The Impact of Streaming Data on Sensemaking with Mixed-Initiative Visual Analytics

Nick Cramer[1], Grant Nakamura[1], and Alex Endert[2(✉)]

[1] Pacific Northwest National Laboratories, Richland, WA, USA
[2] School of Interactive Computing, Georgia Institute of Technology, 85 5th Street NW, Atlanta, GA, USA
endert@gatech.edu

Abstract. Visual data analysis helps people gain insights into data via interactive visualizations. People generate and test hypotheses and questions about data in context of the domain. This process can generally be referred to as sensemaking. Much of the work on studying sensemaking (and creating visual analytic techniques in support of it) has been focused on static datasets. However, how do the cognitive processes of sensemaking change when data are changing? Further, what implication for design does this create for mixed-initiative visual analytics systems? This paper presents the results of a user study analyzing the impact of streaming data on sensemaking. To perform this study, we developed a mixed-initiative visual analytic prototype, the Streaming Canvas, that affords the analysis of streaming text data. We compare the sensemaking process of people using this tool for a static and streaming dataset. We present the results of this study and discuss the implications on future visual analytic systems that combine machine learning and interactive visualization to help people make sense of streaming data.

Keywords: Sensemaking · Streaming data · Visual analytics

1 Introduction

The creation and storage of data from increasing sources creates important challenges for not only the design of technology, but may change the cognitive processes that humans exhibit when analyzing data. Streaming data is becoming more commonly available, as data is more continuously created, sensed, and stored. The speed (or velocity) of how often streaming data updates varies greatly. For example, news updates may happen daily, new email may arrive every few minutes, while packets of network activity or Twitter feeds may occur with sub-second intervals. Nonetheless, streaming data signifies a shift in the persistence of datasets to a model where the data is no longer complete or static throughout the analysis. In turn, this impacts what we know about data analysis – both from a technical system requirements standpoint of how to design and build visual analytic systems, and also from a cognitive, analytical reasoning perspective of what we know about how people reason about data and perform sensemaking.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC 2017, Part I, LNAI 10284, pp. 478–498, 2017.
DOI: 10.1007/978-3-319-58628-1_36

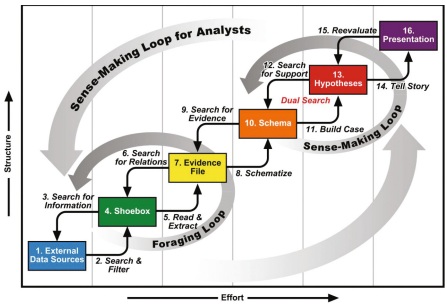

Fig. 1. The sensemaking loop, depicting a notional model for the cognitive stages involved in analyzing and understanding data.

Visual analytic techniques are one way to gain insight into data. These techniques foster sensemaking and discovery through visual data exploration [16,33]. They combine the computational power of analytics with the advantages of interactive visualization to produce insights into data. Incorporating the user into the analysis process creates an analytic discourse between the analyst and the data [10]. Such processes are often described as sensemaking [24]. The "sensemaking loop" (shown in Fig. 1), presented by Pirolli and Card [24], depicts a notional model of cognitive stages that users progress through during a typical exploratory data analysis task. The model was created based on a series of interviews and observations of intelligence analysts performing their professional jobs. This model has been widely adopted by the visual analytics and information visualization community given its applicability to many of the tasks these technologies support.

However, the studies which created this model used static data. Thus, there is an inherent assumption that the data is constant during the immediate analysis session. Further, the design of many visual analytic techniques assert that the dataset remains unchanged throughout the analysis. This raises the important questions of: how do the cognitive reasoning processes of analysts change when data is changing or streaming? what design principles become critical to the success of visual analytic techniques intended to function on non-static data?

In this paper we present the results of a user study examining the impact of streaming data on sensemaking and visual data exploration. We are primarily interested in understanding how introducing new data impacts an ongoing sensemaking task. Our study observed two conditions of the same dataset (static and streaming conditions). For both conditions, the study consisted of 5 one-hour sessions spread over 5 days. The static condition was given the entire dataset at the beginning, while the streaming condition had new data introduced at each session. Finally, we present a set of design guidelines for future visual analytic tools for streaming data.

The primary contributions of this paper are: (i) the results of a user study showing the impact of streaming data on the sensemaking process, and (ii) design guidelines for future visual analytic techniques for streaming data.

Our findings give rise to the notion that streaming data requires tighter coupling between the sensemaking processes of users, and the analytic processes of systems. We show how this can be made possible through interaction techniques such as *semantic interaction* [5], where user interactions with the interface are interpreted by the system to steer the underlying analytic models. We illuminate this need through the results and discussion of the user studying comparing static to streaming data conditions.

2 Related Work

The research presented in this paper is grounded in prior work discussed below.

2.1 Sensemaking and Analytical Reasoning

Analyzing data for the purpose of gaining insight is largely a cognitive process. This process of increasing one's knowledge or understanding about a domain or phenomena through analyzing data has been widely studied. One commonly used concept to describe this process is sensemaking, a cognitive activity of gaining understanding about the world, through the analysis of data [28]. For example, Pirolli and Card depict the cognitive stages of sensemaking in a notional model called the "sensemaking loop" [24]. This model emphasizes the importance of foraging and extracting content from data, as well as synthesizing these pieces of data inter higher-level insights. This complex, iterative process entails generation and testing hypotheses, as well as more low-level data filtering and retrieval tasks. Sensemaking involves internalizing and understanding the information in the context of the person's experiences and prior knowledge. For instance Klein et al. describe the process as refining one's "data-frames" [18], where the refining and augmenting of one's understanding about a phenomena is explained through framing and re-framing. The fluidity of these tasks was more recently emphasized by Kang and Stasko [15], who comment that stages and tasks of sensemaking to not necessarily follow a given order, and people may switch between them at any given time. Zhang and Soergel [37] proposed an iterative sensemaking model to more fully describe the iterative nature of this synthesis process. This process is at times also called "signature discovery" [13].

2.2 Visual Text Analytics

Visual analytic systems have been developed in support of sensemaking for text corpora. These systems tactfully combine statistical and data analytic models with interactive visual interfaces to enable people to reason about their data [33]. Specific to text analytics, several prior examples exist. For instance, Jigsaw [31] provides people with multiple views generated from extracted terms and frequencies.

One common visual metaphor to support sensemaking is a spatial layout, or canvas. The fundamental grounding of this metaphor is the geospatial understanding people have of objects with relative geographic locations between each other (i.e., objects closer together are more similar) [30]. Andrews et al. found that providing analysts with the ability to manually organize information in a spatial workspace enabled them to extend their working memory for a sensemaking task [2]. They found that analysts created spatial constructs that represented knowledge artifacts corresponding to intermediate findings throughout the process (e.g., timelines, lists, piles, etc.). Shipman et al. coined this process of refining intermediate spatial structures over time as "incremental formalism" [29]. They discuss how people were better able to express their knowledge structures spatially because freely organizing information in space does not require people to explicitly specify what the construct means. For example, analysts can create piles or lists without specifying the parameters used to create them. For text analysis in particular, these spatial constructs have been shown to encode significant amounts of semantic information about an analyst's process and insights [9]. Examples of visual analytic applications built to enable users to manually create spatial data layouts that aid their analytical reasoning include Analyst's Workspace [3], the nSpace Sandbox [36], and others.

2.3 Streaming Data Visual Analytics

Streaming data is a growing challenge in the way of data complexity for visual analytics [11,25]. There are areas of related work for streaming data analytics, ranging from algorithmic advances to handle the technical challenges of incorporating additional data during runtime (e.g., [22]), to visualization techniques for showing data changes over time. Mansmann et al. describe the concept of "dynamic visual analytics" and point out that streaming data presents additional challenges over temporal data due to caching and other data storage challenges [19]. A key difference comes through the realization that streaming data (ranging for various speeds of data arrival or updates) has impacts on the human reasoning process, and thus the design of visualizations. Specific to streams of textual data, Rohrdantz et al. have enumerated and discussed several of the challenges [26].

For example, STREAMIT shows users a spatial clustering of text documents, where similarity functions are used to place similar documents near each other [1]. While this is a familiar technique for showing text visually (i.e., [35], this work showed how as new data is imported, it can be added to the existing clusters so that users can observe how the new information maps to existing user-defined clusters. Further, Fisher et al. demonstrate how both interaction and visualization designs specific for streaming data must be carefully considered [12].

Similarly, Stolper et al. have presented the concept of "progressive visual analytics" [32] to describe a visual analytic technique for giving users incremental results for queries of large, complex data. While not specific to streaming data, the key challenge approached by this work is to understand how to show users

incomplete results of queries. With streaming data, the assumption that the data is "not yet compete" (and may never be) may hold true. Thus, some of their findings may be valuable to consider for the design of streaming data visual analytic systems.

3 Streaming Canvas Description

To study the impact of streaming data on sensemaking, we first created a visual analytic prototype that enables the analysis of streaming and static text datasets. The Streaming Canvas is a visual analytics tool for spatially organizing and analyzing textual datasets (see Figs. 2 and 3). The user interface consists mainly of a spatial workspace, or canvas, that enables people to create and organize groups of documents. Specifically, this visual analytic prototype supports streaming, or updating, datasets (i.e., datasets where incremental updates are periodically received).

Fig. 2. The Streaming Canvas lets users group documents into user-defined clusters. The system adds new documents to these clusters based on similarity to existing documents.

3.1 Data Model and Import

The Streaming Canvas models documents using a vector space model similar to the data models used in many other visual analytic tools. We have adopted a model familiar to us from previous work. We provide a brief description here as background. The details of our model are more fully described in [27,35].

Each document is treated as a "bag of words" in that the model considers the presence of words, but not their sequence. The words are extracted and counted for each document, and the resulting counts are used to create a vector space.

Each document is assigned a numeric vector that can be interpreted as a set of coordinates specifying a position in the vector space. Such a vector can also be used as the basis for additional computation; for the Streaming Canvas, this is the key characteristic required of the vector space model.

In our particular model, the features in our vector consist of the top-scored 200 terms in the dataset, which we call topics (extracted and weighted using the entity extraction technique from Rose et al. [27]). These topics form the basis for the vector space, with each dimension corresponding to a topic term.

For the user study, the application augments the vector space model with an additional set of relevance weights for tracking user interest in topics. The application interprets some user interactions as indicators of interest in a document or documents, triggering a boost in the relevance weights for the dimensions most important for the document(s). The application immediately updates document badge sizes based on the modified weights. Subsequent vector space computations such as group assignment of new documents also account for the updated weights.

The application displays documents in groups that are statistical clusters within our vector space model. Our data model makes the group assignments for the initial set of documents on the basis of X-Means clustering [23]. Each group has a centroid vector that is the mean of the vectors for its documents. The application uses multidimensional scaling to convert these n-dimensional centroid vectors to the 2D coordinates in the display space.

The Streaming Canvas is intended to import, process, and display an initial set of documents, to be followed by zero or more increments of additional documents. The description of how streaming data is handled by the system is described in a later section.

3.2 User Interface

The Streaming Canvas is presented as a single-page Web application. The user interface consists of a canvas pane, a reading pane, and a menu bar (shown in Fig. 2). The primary visualization component of this interface is the canvas. The canvas represents documents as small rectangles, grouped based on similarity. The reading pane presents details about selections, including the text of the current document (if any). The menu bar contains additional operations and the application title.

The canvas shows documents appear as rectangular icons against a gray background (we refer to these as "badges"). Documents are clustered into roughly circular groups. The application assigns each document one of four badge sizes, where the size represents content magnitude as measured using the vector space model (normalized for document length). Each such group is labeled with gray text on a white central badge. Each label consists of a small number of prominent features computed based on the text content of the group's documents (shown in Fig. 3).

The application also positions the group badges relative to each other such that proximity correlates to content similarity. The application positions document badges using a force-directed layout, so document badge proximity does not imply content similarity [14].

3.3 User Interactions

The Streaming Canvas supports a number of user interaction in the canvas, and other user interface components. Specifically, the canvas supports:

Selecting a group - The user clicks on a group badge to select a group and all documents belonging to that group. The document badges become highlighted in orange. The document titles are listed in the reading pane.
Selecting a document - The user clicks on a document to select an individual document. That document's title is listed in the reading pane's document list, and selected, causing the document text to be displayed in the lower part of the pane. The document is colored blue in both the canvas and list.
Moving a group - The user can drag a group to any empty space on the Canvas, to move the group (and its documents) there.
Moving a document - The user can drag a document badge from one group and drop it on another group's badge, changing the group assignment of the document.
Panning, Zooming - The user can pan and zoom on the canvas.

The reading pane supports the following user interactions:

Selecting a document - The user clicks on a title in the Document List. The application displays the text for that document in the lower part of the pane.
Highlighting - The user drag-selects a passage in the document text, causing the passage to be highlighted. The application also bookmarks the document.
Searching - The user searches by typing a search query into the search box. The application highlights the resulting documents' badges in orange and lists their titles in the Document List.

The menu bar provides for the following user interactions:

Renaming a group - The user replaces the label via a dialog.
Creating a new group - A new group badge will appear in the canvas. The new group will initially contain no documents.
Bookmarking the current document - The application decorates the document's badge with a green vertical stripe, both in the document list and the canvas.

The user interaction design for the Streaming Canvas follows previously-established *semantic interaction* principles. Semantic interaction is an approach to interaction design for visual analytic tools that tightly couples exploratory user interaction with analytic model steering [8]. In prior work, examples of this coupling include using interactions such as highlighting phrases of text and grouping documents as a means to steer underlying dimension reduction, entity extraction [8], and information retrieval models [4]. User studies of semantic interaction show that this coupling of user interaction with model steering provides a good match between the insights users have during analysis and the parameterization of the model over time [6]. In general, the design decision to use semantic interaction stems from the intended functionality of the system being grouping and spatial organization of documents. For these tasks, specifically,

semantic interaction has been shown to be effective [7]. Finally, the document and feature vectors systematically generated help in the data analysis stages of the study.

The Streaming Canvas implements semantic interactions as follows. Some of the user interactions afforded in the user interface directly correspond to data model updates. Of the user interactions listed above, the two that are coupled to model steering operations are: moving a document from one cluster to another, and bookmarking a document. The resulting model steering impacts document and feature weight vectors to update. These vectors are integrated into the clustering and entity extraction models described in Sect. 3.1, and therefore make them helpful for the spatial organization and grouping functions supported by the system.

In the case of moving a document from one group to another, the vectors involved are those of the document being moved, and the vectors for any documents already in the new group. The highest-magnitude features are selected for each such vector, using a threshold of 1.5 standard deviations above the mean. The weighting coefficient for each of those features is increased by 0.05. No features are down weighted. In the bookmarking case, the only vector involved is that of the bookmarked document. Otherwise the steering computation is exactly the same.

Additionally, as user interacts with the application, the appearance of documents evolves in subtle but noticeable ways. As a document is accessed more, its color may change. Darker shades indicate a greater amount of handling, following a smudge metaphor where more handling of a document makes it less pristine over time. Prior work has shown how this technique can be applied to graphical user interface widgets to indicate more commonly used functions [21].

Document sizes also evolve depending on use. The application interprets some user interactions as indicators of interest in some features in preference to others. It therefore tries to upweight content associated with the preferred features.

3.4 Incorporating Streaming Data

When a new set of documents arrives, the application assigns each new document to an existing user-generated or system-generated group. The application computes a term vector for the new document by extracting terms from new documents using the same method as the initial document import. Weights of these terms are assigned based on the term weights for the current dataset and state of the system. That is, if the user has steered the system to assign certain terms significantly more weight than others, those weights will carry over to the new documents. Then, the document is assigned to the nearest group in the vector space model, where distance is the Euclidean distance from the group centroid to the document term vector. The goal is to assign newly arriving documents to the user- or system-generated groups which most relate.

4 Study Description

The purpose of the study was to explore differences in sensemaking behaviors between user groups faced with an analysis task over a static document collection versus a streaming document collection. We did this via a between-subjects design, testing a streaming versus a static dataset using the Streaming Canvas prototype. Thus, the primary research question this study seeks to understand is *how does streaming data impact the cognitive process of sensemaking in a visual data analysis scenario?*

4.1 Task and Dataset

Each participant was given the same task, stated as "identify suspicious behavior of individuals or organizations contained in this dataset." This task embodies the canonical structure of a sensemaking task, emphasizing open-ended exploration and discovery.

The dataset used for this study consists of 378 short documents modeled after intelligence reports. These documents are a subset of the data from the VAST Challenge 2014 [34]. Documents are typically one or two paragraphs in length, and contain details of some event or action that a person saw that may be of interest to a larger group of analysts. This fictitious dataset describes the events of an island of Kronos where events have unfolded surrounding the kidnapping of 10 GASTech company employees. Two organizations, GASTech and Protectors of Kronos (POK), and their members are of primary interest. The data for analysis is a set of historic news articles spanning from 1982 to 2013 and news articles and blogs covering current events January 19, 20 and 21 of 2014. The dataset has a known ground truth used to evaluate the accuracy of findings by our participants. Also, the dataset includes several "dead ends", consisting of lines of investigation that seem relevant, but ultimately do not result in the correct answer to the task.

4.2 Procedure

Participants were asked to analyze the dataset (either streaming or static) over 5 days. In total, each participant analyzed a collection of 378 text documents with an available time of 4 h 15 min for analysis. The participation time was 1 h per day over a 5 day period with 30 min the first day for tool training and 15 min the last day for post-analysis interview. For analysis time, participants had 30 min on day 1, 1 h on days 2–4, and 45 min on day 5.

Users in both streaming and static conditions analyzed the exact same collection of documents but each group received subsets differently. All users received the same documents in the same sequence. The static group received an initial set of 95 historic documents which spanned approximately 20 years. On day 2, an additional 283 documents were added to the visual analytics tool. On subsequent days, the dataset remained static and no new documents were introduced.

For the streaming condition, users received the same initial set of 95 historic documents on day 1. On day 2, they received an increment of 18 new documents. On days 3, 4, and 5, they received an increments of 83, 139, and 43 documents respectively. The binning of these documents was based on the clean monthly breaks in the temporality of the dataset. Each participant was given a 10 min tutorial on day 1 demonstrating the functionality of the Streaming Canvas software. A one page "cheat sheet" was provided to act as an aid to remind users of the meaning the visual encodings and query syntax options.

On day 1, both static and streaming user groups are asked to "Organize and understand this document set to become familiar with the history of the POK, GASTech, and of the region overall." For subsequent days, streaming users are to "Monitor and explain the activity and events unfolding over the next 3 days" while static users are to "Explain the activity and events which have unfolded over the past 3 days."

4.3 Participant Demographics

We recruited and randomly assigned participants into one of two groups: static or streaming. We recruited 9 volunteers (6 male), aged 21 to 36, with a range of modest data analysis and analysis tool experience. Volunteers had degrees in criminal justice, mathematics, operations research, computer science, civil engineering. We had 3 PhD students, 1 Master's student, 1 Bachelor's student, and 4 post-Bachelors.

A total of 11 people were recruited to participate in the study of which 9 were able to fully participate (5 performing the streaming condition, 4 the static). Partial data from 2 participants was excluded from any analysis because they were unable to participate for the full period of time. Participants with odd numbered identifiers are members of the static data condition while participants with even numbered identifiers are members of the streaming data condition. User01 and User05 (both in the static data condition) did not complete the study and were excluded from the analysis and results.

4.4 Data Capture and Analysis

We collected a variety of data to support the analysis of this study. For each trial, we recorded the audio and video of the participant for the entire duration. We also used screen recording to capture all the user interactions and visualization states of the system throughout the study. Further, the server uses a custom logging facility to capture every call made to server. Some state information from the data model is also logged for events, in order to provide additional context for analysis. Specifically we logged the term weight vectors for the cluster centroids so that we can analyze the content in the cluster compared to the cluster centroid. Throughout the study, one investigator was present to administer the think-aloud approach, taking notes and observing the participants. Finally, at the conclusion of the study, we administered a verbal questionnaire to ask about their findings and process. We analyzed all of this data to more fully understand

the sensemaking process of each participant and condition, as described in the results section below.

5 Results

We present the results of the user study as follows. We describe how the system was used by each condition (streaming compared to static). This includes a description of the user interactions and functionality of the system. Second, we describe how users in the two conditions leveraged spatial constructs, groups, and other affordances to organize and analyze their information in the workspace.

5.1 User Interactions Performed

The participants were tasked with using the Streaming Canvas visual analytics tool to build an understanding of the history and current unfolding events in the dataset. The Streaming Canvas supports a collection of user interaction which allow the user to accomplish this task. Interactions include selecting document full text to read, executing a keyword search, labeling a group of documents, and several more. As the user performs interactions during the course of analysis, these are logged (see Table 1).

Table 1. Overview of interactions performed. Total indicates the number of times each interaction was performed by all users over the duration of the study. Averages indicate how many times, on average, a user in each condition performed the interactions.

Interaction type	Total	Avg. streaming	Avg. static
Get document	8964	1065	910
Search	889	116	78
Move Doc to group	875	146	36
Re-position group	860	114	72
Label group	123	19	7
Bookmark document	118	10	17
Create group	90	14	5
Annotate document	55	4	8

On average, users working in the streaming data scenario performed more interactions of most types compared to users working with static data. Static data users only performed more bookmarking and annotating interactions (see Table 1). These are raw totals and averages. Given the small population size, no statistical significance is computed.

5.2 Spatial Constructs Created

The Streaming Canvas visualization allowed users to create named groups of documents as a mechanism to externalize and aid in their sensemaking. Based on prior work on "incremental formalism", the Streaming Canvas does not require users to explicitly specify the label or analytic reasoning associated with a group [29]. Instead, users can create and modify group labels, and group membership, through a more informal process. Across all users, groupings of documents were created for a variety of reasons but some commonalities emerged. Groups were formed around metadata such as document publication date and news source. Groups were also formed around an entity or topic. Groups for low value documents was created by 4 of the 9 users with labels like "junk", "trash", "less informative", and "too small". Lastly, one user created groups labeled explicitly for evidence to be gathered related to hypotheses.

We analyzed the labels of these user-created groups into concepts. These concepts (shown below) were created based on the meaning that our participants applied to these groups during their investigation. We derived these based on the data collected during the think-aloud protocol, and show examples of labels applied by participants.

User-generated group labels observed during this study include:

Entity-centric: "Karel", "Background on POK"
Topic-centric: "Pollution in Elodis", "Plane that left"

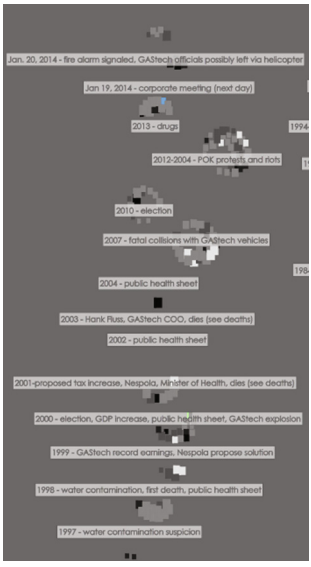

Fig. 3. Study participant User10 made extensive use of document groups representing time periods. Space was used to arrange these groups temporally with recent events at the top ranging to historic groups at the bottom.

Date-centric: "Jan 19 Details", "Late 1/20"
Source-centric: "Breaking", "Voices"
Low Value: "Junk", "Trash", "Less Informative"
Evidence: "H1 Evidence"

In the Streaming Canvas, users could re-position groups anywhere in the two dimensional space as they saw fit. While many users had document groups representing important dates, these groups were often not organized in temporal order within the space. However, a strong use of space to represent time emerged with two users, User2 and User10. In both cases, time was encoded vertically. User10 created a 2 column layout with groups for recent events at the top and groups for older dates descended down (see Fig. 3). User2 organized groups temporally using the vertical space as well but with older groups at the top and recent events at the bottom.

5.3 Grouping Documents

The Streaming Canvas provides several interactions in support of manipulating groupings of documents. Users can create a group, label a group, re-position a group, and move a document into a group. This subset of interactions lets the user organize the documents to fit their mental model and externalize their thinking as a cognitive aid.

The streaming data users performed notably more group-related interactions than the static data users. Looking at document moves into groups day-by-day shows that streaming users performed this interaction more on days where larger numbers of documents are streamed into the visualization (see Fig. 4).

Document Move Count per Day by User

	Day1	Day2	Day3	Day4	Day5	Grand Total
Streaming	**36**	**102**	**240**	**287**	**65**	**730**
User02	8	19	96	124	46	293
User04	4	52	84	45	2	187
User06		3	36	105	12	156
User10	3	28	24	13	5	73
User08	21					21
Static	**7**	**29**	**36**	**23**	**50**	**145**
User03	6	5	23	13	1	48
User05	1		8		49	58
User11		24				24
User09			5	10		15

Fig. 4. Document move interactions per day by user shows notably more for users in the streaming data scenario.

5.4 Interaction Sequences

Exploring the interaction sequences provided an interesting view into user sense-making patterns. Sequence diagrams were generated for all interactions on all days grouped by user and data scenario. Figure 5 gives an example which includes a legend for interaction type encoded in each data point. Each column represents

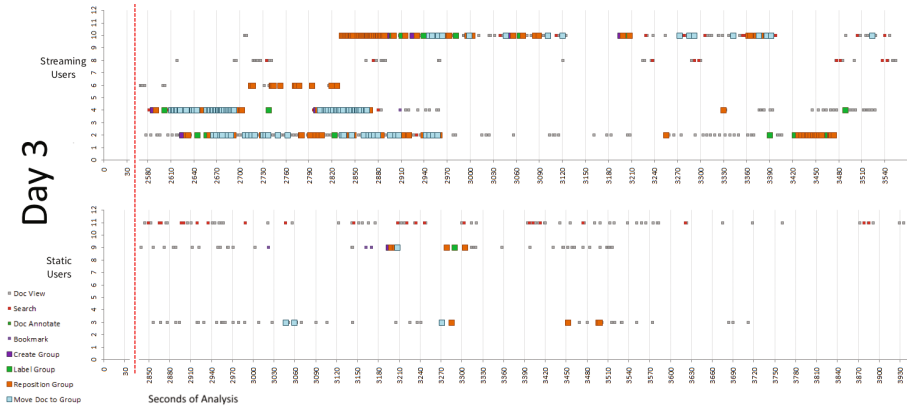

Fig. 5. An interaction sequence diagrams displaying when users performed interactions. User IDs are on the y-axis, with interactions (encoded by color and size) shown across the time of day 3 of the study. (Color figure online)

a study participant with each data point representing an interaction performed at a point in time during the user's analysis session. Time is represented vertically in seconds from start of the analysis session.

We posit that interaction types relate to the two major loops within Pirolli and Card's model of sensemaking [24] (foraging and synthesis) shown in Fig. 1. Triaging interactions such as Search, Read, Bookmark, and Annotate relate to foraging. While grouping interactions such as create, label, re-position, and move document relate to synthesis.

This mapping to the foraging and synthesis loops was included in the interaction sequence diagrams using smaller icons for foraging and larger for synthesis (Fig. 1). Visual patterns in the sequence diagrams show periods of foraging where users focus primarily on searching and reading. While other time periods show intense synthesis behavior with sequences of group manipulation interactions occurring. This is consistent with the sensemaking literature that indicates users performing foraging and synthesis in iteration. However, we found that users performing sensemaking on streaming data perform far more interactions to perform their task.

During user study observation on day 3, User2 expressed as the end of the allotted time approached that they were going to "get things organized for tomorrow". This was a notable statement from a user doing analysis on streaming data. The interaction sequence diagrams show that several of the streaming users performed more grouping-related interactions in the latter portion of their analysis time (see Fig. 5).

5.5 Group Spread

During the course of their analysis, users would group documents together as they saw fit. Document membership to groups would fluctuate over time as users

moved documents from group to group. Membership also changed for users in the streaming scenario when a new increment of documents was added to the visualization. We characterized groups and their change by computing group spread given document membership and our vector space topic model.

We compute a metric for "group spread" to analyze the consistency and cohesiveness of content within a group over time. Group spread is measured by computing the L1 normalized standard deviation from the group member's document vectors. This characterizes whether a group of documents is highly cohesive or more diffuse in the context of our vector space model. A group with more spread suggests member documents are not topically similar and therefore more distant from one another. This measure of group spread can be analyzed to compare groups or look at trend over time. Furthermore, spread can be averaged across all groups for a user then be used to compare users.

We hypothesized that low value groups with labels such as "Junk" or "Trash" would exhibit higher relative measures of spread. For User06, a plot of group spread over time showed that the group labeled "Junk" was created about midway through analysis. Spread for the "Junk" group increased over the remainder of the analysis time to eventually become the most diffuse group which supports our hypothesis (see Fig. 6).

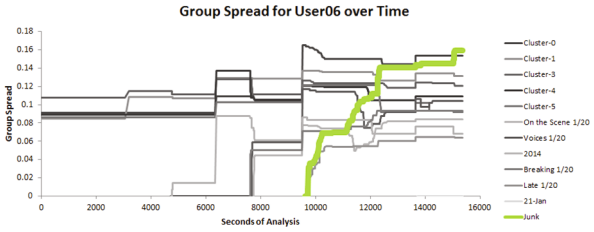

Fig. 6. Group spread for User06's groups over the course of their analysis shows the "Junk" group becoming increasingly diffuse.

An interesting pattern across users is observed when mean group spread is plotted over time for the streaming users. Not surprisingly, when an increment of new documents is introduced there is a jump in mean group spread. Then generally as the user manages group document membership in support of their analysis, mean group spread decreases. Groups become on average more cohesive as analysis progresses (see Fig. 7). This may indicate that the organization of information became more consistent and structured over time. In context of sensemaking, this echoes the iterative progression of internalizing and understanding information. As we found, streaming data has a way of disrupting that process, yet users adjusted over time to compensate for this.

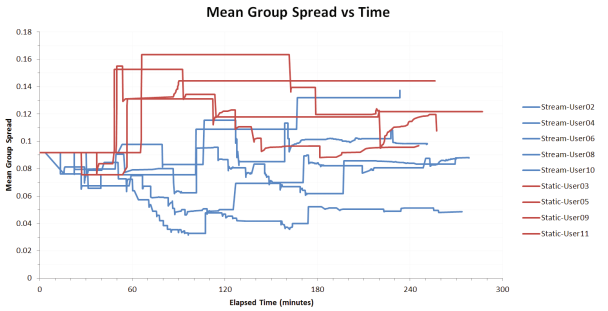

Fig. 7. Mean document group spread over time for users in the streaming versus static data condition. New data was introduced at roughly times: 50, 110, 170, 220.

6 Implications for Design

Users repeated several activities while performing analysis under steaming data conditions. While the Streaming Canvas successfully supported some of these analysis activities, there is certainly room for improvement in both the visualization and interaction design to fully support sensemaking for streaming data. We observed users progressively iterating through three phases. These include (1) thinking about current data, (2) organizing information in anticipation of more data, and (3) integrating new data into their thinking. These observed analysis activities have implication for design of future streaming visual analytics tools.

When users were given a collection of documents, they would spend time consuming the information to make sense of it. Users performed a variety of interactions to explore documents and externalize their thinking. Reading and searching were the most common interactions. Users would take notes in an accompanying Word document as they discovered key information sometimes copying and pasting snippets of text. Users would also bookmark important documents or annotate specific words, phrases or paragraphs. While these features were used as thinking aides, additions should be considered. Users requested the ability to resize the icon representing a document to convey importance, add notes directly to the Canvas, and draw connecting lines between documents. These suggest the need for features which allow a users to further integrate their knowledge into the visual metaphor, beyond the forming of groups and editing of group membership spatially.

As the analytic processes of the users progressed, they started organizing documents by forming groups and appling meaningful labels to create some higher-level structure. Users formed groups which ranged greatly in precision of definition. Some groups had precise definition such as all documents from a news source, a date range, or contained a specific keyword or name. Other groups were more informal where they revolved around a theme such as “POK violence” or “environmental effects”. In anticipation of receiving new documents, users refined document membership within groups and spatially arranged groups.

They made use of the interactive learning aspects of changing group membership of documents to prepare the analytic model in the system for the arrival of new data, as well as to formalize their thoughts about the current increment of data prior to the arrival of new information. This can be seen by the amount of cluster spread measured and shown in Fig. 7, where the mean spread of a cluster centroid across the features of the data decreased leading up to the arrival of new data (Fig. 8).

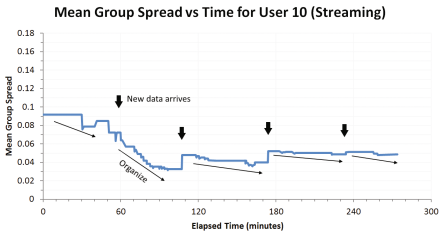

Fig. 8. The behavior of User 10 (from the streaming condition) shows interactions which decreased mean group spread over time before new documents arrive.

With the arrival of new documents, the Streaming Canvas mapped these documents into the user's existing groups which represent an externalization of their mental models. Our technical approach used a nearest neighbor measure to decide group assignments for new documents. While this worked in some cases such a topic-centric groups, document assignment did not always match the user's expectations. For more precisely defined groups such as source or date-centric groups, new documents would "contaminate" these groups as they did not match a user's specific and under-specified meaning for the cluster. The user would then spend time cleaning up these groups and moving documents elsewhere. The system and user would have benefited from support for optionally defining groups with precise queries when the users reached a state of formalism about the meaning of a cluster where they could directly specify it.

7 Discussion

The ways by which people use space as a means for organizing information has been widely studied for situations where the data (or physical objects) do not change [2,9,17,20,29]. From these studies, we learn that people create spatial constructs and fluid spatial arrangements as an inherent part of their process. For example, some of the groupings reflect process-specific artifacts, such as "todo lists", incremental knowledge structures such as groups of important documents, and in the physical example the methods by which a mechanic organizes the parts on the shop floor gives him or her spatial cues to remember how to re-assemble the components. For each of these, the persistence of the information (both in the

spatial location, and as the complete set of the items) provided people the cues to recall aspects of their process or knowledge associated with that information.

However, in this study we observed how data that is changing over time impacts this ability for people to offload parts of their process and working memory into spatial constructs. Our results indicate important distinctions in terms of the canvas usage. Primarily, one of the differences was found in the activities that people perform to "prepare" for the new data. For example, we found users spending more time and effort to organize and label groups so that they would be better able to recall their current state of the investigation when returning the next day to their spatial workspace with additional information.

Our current design decision in the Streaming Canvas was to place the new documents in the labeled group that is most related (based on the term weighting). However, other design alternatives exist. For example, one might choose to create an "inbox-like" view that shows new data, and has the user decide where to place the information. This could be potentially overwhelming when too much new data arrives at any given time increment.

We contend that designers and developers of visual analytic systems for streaming data should take into consideration additional affordances for people to do this "preparation" for new data. The Streaming Canvas allowed for user-defined spatial locations of groups and group labels. This allows people an implicit way to blend the new data with the current investigation state. However, in future iterations of such tools, visual analytic researchers may want to consider other, more explicit techniques for people to inform the system of where they want the new data to visually appear, which data to ignore, and how to prioritize which sub-sets of the new data he or she should read first.

8 Conclusion

Visual data analysis is an effective approach for giving people a greater understanding of their data through techniques such as data visualization and visual analytics. By offloading complex cognitive tasks in part to peoples' perceptual systems, visualizations enable people to think (and interact with) their data. This process is often referred to as sensemaking.

However, much of the work on understanding this cognitive task, and the development of visual analytic systems, has been focused on static datasets. While iteration is assumed and depicted in many popular sensemaking models, the fundamental assumption and study design of much prior research is that all the data to analyze is present at the beginning of a study, and that this set of data does now change. The study presented in this paper seeks to understand how incorporating the assumption that data will update over time impacts sensemaking.

We built a prototype streaming text visual analytics tool, called Streaming Canvas, to test this effect. We compared streaming and static conditions of people analyzing a dataset intended to simulate a sensemaking task. Our results indicate that people in the streaming data condition are more explicit about

tracking their analytic process than people analyzing static data. Streaming data analysts "prepared" their workspace for the arrival of new data, which required them to be more explicit about the status of their investigation. We believe that this, along with other findings, reveals important design guidelines for future streaming data visual analytic tools.

Acknowledgments. The research described in this paper is part of the Analysis In Motion Initiative and the Signature Discovery Initiative at Pacific Northwest National Laboratory. It was conducted under the Laboratory Directed Research and Development Program at PNNL, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy.

References

1. Alsakran, J., Chen, Y., Zhao, Y., Yang, J., Luo, D.: STREAMIT: dynamic visualization and interactive exploration of text streams (2011)
2. Andrews, C., Endert, A., North, C.: Space to think: large high-resolution displays for sensemaking. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2010), pp. 55–64 (2010)
3. Andrews, C., North, C.: Analyst's workspace: an embodied sensemaking environment for large, high-resolution displays. In: 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 123–131 (2012)
4. Bradel, L., North, C., House, L.: Multi-model semantic interaction for text analytics. In: Proceedings of IEEE Conference on Visual Analytics Science and Technology (VAST 2014), pp. 163–172 (2014)
5. Endert, A.: Semantic interaction for visual analytics: inferring analytical reasoning for model steering. Synth. Lect. Vis. **4**(2), 1–99 (2016)
6. Endert, A., Chang, R., North, C., Zhou, M.: Semantic interaction: coupling cognition and computation through usable interactive analytics. IEEE Comput. Graph. Appl. **35**(4), 94–99 (2015)
7. Endert, A., Fiaux, P., North, C.: Semantic interaction for sensemaking: inferring analytical reasoning for model steering. In: IEEE Conference on Visual Analytics Science and Technology (2012)
8. Endert, A., Fiaux, P., North, C.: Semantic interaction for visual text analytics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2012), pp. 473–482 (2012)
9. Endert, A., Fox, S., Maiti, D., Leman, S.C., North, C.: The Semantics of Clustering: Analysis of User-Generated Spatializations of Text Documents (2012)
10. Endert, A., Hossain, M.S., Ramakrishnan, N., North, C., Fiaux, P., Andrews, C.: The human is the loop: new directions for visual analytics. J. Intell. Inf. Syst. 1–25 (2014)
11. Endert, A., Pike, W.A., Cook, K.: From streaming data to streaming insights: the impact of data velocities on mental models (2012)
12. Fischer, F., Mansmann, F., Keim, D.A.: Real-time visual analytics for event data streams. In: Proceedings of the 27th Annual ACM Symposium on Applied Computing, SAC 2012, pp. 801–806. ACM, New York (2012)
13. Jolaoso, S., Burtner, R., Endert, A.: Toward a deeper understanding of data analysis, sensemaking, and signature discovery. In: Abascal, J., Barbosa, S., Fetter, M., Gross, T., Palanque, P., Winckler, M. (eds.) INTERACT 2015. LNCS, vol. 9297, pp. 463–478. Springer, Cham (2015). doi:10.1007/978-3-319-22668-2_36

14. Kamada, T., Kawai, S.: An algorithm for drawing general undirected graphs. Inf. Process. Lett. **31**(1), 7–15 (1989)
15. Kang, Y.-A., Stasko, J.: Characterizing the intelligence analysis process: informing visual analytics design through a longitudinal field study (2011)
16. Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., Melançon, G.: Visual analytics: definition, process, and challenges. In: Kerren, A., Stasko, J.T., Fekete, J.-D., North, C. (eds.) Information Visualization. LNCS, vol. 4950, pp. 154–175. Springer, Heidelberg (2008). doi:10.1007/978-3-540-70956-5_7
17. Kirsh, D.: The intelligent use of space. Artif. Intell. **73**(1–2), 31–68 (1995)
18. Klein, G., Moon, B., Hoffman, R.: Making sense of sensemaking 2: a macrocognitive model. IEEE Intell. Syst. **21**(5), 88–92 (2006)
19. Mansmann, F., Fischer, F., Keim, D.A.: Dynamic visual analytics–facing the real-time challenge. In: Dill, J., Earnshaw, R., asik, D., Vince, J., Wong, P.C. (eds.) Expanding the Frontiers of Visual Analytics and Visualization, pp. 69–80. Springer, Heidelberg (2012)
20. Marshall, C.C.: Spatial hypertext: designing for change. **38**, 88–97
21. Matejka, J., Grossman, T., Fitzmaurice, G.: Patina: dynamic heatmaps for visualizing application usage. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2013, pp. 3227–3236. ACM, New York (2013)
22. O'callaghan, L., Meyerson, A., Motwani, R., Mishra, N., Guha, S.: Streaming-data algorithms for high-quality clustering. In: ICDE, p. 0685. IEEE (2002)
23. Pelleg, D., Moore, A.W.: X-means: extending k-means with efficient estimation of the number of clusters. In: Proceedings of the Seventeenth International Conference on Machine Learning, ICML 2000, pp. 727–734. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2000)
24. Pirolli, P., Card, S.: The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. Proc. Int. Conf. Intell. Anal. **5**, 2–4 (2005)
25. Robertson, G., Ebert, D., Eick, S., Keim, D., Joy, K.: Scale and complexity in visual analytics. Inf. Vis. **8**(4), 247–253 (2009)
26. Rohrdantz, C., Oelke, D., Krstajic, M., Fischer, F.: Real-time visualization of streaming text data: tasks and challenges (2011)
27. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic Keyword Extraction from Individual Documents. Wiley, Hoboken (2010). pp. 1–20
28. Russell, D.M., Slaney, M., Qu, Y., Houston, M.: Being literate with large document collections: observational studies and cost structure tradeoffs, p. 55. IEEE Computer Society (2006). 1109739
29. Shipman, F.M., Marshall, C.C.: Formality considered harmful: experiences, emerging themes, and directions on the use of formal representations in interactive systems. Comput. Support. Coop. Work (CSCW) **8**(4), 333–352 (1999)
30. Skupin, A.: A cartographic approach to visualizing conference abstracts. IEEE Comput. Graph. Appl. **22**, 50–58 (2002)
31. Stasko, J., Goerg, C., Liu, Z.: Jigsaw: supporting investigative analysis through interactive visualization. Inf. Vis. **7**(118–132), 2 (2008)
32. Stolper, C.D., Perer, A., Gotz, D.: Progressive visual analytics: user-driven visual exploration of in-progress analytics. IEEE Trans. Vis. Comput. Graph. **20**(12), 1653–1662 (2014)
33. Thomas, J.J., Cook, K.A.: Illuminating the path: the research and development agenda for visual analytics. IEEE Computer Society Press (2005)

34. Whiting, M., Cook, K., Grinstein, G., Liggett, K., Cooper, M., Fallon, J., Morin, M.: Vast challenge 2014: the kronos incident. In: 2014 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 295–300, October 2014
35. Wise, J.A., Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., Crow, V.: Visualizing the non-visual: spatial analysis and interaction with information for text documents. pp. 442–450. Morgan Kaufmann Publishers Inc, Burlington (1999). 300791
36. Wright, W., Schroh, D., Proulx, P., Skaburskis, A., Cort, B.: The sandbox for analysis: concepts and methods, pp. 801–810. ACM (2006)
37. Zhang, P., Soergel, D., Klavans, J.L., Oard, D.W.: Extending sense-making models with ideas from cognition and learning theories. Proc. Am. Soc. Inf. Sci. Technol. **45**(1), 23 (2008)