# Results and Challenges in Visualizing Analytic Provenance of Text Analysis Tasks Using Interaction Logs

Rhema Linder* Alyssa M. Peña† Sampath Jayarathna‡ Eric D. Ragan§

Texas A&M University

**ABSTRACT**

After data analysis, recalling and communicating the steps and rationale followed during the analysis can be difficult. This paper explores the use of interaction logs to generate summaries of an analyst's interest based on interactions with specific data items in a text analysis scenario. Our approach uses data-interaction events as a proxy for user interest in and experience of information. Logging can produce verbose logs that detail all available readable content, so the discussed approach uses topic modeling (LDA) over different time segments to summarize the verbose information and generate visualizations of the history of user interest. Our preliminary results motivate a discussion on potential benefits and challenges of using interaction data to generate provenance visualizations for text analytics.

**Index Terms:** H.5.2. [Information Interfaces and Presentation (e.g., HCI)—User Interfaces]: Graphical user interfaces (GUI), Interaction styles (e.g., commands, menus, forms, direct manipulation)

## 1 INTRODUCTION

Complex and open-ended data analysis tasks require exploration of data over extended periods of time. A strong understanding of the data often involves the identification of connections among entities or patterns across data attributes. To assist with the inherently complex analyses, analysts often use a variety of visualization and analytics tools to facilitate the process. However, in addition to understanding the data itself, real-world analysis scenarios also require understanding of the analysis *processes* used in the investigation. For example, after performing an intelligence analysis task, a team of analysts must explain how they arrived at a conclusion about a terrorist attack. This would require citing sources and explaining the evidence supporting their hypotheses.

Due to the importance of understanding the history of the analysis process, many visualization and data analysis tools aim to capture *analytic provenance*, which refers to the history of steps taken and changes made throughout the duration of the analysis [9, 14, 20, 21]. Numerous types of provenance information (e.g., the history of data, visualizations, interactions, insights, or rationale) are considered to be important for visual analysis [21].

Interaction logs can be highly effective for understanding the history of data analysis [7]. However, in order for practical use of interaction data to understand analytic provenance, a clear and efficient means of interpreting that information is needed. In this paper, we describe methods that use interaction data from text-analysis activities as proxy for thought processes. We use interactions as means of approximating importance and implicit interest in content, and we apply topic modeling [4] to summarize the information that has been encountered and interacted with.

We show example visualizations generated from the interactions of a proof-of-concept study where we recorded logs data form an open-ended text analytics task. The results reveal research opportunities for finding interactions that best represent user interest and analyzing history in meaningful time segments. Our preliminary results motivate a discussion on potential benefits, challenges, and research space for future provenance visualizations of text analytics processes.

## 2 RELATED WORK

The concept of analytic provenance includes a variety of types of information about the history of data analysis. Researchers have previously discussed interpretations, definitions, and potential uses for use of provenance information for the purposes of visualization and data analysis (e.g., [9, 11, 20]). In a recent review of visualization literature, Ragan et al. [21] organized different perspectives and interpretations of *types* of provenance information and the *purposes* for its use.

Many previous projects support provenance visualization, and every project focuses on different types and purposes of provenance. Here, we describe a few of examples. For instance, the previously mentioned *VisTrails* uses a tree view to visually represent the sequence of actions and changes during a scientific workflow [3]. Other systems also adopt tree-style views to represent history (e.g., [13, 8]).

Some provenance systems aim to provide an overview of topic coverage during analysis. For example, Sarvghad and Tory [25] demonstrated the use of radial, treemap, and sequence-flow diagrams to help users understand data coverage from previous analyses. Text analytics systems infer and show relationships among documents and topics. *CzSaw* [16] and Jigsaw [26] are two examples that do so through various types of visualizations. Our approach differs in that we capture and represent the history of information encountered in the analysis process, not a complete assessment of data coverage. Prior research considers the use of additional annotation interaction to help clarify the process with user-provided notes and input (e.g., [9, 13]), but we seek an approach that does not require additional input from users.

Researchers have shown how processing and visualizing interaction logs can aid both researchers in inferring strategies and analysts in recalling insight. For example, Gotz and Zhou [11] described how the use of common actions could be used to infer the history of meaningful behavior and rationale that lead to insights during analysis, and Dou et al. [7] studied the feasibility and effectiveness of interpreting user interaction logs to understand an analyst's rationale. Lipford et al. [17] found evidence that the interaction visualization can improve recall of certain insights and rationale from the analysis. Also looking at a type of visualization created through interaction alone, a study by Ragan, Goodall, and Tung [23] found that the visual state of the workspace at the end of an analysis was enough to significantly improve memory of the process. Taking a different approach, Brown et al. [5] demonstrated how analysis

*e-mail: rhema@tamu.edu

†e-mail: mupena17@tamu.edu

‡e-mail: sampath@tamu.edu

§e-mail: eragan@tamu.edu

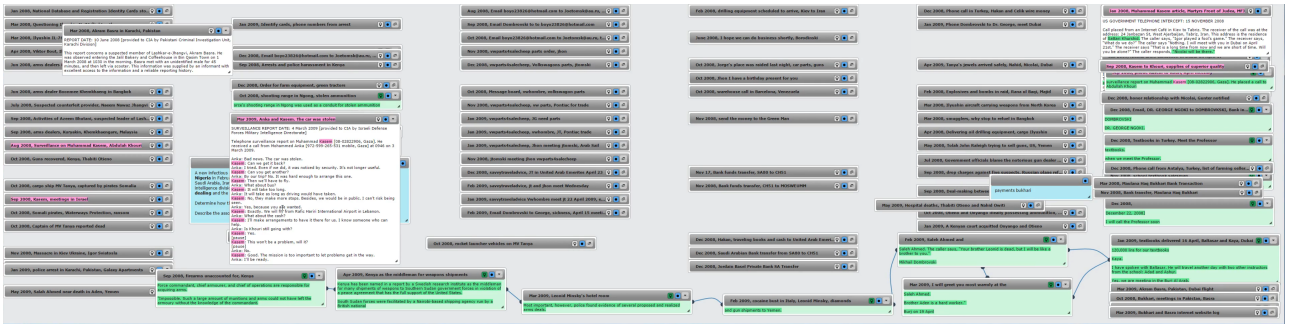

Figure 1: A screenshot of the text exploration tool that shows search results (pink highlights) along with a string of user-highlighted text (dark green highlights), reduced-to-highlight boxes (light green), and documents connected with linking lines. The participant arranged a chain of documents along the bottom and right of the workspace.


Figure 2: A close up of top right of the larger screenshot of the document explorer tool used for the data collection study.

of interaction data was able to determine information about users' strategies as well learn about the users themselves. Like our work, these prior projects took advantage of normal interactions without requiring additional input.

Prior projects have shown strong correlations among user interest and implicit indicators based on interaction [1]. Reading, organizing, and spending time on a document correlated to later ratings of relevance. By weighting the information, Bae et al. created a system for aiding text analysis tasks by automatically annotating multiple user interests [2].

Our approach uses interaction data to summarize analysis themes over time, focusing on textual analysis. For textual data, topic models have drawn interest and been utilized in a wide range of research including the humanities and social sciences [19], large scale social media studies [15], and analyzing political speeches [10]. Topic modeling is an appealing method for text analysis because it organizes words into coherent themes. We investigate topic modeling (Latent Dirichlet Analysis [4]) as a means to make sense of analyst encountered information over time.

## 3 Method

### 3.1 Collecting Provenance Data

To design and test our method for generating provenance summaries, we conducted a study where participants performed a text analysis task (while providing think-aloud verbal updates), and we recorded interaction logs. For the analysis scenario, we selected a task with sufficient complexity and scope to allow the exploration of various topics and hypotheses. To this end, we chose a text analysis scenario from the IEEE VAST 2010 Challenge Mini Challenge #1 [12].

To analyze the data, participants used a document exploration application (as in [22] and [23]) where text documents could be viewed, searched, and spatially manipulated to support organization. In this application, documents were placed in a 2D space and could be manipulated, similar to prior tools [26, 27]. The exploration tool (Figures 2 and 1) logged various actions performed by each user (see Table 1).

**Interactions Captured**

| | |
|---|---|
| *Search* | Search the data set for a word or phrase |
| *Reduce to Highlight* | Reduce the visible text in a document to only the highlighted content |
| *Highlight Text* | Highlight text in a document |
| *Connect Document* | Create a new connection line linking two document or note windows |
| *Collapse Document* | Minimize the document window to only show the its title bar |
| *Open Document* | Expand a collapsed document window to show its full text |
| *Move Document* | Drag a document window to a new location |
| *Mouse Enter* | The mouse position moves over a document window |

Table 1: Types of interactions logged by the document exploration tool.

We recruited six participants for the study, five males and one female. Ages ranged 18 to 30 years old, and all had low to moderate experience in data analysis or visualization. As the data was about weapons dealing, participants were asked to explore the documents to report on the connections and plans involving illegal weapons trade. We recorded log data for their actions and recorded video.

### 3.2 Automatic Provenance Summary Generation

We describe our preliminary methods for generating provenance visualizations automatically using interaction logs. As a case study, we used information and logs from the open-ended text analytics task described in the previous section.

Our method is summarized in five steps: (1) Capture interaction events during a period of analysis. (2) Generate a sequence of text by using the interaction events to establish the encountered text over time. (3) Process the text data using standard information retrieval techniques [18]. (4) Segment the corpus of texts by periods of time and generate topic models per segment. (5) Visualize topics over time to facilitate easier interpretation and pattern recognition.

The first step of our summary generation process is to *Capture Interaction Events* from user interaction with the analysis tool to

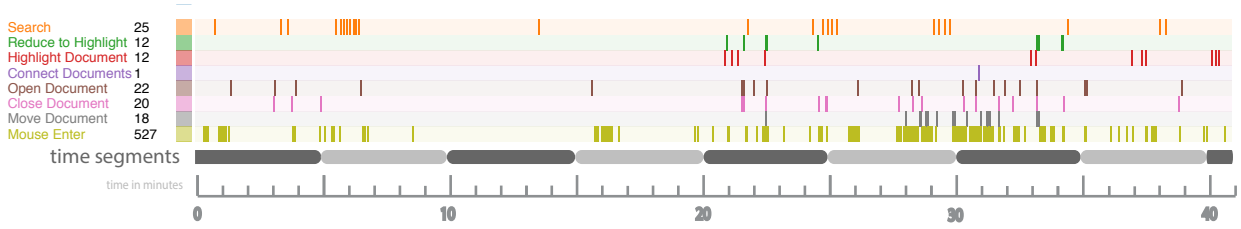


Figure 3: This visualization depicts the interaction history from a participants' document explorer study session. Analysis time is stretched across the horizontal axis. Each thin colored line represents a single action at a given time. The number on the left of each row shows the total number of actions. Beneath the timeline, we show how a five-minute breakdown produced the segments used in the provenance visualization.

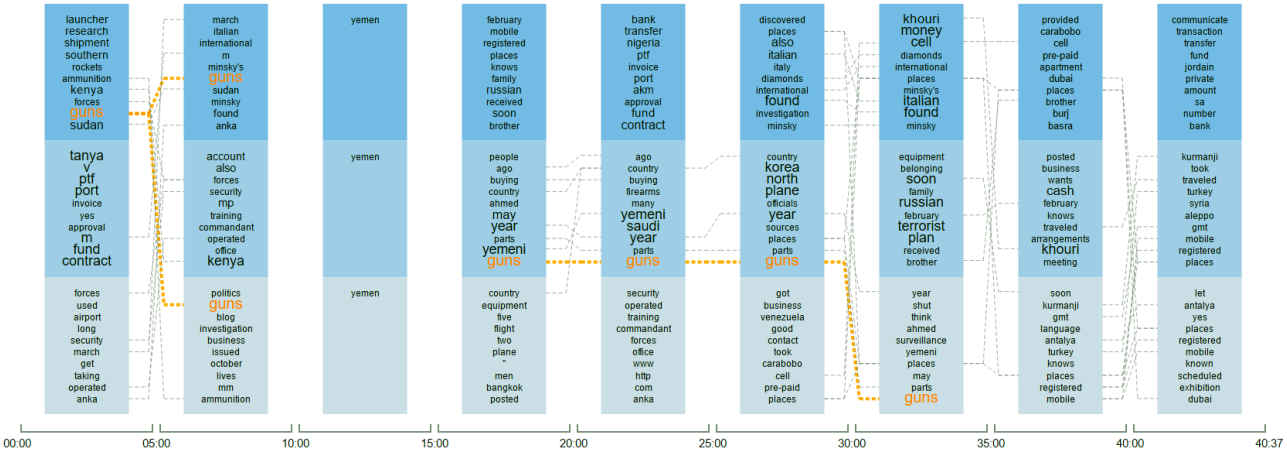


Figure 4: A parallel word cloud design shows the topic segments with the terms in each topic. In this figure, each blue column shows three topics in a time segment. The orange words have been moused-over in order to show all occurrences of the term across all segments. Note that the third column is missing interactions except a single search. This is consistent with our observations of the participant during the study.

create a history of their actions. This history in and of itself explains how the interface was used, but it provides nothing about the content and context of user actions. Ideally, it would be possible to capture and represent the analytic process without requiring additional input or description from analysts.

The second step is to *Generate a Sequence of Text*. Since most interactions involve interacting with a specific document, it was possible we could associate each interaction with a text. Thus, each interaction even can provide a sequence of text documents that the user considered during the analysis. We used these sequences of text as the basis for topic modeling to infer the main stages of analysis over time.

The third step is to *Process the Text* after capturing the interaction history to make it more suitable for topic modeling. We use standard information retrieval methods [18] for tokenization and removing stop words. Additionally, we create another list of stop words of common, less-useful terms from entries (e.g., *report*).

*Segmentation and Topic Modeling* occurs the data is processed. In the end, our goal is to create summaries of the stages and themes of analysis over time. Each record from interaction logs includes associated text, a type, and timestamp data. Simply printing these records would be too verbose for practical interpretation.

To simplify these records, we first use a *segmenter* to break the history down into discrete time segments the summarize with topic modeling. In the results presented here, we break down the 35–40 minute analysis session data into five-minute segments. Once segmented, we use a topic modeler (LDA) to generate a set of topics for each time period. Per the case study, we show three topics and found 15 iterations of LDA to be sufficient. The final results are serialized and saved to provide data for the last step: visualization.

The final step is *Visualization* (see Figure 4). We used *parallel word clouds* [6, 10] as the base representation to visualize the topics over time. In our visualization, the topics are shown in lists of words embedded in blue columns. Each column represents a time segment of user interaction history. Columns show their beginning and end times and their topic model summaries. Within each column, the three topics are sorted based on coherence, as calculated in the Gensim framework [24] and are distinguished with different shades of blue. For a single topic within a column, we show a vertical list of terms ordered by the probabilities of in the model. To highlight important words within topics, individual terms are scaled based on TF–IDF (term frequencyinverse document frequency) scores to decrease the importance of common words words and help representative words remain prominent.

In addition, the design includes linking lines that connect any word to the same word in adjacent time segments. When the user brushes or hovers the cursor over a word, it changes color (orange) and increases for all instances of the word. Linking lines will also be highlighted in orange for increased salience. A "brushed" word slowly transitions back to its default style when the cursor is moved off of the word, causing the word's highlight to persist briefly. This helps a viewer reveal patterns of clusters of repeated words.

## 4 Discussion and Conclusion

### 4.1 Preliminary Results

Our preliminary results are promising. The automatically generated snapshots of analyst interest over time, by our observations, capture meaningful topics. The visualization presents themes based on interactions where users open, looked at, and manipulated content. These topics are connected together to show the flow of topics over time, and for our case study and test data, the effectiveness is clear. While some segment's topic summaries make more sense than other, they are on the whole meaningful.

Our observations are that different participants used different strategies: breadth-first, depth-first, and cyclical processes where a particular theme is repeatedly revisited. We found that exploring with brushing and linking can reveal participants general strategies. For instance, one participant said, "*I feel like I'm doing a depth first search"*, and the summaries showed this with many connections from one time segment to the next, rather than sudden changes in topics and interest.

### 4.2 Implicit Interest from Different Interaction Types

This research raises the importance of discovering the most salient interaction types during analysis. For example, looking at a timeline of interactions (Figure 3), note that some types of interactions are performed more than others. Including and excluding data based on the type of interaction may impact the effectiveness of summaries using the discussed approach. For example, *Mouse Over* events occur many times, and it is likely that participants who used *Mouse Over* were interested in the content of the documents they interacted with. However, an analyst may accidentally *Mouse Over* a document they have no interest in when moving their mouse across the screen. At the same time, we observed that many *Mouse Over* events are meaningful. Participants would move the mouse back and forth between documents, weighing their information and planning what to look for next. On the other hand, *Open* events occur far less frequently, but might be more meaningful than *Mouse Over* events. Performing the *Open Document* action represented committed actions—usually occurring only when a participant thought they were likely to find important material. We find many interaction types could be thought of as having little to much meaning and occur at different frequencies. It is likely that studies and observations are needed for determining which events are meaningful and frequent enough.

### 4.3 Temporal Segmentation Schemes

Another research opportunity is to create methods for segmenting interaction events over time. Our example (as seen in Figures 3 and 4) shows the topic history with five-minute segments. While this is straightforward to implement and understand, it comes with significant drawbacks. In some cases, adjacent columns in the visualization were too similar, adding limited additional information. For example, one participant stopped interacting with the system for about six minutes in the middle of the analysis, opting to read and plan but not move the mouse or click on documents. Using static five-minute segments, this created a gap where only a single search term was captured (see Figure 4, third column from the left).

Better segmentation schemes are needed. After looking at the timelines from different participants, we noticed that searches usually occur in bursts. A single burst of search interaction usually meant that they tried a few queries until the results seemed promising. With these observations in mind, we see an opportunity to create different segmentation schemes that take advantage of a combination of (1) the degree of change in content and (2) implicit time boundaries based on interaction types (e.g., *Open* or *Search*).

### 4.4 Text-Associated Interactions as a Proxy to Thought Processes

This work has begun to explore automated methods for summarizing encountered information during text analysis. While a tremendous amount of text can be generated from interaction logs, reviewing all logs would be impractical. Instead, we believe a summarization approach such as ours can serve as a proxy for understanding analytic provenance and analysts' interests.

While an ideal provenance summary of an analysis would include the internal thought of analysts, this is technically impossible to capture completely without disrupting the analysis. However, associating interaction events with text can provide a view of repeatedly encountered information, which influences the thoughts of analysts. Also, for the purpose of task resumption, we expect that showing analysts provenance visualizations will help them remember their rationals and insights. For researchers looking to understanding others' pathways to insight, we expect that these visualizations will provide views that help them see strategies, such as bread-first or depth-first approaches.

### 4.5 Conclusion

We investigated how topic modeling can be used to automatically generate provenance visualizations from interaction logs. To explore the approach, we collected data from participants who performed a text analytics task. We recorded their interactions and created visual summaries for preliminary evaluation of our method. While far from perfect, the method created surprisingly representative text summaries. We found the visualizations were useful for understanding user interest over time, recurring interest of topics, and aspects of analyst's strategies. The results could be improved by addressing the research opportunities of (1) finding which interaction events are most representative of user interest and (2) developing interaction timeline segmentation schemes. Future work should evaluate the efficacy of using this and similar methods for understanding thought processes in text analysis tasks.


## Acknowledgements

This material is based on work supported by NSF 1565725.


## References


[1] R. Badi, S. Bae, J. M. Moore, K. Meintanis, A. Zacchi, H. Hsieh, F. Shipman, and C. C. Marshall. Recognizing user interest and document value from reading and organizing activities in document triage. In *Proceedings of the 11th International Conference on Intelligent User Interfaces*, IUI '06, pages 218–225, New York, NY, USA, 2006. ACM. 2

[2] S. Bae, D. Kim, K. Meintanis, J. M. Moore, A. Zacchi, F. Shipman, H. Hsieh, and C. C. Marshall. Supporting document triage via annotation-based multi-application visualizations. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, pages 177–186, New York, NY, USA, 2010. ACM. 2

[3] L. Bavoil, S. P. Callahan, P. J. Crossno, J. Freire, C. E. Scheidegger, C. T. Silva, and H. T. Vo. Vistrails: Enabling interactive multiple-view visualizations. In *IEEE Conference on Visualization*, pages 135–142, 2005. 2

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003. 1, 2

[5] E. T. Brown, A. Ottley, H. Zhao, Q. Lin, R. Souvenir, A. Endert, and R. Chang. Finding waldo: Learning about users from their interactions. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):1663–1672, 2014. 2

[6] C. Collins, F. B. Viegas, and M. Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 91–98. IEEE, 2009. 3.2

[7] W. Dou, D. H. Jeong, F. Stukes, W. Ribarsky, H. R. Lipford, and R. Chang. Recovering reasoning processes from user interactions. *IEEE Computer Graphics and Applications*, (3):52–61, 2009. 1, 2

[8] C. Dunne, N. Henry Riche, B. Lee, R. Metoyer, and G. Robertson. Graphtrail: Analyzing large multivariate, heterogeneous networks while supporting exploration history. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1663–1672. ACM, 2012. 2

[9] J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for computational tasks: A survey. *Computing in Science & Engineering*, 10(3):11–21, 2008. 1, 2, 2

[10] S. Gad, W. Javed, S. Ghani, N. Elmqvist, T. Ewing, K. N. Hampton, and N. Ramakrishnan. Themedelta: dynamic segmentations over temporal topic models. *Visualization and Computer Graphics, IEEE Transactions on*, 21(5):672–685, 2015. 2, 3.2

[11] D. Gotz and M. X. Zhou. Characterizing users' visual analytic activity for insight provenance. *Information Visualization*, 8(1):42–55, 2009. 2, 2

[12] G. Grinstein, S. Konecni, J. Scholtz, M. Whiting, and C. Plaisant. VAST 2010 challenge: arms dealings and pandemics. In *IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 263–264. IEEE, 2010. 3.1

[13] D. P. Groth and K. Streefkerk. Provenance and annotation for visual exploration systems. *Visualization and Computer Graphics, IEEE Transactions on*, 12(6):1500–1510, 2006. 2

[14] Z. Hensley, J. Sanyal, and J. New. Provenance in sensor data management. *Communications of the ACM*, 57(2):55–62, 2014. 1

[15] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88. ACM, 2010. 2

[16] N. Kadivar, V. Chen, D. Dunsmuir, E. Lee, C. Qian, J. Dill, C. Shaw, and R. Woodbury. Capturing and supporting the analysis process. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 131–138. IEEE, 2009. 2

[17] H. R. Lipford, F. Stukes, W. Dou, M. E. Hawkins, and R. Chang. Helping users recall their reasoning process. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 187–194. IEEE, 2010. 2

[18] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge University Press, 2008. 3.2, 3.2

[19] J. W. Mohr and P. Bogdanov. Topic models: What they are and why they matter. *Poetics*, 41(6):545–569, 2013. 2

[20] C. North, R. Chang, A. Endert, W. Dou, R. May, B. Pike, and G. Fink. Analytic provenance: process+ interaction+ insight. In *ACM CHI Extended Abstracts on Human Factors in Computing Systems*, pages 33–36, 2011. 1, 2

[21] E. D. Ragan, A. Endert, J. Sanyal, and J. Chen. Characterizing provenance in visualization and data analysis: An organizational framework of provenance types and purposes. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):31–40, 2016. 1, 2

[22] E. D. Ragan and J. R. Goodall. Evaluation methodology for comparing memory and communication of analytic processes in visual analytics. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, pages 27–34. ACM, 2014. 3.1

[23] E. D. Ragan, J. R. Goodall, and A. Tung. Evaluating how level of detail of visual history affects process memory. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2711–2720. ACM, 2015. 2, 3.1

[24] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. 3.2

[25] A. Sarvghad and M. Tory. Exploiting analysis history to support collaborative data analysis. In *Proceedings of the 41st Graphics Interface Conference*, pages 123–130. Canadian Information Processing Society, 2015. 2

[26] J. Stasko, C. Görg, and Z. Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization*, 7(2):118–132, 2008. 2, 3.1

[27] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Information Visualization, 1995. Proceedings.*, pages 51–58. IEEE, 1995. 3.1