Preliminary Guidelines For Combining Data Integration and Visual Data Analysis

Adam Coscia, Ashley Suh, Remco Chang, and Alex Endert

Abstract—Data integration is often performed to consolidate information from multiple disparate data sources during visual data analysis. However, integration operations are usually separate from visual analytics operations such as encode and filter in both interface design and empirical research. We conducted a preliminary user study to investigate whether and how data integration should be incorporated directly into the visual analytics process. We used two interface alternatives featuring contrasting approaches to the data preparation and analysis workflow: manual file-based ex-situ integration as a separate step from visual analytics operations; and automatic UI-based in-situ integration merged with visual analytics operations. Participants were asked to complete specific and free-form tasks with each interface, browsing for patterns, generating insights, and summarizing relationships between attributes distributed across multiple files. Analyzing participants' interactions and feedback, we found both task completion time and total interactions to be similar across interfaces and tasks, as well as unique integration strategies between interfaces and emergent behaviors related to satisficing and cognitive bias. Participants' time spent and interactions revealed that in-situ integration enabled users to spend more time on analysis tasks compared with ex-situ integration. Participants' integration strategies and analytical behaviors revealed differences in interface usage for generating and tracking hypotheses and insights. With these results, we synthesized preliminary guidelines for designing future visual analytics interfaces that can support integrating attributes throughout an active analysis process.

Index Terms—Visual analytics, Data integration, User interface design, Integration strategies, Analytical behaviors.

1 Introduction

From a visual analytics perspective, the rapid growth of data today requires methods to combine information from disparate sources into a unified data representation to facilitate analytical reasoning [1], [2]. From a systems engineering perspective, this process involves data integration, or the task of querying into multiple, often heterogeneous, data sources with potentially differing levels of access and resolving the results into a unified view of the data [3], [4]. A human-in-the-loop perspective promotes "exploratory knowledge discovery in large datasets" where *a priori* knowledge of data is not guaranteed [5]. Yet visual analytics tools such as Tableau present manual data preparation solutions that occur as a separate step from visual analytics operations such as encode and filter [6]. We posit that the data preparation and visual analytics workflow in tools like Tableau has created an expectation in research and design that users' interactions and behaviors are influenced by the integration or analysis process separately.

In response, we raise two open research questions based on the common approach of separating data integration and visual analytics processes in research and design.

1) **Where and how should data integration operations, such as joins, be supported in tandem with visual analytics operations, such as encode and filter?**

Kandel et al. identify breakdowns in analysis workflows that occur in the early stages and when transitioning between tasks, where little research and few tools provide visualization solutions [7]. Theories of information foraging [8] and sensemaking processes [9], [10] that describe the analysis process also maintain a simultaneous and inseparable view of continuously finding and making sense of data. Consider decision-makers integrating columns and rows across multiple spreadsheets using Microsoft Excel to visualize the results as they work. Dimara and Stasko assert there is a gap in visualization tools that support in-situ data integration around maintaining flexibility and flow when performing visual analytics operations [11].

2) **How will incorporating data integration into an ongoing visual analytics process affect user behaviors?**

Pirolli and Card identify potential constraints on analysis during foraging and sensemaking due to time pressures, data overload, and cognitive biases [9]. These constraints may lead to satisficing based on the time and effort people spend finding, gathering, and integrating data [12].

Our aim in this paper is to contribute preliminary guidelines for incorporating data integration into an active visual analytics process, towards fostering better information retrieval that allows people to incorporate their data seamlessly and improve how visualizations are created and used. To do this, we created two interface alternatives inspired by Polestar [13] and ran a two (interfaces) by two (data sets) within-subjects study, recruiting 16 participants and randomly assigning an order to use both interfaces and data sets for performing specific and free-form style visual data analysis tasks. The first interface (SEPARATED) requires manual file-based ex-situ integration of attributes via Microsoft Excel (column concatenation then selection). This more traditional interface represents a simplified data preparation and analy-

- *Adam Coscia and Alex Endert are with Georgia Institute of Technology. Emails: {acoscia6, endert}@gatech.edu.*
- *Ashley Suh and Remco Chang are with Tufts University. Emails: ashley.suh@tufts.edu, remco@cs.tufts.edu.*

Manuscript received April 19, 2005; revised August 26, 2015.

sis workflow common in most research and design that can serve as a baseline for gathering users' analytical behaviors. The second interface (COMBINED) presents automatic user-interface-based (UI-based) in-situ integration combined with common visual analytics operations such as encode and filter (column concatenation via selection). This non-traditional interface removes much of the separation between data preparation and visual analytics operations to investigate how users' strategies and analytical behaviors differ from using the first interface. We then conducted a mixed-methods analysis of users' interactions and behaviors. As a first step in tackling our broad research questions in a productive way, our approach: (1) reduced confounds in relating users' interactions and behaviors between interfaces; (2) helped us elicit rich qualitative insights that raise important questions and lay the foundation for future work; and (3) mirrored real decision-making scenarios, e.g., Dimara and Stasko [11].

We found participants using unique in-situ integration strategies (Sect. 4.3) based on time spent integrating, interactions with different panels, and qualitative feedback. For example, several participants exclusively integrated on the fly on purpose, spending little to no time integrating beforehand. Yet surprisingly, we found that interface and task type did not significantly affect overall task completion time (Sect. 4.1) or the total number of interactions (Sect. 4.2). At the same time, in-situ integration operations sometimes negatively affected the ability to generate and track hypotheses and insights; specifically, participants' analytical behaviors underscored issues of satisficing and exhibiting biased behaviors (Sect. 4.4). With these findings, we synthesized preliminary guidelines for incorporating data integration into visual data analysis: (1) show where and how data are being integrated (Sect. 5.1); (2) use in-situ integration for exploring the space of attributes (Sect. 5.2); and (3) balance manual and automated approaches (Sect. 5.3). We also discuss limitations and future work on eliciting effects in "real-world" data integration and visual data analysis scenarios that can involve extensive planning and custom tool development (Sect. 6).

In summary, our work contributes: (1) a within-subjects user study employing two different visual analytics interfaces for integrating and visualizing attributes across multiple disparate data sources; (2) observations and reflections on user interactions and behaviors with our combined data integration and visual analytics interfaces; and (3) preliminary guidelines for incorporating data integration into visual analytics processes and directions for future work.

2 RELATED WORK

2.1 Data Integration

Issues in data integration include specifying well-structured queries, scaling with the number of sources, resolving the heterogeneity of different file formats and data types, and addressing the privacy and accessibility of each source [4]. Both enterprise and ad-hoc analysis, traditionally utilizing structured data queried from data warehouses with organized schematic definitions, are seeking to incorporate valuable unstructured and semi-structured sources such as PDFs, news feeds, social media, images and video [7], [14], [15], [16]. Further, data sources are increasingly being made publicly available on the Web from government and non-profit organizations such as data.gov and WikiMedia, a movement known as open data [17]. As the scale, availability, and complexity of data sources grows, visual analytics tools should explore ways to incorporate and support the integration process throughout analysis.

However, locating, collecting and integrating data sources remains an open challenge. Heterogeneity between data sets, insufficient semantics to describe data, and errors in data insertion and modification [18] create entity resolution challenges. For example, when labelled data items with identical attributes from different sources could refer to the same real-world entity, deduplication is needed to identify and link those sets of records. Further, there are considerable technical challenges in detecting, linking, removing, and merging entities efficiently [19]. Work in natural language processing (NLP) has studied how entity resolution can be learned and improved over time through user interactions [20]. A number of tools [21], [22] help people model queries into knowledge graphs [23], [24]; however, there are scale limitations for large, complex, and heterogeneous structures [25]. Other automated approaches feature interactive programming interfaces [26], equi-join-able tables [27], and deep learning approaches to entity resolution [28]. As data integration solutions mature, we seek a baseline understanding of user interactions and behaviors in visual analytics tools that incorporate data integration capabilities.

2.2 Visual Analytics

In-situ data transformations in visual analytics tools often seek to resolve issues of heterogeneity, quality, and semantics, i.e., data wrangling [29], [30]. In this study, we focus on a direct manipulation [21], [31] approach to transformations that combine data from separate sources into a single repository, i.e., data integration. From this perspective, we discuss relevant systems and domains that use interactive interfaces to integrate data throughout the visual analytics process.

Systems. A few visual analytics systems and studies have addressed the technical and cognitive limitations of integrating data throughout the analysis process in unique ways. Tableau provides data blending, a technique for dynamically combining data from multiple heterogeneous data sources without any upfront integration effort [14]. Cramer et al. investigate the effects of streaming new data during analysis on sensemaking capabilities, finding an increase in people's explicit focus and reflection on analytic progress [32]. Cashman et al.'s CAVA system allows users to interactively augment related attributes from knowledge graphs into an existing data set and visualize them during exploration and analysis tasks [33]. Similarly, Latif et al. utilize EventKG in a visualization system to automatically integrate relevant event information and relationships for historical figures in an existing data set [34]. A goal that cuts across these examples is to help people explore differently structured data across independently located sources without interrupting the visual analytics process [35], [36]. Our aim is to empirically describe user interactions and behaviors when coupling both data integration and visual analytics operations (UI) and processes (workflows).

Domains. Data integration challenges can be found embedded in various communities that engage in visual data analysis. Kandel et al. identify data integration as a challenge for analysts in analytics, biology, datamart, finance, healthcare, insurance, marketing, media, retail, social networking, sports, and web development [7]. Zheng et al. find that urban computing studies routinely integrate and visualize traffic data for public safety and security applications from large, open data sources that are often unstructured, noisy, and heterogeneous [37]. In recent work by Dimara and Stasko [11], recent visualization software is criticized by decision-makers for its inability to restructure, integrate, and forage for new data on the fly. Instead, flexible data software like Excel is preferred by decision-makers, as data is often unstructured and incomplete before analysis begins. Adding new attributes to an otherwise fixed data set towards improving model performance is an active area of research in machine learning, commonly referred to as data augmentation [38]. These challenges have inspired us to ask how the data integration process affects the visual analytics process commonly used across these domains.

3 Study Design

We conducted a user study to empirically describe user interactions and behaviors when data integration and visual analytics processes are combined. Our goals were:

- Investigate if and how people use visual analytics operations such as encode or filter to help them integrate new attributes in-situ.
- Investigate if and how users' strategies and analytical behaviors during visual data analysis differ between ex-situ and in-situ integration approaches.

We built two interfaces with different approaches to data preparation and analysis: the Separated interface, with manual file-based ex-situ integration as a separate step from visual analytics operations; and the Combined interface, with automatic UI-based in-situ integration merged with visual analytics operations. Both simplify integration to column concatenation and selection to reduce confounds in observing users' analysis strategies and interactions. This improved our ability to directly compare interactions and behaviors between these two different interfaces. To further distinguish in-situ and ex-situ integration strategies in the Combined interface, we define *primary* and *secondary* integration processes and attribute interactions in Sect. 3.2.2. We utilized a two (interfaces) by two (data sets) within-subjects experiment design, exposing participants to both interfaces and data sets to foster reflection on how their analysis process differed between conditions. 16 participants performed specific and free-form style visual data analysis tasks with each interface and data set in a random order. We logged all mouse events and captured both the screen and audio of participants as they followed a think-aloud protocol [39] and described their experience in a post-study semi-structured interview. In this section, we describe the data sets and tasks curated, the experimental systems developed, and the procedure employed in this study.

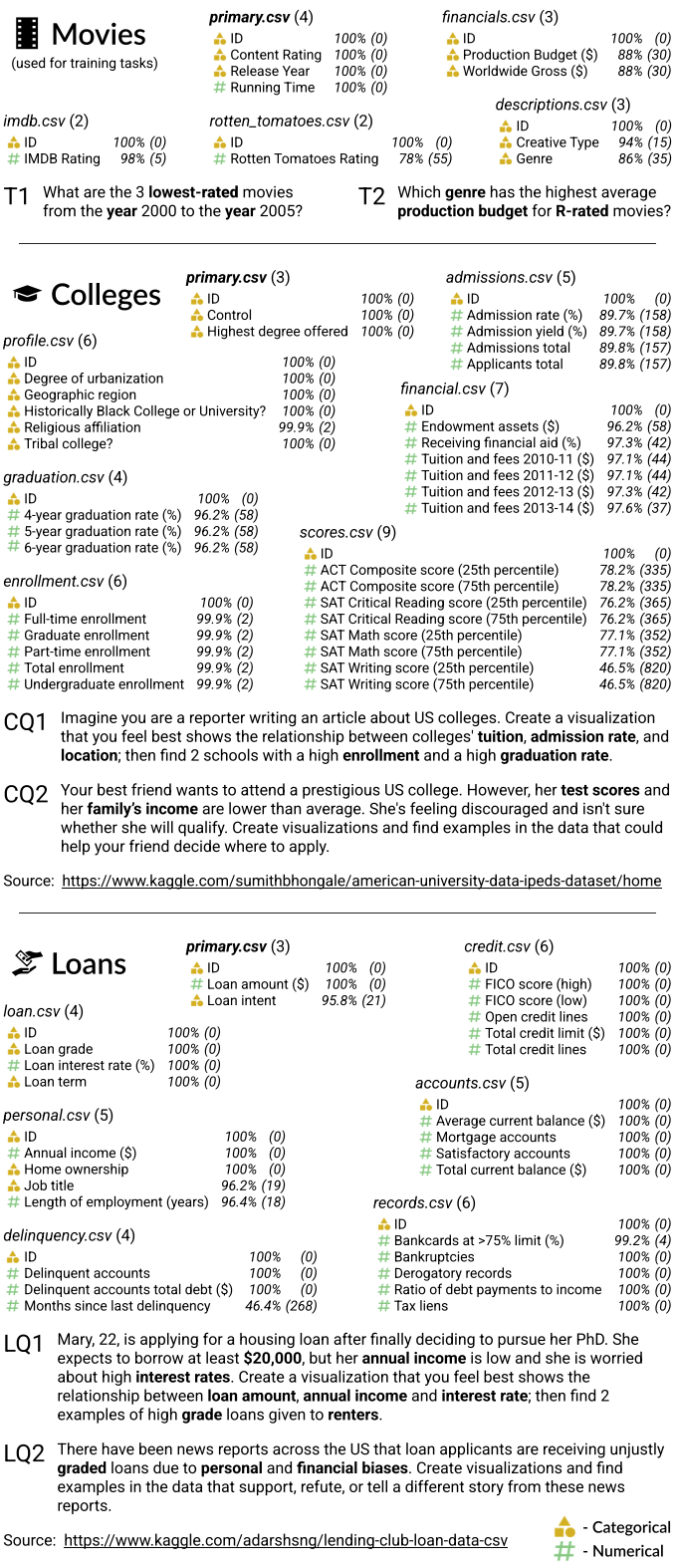

Fig. 1. Data sets and tasks used in the study (Sect. 3.1). All data sets list each attribute's data type, percent complete (in italics), and total number of missing records (in parentheses), grouped under the files they are in.

3.1 Data Sets and Tasks

We modified two publicly available tabular data sets on Kaggle relating to colleges and loans and generated a training movies data set in the style of the other two data sets, shown in Fig. 1. The modifications to the colleges and loans data sets include reducing the number of numerical attributes in

the colleges data set and the number of records in the loans data set to: (1) make it easier to find categorical attributes; and (2) reduce the total number of data points to inspect visually, respectively. While these changes limited ecological validity, they also reduced potential confounds in participant feedback when scaling integration to large data sets.

The final colleges data set contained all 1534 records from the original and comprised one primary key (*ID*), seven categorical attributes, and 26 numerical attributes out of the 145 original attributes. Each record of the colleges data set represents one U.S. college. The final loans data set contained a subset of 500 records from the original and comprised one primary key (*ID*), five categorical attributes, and 21 numerical attributes. Each record of the loans data set represents one loan application. The final movies data set contained 250 records and comprised one primary key (*ID*), four categorical attributes, and five numerical attributes. Each record of the movies data set represents one movie. For all data sets, the attributes were distributed between the *primary.csv* file loaded into the interface to start and all other CSV files, with a one-to-one mapping of records between every file using the attribute *ID* as a relational key. The records in each file except *primary.csv* were randomly sorted, any values missing in the original data sets for a given record or attribute were kept, and any potentially identifiable information, such as the names of colleges or loan applicants, was removed.

We created two tasks for the colleges and loans data sets (*CQ1*, *CQ2*; and *LQ1*, *LQ2*), labeled by data set (*C* and *L*) and task type (*Q1* and *Q2*), as well as two specific training tasks (*T1* and *T2*) for the training movies data set. The task descriptions for each data set can be read in Fig. 1. Participants were instructed to discover insights from the data by locating attributes of interest and browsing for patterns visually, then summarizing the uncovered relationships between visualized attributes [40]. The first task of the colleges and loans data sets, henceforth called the *specific task* and labeled *CQ1* and *LQ1*, required a specific visualization and data points to answer successfully. The second task of the colleges and loans data sets, henceforth called the *free-form task* and labeled *CQ2* and *LQ2*, featured open-ended analysis and assignments to interpret the data. Every task description listed at least one relevant attribute that was not in *primary.csv*, thus requiring participants to locate attributes in other files. For example, in *CQ1*, none of "tuition", "admission rate", "location", "enrollment", or "graduation rate" are located in *primary.csv*.

3.2 Experimental Systems

We developed both SEPARATED and COMBINED interfaces inspired by PoleStar [13] (Fig. 2). With these interfaces, participants encoded data via drop-down menus and manipulated it via select, arrange, change, filter, and aggregate operations [6], [40]. The primary difference between these two interfaces is how they structure the data integration process. The COMBINED interface has data integration functionality (column concatenation) embedded directly into drop-downs (column selection), traditional user interface (UI) controls used for visual data analysis found in tools such as Tableau. In comparison, the SEPARATED interface has the same visualization UI for column selection, but data integration happens outside of the interface, where users manually combine attributes into *primary.csv* using Excel via column concatenation.

3.2.1 Separated Interface

This baseline interface shown in Fig. 2 features six panels for performing visual data analysis: an Attributes panel for displaying the attributes and their data types from *primary.csv*; an Encode panel for specifying the data-visual mapping; a Filter panel for applying categorical and numerical filters to the data; a Visualization panel that allows users to hover and click on the data marks; and an Elaborate panel that shows a table of the records corresponding with hovered and clicked on data marks. A Navigation bar at the top of the screen shows participants the current task description. To integrate data, users navigated the operating system's file browser to open CSV files in Excel that collectively contained the data sets used for the study (see Sect. 3.1 for a description of the data and Fig. 1 for a representation of the file structure). Participants were allowed to concatenate columns between files using any operations available in Excel. The results of their operations, including the integrated attributes, were loaded into the interface through a single file, *primary.csv*, via the Refresh button in the Attributes panel. We chose Excel as a baseline for manual data integration for two reasons: (1) it mirrors real decision-making scenarios described by Dimara and Stasko [11]; and (2) its familiarity in our target participant group.

3.2.2 Combined Interface

In contrast to the SEPARATED interface, we designed a second interface shown in Fig. 2 that incorporates integration operations directly into the Attribute, Encode, Filter and Elaborate panels without revisiting separate files or tools, e.g., Excel. To help participants focus on how in-situ integration affected their analysis process, we sought to reduce potential confounds, such as join errors, that could arise in feedback. We achieved this by merging the separate files for each data set used with the SEPARATED interface into a single file and schema when using the COMBINED interface. Then, when participants used a panel to add an attribute to their analysis, a lookup was performed. By exposing the operation as a join in the interface and to the participant, we believe this study design decision helped keep feedback focused on how the combination of integration and visualization operations affected analysis.

Primary and secondary attributes. While we know that ex-situ integration operations can affect users' analysis processes during visual data analysis [9], it is unclear if analogous in-situ integration operations uniquely affect users' workflows and analytical behaviors. To distinguish ex-situ and in-situ integration, we separated the *primary* integration process of adding attributes into the Attributes panel, available in both interfaces, from the *secondary* method of integrating attributes on the fly in the Encode and Filter panels, available only in the COMBINED interface. Thus, we define *primary* attributes as those attributes that are present in the Attributes panel when they are used in any panel, and *secondary* attributes are those attributes that are not present in the Attributes panel when they are used in any panel. Fig. 2 describes the location of *primary* and *secondary* attributes relative to

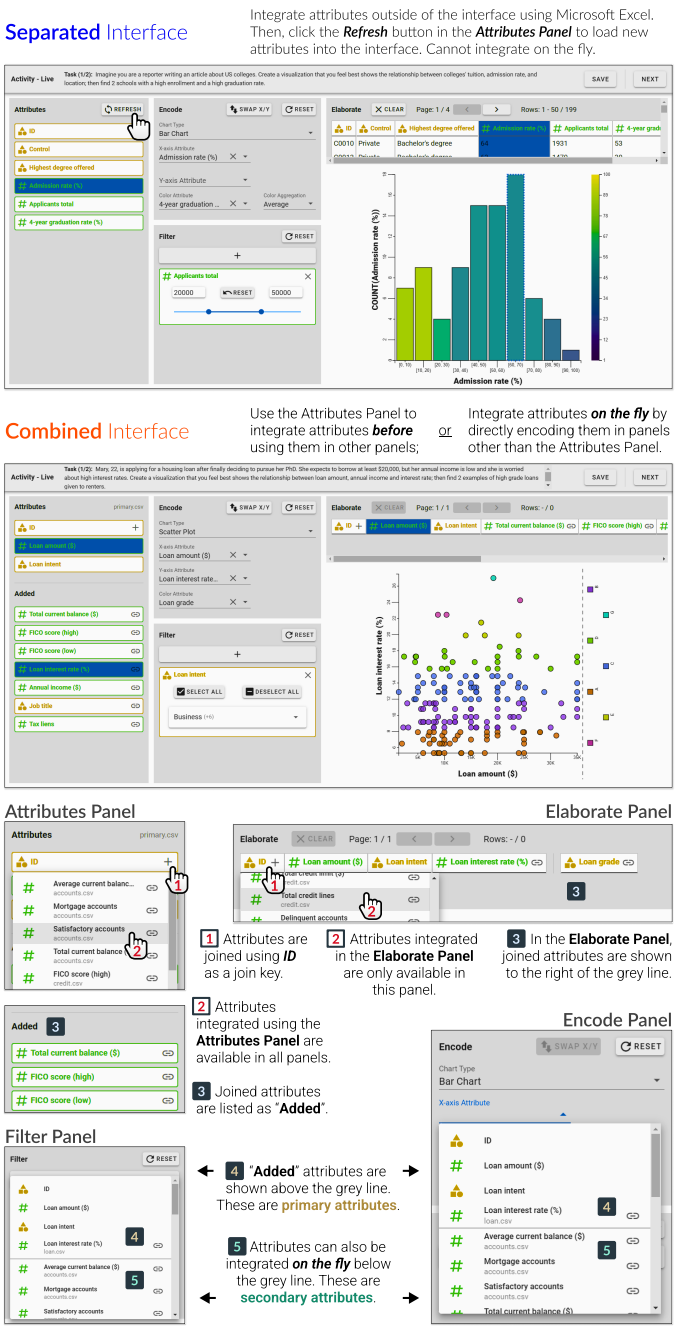

Fig. 2. Our two interfaces: (1) SEPARATED (top; showing task *CQ1*); and (2) COMBINED (bottom; showing task *LQ1*) (Sect. 3.2).

the COMBINED interface panels. For example, a user could visualize an attribute from outside *primary.csv* on the fly in the Encode panel (i.e. *secondary* attribute interaction), then intentionally add it to *primary.csv* and their Attributes panel and encode it once again (i.e. *primary* attribute interaction). They could also mostly integrate and visualize attributes on the fly (i.e. *secondary* attribute interactions), a unique integration strategy, or mostly integrate attributes before visualizing them (i.e. *primary* attribute interactions), similar to the SEPARATED interface. We describe the results from investigating these interactions in Sect. 4.2.

Attributes panel. In the Attributes panel (Fig. 2), participants can integrate attributes into *primary.csv* via a drop-down menu accessed by clicking on the join key (*ID*). Each item in the menu is an attribute labeled with data type, source file, and an indicator that the attribute is not originally from *primary.csv*. Once clicked, attributes are joined into *primary.csv* and displayed under the "Added" header.

Encode and Filter panels. In the drop-down menus of the Encode and Filter panels (Fig. 2), "Added" attributes from the Attributes panel can be used directly (i.e. *primary* attribute interactions). Attributes not in *primary.csv* can also be used (i.e. *secondary* attribute interactions) and thus "seamlessly" integrated on the fly through visual analytics operations. "Added" attributes are shown above the horizontal dividing line while all other attributes available to integrate on the fly are shown below it. Integrating an attribute on the fly does not add it to the Attributes panel; however, once an attribute is added to the Attributes panel, it moves above the dividing line and can no longer be integrated on the fly.

Elaborate panel. In the Elaborate panel (Fig. 2), participants access a drop-down menu by clicking on the join key (*ID*) to add attributes from separate files directly to that panel without adding them to the Attributes panel, similar to the Encode and Filter panels. Attributes added on the fly are shown to the right of the vertical dividing line. If an attribute added on the fly is then added to the Attributes panel, it moves to the left of the vertical dividing line.

3.3 Study Procedure

We recruited 16 participants (P1 – 16) via recruitment emails to university mailing lists. Seven self-identified as male, and nine self-identified as female. All of the participants held or were pursuing undergraduate or higher degrees in fields spanning Computer Science (8), Analytics (4), Human-Computer Interaction (2), Human-Centered Computing (1), and Industrial Design (1). Participants self-reported prior engagement with a wide range of visual data analysis tools including Tableau (15), Python/Matplotlib (11), R/ggplot2 (6), Microsoft Power BI (4), D3.js (2), SAS (2), and AWS Quicksight (1). All participants also had prior experience with using Excel for visual data analysis: 11 participants self-reported moderate experience; the rest self-reported as having either a lot or a little experience.

After obtaining consent, participants self-reported their demographics in a pre-study survey. Then one of each of the two possible interfaces and data sets was randomly chosen to control for ordering effects. The researcher conducting the session explained aloud the various features of the interface. Participants were then asked to complete a training task (*T1* or *T2*) with the movies data set that would not be evaluated. After, they were asked to complete a specific (*Q1*) task, then a free-form (*Q2*) task, that would both be evaluated. Participants then had the other interface explained aloud to them and were similarly asked to complete a training task (*T1* or *T2*) with the movies data set, a specific (*Q1*) task, and finally a free-form (*Q2*) task using the other data set and the other interface. Participants were numbered in this paper according to this counterbalancing effort: participants P1 – 8 used the COMBINED interface with the loans data set and the SEPARATED interface with the colleges data set; and participants P9 – 16 used the COMBINED interface with the colleges data set and the SEPARATED interface with the loans

data set. Finally, the participant and researcher engaged in a semi-structured debrief interview discussing the participant's experience during the session.

We asked participants to use a think-aloud protocol [39] as they worked and to summarize how they arrived at their results after completing the tasks. We recorded the screen and audio (participant and researcher) as well as interaction logs of all mouse events. Each session took place in-person and lasted between two and three and a half hours, with a mean time of two and a half hours. Each participant was compensated $30 USD via an Amazon gift card.

4 Study Results

We used event logs and video recordings to uncover quantitative and qualitative patterns in participant's interactions and behaviors. Following Dragicevic [41], we interpreted effect sizes such as sample means using bootstrapped 95% confidence intervals (CIs) with 1000 resamples to represent uncertainty (Fig. 3, Fig. 4). For a given confidence level and sample size, CI width increases with increasing variability; results are generally significant if CIs do not overlap. We also provide absolute counts as bar charts (Fig. 3, Fig. 5). We further evaluated both the video recordings of each session and the audio recordings of participants' think-aloud protocol, debrief interview, and the researcher's notes, and conducted inductive thematic analysis [42], identifying emergent themes that were discussed amongst all authors. We acknowledge that a think-aloud protocol carries the potential to affect time spent on tasks and interactions during studies and consider this in our subjective interpretations. In this section, we present preliminary observations and open questions from our mixed-methods analysis around time spent (Sect. 4.1), interactions (Sect. 4.2), integration strategies (Sect. 4.3), and analytical behaviors (Sect. 4.4).

4.1 Time Spent

When presented with both in-situ and ex-situ integration capabilities, will participants spend time differently between interfaces and tasks?

We needed to temporally separate when participants were integrating and visualizing data, especially in-situ where these operations are combined. From Pirolli and Card [9] we know that users locate sources of data, integrate these data, and then analyze them, in a foraging and sensemaking loop. We designed both of our interfaces with an Attributes panel for users to keep track of what they have integrated in the loop. Thus, as a first step, we can use the set of operations that result in an attribute being added to the Attributes panel as a proxy for intervals of data integration in both interfaces. In the Separated interface, it is the time between when Excel is first made the top-level, primary window on the screen, at least one attribute is added to the Attributes panel, and the next time that the participant creates or modifies a visualization. In the Combined interface, it is the time spent between when the Attributes panel drop-down menu is first opened, at least one attribute is clicked on, and the next time that the participant creates or modifies a visualization. We also defined a proxy for when the participant started their analysis as the moment they first assigned an attribute to an encoding channel in the Encode panel. We reviewed our video recordings and applied these definitions when manually recording the time intervals of integration.

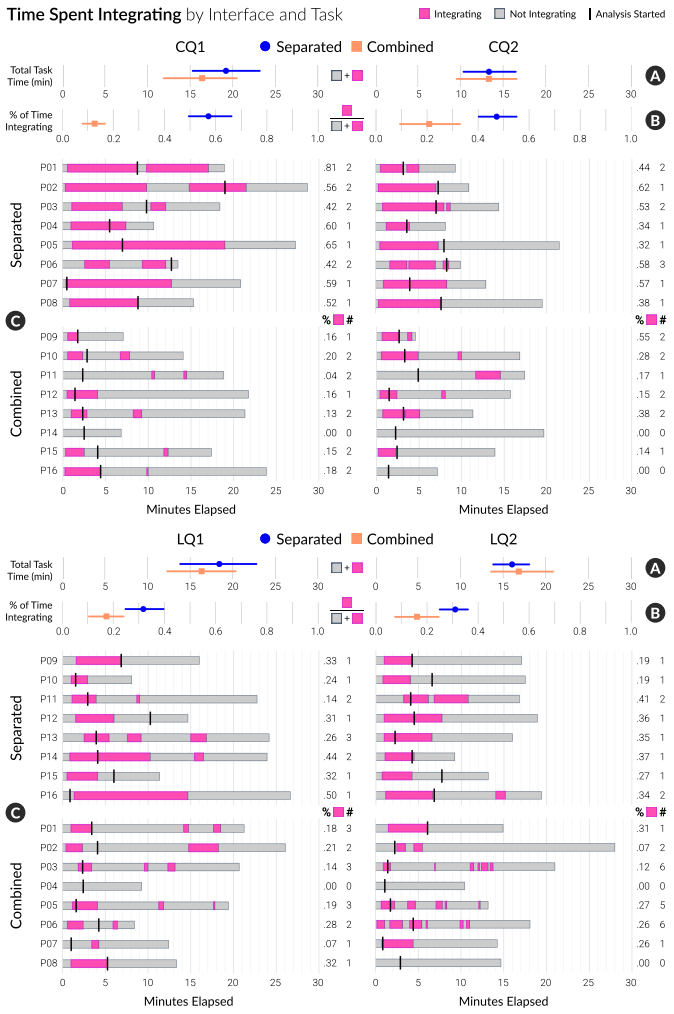

Fig. 3. Bootstrapped 95% CIs around the mean estimations of total task completion time **(A)** and percent time spent integrating between interfaces **(B)**, as well as the time spent integrating organized by interface and task **(C)** (Sect. 4.1). Each estimate represents eight participants with 1000 resamples. The darker pink bars represent intervals of data integration and the vertical black lines are a proxy for showing when analysis started. The percent (%) of time spent integrating and number (#) of intervals of integration are shown to the right of each bar.

Observations. Comparing CI width and overlap of task completion time in Fig. 3A, we found no significant difference between interfaces or tasks. The percentage of time spent integrating in Fig. 3B was significantly different between interfaces; between 25% to 65% with the Separated, while only 10% to 30% with the Combined. P7 described this difference during their session: *"In the first [Separated] interface, I had to go through 8 different files to select the attributes, sort them, and add them to primary.csv. That was tedious and difficult compared to the second [Combined] interface, where those operations were much smoother."*

Breaking down when participants integrated throughout the task in Fig. 3C, we see different user behaviors emerging. When using Separated interface, participants mostly integrated in one or two intervals and usually before starting analysis, whereas with the Combined interface, integration

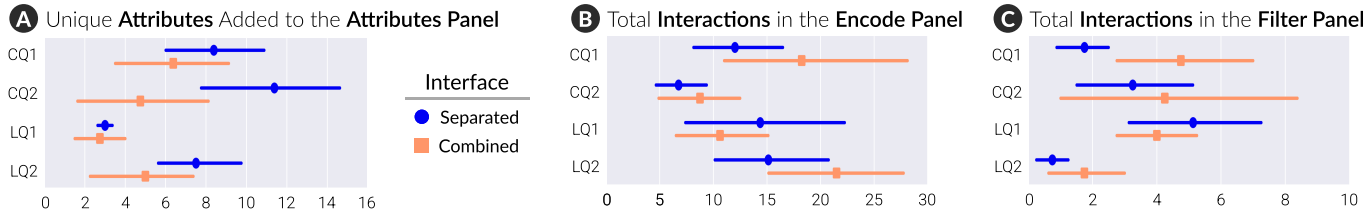

Fig. 4. Bootstrapped 95% CIs around the mean estimations of **(A)** unique attributes added to the Attributes panel as well as total attribute interactions between interfaces in the **(B)** Encode and **(C)** Filter panels (Sect. 4.2). Each estimate represents eight participants with 1000 resamples.

usually happened in two or more intervals any time throughout the analysis. For P9, using the SEPARATED interface mirrored processes they followed with other visual analytics tools: *"When I try to work on visualization, I think of it as a two-step process: I find the attributes first, then make the visualizations. Otherwise it's a lot to keep track of and think about... I'm just in the habit of making my list before visualizing... I think of the tasks as separate... I think my experience in Tableau makes me expect to have to connect data in sheets first."* Unique strategies also emerged; e.g., P4, 8, 14, 16 exclusively integrated attributes on the fly (i.e. no time spent integrating and no pink bars in Fig. 3C), while P3, 5, 6 used the Attributes Panel to integrate one attribute at a time up to 6 times (i.e. many small pink bars in Fig. 3C). P8 switched to this strategy during the free-form task (*Q2*) saying: *"It was very quick and efficient to test and move attributes."* Differences between task types (*Q1* and *Q2*) were inconclusive.

When should data integration operations be combined with visual analytics operations? Overall, we were surprised to see that both specific and free-form tasks took roughly the same time to complete with either interface, even though there are fewer integration operations to perform in the COMBINED interface. We also saw participants spend more time integrating data before analyzing it using the SEPARATED interface, whereas integration happened less often and more frequently throughout analysis with the COMBINED interface. This could suggest that in-situ integration helps users stay focused on analysis tasks longer than ex-situ integration. At the same time, some participants chose when and how long to integrate regardless of total task time based on previous experience with tools like Tableau, where integration usually happens separately from analysis. Users may require a balance between manual and automated integration.

Open questions. In addition to the findings above, our study revealed open questions, listed below, that encourage future research in this area:

1) Are users spending more time thinking about their analysis when using in-situ versus ex-situ integration?
2) Are users able to quickly transition between integration and analysis with in-situ integration, or does new data being introduced interrupt the flow?
3) What factors in a user's prior experience affect time spent? How do they correlate with task requirements and interface design?
4) Will time spent be affected if users can choose between in-situ and ex-situ integration on the fly?

4.2 Interactions

Will in-situ or ex-situ integration reveal differences in how users perform visual analytics operations?

Observations. Comparing CI width and overlap of estimations in Fig. 4, we found participants integrated slightly more unique attributes into their analysis using the SEPARATED interface overall (Fig. 4A), yet interacted with slightly more attributes in both the Encode and Filter panels of the COMBINED interface (Fig. 4B and Fig. 4C), though not enough to be significant. P15 demonstrated both of these patterns by (1) mostly using the Attributes panel to integrate and organize lots of unique attributes (*"Since there were a lot of attributes, having them at the top of the drop-downs was useful"*) while (2) also exploring new attributes on the fly in the Encode and Filter panels because it was easy (*"In the [COMBINED] interface, I didn't have to add the attribute to the Attributes panel, I could just add the attribute to the panel I wanted it in"*).

Based on whether attributes were integrated ex-situ (*primary*) or in-situ (*secondary*) in the COMBINED interface, we observed three distinct patterns of interaction in Fig. 5 and compared them with the intervals of data integration in the COMBINED interface (Fig. 3C). $9/16$ participants (P1, 3, 5, 6, 7, 10, 12, 13, 15) interacted mostly with *primary* attributes, similar to how the SEPARATED interface is used; $6/9$ of these participants (P1, 7, 10, 12, 13, 15) also spent a majority of their time integrating with the Attributes panel before starting their analysis. $4/16$ participants (P4, 8, 11, 14) interacted mostly with *secondary* attributes; $3/4$ of these participants (P4, 11, 14) also rarely used the Attributes panel, in contrast to how the SEPARATED interface is used. $3/16$ participants (P2, 9, 16) interacted with a mixture of *primary* and *secondary* attributes with no clear preference. P9 based their interactions on the task: *"If it's a specific task and we need attributes 'x', 'y', and 'z', then we can directly use the Encode panel. But if I have a free-form task, then I want to shortlist my attributes first [in the Attributes panel] and explore those."* P2 initially used the Attributes panel, then switched to integrating with the Encode and Filter panels, saying *"I don't think adding attributes to the Attributes panel made a difference... that is not where I use my attributes."* Attribute panel usage was inconsistent for these three participants and the four remaining participants (P3, 5, 6, 8) with mostly *primary* or *secondary* attribute interactions. For example, P8 spent time using the Attributes panel before starting analysis in *LQ1*, then did not use the Attributes panel at all in *LQ2*. Yet we found that they mostly interacted with *secondary* attributes in the Encode and Filter panels across both tasks.

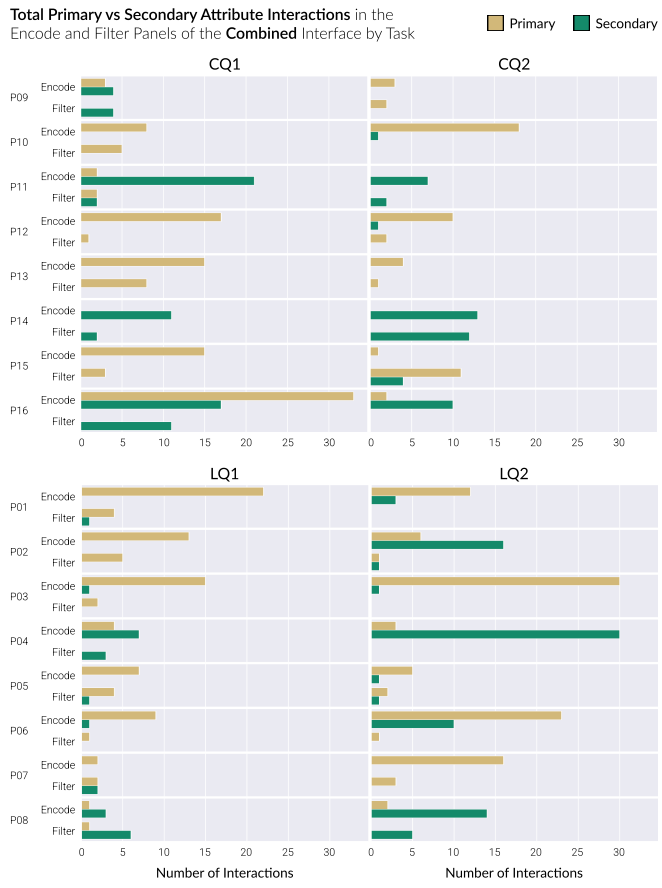

Fig. 5. Total counts of *primary* and *secondary* attribute interactions in the Encode and Filter panels of the COMBINED interface (Sect. 4.2). See Sect. 3.2.2 and Fig. 2 for definitions.

Where should data integration operations be combined with visual analytics operations? Overall, while attributes took less operations to integrate in the COMBINED interface, some participants integrated slightly more unique attributes with the SEPARATED interface. We also did not see large differences in total interactions with attributes in either the Encode or Filter panels between interfaces. In-situ integration taking less operations may not strongly affect how many attributes users interact with compared with ex-situ integration, suggesting a more balanced approached between in-situ and ex-situ. Yet considering that the Attributes panel is a common way to incorporate new attributes during analysis and is our proxy for intervals of data integration in both interfaces, we were surprised that Attributes panel usage in the COMBINED interface was split. Similarly, while *primary* attribute interactions are more common in tools like Tableau, *primary* and *secondary* attribute interactions in the COMBINED interface were also split and users did not converge on a single strategy for using integrated attributes during analysis. Participants further demonstrated preferences, e.g., those with a preference towards *primary* attribute interactions used the Attributes panel more often, and analogously less often for those with more *secondary* attribute interactions. Participant preferences for integration strategies (Sect. 4.3) suggest that in-situ integration helps users during visual data analysis in unique ways, despite little difference in interactions. We describe participants' strategies for using in-situ integration in the next section.

Open questions. In addition to the findings above, our study revealed open questions listed below:

1) If users can choose between in-situ and ex-situ integration on the fly, will constraints based on time, task requirements, or interface design affect users' interactions with a large coverage of attributes?
2) Similar to attribute coverage, if less operations does not lead to more interactions overall, will other constraints influence integration method preference?
3) Are users adapting their analysis process to new ways of integrating on the fly? How does prior experience affect whether users choose in-situ or ex-situ integration?

4.3 Integration Strategies

How should data integration operations be supported in tandem with visual analytics operations? Based on participants' time spent and interactions, as well as video recordings and qualitative feedback, we identified four distinct integration strategies across both interfaces. Most participants used a single strategy for each interface across both tasks, while a few switched strategies between tasks.

S1 – "Integrate attributes first" – both interfaces. Several participants integrated attributes before starting analysis and rarely integrated more afterwards. **S1** is characterized by more time spent integrating attributes into the Attributes panel than other strategies and, in the COMBINED interface, more *primary* attribute than *secondary* attribute interactions. The time spent adding attributes to the Attributes panel was usually apportioned to one integration session in the beginning of each task collecting a large subset of attributes, with participants rarely integrating after their initial collection. With the SEPARATED interface, this strategy was used almost exclusively since attributes could only be added to the Attributes panel via Excel. When using the COMBINED interface, $7/16$ participants ($P1, 7, 8, 9, 12, 15, 16$) utilized this strategy at least once and $4/16$ used it exclusively ($P7, 9, 12, 15$). P1 used this strategy to make a plan before analyzing the data: *"Based on the question, I didn't need to worry about all attributes. I wanted to pick attributes that would help me answer the question... I wanted to have in my head, 'what makes sense to visualize?', before starting analysis."*

S2 – "Don't think about integrating attributes" – COMBINED interface only. Some participants integrated on the fly and rarely shortlisted attributes before using them. **S2** is characterized by little to no time spent adding attributes to the Attributes panel at any point during analysis and, in the COMBINED interface, mostly *secondary* attribute interactions. Instead, participants used attributes directly in the panels they were interested in; e.g., the Encode or Filter panel. **S2** was not possible when using the SEPARATED interface. When using the COMBINED interface, $6/16$ participants ($P2, 4, 8, 11, 14, 16$) utilized this strategy at least once and $3/16$ used it exclusively ($P4, 11, 14$). P6 reflected on how they would use this strategy after the study: *"Adding attributes in the Attributes panel was confusing. At first I used it to see what attributes were there, but you could do the same operation in the other panels... The more you know the attributes, the less you would need the Attributes panel."*

S3 – "Integrate attributes as needed" – both interfaces. A few participants used their analysis process to inform them when to integrate, instead of all up front or on the fly. **S3** is characterized similarly to **S2**, including more time spent adding attributes to the Attributes panel and, in the COMBINED interface, mostly *primary* attribute interactions. However, that time was apportioned more evenly throughout the task in an "analyze as you go" fashion, with participants returning two or three times to integrate smaller subsets of attributes as needed. While this strategy was possible with the SEPARATED interface, no participants engaged with it when using the SEPARATED interface. When using the COMBINED interface, $6/16$ participants ($P1, 2, 5, 6, 10, 13$) utilized this strategy at least once and $2/16$ used it exclusively ($P10, 13$). P13 describes this strategy as an extension of their analysis process: *"I do that normally. I wanted to make changes then see what happens... what if I come across something interesting as the visualization changes?"* For them, integrating helps them generate insights: *"I think the free-form task was more about getting insights. I was seeing if adding or deleting anything was changing the visualization."*

S4 – "Integrate attributes one-at-a-time" – both interfaces. A handful of participants focused on a small number of attributes in detail, adding them to the interface one at a time. **S4** is characterized by less time spent adding attributes to the Attributes panel than the other strategies and, in the COMBINED interface, mostly *primary* attribute usage. To achieve this, participants briefly and sporadically added a few attributes to the Attributes panel as many as six times throughout a task. Only P13 used this strategy with the SEPARATED interface, while working on LQ1, returning to Excel multiple times for integrating one to two attributes at a time. When using the COMBINED interface, $3/16$ participants ($P3, 5, 6$) utilized this strategy at least once and one used it exclusively (P3). P3 used this strategy to organize their Attributes panel: *"At first, I tried putting in the specific attributes into the Encode panel, but that wasn't such a good idea. I want to have all of the attributes organized together in the Attributes panel instead."* P5 used it to organize the Elaborate panel: *"I wanted to see the order of attributes I added in the Attributes panel in the Elaborate table as well."*

Switching strategies. Some participants changed their strategies between the specific task (*Q1*) and free-form task (*Q2*). It is not clear what effects the task type had on why participants switched tactics. Four participants ($P1, 2, 5, 6$) started with **S3**: P1 switched to **S1**; P2 switched to **S2**; and $P5, 6$ switched to **S4**. Two participants ($P8, 16$) started with **S1**, then both switched to **S2**.

Ex-situ integration operations. Two ways of integrating attributes ex-situ emerged: (1) copy-and-paste; and (2) the Excel function *V-LOOKUP*. $13/16$ participants copied and pasted columns from one file to the next. P15 explains that *"it was similar to how I use other data analysis tools. I would copy and paste data from one spreadsheet to another."* Participants varied in whether and how they would validate the success of their joins. For example, P11 told us that *"when I was copy-pasting, I assumed that the rows from the primary table were all there in the secondary tables."* $3/16$ participants ($P2, 12, 16$) used *V-LOOKUP* to populate a new column with values based on a look-up with the primary key (*ID*) column, ensuring a correct join. P1 mentioned using *V-LOOKUP* as preferable to copy-and-paste but did not use it, citing a lack of time and familiarity: *"Sorry it's a bit slow [copy-and-paste], I'm not very good at Excel... another good way [to integrate] would be to use a V-LOOKUP to match the IDs... But looking at the IDs, they look like they match up."*

4.4 Analytical Behaviors

How will incorporating data integration into an on-going visual analytics process affect user behaviors? We describe observed analytical behaviors related to satsificing and exhibiting bias from a sensemaking perspective.

Satisficing. We observed patterns of satisficing, a cognitive heuristic for choosing a satisfactory or "good enough" option from alternatives [12]. Pirolli and Card explain that "time pressures and data overload work against the individual analyst's ability to rigorously follow effective methods for generating, managing, and evaluating hypotheses" [9]. P5 managed the constraints of integration by prioritizing insight generation: *"I got less time to decide on which attributes to use, and I spent more time on the data pre-processing. I would prefer the [COMBINED] interface more. In visual data analysis, it's more important to gain insights."* P9 told us that they considered not introducing attributes into their analysis as a consequence of their sensemaking: *"It was a lot of operations to just add a single variable. If I was 50/50 about whether to include an attribute, then I may not include it in my analysis."*

Conversely, from a data-frame theory perspective, Klein et al. suggest that sensemaking is the balance of fitting data to a frame and the frame affecting how data is interpreted [10]. When asked why they satisficed, P9 attributed the difference between interface designs to their trust in the data: *"In terms of accuracy and insights the [SEPARATED] interface was better. But the simplicity of the [COMBINED] interface was better... I think it all comes down to how much you trust the data. If you trust it, the [COMBINED] is better. But if the data isn't clean, the [SEPARATED] is better."* Having integration be a simpler and more seamless part of visual data analysis could improve both the generation and coverage of hypotheses afforded by access to new attributes. At the same time, removing the seams may affect the balance of fitting data to frames and frames to data that proceeds under the constraints of time pressures and data overload.

Exhibiting bias. We identified potential examples of cognitive bias due to the separation of data preparation and analysis. For example, participants visualized the same set of attributes in similar ways, often saying they were "familiar" with these attributes, and may have been exhibiting confirmation bias [43]. We expected participants to consider using different combinations of attributes to change their perspective, particularly considering the ease of immediately viewing integrated attributes in the COMBINED interface. However, many of these participants instead claimed that "the data just wasn't showing them what they wanted to see" and continued attempting to preserve their frames by rearranging the encodings of the same attributes. Klein et al. suggest that initial anchors, in this case the data that these participants first integrated, can have profound effects on performance during sensemaking tasks [10].

Similarly, we found a tendency for some participants to exhibit anchoring effects [43], [44] by sticking to a smaller number of attributes for a longer amount of time before attempting to branch out and find more information, if at all. Often these participants, when asked how their analysis was going, would tell us that they were "trying to make it work". However, they rarely used the integration features of either interface to overcome these issues, instead focusing on trying different combinations of encodings and filters to solve the problem. Both of these examples suggest that such biases could persist even when the interface design affords the opportunity to integrate data with a single click.

5 Preliminary Guidelines

Towards preliminary guidelines, many participants cited commonly raised concerns around affordances, direct manipulation of the data [45], and fluid interaction [46], such as the responsiveness of the system and lack of load times, the layouts of the drop-down menus and panels, the usefulness of dynamic querying [47], and the user experience. With these in mind and our findings, we synthesized three guidelines for designing future visual analytics interfaces that can support integrating attributes throughout an active analysis process: (1) show where and how data are being integrated (Sect. 5.1); (2) use in-situ integration for exploring the space of attributes (Sect. 5.2); and (3) balance manual and automated approaches (Sect. 5.3).

5.1 Show where the data comes from

The transparency of *how* and *what* data are integrated is essential for in-situ data integration within a visual analytics system. For example, P1, 8, 12 all used strategy **S1** in the Combined interface and asked for more access to the raw data. P1 specifically wanted access to verify the quality of the data: *"I liked [Combined] for the ease of use, but I also like to see the actual data in the [Separated]. I would go back and use the tool to verify the quality of the data."* On the other hand, P5 felt that the Combined interface would process the raw data better than they could, based on their experience with similar visual analytics tools: *"When I copy-and-pasted data in the [Separated] interface, I had to manage column names and there couldn't be manual errors, and I feel like the [Combined] interface would do a better job of overcoming those... I've used Tableau and Power BI. I was told that those interfaces mapped data correctly, so I assumed this one did as well."* While our study intentionally controlled for data quality, future interfaces should clarify the limitations of how and what data are integrated. For example, Cashman et al. use a pop-up window in their CAVA system [33] to display how a join will be performed before it is ultimately integrated into the data set. Analysis outcomes that follow from "anonymous" integration could be dangerous if not carefully evaluated.

However, we found strong evidence that showing the user all of what attributes can be integrated can negatively affect analysis in several ways. For example, P2 expressed concerns of cognitive overload from having too many attributes to think about: *"There were so many variables [in the drop-downs] that I missed some. If the attributes were laid out [in the Attributes panel] like in Tableau then maybe I would have seen it..."* Too many attributes can also negatively affect task completion time. With the Combined interface, P10 cited too many attributes in the drop-downs: *"In the [Combined] interface, having too many attributes in the drop-downs in the Encode and Filter panels took more time than expected to look through."* With the Separated interface, P5 took time to traverse multiple files: *"The attributes were spread across a lot of files. Traversing the files to find the right attribute took a while."* Further, making attributes more visible may contribute to satisficing. P11 satisficed based on the task: *"Sometimes I wasn't sure where to look... When doing the tasks, I would often look for just the attributes I felt like were relevant to the task. I ignored the rest because I had to go through the tables to find them."* P6 attributed this to interface design: *"With the specific questions, I knew where to look in the tables with the [Separated] interface, compared with the [Combined] interface which was harder for this."* However, when the number of attributes felt manageable, P9 liked the drop-downs for finding relevant attributes: *"I didn't know where the attributes were in the system, but that wasn't a problem. I had to skim through the attributes in the drop-down [of the Combined interface], but the number of attributes was not so great. If there was more, it would have been more difficult."* Thus, designers should carefully consider how the number of attributes in the data may influence users' time spent, interactions, and analytical behaviors by balancing how much is shown to the user at once with how easy it is for users to find relevant attributes to integrate during analysis.

5.2 Use in-situ integration for exploration

The issues faced by participants using the Separated interface are common to data integration as an entry point to analysis. There is some evidence that the overhead cost of integration outside the interface could prevent users from finding relevant attributes. For example, P2 described trading off between analyzing and integrating, causing conflicts where the interface lacked support. They attributed mistakes they made to managing multiple tables manually across several windows with the Separated interface: *"I copied the values into the wrong file because so many windows were open. That wasted my time."* Instead, we saw more evidence that encode and filter operations were useful for exploring the attribute space to find and integrate new attributes on the fly in the Combined interface. P11 found the grouping of attributes in a single drop-down helpful because it allowed them to see all of the available data: *"I liked having all of the attributes that were relevant to the tasks in one place."* P14 felt the Combined interface helped them find more attributes: *"It was more convenient to see all of the attributes in the [Combined] interface. For example, what if there was an attribute hiding in a table that I missed?"* P5 used the names of the files as a proxy for determining the semantic relevance of an attribute to the task: *"I didn't know what all attributes were [in the Combined interface], but I checked the names of the files for the attributes in order to choose which attributes to use."* Thus, in-situ data integration for quickly encoding new attributes in the visualization could help users maintain their focus on performing visual data analysis. In the Combined interface, participants can immediately observe the results of integration in their visualizations and generate new ideas and strategies to

explore [14]. This can also allow participants to evaluate the quality of the integrated information visually [15]. An "undo" feature would further promote principles of direct manipulation that preserve the flow of the analysis process [46]. When combined with visual analytics operations, in-situ integration may help users maintain an active and continuous analysis process.

5.3 Balance manual and automated approaches

The differences in users' interactions and behaviors between ex-situ and in-situ integration provides some evidence for when manual approaches should be used over automated approaches. P2 had difficulty remembering relevant attributes when the integration process was automated: *"In Tableau and Power BI, you have to manually create tables when joining tables... but since I wasn't the one doing the joins [in the* COMBINED *interface], it was harder to remember the attributes that were available to me. I would have remembered them if I had to manually join the attributes."* P16 also preferred manual interactions for understanding their data: *"I liked the [*SEPARATED*] interface a lot even though it involved a lot of manual interactions with the CSVs... In the [*SEPARATED*] interface, you are really visualizing your data, you understand how your data will be visualized. In the [*COMBINED*] interface, you have to understand how the different panels work together instead."* Yet there is evidence that automated approaches may improve the analysis process. For example, despite tasks not being timed, the distribution of time spent with the SEPARATED interface affected the performance of P5: *"I think most of the time was spent doing data preparation and I felt rushed."* P9 also spent longer with the SEPARATED interface because they had to separate analysis from integration: *"It takes a long time to do manual integration. For example, when I open a CSV [file], I have thoughts about what it may contain, then I see the attributes. It's not the same operation to find the attribute and use the attribute, unlike in the [*COMBINED*] interface."* P2 elaborated on how maintaining context affected both their time spent and insights generated: *"Adding attributes [to the* SEPARATED *interface] was a pain. I had to open tables and copy and remove incorrect values. I had to look at the names of the files to guess where attributes would be... [my time] was mostly spent cleaning and arranging data. I didn't have a lot of time to focus on how I could improve the visualization."*). P9 explains how their experience differed between interfaces: *"In the [*COMBINED*] interface, I can explore more attributes in a shorter amount of time. In the [*SEPARATED*] interface, there's more overhead that takes time away from the task of visualizing data."* This suggests that designers should consider a minimal but fluid design [46] for in-situ integration when time spent and interactions should be minimized, otherwise opt for manual approaches. One potential in-situ solution could be the automation of integration steps that do not require as much human input. Data blending techniques such as those provided by Tableau [14] and Google Data Studio exemplify this idea by maintaining a human-in-the-loop control while abstracting away the more technical details of integration. This may help users reduce the number of concurrent processes to manage while helping them maintain context.

6 LIMITATIONS AND FUTURE WORK

Beyond the preliminary state of our guidelines, we contribute open research questions and avenues for future work investigating more types of integration, task requirements, and users' prior experience towards "concrete" guidelines.

Types of integration. To elicit initial observations, we directly compared ex-situ and in-situ integration by intentionally simplifying and limiting integration to column concatenation and selection. Because of this, we could not study technical challenges associated with more complex data integration, such as performing deduplication and entity resolution described in Sect. 2.1, as well as issues of latency in how long it takes to resolve joins and data quality across heterogeneous sources. Column concatenation itself is very simple compared to "real" data integration that can involve extensive planning and even custom tool development when the semantics of data is complicated and/or involved semi-structured data such as text. Additionally, we ensured a one-to-one mapping of recordings between every file and described all files and attributes to participants up front in Sect. 3.1. This choice did not allow us to investigate how the quality of data from potentially unknown sources affects users interactions and behaviors, e.g., introducing attributes from multiple websites or APIs, potentially on the fly as they become available, and throughout the analysis process as users request them. Given these limitations, it is unclear whether users will consider the validity of a data source when data is not presented up front. The way forward might require adopting (even developing) a more sophisticated approach to comparative evaluation not based on direct "apples-to-apples" comparison, e.g., focusing on single-table data wrangling with more complex integration tasks compared with equivalent field calculations wrapped in a UI-based approach.

Task requirements. We chose tasks following Fekete et al. [48], who argue that visualizations are often best for exploratory tasks where the goal is to make discoveries or generate insights about data. Our methodology did not allow us to investigate effects on task performance, including what markers indicate the end of a task and when or if participants decide to stop integrating. Future work could model factors that determine how much data will be integrated and when based on task requirements and/or UI design. It is unclear whether there is a threshold across which user interactions and behaviors change. Additionally, while we found that our definition of time spent integrating closely matched qualitative feedback from participants, participants could have been using the time between integrating and visualizing data to consider what visualization to create that best utilizes the integrated attributes; this time may be significant. We acknowledge that other measurements for marking the start and end of integration are also reasonable. Finally, we limited the number of attributes in the data set, unlike real analysis scenarios where users may have *a priori* knowledge of attributes they want and/or risk cognitive overload while browsing for relevant attributes out of many. Our preliminary guidelines suggest not to overload users with every attribute that can be integrated; future work should isolate these effects and systematically describe them as a factors in the data integration process during visual data analysis.

Users' experience. Our participants comprised a fairly homogeneous user group (all recruited from the same academic institution with varying levels of experience conducting data analysis, creating visualizations, and using data analysis software) to reduce confounds in comparing feedback. Because of this, it is unclear whether incorporating data integration into visual analytics processes at all will depend on familiarity with software and analysis practices, particularly where data integration plays a major role, either as part of the analysis process or off-loaded to others. For example, decision-makers with existing routines may or may not conduct analysis differently [11] when integration on the fly is possible. Participants also varied in how much time they spent on tasks depending on their prior domain knowledge of the data sets, skill in performing visual data analysis, and comfort with learning and using the interfaces. For example, P16 avoided attributes based on their domain experience: *"In the loan question, I wasn't familiar with some of the terms, so I didn't really touch some of the files because I wasn't comfortable using them."* Additionally, when participants felt that a task description was ambiguous or open to interpretation, they took longer to prepare data. P15 reflected: *"I feel like the difference in the quality of my analysis was less about the interface I was given and more about the task I was given."* Thus future studies should investigate how task requirements and existing domain knowledge impact integration during visual data analysis. Our results indicate that it may be neither preferable nor realistic to start with a single file full of attributes.

7 Conclusion

This paper presents preliminary results and guidelines when combining data integration and visual analytics processes. We developed two visual analytics interfaces: one that presents manual file-based ex-situ integration (Separated); and one that presents automatic UI-based in-situ integration (Combined). We conducted a within-subjects user study with 16 participants and a mixed-methods analysis of participants' interactions and behaviors.

Where and how should we support data integration operations in tandem with visual analytics operations? Participants spent time integrating before analysis, on the fly, and switching between strategies. The time spent on tasks and interactions between interfaces was also not significantly different. In-situ integration could enable analysts to explore attributes faster than analogous ex-situ strategies, leaving more time for analysis tasks. How will incorporating data integration into an on-going visual analytics process affect user behaviors? We observed participants using integration operations to generate and track hypotheses and insights as well as patterns of satisficing and bias in participants' analytical behaviors. Supporting integration operations in visual analytics tools will also require transparency up front about what and how data are integrated as well as balancing both automated and manual approaches.

Acknowledgments

This work was supported in part by the National Science Foundation grant IIS-1813281 and DRL-2247790.

References

[1] J. J. Thomas and K. A. Cook, "A visual analytics agenda," *IEEE Computer Graphics and Applications*, vol. 26, no. 1, pp. 10–13, 2006.

[2] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, "Visual analytics: Definition, process, and challenges," in *Information visualization*. Springer, 2008, pp. 154–175.

[3] M. Lenzerini, "Data integration: A theoretical perspective," in *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2002, pp. 233–246.

[4] A. Doan, A. Halevy, and Z. Ives, *Principles of Data Integration*, 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2012.

[5] A. Endert, M. S. Hossain, N. Ramakrishnan, C. North, P. Fiaux, and C. Andrews, "The human is the loop: new directions for visual analytics," *Journal of Intelligent Information Systems*, vol. 43, no. 3, pp. 411–435, Dec 2014. [Online]. Available: https://doi.org/10.1007/s10844-014-0304-9

[6] J. S. Yi, Y. ah Kang, J. Stasko, and J. A. Jacko, "Toward a deeper understanding of the role of interaction in information visualization," *IEEE transactions on visualization and computer graphics*, vol. 13, no. 6, pp. 1224–1231, 2007.

[7] S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer, "Enterprise data analysis and visualization: An interview study," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2917–2926, 2012.

[8] P. Pirolli and S. Card, "Information foraging." *Psychological review*, vol. 106, no. 4, p. 643, 1999.

[9] ——, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *Proceedings of international conference on intelligence analysis*, vol. 5. McLean, VA, USA, 2005, pp. 2–4.

[10] G. Klein, J. K. Phillips, E. L. Rall, and D. A. Peluso, "A data-frame theory of sensemaking," in *Expertise out of context: Proceedings of the sixth international conference on naturalistic decision making*. New York, NY: Lawrence Erlbaum Assoc Inc, 2007, pp. 113–155.

[11] E. Dimara and J. Stasko, "A critical reflection on visualization research: Where do decision making tasks hide?" *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[12] R. J. Heuer, *Psychology of intelligence analysis*. Center for the Study of Intelligence, 1999.

[13] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer, "Voyager: Exploratory analysis via faceted browsing of visualization recommendations," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 649–658, 2015.

[14] K. Morton, R. Bunker, J. Mackinlay, R. Morton, and C. Stolte, "Dynamic workload driven data integration in tableau," in *Proceedings of the 2012 ACM SIGMOD international conference on management of data*, 2012, pp. 807–816.

[15] M. Van Kleek, D. A. Smith, H. S. Packer, J. Skinner, and N. R. Shadbolt, "Carpé data: Supporting serendipitous data integration in personal information management," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 2339–2348. [Online]. Available: https://doi.org/10.1145/2470654.2481324

[16] J. Hendler, "Data integration for heterogenous datasets," *Big data*, vol. 2, no. 4, pp. 205–215, 2014.

[17] C. Bizer, "The emerging web of linked data," *IEEE Intelligent Systems*, vol. 24, no. 5, pp. 87–92, 2009.

[18] A. Gal, "Uncertain entity resolution: re-evaluating entity resolution in the big data era: tutorial," *Proceedings of the VLDB Endowment*, vol. 7, no. 13, pp. 1711–1712, 2014.

[19] P. Konda, S. Das, A. Doan, A. Ardalan, J. R. Ballard, H. Li, F. Panahi, H. Zhang, J. Naughton, S. Prasad *et al.*, "Magellan: toward building entity matching management systems over data science stacks," *Proceedings of the VLDB Endowment*, vol. 9, no. 13, pp. 1581–1584, 2016.

[20] Y. Li, J. Li, Y. Suhara, A. Doan, and W.-C. Tan, "Deep entity matching with pre-trained language models," *arXiv preprint arXiv:2004.00584*, 2020.

[21] P. Hoefler, M. Granitzer, E. E. Veas, and C. Seifert, "Linked data query wizard: A novel interface for accessing sparql endpoints," in *LDOW*, 2014.

[22] A. Mohamed, G. Abuoda, A. Ghanem, Z. Kaoudi, and A. Aboulnaga, "Rdfframes: Knowledge graph access for machine learning tools," 2020.

[23] P. A. Bonatti, S. Decker, A. Polleres, and V. Presutti, "Knowledge graphs: New directions for knowledge representation on the semantic web (dagstuhl seminar 18371)," in *Dagstuhl Reports*, vol. 8. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[24] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. d. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier *et al.*, "Knowledge graphs," *Synthesis Lectures on Data, Semantics, and Knowledge*, vol. 12, no. 2, pp. 1–257, 2021.

[25] Y. Lou, M. Uddin, N. Brown, and M. Cafarella, "Knowledge graph programming with a human-in-the-loop: Preliminary results," in *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, 2019, pp. 1–7.

[26] M. Dallachiesa, A. Ebaid, A. Eldawy, A. Elmagarmid, I. F. Ilyas, M. Ouzzani, and N. Tang, "Nadeef: a commodity data cleaning system," in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, 2013, pp. 541–552.

[27] E. Zhu, Y. He, and S. Chaudhuri, "Auto-join: Joining tables by leveraging transformations," *Proceedings of the VLDB Endowment*, vol. 10, no. 10, pp. 1034–1045, 2017.

[28] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, and V. Raghavendra, "Deep learning for entity matching: A design space exploration," in *Proceedings of the 2018 International Conference on Management of Data*, 2018, pp. 19–34.

[29] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. Van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono, "Research directions in data wrangling: Visualizations and transformations for usable and credible data," *Information Visualization*, vol. 10, no. 4, pp. 271–288, 2011.

[30] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer, "Wrangler: Interactive visual specification of data transformation scripts," in *Proceedings of the sigchi conference on human factors in computing systems*, 2011, pp. 3363–3372.

[31] M. Kahng, S. B. Navathe, J. T. Stasko, and D. H. Chau, "Interactive browsing and navigation in relational databases," *arXiv preprint arXiv:1603.02371*, 2016.

[32] N. Cramer, G. Nakamura, and A. Endert, "The impact of streaming data on sensemaking with mixed-initiative visual analytics," in *Augmented Cognition. Neurocognition and Machine Learning*, D. D. Schmorrow and C. M. Fidopiastis, Eds. Cham: Springer International Publishing, 2017, pp. 478–498.

[33] D. Cashman, S. Xu, S. Das, F. Heimerl, C. Liu, S. R. Humayoun, M. Gleicher, A. Endert, and R. Chang, "Cava: A visual analytics system for exploratory columnar data augmentation using knowledge graphs," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1731–1741, 2021.

[34] S. Latif, S. Agarwal, S. Gottschalk, C. Chrosch, F. Feit, J. Jahn, T. Braun, Y. C. Tchenko, E. Demidova, and F. Beck, "Visually connecting historical figures through event knowledge graphs," in *2021 IEEE Visualization Conference (VIS)*, 2021, pp. 156–160.

[35] G. Smith, M. Czerwinski, B. Meyers, D. Robbins, G. Robertson, and D. S. Tan, "Facetmap: A scalable search and browse visualization," *IEEE Transactions on visualization and computer graphics*, vol. 12, no. 5, pp. 797–804, 2006.

[36] J. Bernard, M. Steiger, S. Widmer, H. Lücke-Tieke, T. May, and J. Kohlhammer, "Visual-interactive exploration of interesting multivariate relations in mixed research data sets," in *Computer Graphics Forum*, vol. 33. Wiley Online Library, 2014, pp. 291–300.

[37] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: concepts, methodologies, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 3, pp. 1–55, 2014.

[38] J. M. Kanter and K. Veeramachaneni, "Deep feature synthesis: Towards automating data science endeavors," in *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2015, pp. 1–10.

[39] K. A. Ericsson and H. A. Simon, *Protocol analysis: Verbal reports as data*. the MIT Press, 1984.

[40] M. Brehmer and T. Munzner, "A multi-level typology of abstract visualization tasks," *IEEE transactions on visualization and computer graphics*, vol. 19, no. 12, pp. 2376–2385, 2013.

[41] P. Dragicevic, "Fair statistical communication in hci," in *Modern statistical methods for HCI*. Springer, 2016, pp. 291–330.

[42] R. E. Boyatzis, *Transforming qualitative information: Thematic analysis and code development*. sage, 1998.

[43] E. Wall, L. M. Blaha, C. L. Paul, K. Cook, and A. Endert, "Four perspectives on human bias in visual analytics," in *Cognitive biases in visualizations*. Springer, 2018, pp. 29–42.

[44] I. Cho, R. Wesslen, A. Karduni, S. Santhanam, S. Shaikh, and W. Dou, "The anchoring effect in decision-making with visual analytics," in *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2017, pp. 116–126.

[45] E. L. Hutchins, J. D. Hollan, and D. A. Norman, "Direct manipulation interfaces," *Human–computer interaction*, vol. 1, no. 4, pp. 311–338, 1985.

[46] N. Elmqvist, A. V. Moere, H.-C. Jetter, D. Cernea, H. Reiterer, and T. Jankun-Kelly, "Fluid interaction for information visualization," *Information Visualization*, vol. 10, no. 4, pp. 327–340, 2011.

[47] C. Ahlberg, C. Williamson, and B. Shneiderman, "Dynamic queries for information exploration: An implementation and evaluation," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1992, pp. 619–626.

[48] J.-D. Fekete, J. J. v. Wijk, J. T. Stasko, and C. North, "The value of information visualization," in *Information visualization*. Springer, 2008, pp. 1–18.

Adam Coscia is a PhD student at Georgia Tech's School of Interactive Computing and a member of the Visual Analytics Lab. His research interests include Visual Analytics, Human-computer Interaction (HCI), Large Language Models (LLMs), and Explainable Artifical Intelligence (XAI). He received his B.S. in Physics. He won the President's Fellowship for top incoming PhD students.

Ashley Suh received her MS in Computer Science from Tufts University where she is currently pursuing a PhD. Her research interests include information visualization, visual analytics, and human-centered AI.

Remco Chang received his PhD in computer science from the University of North Carolina Charlotte. He is an associate professor in computer science with Tufts University. His research interests include visual analytics, information visualization, human computer interaction, and databases.

Alex Endert is an associate professor at the School of Interactive Computing, Georgia Tech. He directs the Visual Analytics Lab, which explores novel user interaction techniques for visual analytics, for domains including intelligence analysis, cyber security, manufacturing, decision making, and others.