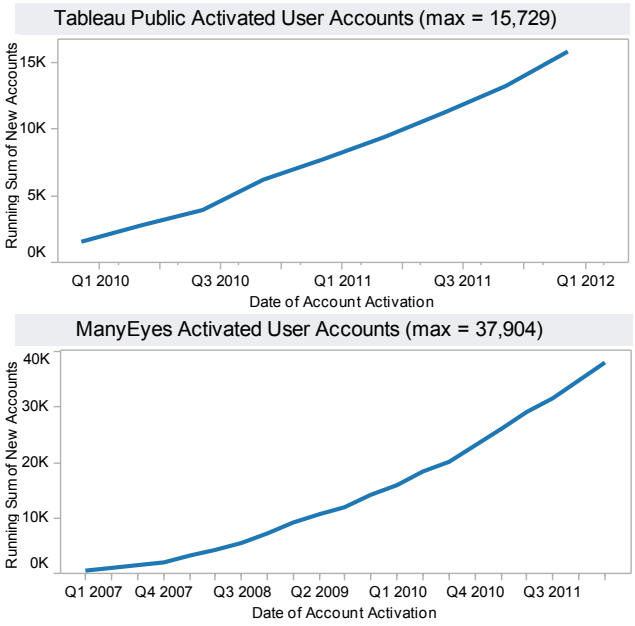

Figure 4: Tableau Public and Many Eyes Accounts

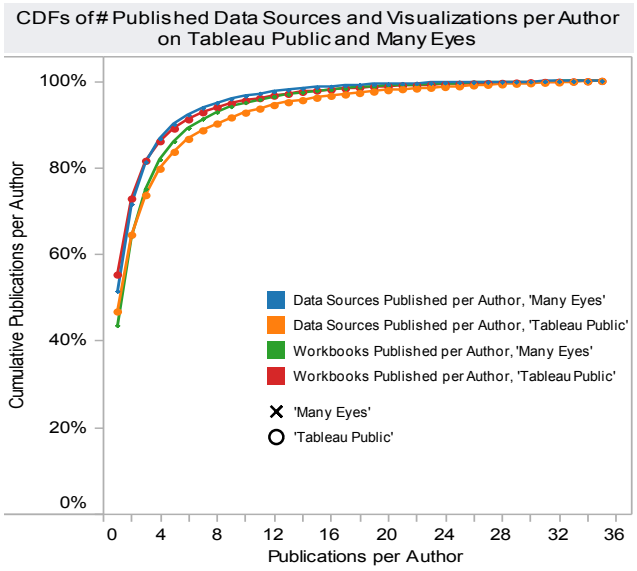

Figure 5: CDFs of Tableau Public and Many Eyes per Author Publications of Data and Visualizations

(a) Tableau Public: 36% publish 2–4 data sources and 35% publish 2–4 workbooks
(b) Many Eyes: 39% publish 2 - 4 data sources and 42% publish 2–4 visualizations

3. Prolific users:
   (a) Tableau Public: 17% publish 5 or more datasets and 10% publish 5 or more workbooks
   (b) Many Eyes: 10% publish 5 or more datasets and 14% publish 5 or more visualizations

The user-base is thus dominated by one-time and light users.

What is the author retention/churn?

Figures 6 and 7 further study the trends of author churn. The figures show a cohort analysis involving segmenting the Tableau Public and Many Eyes author groups based on their account activation date and shows how these specific, unchanging groups of users behave over the same periods in their respective user life-cycles. We group users into cohorts based on the quarter in which they published their first workbook and track their publications. For both systems we trace their cohort publication activities for 9 quarters.

First, Figure 6 shows that by the end of the 9th quarter, Tableau Public has over 3,000 actively-publishing authors and Many Eyes has 1,800 active accounts. Additionally, over 10,000 new workbooks were published on Tableau Public in the 9th quarter (6,000 on Many Eyes). Even though the total number of users and workbooks is greater in Tableau Public than Many Eyes, both systems show a strikingly similar pattern in terms of workload distribution between new and returning users (see the last pair of graphs in Figure 6). For example, in the last quarter, 66% of the actively publishing authors on Tableau Public were new and had contributed 53% of the published workbooks that quarter. For Many Eyes, 83% of the active user accounts were created in the 9th quarter; they contributed 71% of the new workbooks. Overall, both systems exhibit the trend of high author turnover. Looking at the percent of actively publishing accounts by new authors for each quarter, Many Eyes averages at 84% and Tableau Public averages at 71% (not taking the first quarter into account since all accounts at that point are new).

To better understand this trend, Figure 7 shows the percentage of initial authors in each cohort that publish again in later quarters. As expected (by design of the graph), for the first quarter (i.e 0) 100% of the active authors in each cohort publish an initial workbook. We then see a significant drop off in the next quarter; less than 30% of these initial authors for each cohort on Tableau Public publish again. Many Eyes has a similar retention trend for their initial authors: less than 19% return again to publish.

Low retention after initial use is, naturally, common for free, Web-based services. For example, according to a 2009 Nielsen report [16] only 40% of Twitter users returned to use the site after the first month. However other websites like MySpace and Facebook achieved retention rates closer to 60%. This result was measured for all three systems at the same point in their respective user growth curves.

Are authors limited by the size of their accounts?

According to Figure 8 (top), we see that 90% of user accounts use less than half of their 50MB quotas. Since each account contains datasets and workbooks consisting of a collection of visualizations, we further study the sizes of workbooks published on the site (see bottom of Figure 8) to see if the sizes of the visualizations are a limiting factor. The figure shows that 90% of all workbooks are less than 762KB in size – which means that most authors can publish multiple workbooks to their accounts and still be well under the 50MB quota.

Are users limited by the query processing times?

According to a study published on Web users' tolerable wait-

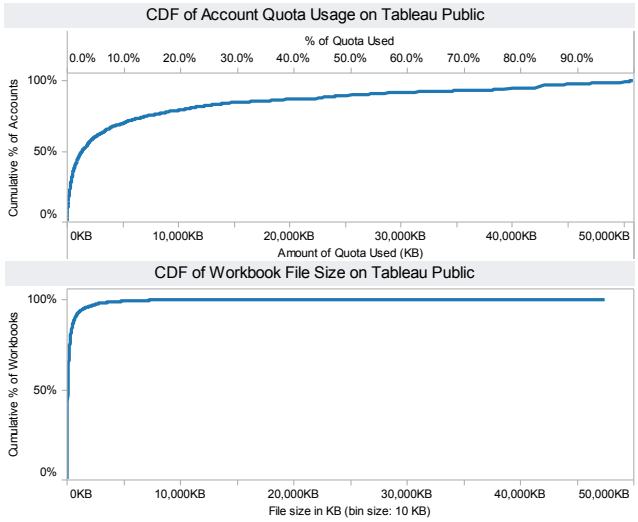

Figure 8: CDFs of Account Quota Usage (top) and Workbook File Size (bottom) on Tableau Public

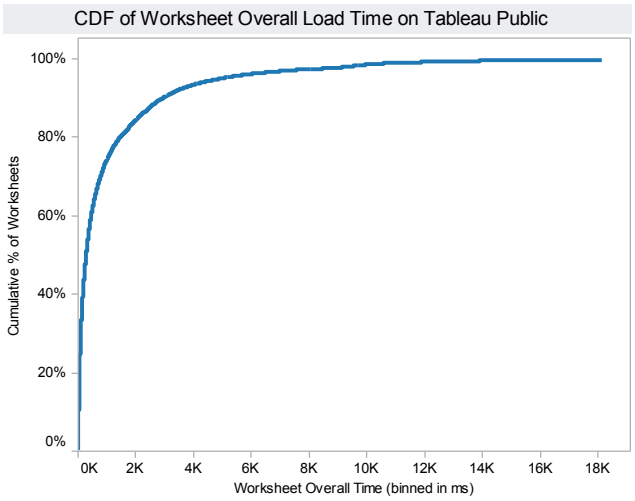

Figure 9: CDF of Worksheet Load Times on Tableau Public

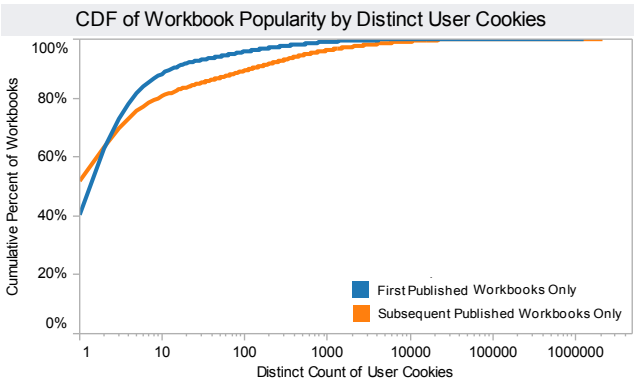

Figure 10: CDF of Workbook Popularity on Tableau Public (max # user cookies = 20.9 Million)

ing time [19], 2 seconds is considered an acceptable waiting time for loading Web pages. In Figure 9 we see that 84% of all visualizations on Tableau Public take less than 2 seconds to load (includes both query and rendering time) and 98% are under 10 seconds (the accepted limit for keeping a user's attention focused on a given task [17]). Although attitudes and expectations change over time, the basic capability of human attention has not changed over the decades [11, 17]. Thus, our results indicate that the majority of load times should not negatively impact Tableau Public's users.

Discussion: The demographic results thus show a continued growth in users but a low retention rate of these users. The overwhelming majority of users are either "one-time users" or "light" users. At the same time, users do not appear to be hindered by constraints on the size of their accounts or the size of their visualizations. Similarly, query performance is below well-known thresholds for user attention. A few direct implications of these results are that (1) online visual analytics systems today have a user-base primarily comprised of users with little to no experience. At the same time, (2) while attracting new users to these systems is not a problem (Figure 4), retaining them beyond their first visualization appears to be a critical challenge, which appears not to be due to performance nor account-size limitations. There must be other more fundamental causes (perhaps relating to usability or the fact that users tend to not be regular visualization creators) that lead users to abandon the site. Finally, these systems focus strictly on small-data users. It would be interesting to see if the above trends would change if the systems had support for big-data users.

3.2 User Interaction and Collaboration

Since both systems are designed for sharing visualizations and collaboratively analyzing data, we explore the frequency of viewership, collaboration, and sharing in this section.

3.2.1 Users who simply view and interact (read-only)

Based on a distinct count of user cookies, we found that there are around 20.9 million unique visitors to Tableau Public. The visitors are thus *several orders of magnitude* more numerous than the authors (only 16,000 authors), Additionally, we found that the top 50% of all Tableau Public traffic is attributed to 98 distinct workbooks (or 0.2% of all workbooks). For the results presented in this subsection, we did not have access to the equivalent traffic and viewership information for Many Eyes.

Figure 10 shows the distribution of workbooks by their viewing popularity. In this graph, we split the workbooks into two groups to compare their relative popularity distributions: those that were an author's first publication and those that were a later publication. Since first-time publications make up a sizable fraction of the overall total number of publications (29%), we observe the viewership trends of this group of workbooks in comparison to the trends of subsequent published workbooks. In this figure, we see that 42% of workbooks that were an author's first publication on Tableau Public are only viewed by a single user. Interestingly, we see that, likewise, 53% of subsequently published

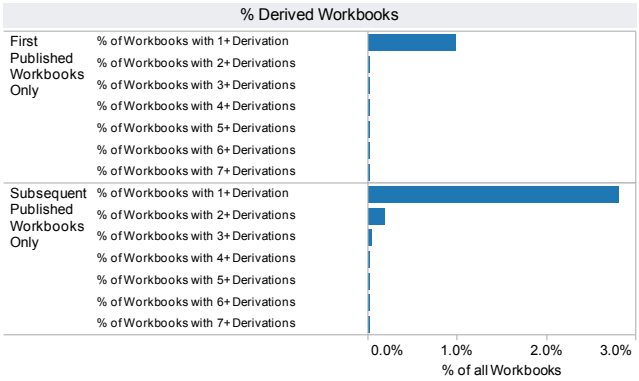

Figure 11: Tableau Public Workbooks Derived by Other Authors

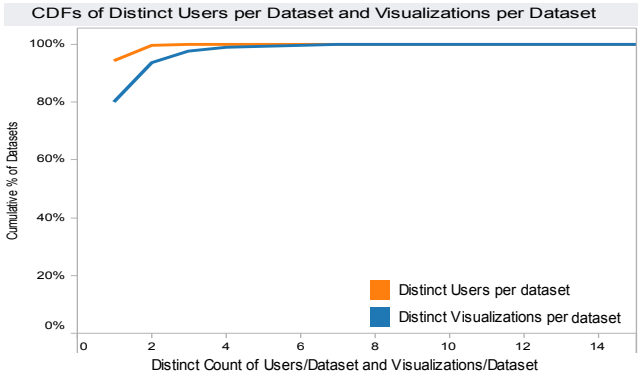

Figure 12: CDFs of Users per Dataset and Visualizations per Dataset for Many Eyes

workbooks are viewed by one user. As expected, the curve for the most popular workbooks that were an author's first publication is sharper than workbooks that were not first publications (*i.e.* popular workbooks tend to not come from first-time authors): At the 90th percentile, we see almost an order of magnitude difference in viewership with only 13 unique users for first-time publications compared to as many as 120 for workbooks that were not the author's first publication. The top 1% of first-time publications received at least 1,500 views, with a maximum viewership of over 1.2 million. In contrast, the top 1% of subsequent publications received at least 10,000 views with a maximum viewership of 2.1 million.

3.2.2 *Users who view and collaborate (read-write)*

To get a sense of the degree of collaboration between authors, where multiple authors edit the same visualization, on Tableau Public, we explore how often authors take existing content and evolve it for their own analytical needs (*e.g.* by changing the visualization content to explore some other dimension or measure) and then republish it with their insights. In our approach, we traced the provenance of workbooks that were created by one author and edited and republished by a different author (called a *derivation*). We initially found that few authors directly collaborated (only 2.4% of all workbooks published on Tableau Public contain visualizations that were derived from other workbooks). Since so few workbooks are derived, we tested to see if this was due to the fact that a lot of authors (53%) simply publish a single workbook and never return. Figure 11 shows the breakdown of workbook derivations grouped by whether or not it was the author's first publication. We see that a workbook is three times more likely to be derived if it is not the author's first publication. However, the probability of derivation remains minuscule.

Unfortunately, no such equivalent derivation information is available for Many Eyes. However, in order to get a sense of the degree of influence one author's contributions have on other authors, we show in Figure 12 how often authors reuse datasets uploaded and shared by others for their visual analysis in Many Eyes. In this figure, we see that only 6% of datasets are used by multiple authors and that 20% of datasets are used in multiple visualizations. Some users are publishing multiple visualizations for a given dataset. We similarly cannot plot Figure 12 for Tableau Public because published workbooks make a copy of the data being visualized.

Overall, however, the frequency of reuse of other authors' data on Many Eyes is consistent with the derivation results presented for Tableau Public.

Discussion: The clear conclusion from the above results is that online visual analytics systems are read-heavy today: Orders of magnitude more people are viewers compared to authors. Additionally, as is typically the case for database access patterns, viewership is skewed toward a small fraction of *hot* visualizations. Furthermore, as expected, first-time publications, which account for a large fraction of all publications, are less likely to be shared, derived, or viewed by a large audience than subsequent publications. At the same time, however, some first-time publications can be extremely popular. Also, in general workbooks are not likely to be derived from other workbooks and republished. Hence, true collaboration remains limited between users. Incentivizing and supporting collaborations remain critical challenges for today's online, visual data analytics systems.

3.3 How do users do visual analysis on one dataset?

We now focus on the details of the visualization contents found on Tableau Public and (when possible) Many Eyes. We first examine how users explore a single dataset and then how they integrate and explore multiple datasets.

How big is the data that drives a visualization?
Today's online visual analytics systems are designed for small data. Most of these systems put a bound on the size of datasets that can be processed. Recall that, in Tableau Public, each user only gets a 50MB account and a visualization can only operate on 100K rows. Similarly in Many Eyes data sizes are limited to 5MB. Given these restrictions, as expected, we see in Figure 13 that the median number of rows in a visualization is low. On Tableau Public, for example, 53% of all data sources contain less than 1,000 rows, 74% have 5,000 or less, and 90% contain 22,000 or less. The bottom 25% have less than 81 rows and the smallest dataset contains a single row. This trend is stronger for Many Eyes where 63% have less than 100 rows, 90% contain less than

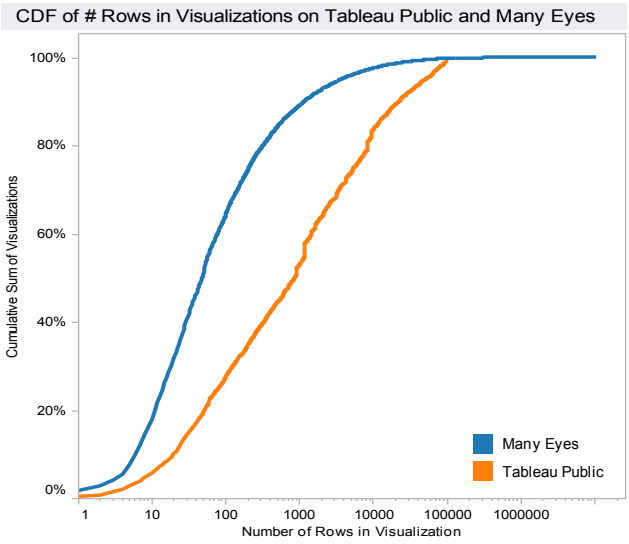

Figure 13: CDF of Number of Rows in Visualizations

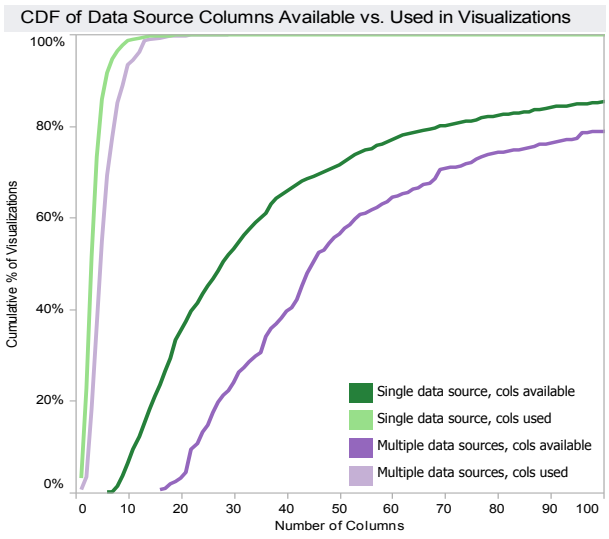

Figure 14: CDF of Number of Columns in Visualizations with One vs. Multiple (Joined) Data Sources

1,000 rows, and 99% have less than 18,000 rows.

Interestingly, in Tableau Public, a few special accounts are allowed to go beyond the 100K limit. We see that these accounts (along with some accounts on Many Eyes) visualize more than an order of magnitude more data, which seems to imply the need for the online visualization of bigger data too.

How many columns are used in the visualization of a single data source?

Figure 14 shows the breakdown of data columns used versus available in visualizations with a single data source versus multiple (joined) data sources on Tableau Public (no equivalent information was available for Many Eyes). First, we see that 50% of visualizations with a single data source use at most 3 columns and 90% use at most 7. As expected, the distribution of the columns available is much broader, indicating that there are many more columns available that are not being leveraged by the visualization (max columns available = 1200, max columns used = 127). For example, 50% of single data sources contain 28 or more columns. Furthermore, Table 2 shows the breakdown of the most common visualization types used for a given number of columns. The values denoted with a '∗' in Table 2 show that a second visualization type was within 5% from the top choice for that given number of columns. For single data sources, we see that the text table is the most common type when there is only one data column present in the visualization. As the number of columns increases, we see a shift in visualization techniques used: bar views become the dominant technique for 2–5 columns and maps are the most popular for 6–8 columns. This behavior is not too surprising since map views have a minimum requirement of two geographic dimensions (*i.e.* latitude and longitude). The maximum number of columns available is 1,244 and the maximum used columns is 127.[1]

[1]The workbook with 127 columns has a poorly designed dataset that contains a column for each day over 4 months.

How are single datasets visualized?

We further study the visualization types that users prefer on Tableau Public and then in comparison with Many Eyes. While these two systems share many of the same visualization types (*e.g.* bar view, text table, map view, etc.) there are a few differences worth noting, in particular with regard to text data. The large number of text views (Figure 17, right) is due to the variety and quality of text visualization views on Many Eyes, most which are not available anywhere else (Tableau Public's text table is just a table, unlike the rich interactive text views on Many Eyes). Similarly, bubble views are attractive but also rather uncommon in visualization and spreadsheet software. If we disregard these two, which are not available in Tableau Public, the most common visualization type that exists in both systems is the bar chart. The view types that exist in both systems appear in approximately the same order in both: bar chart, map, line chart, pie chart, and area chart. The only exception is the scatterplot, which is fairly common in Tableau Public but rather unpopular in Many Eyes.

The left half of Figure 17 shows that the most common visualization techniques with a single data source on Tableau Public are the bar view (37%), text table (19%), and line view (15%). This result is consistent with Table 2, in which the bar and text table dominate for visualizations containing 1 to 5 columns, and Figure 14 where 55% of visualization of a single data source use 5 or fewer columns. On Many Eyes the most common ones are the text view (40%), bubble view (13%) and bar view (9%).

Overall, we thus see consistent results for the two most common visualization types used on Tableau Public; the bar view and text table are the most common. Many Eyes, with its stronger focus on text data, has more popular text views than any other type. The comparable visualization types in both systems have roughly the same order of popularity.

What are the most common column data types used in visualizations with a single source?

Table 2: Most Common Visualization Types vs. Number of Columns in Visualization

Data Source(s) Used	Number of Columns in Visualization													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
One	Text (69%)	Bar (52%)	Bar (47%)	Bar (30%)	Bar* (29%) Map	Map* (29%) Bar	Map* (27%) Bar	Map* (28%) Bar	Bar* (24%) Map	Line (41%)	Map* (25%) Bar	Bar (30%)	Bar (38%)	Bar (31%)
Multiple	Text (92%)	Bar (49%)	Bar (48%)	Bar (43%)	Map* (23%) Text	Map* (31%) Text	Map (36%)	Text (33%)	Map (46%)	Map (45%)	Scatter (50%)	Scatter/Bar (32%)	Circle (81%)	Heat Map (37%)

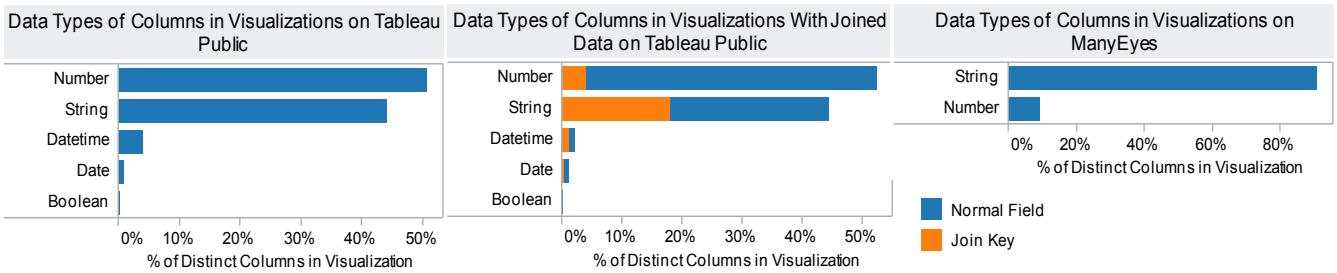

Figure 16: Data Types in Visualizations on Tableau Public and Many Eyes

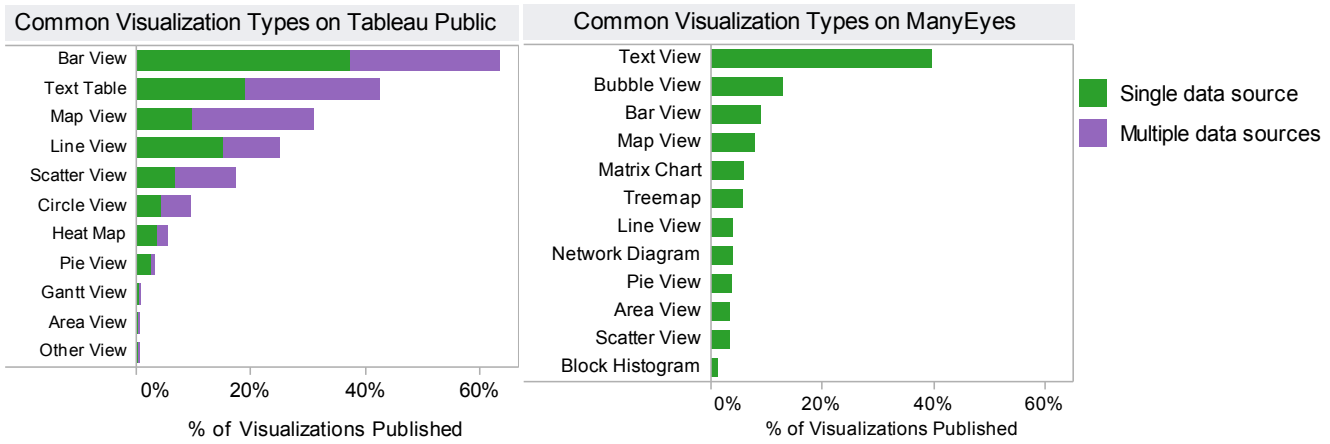

Figure 17: Common Visualization Types on Tableau Public (left) and Many Eyes (right)

The left-most graph of Figure 16 shows that `Number` (51%) and `String` (44%) are the most common data types in visualizations of a single dataset. It is interesting that their use is fairly balanced, while intuition would indicate that numbers might be more common. The `Number` data type includes both integers and reals. Finally, in the right-most graph of Figure 16, we see that 91% of columns on Many Eyes are `String` types. This finding is consistent with the previous one regarding the dominance of text-based visualizations on Many Eyes.

Discussion: In summary, most visualizations have modest data sizes, and seem to not be limited by the 100K rows restriction, although some users with special privileges visualize datasets with more than 1M rows. There is thus potential in these systems to support an entirely different class of users with much greater visualization requirements. Furthermore we see that as the number of columns used increases, so does the complexity of the visualization type (*e.g.* maps require more columns than other types like bar views.) Additionally, visualizations of single datasets tend to use many fewer columns than available. One explanation for this gulf can be drawn from the use of map visualizations in Tableau Public; 62% of such visualizations rely on a Tableau-supplied geocoding database for translating location names into latitude and longitude, since many data sources do not include this necessary context. Similarly, users may disregard opaque data that lacks any context, such as alpha-numeric identifiers, as well as data whose context is merely implied, such as timelines. Finally, visualizations on Tableau Public and Many Eyes contain columns of type `Number` or `String`; the split is very even between these two data types for Tableau Public, and Many Eyes is dominated by `Strings` due to the prevalence of text-based visualizations.

3.4 How do users integrate multiple datasets for their analysis task?

In this section we study the trends in data and visualiza-

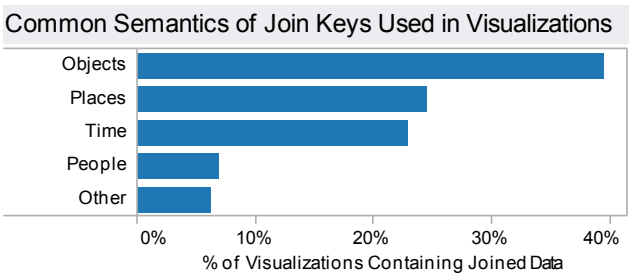

Figure 15: Semantic Entities in Visualizations with Multiple (Joined) Data Sources

tion on Tableau Public in the context of blending data from multiple data sources. We omit Many Eyes from this section because the platform currently does not support blending data.

What are the common semantic entities found in blended data?
To answer this question we manually categorized all of the join keys for the 2,400 visualizations that have blended data to get a sense of the most popular semantic entities. This process entailed inspecting the column name, data type, and data values of each join key. In the case where the column name was in a foreign language, we used Google Translate on the name and (in some cases) values of that column. If we were still unsure, we opened the workbook to inspect the visualization that was associated with that join key. Figure 15 summarizes the semantic entities of the join keys in five different categories: people, places, time, objects, and other. The people category contains any information pertaining to people, including names and demographics. The places category is restricted to geolocations and other identifying characteristics such as zip codes, regions, states, countries, continents, etc. As expected, the time category refers to dates and date times and objects refer to any physical entity that is not a person, place, or time. Objects consist mainly of opaque identifiers like alphanumeric product codes as well as more well-known, descriptive entities such as "university", "department", or "team". Finally, Figure 15 shows that visualizations of multiple data sources tend to join on objects (39%), places (24%), and time (23%).
How many columns are used in visualizations with blended data sources?
From the CDF in Figure 14, we see that visualizations with columns from multiple (joined) datasets tend to be more complex than those containing columns from a single dataset. For example, 45% of the blended views contain 5 or more columns, while only 14% of views with columns from a single dataset contain 5 or more attributes. Furthermore, we see a familiar trend as with single data sources: there is a sizable gulf between the number of columns used and the number of columns available in the blended data sources. Additionally, Table 2 shows that, like for single data sources, that visualizations containing a single column tend to be text tables (92%). We also see that the bar view dominates for visualizations containing 2–4 columns and map views for 5–7 columns. This finding is consistent with the distribution of visualization types for single data sources. For blended datasets, the maximum available columns is 1,255 and the maximum used columns is 29.
What are the most common column data types used in visualizations with blended data?
Figure 16 shows that, as in visualizations with a single data source, `Number` (52%)and `String` (44%) are the most common data types overall. Additionally, the stacked orange bars represent the data types of the join keys, and `String` types are the most common.
How are multiple (joined) datasets visualized?
Figure 17 shows that the most common ways to visualize blended data is with a bar view (26%), text table (23%), or map view (21%). Compared to the distribution for single data sets (recall bar views made up 37%, 19% for text tables, and 15% for line views), we see fewer bar views and more text tables and maps. This result is consistent with Table 2, in which the text table and map view dominate for visualizations containing 5–10 columns, and Figure 14 where 45% of blended visualizations use 5 or more columns.

Figure 17 also shows the visualization types with higher percentages of blended views: (in order) map views, scatter views, and text tables. Map views are a special case in Tableau Public, because prior to Tableau version 7 (*i.e.* before January of 2012), filled maps required tricks involving polygon shapes that were placed using blending. This inflates the number of blended views using maps somewhat, though there are also many other use cases where maps can be used as part of blended views. For example, a common blending pattern for maps is to join on a secondary datasource containing detailed latitude/longitude values. Scatter views are generally used for visualizing correlations, and for authors on Tableau Public, this visualization type is useful for showing correlations between measures from two different data sources. Finally, text tables are often used as a trial/debugging tool for checking out the resulting values from the join operation (*e.g.* how many `Null` values appear?).

Discussion: In summary, data blending occurs primarily by combining multiple attributes about the same uniquely identified entities from different data sources. This type of blending is more common than simply placing multiple entities at the same location or at the same point in time, although the latter two dominate when considered together. This finding is especially interesting for data integration tools. For example, a recent tool provides recommendations of potentially useful data to integrate with a given database [20]. This tool does not consider joining on place or time. It only considers extending semantic entities with additional attributes. With our study, it becomes clear that such a tool would ignore more than half of all blending scenarios. Additionally, blended visualizations tend to be more complex (*i.e.* use more columns and have more columns available) than unblended ones. However, the distribution of the most common visualization types for a given number of columns is similar for blended visualizations and those using only single datasets. We also see different trends in visualization techniques for those containing blended data versus single data. Blended views tend to be more prevalent in map views, scatter views, and text tables; these visualization types tend to be more complex (*i.e.* use more columns).

4. RELATED WORK

To our knowledge, there have not been any prior formal measurement studies of online visualization tools in the literature. However, the Guardian [8] published an informal analysis of the Many Eyes system. They studied the provenance of the data sources, and reported that the US Census Bureau was one of the most widely used sources. They also presented the most common topic tags for visualizations, most active users, and the number of data sets uploaded per user.

In this paper, we focus our study on Tableau Public and Many Eyes. There exist similar Web-based collaborative and visual data analysis systems. For example, Fusion Tables [12, 13] is similar to Many Eyes and Tableau Public in that it enables users to upload data and visualize it in a variety of ways. However, it offers a different model of sharing that does not require authors to make their data public, and supports a subset of Tableau Public's interactive query capabilities for visualizations. This paper thus presents a study from two systems that are good representatives of this class of systems.

5. CONCLUSIONS

In this paper, we studied four primary dimensions of two popular online visual analytics systems: (1)what types of users are leveraging these systems and what are their workloads, (2)how are users collaborating and interacting with the published content, (3)what are the trends for doing visual analysis over a single-dataset, and (4)how do users analyze data joined from multiple sources. First, we found that such systems today need to effectively support primarily novice users with small datasets. These findings also point to the lack of online, visual analytics tools that would better support users with larger datasets and more sustained data analysis, visualization, and sharing needs. Second, we showed that authors tend to bring their own data and do not leverage the contributions of content from other authors. However, collaborations primarily take the form of amplifying an author's impact by attracting large numbers of viewers – we measured that Tableau Public attracts over 20 million unique visitors. Finally, we discovered that most visualizations of single-dataset (and multi-datasets) tend to use far fewer columns than available. Since both systems have a large repository of potentially useful data sets, we need tools that can help connect users to other good quality data to aid them in their analysis, especially in the case where additional context is needed in order to take advantage of columns that would otherwise go unused due to their opacity.

6. REFERENCES

[1] Gapminder. `http://www.gapminder.org/`, 2012.

[2] iCharts. `http://www.icharts.net/`, 2012.

[3] Many Eyes. `http://many-eyes.com/`, 2012.

[4] Microsoft Windows Azure Marketplace. `https://datamarket.azure.com/`, 2012.

[5] Open Data Protocol. `http://www.odata.org/`, 2012.

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

[7] Tableau Software. `http://www.tableausoftware.com/`, 2012.

[8] The Guardian Datablog. `http://www.guardian.co.uk/news/datablog/2011/mar/17/visualise-data-trends#data`, 2012.

[9] ViewShare: Interfaces to our Heritage. `http://viewshare.org/`, 2012.

[10] T. Berners-Lee. The year open data went worldwide. The Huffington Post, `http://www.huffingtonpost.com/tedtalks/tim-berners-lee-the-year_b_490726.html`, 2010.

[11] S. Card, G. Robertson, and J. Mackinlay. The information visualizer, an information workspace. In *Proceedings of the ACM Computer Human Interaction (CHI) Conference*, pages 181–186, 1991.

[12] 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 *Proceedings of the International Symposium on Cloud Computing*, pages 175–180, 2010.

[13] H. Gonzalez, A. Halevy, C. Jensen, A. Langen, J. Madhavan, R. Shapley, W. Shen, and J. Goldberg-Kidon. Google fusion tables: Web-centered data management and collaboration. In *Proceedings of the ACM SIGMOD Conference*, pages 1061–1066, 2010.

[14] P. Hanrahan. Analytic database technologies for a new kind of user: the data enthusiast. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 577–578, 2012.

[15] J. Heer and M. Agrawala. Design considerations for collaborative visual analytics. In *Proceedings IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 171–178, 2007.

[16] D. Martin. Twitter Quitters Post Roadblock to Long-Term Growth. `http://blog.nielsen.com/nielsenwire/online_mobile/twitter-quitters-post-roadblock-to-long-term-growth/`, 2009.

[17] R. Miller. Response Time in Man-Computer Conversational Transactions. In *Proceedings of the AFIPS Fall Join Computer Conference*, volume 33, pages 267–277, 1968.

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

[19] F. Nah. A Study on Tolerable Waiting Time: How Long Are Web Users Willing to Wait? In *Behaviour and Information Technology*, volume 23, 2004.

[20] 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 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 817–828, 2012.

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

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

[23] K. Wesley, M. Eldridge, and P. Terlecki. An Analytic Data Engine for Visualization in Tableau. In *Proceedings of the ACM SIGMOD Conference*, pages 1185–1194, 2011.

[24] W. Willett, J. Heer, J. Hellerstein, and M. Agrawala. Commentspace: Structured support for collaborative visual analytics. In *ACM Human Factors in Computing Systems (CHI)*, pages 3131–3140, 2011.