From Data Dump to Digestible Chunks: Automated Segmentation and Summarization of Provenance Logs for Communication
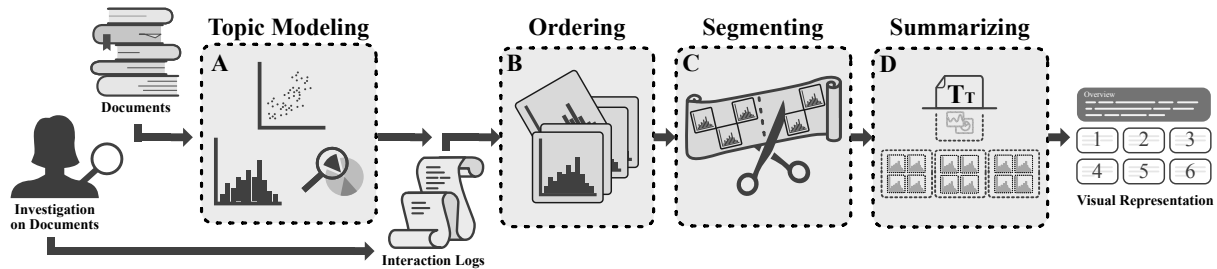
Anonymous Authors

Fig. 1: An abstraction of our segmentation and summarization methodology for interaction histories in analysis scenarios. Essentially, our technique uses A) topic models of the underlying dataset and B) the order of interactions with these documents to identify C) key breakpoints to segment an investigation along. D) We then summarize each segment as textual cards to efficiently tell the story.

Abstract—Communicating one's sensemaking during a complex analysis session to explain thought processes is hard, yet the vast majority of intelligence occurs in collaborative settings. Team members require a deeper understanding of the work being completed by their peers and subordinates, but little research has fully articulated best practices for analytic provenance consumers. This work proposes an automatic summarization technique that separates an analysis session and summarizes interaction provenance as textual blurbs to allow for meta-analysis of work done. Focusing on the domain of intelligence analysis, we demonstrate our segmentation technique using five datasets, including both publicly available and classified interaction logs. We shared our demonstration with a notoriously inaccessible population of expert reviewers with experience as United States Department of Defense analysts. Our findings indicate that the proposed pipeline effectively generates cards that display key events from interaction logs, facilitating the sharing of analysis progress. Yet, we also hear that there is a need for more prominent justifications and pattern elicitation controls to communicate analysis summaries more effectively. The expert review highlights the potential of automated approaches in addressing the challenges of provenance information in complex domains. We emphasize the need for further research into provenance communication in other domains. A free copy of this paper and all supplemental materials are available at OSF link to be provided after review.

Index Terms—Analytic Provenance, Workflow Design, Communication/Presentation, Storytelling

✦

1 INTRODUCTION

Exploratory data analysis is hard. Analytics and software support are commonly used to facilitate data exploration, but practical cases of analysis involve incomplete or messy data that still rely on human decision-making for investigation. In real-world contexts, data analysts do not conduct work in isolation. Understanding of data is sought for a reason, and findings need to be communicated to others—be it superiors, stakeholders, or peers. In some cases, a new analyst might take over for another analyst, or multiple analysts might work together to share insights. However, people think differently and adopt different strategies in open-ended analyses, and it would be unlikely for any two analysts to explore the same dataset in exactly the same way. Consequently, understanding the data analysis work of others requires more than only the final "findings" for collaborators to have confidence and trust in each others' process.

Effectively capturing and sharing knowledge of the analysis process (sometimes called *analytic provenance*) is a known challenge. Self-documentation and reporting are onerous, and they distract from the primary job of analysis. Fortunately, ample research has demonstrated the utility of automatic software logging to capture the history of the data analysis workflow. Software logs can record the sequence of user

interactions as well as the history of data accessed over time, and prior work has found that review of such provenance records substantially improves understanding of prior analyses [45]. This approach can capture the analysis history automatically while eliminating (or greatly reducing) the need for manual human documentation, and the history can be reviewed by others to meet the needs of collaborative settings or reporting. The downside, however, is the large scope of software logs and the large quantity of captured information. Major challenges remain for how to transform analytic provenance information into an easily-understandable format. While researchers have designed many visualization solutions to facilitate deep review of analytic provenance [13, 66], our work addresses the practical need to summarize the workflow for communication.

This paper presents a unique way of processing analytic provenance information with an emphasis on summarization for easier human interpretation. Interested in simplifying analysis sessions into more easily understandable chunks, our approach breaks down interaction sessions into segments, summarizes the associated data interests and interaction activity, and generates a set of cards with visually-annotated natural language summaries. Because different communities and analysis domains have different priorities or preferences for summarization, we focused our design on the intelligence analysis domain as a concrete basis for understanding user needs. Broadly speaking, intelligence analysis involves making sense of large amounts of different data types from various sources, including textual data, images, and network data. This paper presents our technique along with a case study in which we worked with experienced professional analysts to realistically recognize expectations and practical needs for summarizing analytic provenance.

• Anonymous Author emails.

2 RELATED WORK

Across many domains, exploratory data analysis does not start and end with a single person. Often there is a need to support collaborative work. Communicating findings and processes can be aided by provenance, or the story of how conclusions develop over time. In this section, we discuss some of the scenarios in which collaborative analysis is conducted, the tools used, and the ways others have communicated work complete.

2.1 Collaborative Visual Analysis

Understanding complex issues often requires deliberation and discussion with others to fully comprehend. Heer et al. [25] were an early example to recommend a set of features that make analysis more collaborative. Since then, there have been numerous examples that have incorporated support for team collaborations. For example, Dunne et al. [13] display all the events involved in analyzing archaeology data to more completely express investigation steps and the types of data being explored over time by multiple users. Expressing work complete has also been beneficial in the intelligence domain. Both Goyal and Fussell [20] and Zhao et al. [66] designed tools that essentially share group understanding as either bar charts or knowledge graphs to aid in communicating understanding. Often, in this domain, they rely on "coverage-based" techniques that show what kinds of data have received attention to more efficiently describe the investigation status. Yet, understanding the order of analysis can still be helpful for communication.

2.2 Analytic Provenance

Collaborative visual analytics often lacks support for sequences of events because it is difficult to derive meaning from all the steps in an efficient way [3]. Often, in these exploratory settings, the signal representing one's final result is crowded by a cacophony of smaller events and details that would help reproduce similar results but would be impractical for others to review. The sum of these smaller events describes analytic outcomes, behaviors, and strategies. One solution is to analyze the event data to extract meaningful patterns and make it easier to digest; we call this analytic provenance [39, 44, 60].

Analytic provenance can be represented in multiple ways. Often, the way provenance information is presented can dictate its use [44]. For example, provenance data can be used for personal retrospectives of work completed in a day or more holistic representations, referring to a meta-review of perhaps an entire department. Retrospectives are important for various contexts: including recommendations about past notes and contexts of behaviors [53] and even for refreshing user memories [46].

Essentially, Gotz and Zhou [19] recognized that provenance information has a hierarchy of semantic intent. Provenance at the lowest level (i.e., event-based provenance) is helpful for transparency and the reproduction of tasks. Still, provenance at the highest level (i.e., task-based provenance) is helpful for communicating task completion. Further experimental evidence was found by Block et al. [4], where the higher-level provenance (coverage-based provenance) was more helpful than a list of events (event-based provenance) when handing-off a collaborative analysis task.

Yet, questions remain regarding representing these details in different contexts and domains. Lemieux describes the variety of purposes served by provenance in different fields and representations it can take [31]. One recommendation is to separate and share semantically meaningful pieces of the sequence and use a visual analytics tool to identify meaningful patterns [30]. Yet these techniques are commonly focused on providing access to provenance data while still asking the user to complete the analysis. They expect users to identify interesting features and, ultimately, complete the meta-analysis of an analysis process. Yet, analysts only have so much time and need help parsing analysis history.

2.3 Visual Analytics for Narrative and Presentation

To make things digestible to new users or to better understand what has been done, some guidance and separation can help clarify the information. For example, Storyfacets [42], VisInReports [51], and InsideInsights [34] ask users to select meaningful pieces from an analysis session and aids in the presentation of these distinct parts. Many are based on the earliest CLUE model by Gratzl et al. [21], which recommends the need for provenance information to be both about analysis, formatting, and presenting information to aid in collaborative settings. They all recommend that analysis tools should help analysts tell the story of their work. Fundamentally, the goal of storytelling is to target and communicate to a less familiar audience.

Yet, there is a gap between the skills needed to be an expert in visual analytics and the skill set needed for effective storytelling [8]. Analysts are already burdened with completing the analysis, but some techniques exist to aid in the development of stories automatically. Bryan et al. [7] and Zhao et al. [67] both have recommended frameworks for considering how to tell a data story. Kouril et al. [28] has recently designed a technique that automatically generated miniature "documentaries" based on the findings and notes made by molecular biologists. Veras and Collins describe how the minimum description length principle can help make visualizations more semantically meaningful for hierarchical datasets [55]. A common theme across these works is the need to minimize the amount of content and identify the essence of the data or the analysis. What is not often considered are the downstream effects of these narratives for people within a specific work environment. We explore a technique for segmenting and simplifying analysis histories for communication within an organization. Still, there are outstanding questions about how to represent these summaries.

2.4 Text Summarization

Textual reports commonly communicate information within multiple professions, including nurses' notes [9, 12, 58], between rocket control shifts [43], and collaborative investigations [66]. Reading is a standard skill set and does not require complex instructions to help users navigate information. Written words are common across domains, with examples ranging from emails, informal text messaging, internal reporting, and writing for external publications. Natural language processing (NLP) techniques can now generate coherent sentences (though major problems still exist [26]), and we can ask these models to summarize the content. In the NLP domain, these summaries are often evaluated with automated scoring functions (e.g., BLEU, ROUGE, etc.), but these all consider the summarization to be on underlying text data [14].

In the realm of analytic provenance, the goal of summarization is not to shorten text corpora but to communicate interaction histories. Interaction histories are their own unique challenge because the conversion of interactions to textual summaries must first be translated into a common transition language. Some work has attempted to use narrative theories like Rhetorical Structure Theory to frame the use of W3C provenance data to generate narratives at various levels of detail [11]. Norambuena et al. [38] also describe how narratives can help express sensemaking processes. For this reason, our work recognizes existing solutions that can be adapted to preparing natural language summaries tailored to the needs of any specific domain. In the example implementation presented in this paper, we aim for simplified natural language summarization in order to focus on a small set of descriptive characteristics of analysis activity.

3 DESIGN OVERVIEW

In the literature about collaborative sensemaking and provenance supports, it is common to incorporate aspects of narrative visualizations and presentation support features as part of visualization tools. These techniques draw on the aspects of different narrative structures [50] or design an environment that integrates a presentation support feature into the analysis [21, 34, 42]. While considering the narrative structure of data presentation in a report is an important consideration, as seen in these examples, building these narratives still takes time and requires additional steps to call attention to key moments in an analysis session to prepare a presentation [21, 34, 37, 42, 51]. To help address these time concerns, there is value in exploring automated storytelling techniques. These considerations motivate the methodology described in this paper.

3.1 Design Rationale and Goals

When asked the colloquialism, "How does one eat an elephant?," we reply with "One bite at a time." With this metaphor in mind, we turn our attention to the natural human behavior to break large, complex, overwhelming problems down into manageable chunks to apply our mind's often serial processes. This is to say that humans perceive more, have better memory, and communicate events better when that information is segmented [65]. Psychological research has shown how the concretizing of events into smaller segments helps in understanding those events [63, 64]. According to neuroimaging studies, when observing events in a movie, actions taken by a subject that are perceived as anomalous are more likely to be considered an event boundary and described as a separate episode in one's memory [62]. This provides evidence to a theory of mind where human perceptual systems maintain a meta-narrative of an actor's goals and behaviors and naturally segment these unexpected events to assist in understanding and communication. Evidence of the benefits of segmentation can also be found in human-computer interaction research. They show how providing aggregations of interaction events, and dynamic binning of information can be beneficial when reviewing temporally relevant event sequences [57]. We also see design recommendations for explaining data using story backbones and layouts found in comic books to aid in communication [2]. Segmentation is a natural way that humans perceive episodic behaviors, so having a technique that can separate event sequences into appropriate chunks would be useful.

At the same time, there are also calls for meaningful summarization of interaction information. On the one hand, detailed representations of provenance information can aid in visual analytics when the goal is to describe steps transparently. To illustrate this point, consider, SenseMap [36], a browser extension that captures and displays pages visited as an expanding tree. This tool can capture an exhaustive list of page visits, but as more pages are visited, the tree continues to expand and can become overwhelming to review. These types of tools—which aim to capture and facilitate a review of all the behaviors in a session—are especially helpful for transparency purposes or personal reflection; however, to be useful for other users who are trying to understand the gist or overall flow of the analysis, the complete history is often overwhelming. Simplification is a priority for effective communication and integration into practical work styles.

Other methods propose a succinct overview of provenance information by showing what information was interacted with over time. This focus on what was covered can be helpful for explaining work done by researchers [32], showing what information has already been analysed [48], or encoding visit history in the visualization's default view [15]. Some techniques, like Zhao et al. [66], include an interactive timeline with a knowledge graph that shows what a team knows over time. Yet, when comparing provenance visualization techniques for hand-off communication, Block et al. [4] found that providing summaries can be more efficient because providing too much detail creates a new, secondary sensemaking task for the user to try to analyze the tasks of a prior user. While there are places for deeper, more holistic representations of the progression of analysis, we focus our method on representations that inherently simplify analysis workflows. Summarization is not only helpful for collaborators (i.e., people working toward the same goals), but also for managers who supervise multiple individuals. Managers are not expected to complete the same work as their team and are more concerned with work coordination and worker performance. Summarization helps provide information at a high level, thus making it human-understandable. Adjusting what is summarized at different levels of detail can be especially helpful for both managers and collaborators. Our goal, as we describe next, is to communicate an overview of the flow of analysis and progression of topics/items of interest in summarized chunks at different levels of detail to help reduce overwhelm and cognitive demands.

The five goals for our overview technique are to:

G1 Temporally segment analysis sessions into digestible chunks as episodes in the analysis narrative.

G2 Summarize interaction logs to communicate processes efficiently with an appropriate amount of meaningful semantic information.

G3 Support information hand-off such as for analyst-to-analyst and analyst-to-manager communications.

G4 Enable identification of different analyst behaviors and patterns, as would be desirable for collaboration, managerial review, and mentoring purposes.

G5 Mimic simple and familiar reporting structures commonly used in slideshows and text-based reports (examples include plain natural language, bulleted lists, and primitive data visualizations).

3.2 Data Properties

The analysis applications used in our design primarily focused on these situations where a user makes queries that return text documents and all the interactions associated with this review. The core features of our system rely on two key components: 1.) the corpus of documents being analyzed and 2.) the timestamped events of user interactions with their analysis tools. Often these document corpora can be massive. But they contain valuable context information for analysis sessions. They can be used to define relevant terms and themes associated with an analysis session that are not often captured in interaction histories. Linking the document set with interactions can help describe what areas of the text corpora receive attention, for how long, and how it was discovered. Interaction histories are time-ordered lists of events that contain traces of semantic information (e.g., search terms, document identifiers, highlighted text, etc.) but often are missing contextual information (e.g., the major topics associated with a document or the types of content returned from a search). It is not unusual for these interaction histories to register hundreds of interactions in an hour, and reviewing these without automated extraction techniques is unreasonably time intensive. Yet, most commercial software logs interactions to aid in debugging or audit performance. It is this common combination of documents and interactions that we focus our attention on generating an automated segmentation and summarization of events. Further information about the datasets used in our later demonstration is provided in Section 6.2.

4 VISUAL ANALYTICS PIPELINE

The contributions of this paper are essentially twofold. We recommend a methodology for segmenting and summarizing analysis sessions and demonstrate the applicability of our technique by collecting feedback from experts from the intelligence domain. In this section, we describe the specific components required for our demonstration and the visual interface. We re-iterate that the specifics of our approach are not strictly applicable to the intelligence domain but rather serve as a laboratory for understanding the impacts of segmentation and summarization for communication. We recommend further exploration into domain-specific designs.

4.1 Data Processing

At a high level, our technique uses documents and interactions on those documents to generate a series of segments and summarize an investigation. We want to segment analysis sessions into chunks automatically to make them more digestible (**G1**) and then summarize those segments to explain the story of someone's activity briefly and quickly (**G2**). To support information hand-off (**G3**) and the identification of different analysis styles (**G4**), we chose a methodology that would convert semantically poor interaction logs to something more familiar and digestible (**G5**): Textual Reports. At its heart, our method (As seen in Fig. 1) is essentially four phases. 1. vectorization of document content, 2. synchronization of document visits with the order from interaction histories, 3. segmenting the history based on document content interacted with over time, and finally, 4. summarization for each segment of identified time. At a high level, we describe this technique and its potential benefits for one community. Yet, every domain interested in this approach will have unique data challenges. Therefore, we demonstrate the fundamentals of the technique by implementing a basic rendition of our pipeline without fine-tuning or selecting the best components, as these will adapt to every scenario. Below we describe

Provenance Segment Visualization

Session Summary

They focused on **27** identified topics in this analysis session, exploring **45%** of the dataset. The topics that received the most attention were *nigeria*, *george*, **and** *dubai*. They started searching for *nigeria*, before transitioning to *weapon* and finally looking for *dubai*. In Segment *#1* they opened the most different documents. Segment *#1* was the longest period where they **opened more documents than usual**. They conducted **68** searches throughout their session, especially in segment *#1*.

#1	19:33 minutes

The user made **10 different** searches, including "**disease**", "**Nigeria**", and "**gun**". They explored **10 new** documents, which were all created in **August 2008**. The most referenced people in the viewed documents were **Minsky** and **Leonid Minsky's**. The most referenced locations in the viewed documents were **US** and **Kenya**. Of the **10** documents open, each document recieved attention for an average of **1:57** minutes.

#2	10:09 minutes

The user searched for "**Dr. George**". They explored **10 different** documents, **2** of which had **not previously been read**. Those documents were created between **August 2008** and **December 2008**. The most referenced people in the viewed documents were **Mikhail Dombrovski** and **George**. The most referenced locations in the viewed documents were **US** and **Russia**. Of the **10** documents open, each document recieved attention for an average of **1:01** minutes.

#3	13:27 minutes

The user made **19 different** searches, including "**weapon**", "**nigeria**", and "**russia**". They explored **11 different** documents, **10** of which had **not previously been read**. Those documents were created between **December 2007** and **March 2009**. The most referenced people in the viewed documents were **Minsky** and **Leonid Minsky's**. The most referenced locations in the viewed documents were **Yemen** and **Russia**. Of the **11** documents open, each document recieved attention for an average of **1:13** minutes.

#4	8:56 minutes

The user made **3 different** searches, including "**dubai**", "**textbook**", and "**basra**". They explored **7 different** documents, **6** of which had **not previously been read**. Those documents were created in **November 2008**. The most referenced people in the viewed documents were **Lashkar-e-Jhangvi** and **Karachi**. The most referenced locations in the viewed documents were **Antalya** and **Turkey**. Of the **7** documents open, each document recieved attention for an average of **1:16** minutes.

#5	12:37 minutes

The user made **5 different** searches, including "**bukhari**", "**death**", and "**hospital**". They explored **17 different** documents, **11** of which had **not previously been read**. Those documents were created between **March 2008** and **September 2008**. The most referenced people in the viewed documents were **Maulana** and **Haq Bukhari**. The most referenced locations in the viewed documents were **Basra** and **Gaza**. Of the **17** documents open, each document recieved attention for an average of **0:44** seconds.

#6	12:16 minutes

The user made **8 different** searches, including "**khouri**", "**dubai**", and "**nahid**". They explored **16 different** documents, **7** of which had **not previously been read**. Those documents were created between **January 2008** and **November 2008**. The most referenced people in the viewed documents were **Nicolai** and **Lashkar-e-Jhangvi**. The most referenced locations in the viewed documents were **Dubai** and **Moscow**. Of the **16** documents open, each document recieved attention for an average of **0:46** seconds.
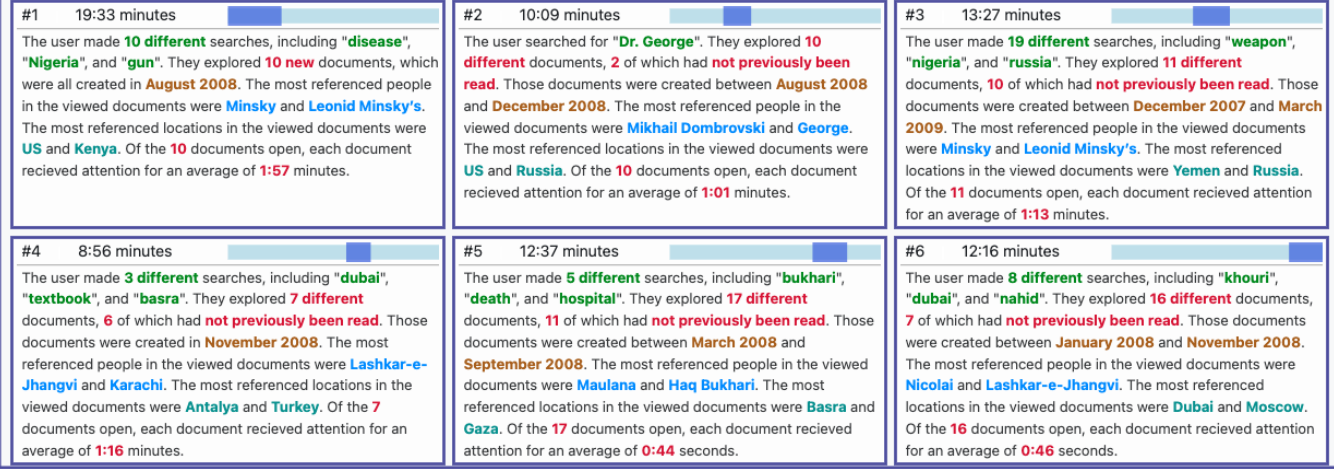
Fig. 2: The interface we designed to help communicate analytic provenance information to new users. Notice that it contains an overarching summary (in yellow) and individual summaries (the set of cards) for segments of time to help explain the analysis process. In this paper, we present the visualization approach and demonstrate the automated technique on five different datasets.

the steps we took to prepare a non-specific version of this conceptual technique.

In our specific implementation, we calculate a Term Frequency-Inverse Document Frequency (TF-IDF) vector for each document. We then loop over the events in an interaction log and replace any document visit with its TF-IDF vector. For any semantically meaningful events (e.g., query text, highlighted information, notes taken, etc.), we consider the textual content of these interactions as their own documents. Once the pre-processing has been completed, we have a vectorized form of the document and interaction content over time. To divide these interaction histories into segments, we use iterative binary segmentation, an offline-time change point detection algorithm [54]. At each iteration, the algorithm selects a point in time that maximizes the difference between the aggregated vectors of the two new segments that would be created by slicing the existing segment at that point. This is repeated iteratively until either a set number of segments are created or until there is no new segmentation point that would result in sufficiently different segments. Based on internal testing, we chose 10 slices resulting in 11 segments for this basic implementation. The choice of segmentation algorithm and its parameters may vary greatly between applications. Our specific selections merely serve as an example.

Once the interaction histories are segmented, we can generate summarizations. To summarize the content of the documents and textual user interactions within each segment, we identify the top words from the aggregated vectors to display as keywords. Additionally, we use named entity recognition from a spaCy English trained pipeline[1] to extract the most referenced people and locations from the documents interacted with in each segment. These are then prepared for the visualization component (described in Section 4.2). In our case, we choose to represent these patterns as text because it does not rely on visualization literacy and can explain potential patterns more explicitly without relying on taxing user exploration and analysis.

We also implement an overview that does not consider the segmentation. To generate this feature, we consider the entire interaction history for anomalies and calculate activity measures for each segment. This

set of superlatives can then be used to describe key behaviors, identify anomalous segments or call attention to the vital topic changes over time. Ultimately, we demonstrate this algorithm with a basic set of components to better understand what recommendations would benefit a tool used to summarize work in the wild. The code used to process and display our technique can be found at [to be provided after review]

4.2 Visual Interface

To represent and allow for superficial exploration of the segmentation and summarizations produced by our pipeline, we developed an associated interface. The interface shows a summary of one analyst's history doing an investigation using standard web technologies and D3 for card rendering and miniature bar charts. The interface essentially consists of two parts. Aside from a basic control panel for selecting parameters to display, there is an overview brief at the top and a series of cards showing interaction information at different points in time. We describe each component in the following subsections.

4.2.1 The Overview

In the yellow box at the top are a series of sentences describing stand-out periods or key features of an analysis history that serves design **G2** by providing a broad overview of session highlights and calling attention to segments of interest. In the visualization we present, this overview is calculated outside the segmentation algorithm and dynamically adjusts the content based on the active session. This is similar to the work by Packer and Moreau [41], who use W3C provenance data and fill-in sentence templates to explain the provenance as sentences. To fill in the gaps in the sentences, we calculate a series of statistics, including the most frequent search terms, the number of keywords, the percent of the dataset reviewed, and the average rate of interaction across the entire session. Since these statistics are calculated after the segmentation is complete, we can also calculate superlatives for each segment and call attention to them to better guide the user's exploration.

4.2.2 Prose Mode

Below the overview are a set of cards summarizing segments of a session. The cards go in order of time, starting at the top left, and each

---

[1]spaCy available at https://spacy.io/models/en#en_core_web_md
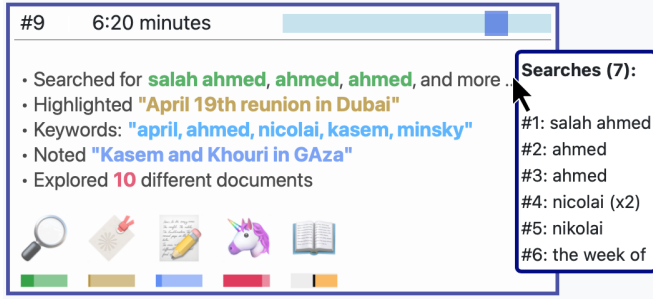
Fig. 3: A single card in list mode shows the information from a segment more completely, especially to aid debugging or for more detailed review or interaction provenance. The tooltips help show the various data more thoroughly. For example, we can see the list of searches completed in this segment when hovering over the list.

card has a blue time bar at the top representing what period of the session is summarized. The colors in these cards are associated with different types of information (i.e., green is for search information, and red is about the documents opened). To help users parse an analysis session's content, we separated an interaction history into smaller segments and summarized them. These 11 segments were pre-processed (as described in Section 4.1) and rendered to the screen to show various information like the searches complete, documents open, and entities related to documents open.

4.2.3 List Mode

There are two modes to the interface, and we have a toggle to adjust the view of the data. In prose mode, users can read a natural language summary of the events in a segment, whereas, in list mode (see Figure 3), the data is provided as a list with tooltips to show more details as you mouse over items. For example, the keywords are extracted from which documents were opened during a particular time segment. The paragraphs tell the story, and the list view shows all the data and requires mouse-over interactions to reveal details.

5 USAGE SCENARIO

Our tool allows a user to quickly grasp the content and focus of a previous work session. This can be used to refresh oneself on prior work, ease the hand-off of projects, or review the work of a peer or employee. Here, we present usage scenarios for each of these cases to contextualize the features of the tool.

Consider Jack, a researcher who is conducting a literature review along with his colleague Jill, under the supervision of their manager, Madison. Jack spends one work session accessing dozens of documents that contain common threads, such as discussion topics or references to items or figures. During the morning work session, Jack searches for key terms, highlights items of interest, and takes notes on his conclusions and later questions before taking a break for lunch.

When Jack returns, he is able to review his research session by exploring the list view. Here, his sequence of actions is more transparent. The interface adapts to the number of interactions from the morning, and without having to scroll the page, he can recall the list of searches conducted as well as where he wrote more notes. This summary of his session serves to quickly remind him of the work he has already done and reflect on other directions he can research in the afternoon. This demonstrates a critical value of our recommended pipeline: by showing analysis processes as segmented cards, familiarization with past work is easier to do, because the analysis is broken into stages and summarized.

The next day, Jack is unavailable, and Jill is tasked with continuing the literature review. To pick up where Jack left off, Jill would need to understand, at a high-level, the status of their shared project and any potential areas where she could contribute. Seeing Jack's investigation session segmented and summarized in our tool is helpful for inferring the main topics already explored and the general flow of logic. By

reading the summary at the top, Jill can see what topics received the most attention from her colleague and the segments with significant activities. Looking at the summary cards, she can view the various segments that made up Jack's work sessions, referring to the keywords and entities involved with the documents accessed in each of the segments. Even without referring to the more detailed view, she is able to familiarize herself with Jack's work in a few minutes and set a trajectory for herself to continue the literature review. This shows another important aspect of our technique: by having different levels of summarization, individuals interested in understanding the process more holistically can do so without zooming into all the specific events.

When the literature review is complete, Madison is tasked with reviewing the work and reporting on the performance of her team. Using the segment visualization tool, she is able to see the processes used to arrive at the final result. This can inform the edits she makes. For example, she can tell what terms were commonly searched for by comparing the two session reports and more confidently report on how thorough she feels her team has been in their review, all without having to ask Jack or Jill to prepare a report or arrange a one-on-one meeting. Perhaps more crucially, this tool can also inform future training. Looking at the steps taken by her two workers gives her insight into the work process of Jack and Jill, and she could use this information to offer constructive feedback not just on their results, but on their process and techniques. This scene represents how the segmentation and summarization of past work can help automate reporting, especially reports describing business processes.

6 CASE STUDY WITH INTELLIGENCE ANALYSTS

The inherent nature of provenance communication is that it must match the needs of the specific audience or users [33], so the specifics for the implementation of our proposed segmentation method must be tailored to a particular purpose. As a demonstration of our technique, we conducted a case study in the domain of intelligence analysis. Of the large number of domains interested in provenance support [60], the intelligence community relies heavily on reporting to share information with decision and policy makers [1]. We look at a scenario of intelligence work using provenance summaries of analysis for *pass-downs*, which refers to one analyst communicating complete and outstanding work to a new analyst. In practice, pass-downs are often static, manually generated text documents, and interaction and automation could save analysts time and help avoid mistakes. There are even recent efforts to generate tailored information reports for analysts of many kinds [27]. This fertile context of multiple analysts collaborating or managers reviewing analyst teams makes an ideal focus for reviewing our tool.

6.1 Iterative Design Reviews

Early in the development of the automated pipeline, we worked with members of the United States Department of Defense (DoD) to present our problem, show initial prototypes of the design, and identify beneficial features. Over a year of conversation, we developed a better understanding of the culture within the intelligence community and modified our design to reflect those changes. For example, we removed an early feature where the cards could be manually segmented, combined, and dynamically adjusted. We learned from these conversations that analysts already have an overwhelming task, and they would have little interest in manually determining which segments should be summarised. We found these discussions to be invaluable for our chosen domain to better understand what analysts actually value from an automated provenance summarization technique.

6.2 Test Datasets

While our external collaborators could not share the specifics of the datasets and tools they currently use, we could identify proxy analysis interaction log datasets completed on both fictional and real-world data. These logs shared a similar structure to the interaction information captured by intelligence tradecraft and tools. We discuss the five datasets and the process of adapting our technique accordingly.

6.2.1 Fabricated Intelligence Analysis Datasets

Three of the provenance datasets were based on interaction logs from sample analysis sessions conducted with fictitious text document data [35]. The VAST Challenge [2] is an annual competition within the visualization community designed to provide participants with an opportunity to work on real-world data sets and tackle challenging problems in visual analytics. It is important to note that all of these documents are fabricated and designed to mimic the kinds of textual reports seen by intelligence operators. The documents encountered often include newspaper clips, email correspondence, bank transactions, and more, including the occasional misspelling of names and incompleteness of data experienced by intelligence operators. Based on previous VAST Challenge datasets, prior records of provenance log datasets from Mohseni et al. [35] recorded the interaction history from students using an analysis tool to review three separate textual data associated with the VAST mini-challenge. Each of the three sample datasets considers the interactions complete on 102–158 text documents. The first dataset (from 2010 mini-challenge 1) considered a potential connection between illegal arms trading around the globe and the spread of a mysterious disease. The second (from 2011 mini-challenge 3) referenced the goings-on of terrorists in a fictional city. Finally, the third (2014 mini-challenge 1) traced news coverage about the kidnapping of individuals from a fictional company and identified likely suspects. The interaction logs from the datasets comprised interactions completed within a visual analysis tool for searching and viewing text documents. This tool consisted of an explorer with moveable windowed documents that allowed users to read, search for exact terms, highlight text, and write their own notes. The logs were triggered for every action listed, as well as mouse hovering over different documents, the dragging of documents, and any verbal think-aloud. The system tracked the time associated with each event and stored these in a file. The amount of interactions in each log range from 622 to 1633, with a mean of 1051.83 interaction events. In total, twenty-four participants used this tool for about 150 minutes to solve one of the three mysteries (i.e., eight participants per mystery).

6.2.2 Language Analysis of Conversation Records

To further demonstrate our technique's applicability, we also applied our approach to the interactions completed on a dataset collected independently by external collaborators. In this case, they had interactions from participants mimicking the work completed by language analysts in an unclassified setting. They developed an independent analysis tool that helped find audio recordings and imperfect transcripts from a publicly available dataset of former United States President Richard Nixon's phone recordings from his time in office[3]. To better understand how individuals work with large corpora of audio recordings, they asked participants to identify when events took place, decode code words and resolve different data discrepancies while tracking their interactions. The tool had an advanced search function allowing for not only keyword identification, but also could be refined by certain speakers, tape recording locations, and ranges of dates.

For this dataset, eight participants conducted the sample analysis. Many participants had training as professional analysts, but some had little to no analysis experience. Their analysis tool logged interactions for any searches completed, any transcripts opened, and when an audio file was played. The transcripts were exceptionally long since many tapes contained the beginnings and ends of multiple conversations upwards of two hours in length. For context, participants commonly performed a search, reviewed the list of results, opened a transcript, and spent a long time playing back the associated audio to find the information of interest. Therefore, the interaction logs included far fewer interactions than the fabricated datasets described prior. This specific context contrasts with the earlier three datasets because it

referred to long audio files with often incorrect transcription instead of only relying on text and misspellings.

We showed that with a custom pre-processing step, we could translate the interactions into a format expected by our tool for summarization. Our representation shows the search terms used over time but was especially limited in the meaning of the keywords identified from the dataset. We believe this is due to the limited number of interactions available to parse the entire session, resulting in many cards with less information. Yet, the keywords surfaced by the pipeline appeared to be associated with the task, including the code words and important figures from the era, instead of generic terms from the transcripts. This implementation helps show that our pipeline can be adapted and fine-tuned to accept interaction logs from a custom-built tool.

6.2.3 Operational Intelligence Dataset in Applied Context

One additional dataset context was based on a real-world intelligence analysis setting. As expressed by our external intelligence expert collaborators, many tools developed through academic research are limited to proxy data without demonstration on the actual real data of interest to specialized groups. This collaboration intentionally confronts this problem by bridging it to a similar problem with sensitive domain data. For this component of the case study, someone external to the core research team used our technique and the software to process their own sensitive dataset from an applied analysis setting and reported on their experience. While the level of detail we can describe the dataset is limited due to its sensitive nature, we describe the main points important for understanding the similarity to and differences from the previously described data sets.

The implementation used an existing dataset designed for evaluating the search performance of intelligence tools and compliance checking among intelligence community offices to improve internal tool development and inform analysis strategies. The interaction dataset was previously logged with an application deployed on a DoD system to look at query metadata for query compliance, a formally defined requirement of the intelligence community [1]. The external collaborator selected a single office's 24-hour search activity to generate a static demonstration for their discussion based on actions involving queries for information. This consisted of a small collection of people making queries with rich metadata information about what was being searched for and the associated justification. The query logs also included various types of metadata recorded for compliance purposes. For example, they included timestamps, document sources, and information relevant to access restrictions. In practice, such metadata is often captured but not directly visible for review. For the demo, they included the query string, the number of results (yet, not the resulting documents associated with a query), the date and time of the query, and the designator requesting the information.

Due to the sensitive nature of the data context, the data did not include links to the underlying corpus of documents being searched in the interaction logs. This made automatic segmentation difficult since the binary segmentation (described in Section 4.1) relies on the computed difference in document topics over time. Instead, our external collaborator used only the query metadata for segmentation and could account for data similarity or differences across segments. This demonstrates a practical challenge in summarization for analysis cases where data privacy or access restrictions might be in place. However, the approach can be adapted to operate on a subset of available data features to circumvent such limitations.

After applying the technique and generating sample analysis summaries, two associates (also external to the authors and core research team) demonstrated the provenance output for use in the meta-analysis task in reviewing the analysis workflow to gather impressions of this technique and interface from additional experts. We report the insights from this exercise in Section 7.2.

6.3 Expert Review

To evaluate the technique and implementation, we conducted an expert review with five domain experts in intelligence analysis. The experts had several years of experience within the intelligence community and

---

- Informed Consent	B. Binary Classification x4	C. Qualitative Review
Context and tutorial	*Analysis Style* *Exploratory | Targeted*	*1.Missing Info.* *2.Audience Type*
A. Open Exploration	*Information Tracking* *In tool | In head*	*3.Info. Types* *4.Info. Quantity*
What were the *3 main phases?*	*Query Types* *Trial-Error | Systematic*	*5.Time Chunking* *6.Open Comments*
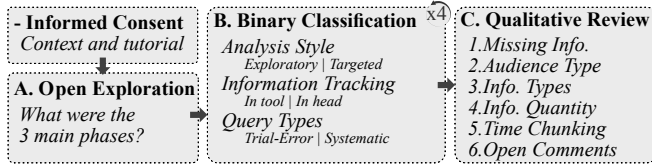
Fig. 4: A visual representation of the phases of the expert review used to evaluate the visualization and automatic summarization technique. A) Participants openly explored the interface with think-aloud. They then B) classified different analysis sessions before finally being C) asked questions to better understand their preferences.

dealt with the passing of intelligence information throughout their careers. We identified this group because of their experience with passing information and the similarity to the kind of interactions available in the test datasets (see Section 6.2). Three of the five experts had experience managing other analysts, and they all served in many capacities, including a watch-floor officer, mission manager, and discovery analyst. The goal was to understand if and how domain experts could use provenance summarizations to understand the flow of an analysis completed by another person.

This study was approved by our organization's Institutional Review Board (IRB). This study consisted of a semi-structured interview with domain experts in intelligence analysis. The study had three phases (see Figure 4). After receiving informed consent and describing the scenario of use, we asked participants to inspect the summarization for one person's interactions to get more comfortable with the interface and ask questions about the available features. Participants reviewed and explored the sample analyses for approximately 10 minutes.

Next, participants conducted an analysis review task that required interpreting other analysts' strategies. For this task, participants were asked to classify each analysis record according to three analysis styles based on previous discussions with domain experts in intelligence. Participants were provided a set of definitions for different analysis behaviors. They then reviewed and classified four sample analysis records in the tool (they reviewed one at a time with order randomized for each participant). First, participants had to determine whether the analysis was either *exploratory*, where it involves a more broad review of many different topics, or more *targeted*, where it focuses on a set of specific kinds of information or a specific target. Next, participants determined how information was tracked during an analysis session. Was it in the analysis *tool*, where they used the document viewer tool to leave records of their interests and findings, or in their *head*, where they left little record of their analysis in the tool until they were more confident about documenting findings? Finally, participants decided on query behaviors, determining if the searches were more *trial-and-error*, with bursts of activity and ad-hoc search frequency, or *systematic* and with more focus on what they searched for and the frequency how they used searches. Once they had classified the styles, we asked participants to rank the challenge of each classification task.

Finally, we asked participants to comment on the features we wanted to better understand. We asked about the information they would expect to see to better understand user behaviors. We also asked them to consider the intended audience of the tool and the modifications they would make for two likely use cases for provenance information (i.e., collaborating peers and team managers). We then asked about the quantities of information on a card and the chunking among the cards before opening up for any final generic comments as well. All of these later questions were considered through the lens of different audience types to better understand how peers' and managers' needs for provenance may differ. We discuss the details of user responses in Section 7.1.1 and Section 7.1.2.

6.4 Demonstration in an Operational Intelligence Context

In addition to the semi-structured interview described above, we also include discussions about the feedback we heard from an external col-

laborator and their experience adapting our technique for their specific context. They reported minimal issues creating brief data summaries as described in Section 6.2.3. The external collaborator held two informal occasions to collect feedback on the approach. First, there was a poster session where an external collaborator provided a live demonstration and answered more in-depth questions with 4 people. Second, there was a 1 hour and 45-minute meeting with about 10 individuals. Both occasions invited lots of interest from individuals focused on operations in general, including general technologists, technical analysts, and language analysts. Generally, they did not have specific opinions about the features, but the idea of the meta-analysis was seen as helpful as it does not have much support currently. In a later meeting, we met with this external collaborator and asked them to recall what was the experience of adapting our technique and also describe reactions to the tool. We recount the second-hand descriptions of this discussion in Section 7.2

7 RESULTS

In this section, we review the tasks, feedback, and additional comments received on the original interface and our technique's external application to and validation within a classified environment.

7.1 Results from Expert Review

We presented our technique's rendered visualization to former and current analysts associated with the DoD. This population has first-hand experience with collaborative investigation breakdowns and available support. Having an example, as we presented to them, is helpful for eliciting rich feedback and creative ideas from users close to an issue they handle daily: reviewing the analysis of their peers. We heard that tools to support the review of analytic provenance are not used, either because of their limited feature set or the time it takes to arrive at a meaningful conclusion. To better support these teams, we recruited five analysts to review our tool, complete a classification task and provide qualitative feedback through a semi-structured interview.

7.1.1 Analysis Review and Classification Results

With the intention of exposing our experts to ways they could use our tool's features in a meta-review, we asked analysts to classify a session's analysis styles. We asked for binary classifications for three analyst styles: analysis type, information tracking, and query type. Experts used our interface to determine if a session was more exploratory or targeted, tracked information in the tool or in one's head, and if the searches were more ad-hoc or systematic. We also asked participants to rank the difficulty of each classification. Ground truth was classified by one author, who spent the most time reviewing the test datasets. The author selected four sessions from one dataset with an equal mix of each binary quality (i.e., 2 sessions were exploratory, 2 were targeted, 2 tracked information in the tool, etc.) We calculate a performance score by comparing the classifications made by participants in their short 10-minute introduction to our tool compared to the ground truth determined by the author.

Overall performance was high at 76.6%, signaling that our interface could help identify these qualities. Performance was higher than random guessing (50%) but clearly was still challenging. Looking across the style categories, we see that the highest score relates to *information tracking* classification (90% correct). This high score was expected because the interface directly lists highlights and notes made during a session. It is likely that participants could use this feature to easily determine a classification. The lack of a perfect score can be attributed to one session that only used highlighting and did not leave notes. There was ambiguity in our definition as to whether this counted as using the tool to track information. Participants also scored highly (80%) on the *analysis type* classification, though participants consistently considered this the most difficult to determine. Our participants said that classifying the entire session into a single category was challenging because, often, a single session can have periods of both types of activity. Finally, analysis of *query behavior* was also considered hard to determine by participants and also had the most errors with a score of 60% correct compared to the ground truth. While better than guessing, this does suggest that our definition for query types may have been too loose, or

the representation of analysis with summary segments might make it difficult to process details of specific action behaviors. We expected analysts to look at the frequency of search events across the session to determine if the queries were systematically dissecting a topic over time or occurring in bursts. While this information was available to the participant, it required them to inspect the number of searches conducted in each segment of the session and determine if any segments had bursts of unusual search activity. Ultimately, these performance scores suggest that the segmentation of interaction behaviors and their associated summaries are helpful for recognizing analysis behaviors and helped inspire further conversation about what would be beneficial from a methodology that segmented and summarized interaction history.

7.1.2 Qualitative Feedback

From our expert reviews, one of the first things our interface was recognized for was its ability to inspire questions about meta-analysis in general. All of the experts were visually excited by the tool's potential and emphasized the benefits provenance visualization could have across their organization. Experts mentioned a more diverse set of audience members that could benefit from the tool than we considered initially, including higher-level management, analysis trainers, trade-craft designers, team-level management, collaboration, and even self reminders too. This shows one anecdote into how our example helped to elicit more creative thoughts and directions for provenance information.

Looking more closely at the information on the interface, we see universal agreement that interactions with semantic information (e.g., search terms and note content) were considered especially insightful for understanding what was happening (**G1**). We also heard that the tool overemphasized the *steps* of the analysis process and wanted to focus more on the *why* for the progression of the analysis. That is, some analysts were less interested in the details of how much time and how many documents, and they were more interested in following the flow of data topics. Specifically, two experts stated disinterest in the number of documents opened and the average length of time spent on each document in the textual prose. For further clarifying rationale, though it is possible to match behavior patterns to user intentions [61], automatically abstracting user intention without a sense of ground truth (e.g., think-alouds) will always be imperfect. In line with the common mantra of *overview first, zoom and filter, details on demand* [52], one expert stated, "I want a higher level of summarization and the ability to drill down [into more detail]." We note that our demonstration and review focused on collecting feedback on the summarization view, though enabling link-following to access associated source information is an obvious, required feature.

Additionally, three experts looked at the text in the prose mode and mentioned that they assumed more variability for the types of information communicated in the text. While many appreciated that the interface explained steps in prose, they noted interest in seeing more variety, unique patterns in a period of time, or details not seen elsewhere. We set out to use text for our technique because it mimics a familiar reporting structure (**G5**) in the intelligence community. Still, expert feedback helps show that there is nuance to how text should be used, and the deployment for any real use case would require a more tailored adaptation of summarizations.

Related to the textual component, all experts found the full session summary to be useful for getting a feel for what aspects to pay attention to and the key features of the session (**G4**). Yet, while some of our questions could be answered with the information in the summary section, we saw that most experts relied almost exclusively on the list mode to answer our questions. The list mode provided details on demand with tooltips and had information spaced consistently, making scanning easier. We believe this is because the experts were all trained analysts and wanted the most amount of detail to answer our questions. Because the experts spent the majority of their time reviewing the cards, eventually, they all asked about how the segmentation was being computed. As one expert put it, the "cards emphasize the idea that each segment is distinct and different from the others." Echoed by others, essentially, our experts recommended explicit representation of

the reason for segmenting one time segment from the others, and this reasoning must be explained more explicitly. Segmentation algorithms with simple rules are easy to explain and would likely be better for episodic summaries (**G1**, **G2**). Thus, this helps shed insights into the kinds of segmentation techniques that should be considered when presenting interactions as a series of cards.

Following the expert review, we note two major changes we made to the tool. First, we implemented a slider to adjust the number of cards to segment the analysis session into, thus allowing the reviewer the choice of the granularity for the amount of temporal segmentation and summarization. Second, to help emphasize the flow and connection of key data topics throughout the analysis, we added a brush-and-link feature that helps to reveal other segments that contain the same keyword when mousing over the cards.

7.2 Feedback from Operational Intelligence Context

The applied version of the tool reviewed in the sensitive intelligence context was well received warmly when presented to domain analysts, team managers, and upper-level management. The external collaborator said, "The basic elements happily surprised the people on the high side." The interface inspired more questions about what was possible to learn from the provenance records. For example, attendees considered how access to visualized workflows might influence analyst training. The ability to see workflows invited suggestions for detecting different analysis styles. As one observer noted, when a world event occurs, the ability to see workflows could help identify if analysis strategies adapt or stay the same using a meta-analysis technique. According to our external collaborators, the current training program relies on a collection of "good" analyst suggestions. This definition of good is also debated because it is based on semi-reliable metrics, such as how often an analyst's work is included in another analyst's report. However, this does not necessarily indicate quality. Visualizing behaviors from a meta-analysis perspective could better inform training programs for analysts instead of relying on inductive reasoning and semi-meaningful metrics. Ultimately, because few tools support meta-analysis of provenance interactions, the feedback indicated the reviewers found the tool to be thought-provoking and potentially valuable for helping to uncover ecosystem-wide challenges that are symptoms of different internal problems.

To help uncover these ecosystem-wide challenges, we also heard a need for adaptability. Because there is limited exposure to tools that support meta-analysis of work behavior, we recommend developing tools with flexibility at its core. This way, users can adapt their existing systems or work with data of various types. According to the enterprise objectives discussed in the United States National Intelligence Strategy [40], there is interest in supporting more meta-analysis of work practices across the intelligence community to help identify efficiencies and improve human experiences working with digital tools. Yet, a few individuals who saw the applied version of our technique mentioned being unfamiliar with the steps involved in completing a meta-analysis. Some could imagine a context where a user may want to look at what one analyst has done and the selection terms they used. In contrast, someone else may care about the whole organization and wants to find the abnormal actions that might be happening. Having a pipeline or tool that can dynamically wrangle these different data sources would be helpful. Additionally, if the ecosystem can consume interaction provenance from various sources, it would likely serve a more diverse set of users. Therefore, we also suggest that the kinds of data included in a provenance representation or pattern identification methods should support a variety of user types, visual representations, and data scenarios.

8 DISCUSSION

The primary contribution of our research is the presentation of segmentation and summarization methods for generating visual provenance support for communication and collaboration, and the case studies contribute new knowledge about important considerations for how to implement this approach. Here, we discuss lessons learned from the

case study with the intelligence analysis context, and we consider implications for broader applications. Based on the design iterations and follow-up correspondence with experts following the expert review studies, we identified a handful of key takeaways and design recommendations for adopting this technique. While partially grounded in comments provided by experts regarding the interface they worked with, we also make recommendations about provenance representation techniques as a whole. Our work most generally serves to strengthen the need for more communicative tools in the visualization and data science literature [10].

8.1 Takeaway 1: Use Meaningful Segmentation and Hierarchical Segmentation

Ideally, segmentation and summarization should be inherently interactive. Many experts commented on the segmentation approach provided by our tool, requesting additional control over how to segment the data. They recommended that the boundary between each segment should be "meaningful," not only based on keyword extraction but also on bursts of activity, as an example. Additional techniques exist for segmentation, including segmentation based on the time between events [6], pattern mining based on the order of event types [18], and Markov modeling for identifying the likelihood of different event transitions [29]. Yet, there is also recent evidence that individuals disagree on the "best" segmentation of events [49].

As Wang et al. [57] identified, only allowing control of temporal resolution is not sufficient because it does not allow users to select a specific segment, narrow their scope, and learn more about specific periods of the investigation session. This anecdote demonstrates two key recommendations: 1) ensure the user understands the segmentation method being applied and potentially provide additional ways to chunk investigation sessions, and 2) when summarizing information for users, these summaries should be hierarchical and allow users to zoom into specific segments for further investigation.

8.2 Takeaway 2: Pattern Identification for Different Audiences and Purposes

Our demonstration and expert review sheds light on broader uses about the use of provenance data. There is a desire to involve more fuzzy nuances to what should be included in the representation (e.g., rapid or repeated searches in relatively short bursts of activity should be collapsed). In particular, we heard from our experts that in certain use cases, they would want different kinds of data. For example, team managers need fewer details about the process and more about the general approach. As provenance data is scaled, there is a need to better support adjusting levels of detail and soft pattern elicitation to more effectively identify areas of interest and support insight [17]. Knowing the right patterns to elicit from an interaction log is hard to predict, though. Suggesting variables of interest or similar visualizations like the voyager tool by Wongsuphasawat et al. [59] could be beneficial because it would guide deeper exploration into the data.

Alternatively, the kinds of information displayed could be user adjustable. As one expert stated, "I just want a button to turn off different kinds of information." When we added brush-and-link functions to our tool, our experts said it was especially helpful in seeing patterns over time because it provides another way of considering data and its relationships. Much like different scientific disciplines may be familiar with certain design conventions [16], there are opportunities to adjust a summarization technique to display different information for different audiences. Having tools that adjust their visualizations for different viewers is a core component of Viégas and Wattenberg [56] concept of Communication-Minded Visualization. Considering the type of audience and their expected use of the information should influence what kinds of behavior patterns are made available. Yan et al. [61] describe a technique for automatic pattern identification from interaction logs. While this technique does not actively consider the type of user reviewing the information and their information needs, they do emphasize how higher-level explanations about why someone did something can be derived from low-level event sequences. There is also a need for more computer-supported pattern elicitation, including

human-in-the-loop approaches that consider the interactions of a user to suggest additional patterns involving similar data [24,67]. This focus on pattern elicitation from lower-level events holds promise for future provenance visualization techniques.

8.3 Takeaway 3: Similarity of Narrative Visualization and Provenance Visualization

Participants generally wanted the prose mode to carry more information than it did, especially more to explain *why* for various segments. While our implementation opted for a simplified base solution, integrating more advanced NLP techniques would likely better meet user expectations for what would be written on each card. Textual blurbs would also benefit from considerations from narrative storytelling. Models emphasizing narrative structures like discovery, rescue, or mystery, could likely increase engagement and benefit comprehension [22]. Since provenance data is essentially hierarchical, it could also benefit from the *minimum description length* principle to aid in summarization [55]. There are also applications for considering temporal summary images that include a temporal component with comic strip-style snapshots and annotations for communicating analysis over time [7]. Fundamentally, interaction histories carry an analogy to storytelling and narrative techniques because they are essentially episodes over time.

8.4 Takeaway 4: Adaptability of Provenance Processing

In addition to the adjustable visualizations for different users, we also recommend the need for flexible data-wrangling tools. Generating a universal provenance specification for interoperability has been suggested for years [5,6,21,23]. The experts immediately identified many different types of users who could be supported by this and the variety of data sources they would draw from to generate visualizations. There was an interest in better understanding how to conduct a meta-review. Many recognized how a tool that can dynamically adjust the kind of details included was recognized as valuable. Generally, there was a need for pipelines that could flexibly handle interaction logs with various forms. We demonstrate how relatively small data transformations could be used to allow our technique to interpret five different kinds of interaction logs. In the wild, legacy systems are commonplace, and building techniques that are agnostic to the variety of data forms yet still interpretable is immediately valuable.

9 CONCLUSION

We present an algorithmic technique paired with a new representation to better understand user reactions to provenance segmentation and summarization. Our demonstration with intelligence analysts helped us understand the features of interest related to a segmentation and summarization approach, not to confirm our pipeline was effective. Using our demonstration as a design artifact, we elicit more realistic user needs for meta-analysis of provenance information. Our demonstration generates creative interest in the concept of meta-analysis within the intelligence community. Our experts found the emphasis on time to be less meaningful unless they were to spend more time reviewing the work of their peers. Instead, they requested higher levels of data summarization (e.g., topic modeling, action rationale, and final results). These higher-level summaries are a known challenge to extract from interaction events but represent the goal of analytic provenance et large [47]. Based on our case study, we call attention to the need for segmentation to have meaningful boundaries and ensure summarization occurs hierarchically so it can be drilled into for additional details. We call for future work in the directions of additional pattern-matching metric elicitation and interactive filtering to provide brief overviews of work completed for different users. This inspires future work to better understand how elements like the level of detail, amount of information, and audience types may impact users' understanding of a provenance history.

implementing our approach and recounting feedback from sensitive, applied environments.

REFERENCES

[1] *Procedures Governing the Conduct of DoD Intelligence Activities: Annex Governing Signals Intelligence Information and Data Collected Pursuant to Section 1.7(c) of E.O. 12333*. Department of Defense, Jan 2021. SIGINT Annex. 5, 6

[2] B. Bach, Z. Wang, M. Farinella, D. Murray-Rust, and N. Henry Riche. Design patterns for data comics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–12. Association for Computing Machinery, Apr 2018. doi: 10.1145/3173574.3173612 3

[3] P. Bao. *Sharing Insight Provenance in Collaborative Visual Analytics*. PhD thesis, Northwestern University, Jul 2013. 2

[4] J. E. Block, S. Esmaeili, E. D. Ragan, J. R. Goodall, and G. D. Richardson. The influence of visual provenance representations on strategies in a collaborative hand-off data analysis scenario. *IEEE Transactions on Visualization and Computer Graphics*, p. 1–11, 2022. doi: 10.1109/TVCG .2022.3209495 2, 3

[5] A. Bock, E. Axelsson, J. Costa, G. Payne, M. Acinapura, V. Trakinski, C. Emmart, C. Silva, C. Hansen, and A. Ynnerman. Openspace: A system for astrographics. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):633–642, Jan 2020. doi: 10.1109/TVCG.2019.2934259 9

[6] M. A. Borkin, C. S. Yeh, M. Boyd, P. Macko, K. Z. Gajos, M. Seltzer, and H. Pfister. Evaluation of filesystem provenance visualization tools. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2476–2485, Dec 2013. doi: 10.1109/TVCG.2013.155 9

[7] C. Bryan, K.-L. Ma, and J. Woodring. Temporal summary images: An approach to narrative visualization via interactive annotation generation and placement. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):511–520, Jan 2017. doi: 10.1109/TVCG.2016.2598876 2, 9

[8] S. Chen, J. Li, G. Andrienko, N. Andrienko, Y. Wang, P. H. Nguyen, and C. Turkay. Supporting story synthesis: Bridging the gap between visual analytics and storytelling. *IEEE Transactions on Visualization and Computer Graphics*, 26(7):2499–2516, Jul 2020. doi: 10.1109/TVCG. 2018.2889054 2

[9] A. M. Chircu, J. L. Gogan, S. R. Boss, and R. Baxter. Medication errors, handoff processes and information quality: A community hospital case study. *Business Process Management Journal*, 19(2):201–216, Jan 2013. doi: 10.1108/14637151311308286 2

[10] A. Crisan, B. Fiore-Gartland, and M. Tory. Passing the data baton: A retrospective analysis on data science work and workers. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1860–1870, Feb 2021. doi: 10.1109/TVCG.2020.3030340 9

[11] A. Cunningham, M. Nowina-Krowicki, J. Walsh, J. Chung, S. Wark, and B. H. Thomas. Provenance narratives and visualisation to support understanding and trust. In *El Sawah, S. (ed.) MODSIM2019, 23rd International Congress on Modelling and Simulation.*, pp. 421–427. Modelling and Simulation Society of Australia and New Zealand, Dec 2019. doi: 10. 36334/modsim.2019.D1.cunningham 2

[12] K. Dracup and P. E. Morris. Passing the torch: The challenge of handoffs. *American Journal of Critical Care*, 17(2):95–97, Mar 2008. doi: 10.4037/ ajcc2008.17.2.95 2

[13] 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*, CHI '12, p. 1663–1672. Association for Computing Machinery, May 2012. doi: 10.1145/2207676. 2208293 1, 2

[14] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, Apr 2021. doi: 10.1162/tacl_a_00373 2

[15] M. Feng, C. Deng, E. M. Peck, and L. Harrison. Hindsight: Encouraging exploration through direct encoding of personal interaction history. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):351–360, Jan 2017. doi: 10.1109/TVCG.2016.2599058 3

[16] S. R. Gomez, R. Jianu, C. Ziemkiewicz, H. Guo, and D. Laidlaw. Different strokes for different folks: Visual presentation design between disciplines. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2411–2420, Dec 2012. doi: 10.1109/TVCG.2012.214 9

[17] D. Gotz. Soft patterns: Moving beyond explicit sequential patterns during visual analysis of longitudinal event datasets. In *Proceedings of the IEEE VIS 2016 Workshop on Temporal& Sequential Event Analysis*, 2016. 9

[18] D. Gotz, F. Wang, and A. Perer. A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. *Journal of Biomedical Informatics*, 48:148–159, Apr 2014. doi: 10. 1016/j.jbi.2014.01.007 9

[19] D. Gotz and M. X. Zhou. Characterizing users' visual analytic activity for insight provenance. *Information Visualization*, 8(1):42–55, Jan 2009. doi: 10.1057/ivs.2008.31 2

[20] N. Goyal and S. R. Fussell. Effects of sensemaking translucence on distributed collaborative analysis. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, p. 288–302. Association for Computing Machinery, Mar 2016. doi: 10. 1145/2818048.2820071 2

[21] S. Gratzl, A. Lex, N. Gehlenborg, N. Cosgrove, and M. Streit. From visual exploration to storytelling and back again. *Computer Graphics Forum*, 35(3):491–500, Jun 2016. 2, 9

[22] S. J. Green, K. Grorud-Colvert, and H. Mannix. Uniting science and stories: Perspectives on the value of storytelling for communicating science. *FACETS*, 3(1):164–173, Oct 2018. doi: 10.1139/facets-2016-0079 9

[23] P. Groth and L. Moreau. Prov-overview, Apr 2013. 9

[24] W. Han and H.-J. Schulz. Providing visual analytics guidance through decision support. *Information Visualization*, 22(2):140–165, Apr 2023. doi: 10.1177/14738716221147289 9

[25] J. Heer, F. B. Viégas, and M. Wattenberg. Voyagers and voyeurs: Supporting asynchronous collaborative visualization. *Communications of the ACM*, 52(1):87–97, Jan 2009. doi: 10.1145/1435417.1435439 2

[26] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):248:1–248:38, Mar 2023. doi: 10.1145/3571730 2

[27] K. Kershaw. Creating a "tldr" for knowledge workers, Aug 2022. 5

[28] D. Kouřil, O. Strnad, P. Mindek, S. Halladjian, T. Isenberg, M. E. Gröller, and I. Viola. Molecumentary: Adaptable narrated documentaries using molecular visualization. *IEEE Transactions on Visualization and Computer Graphics*, 29(3):1733–1747, Mar 2023. doi: 10.1109/TVCG.2021. 3130670 2

[29] B. C. Kwon, V. Anand, K. A. Severson, S. Ghosh, Z. Sun, B. I. Frohnert, M. Lundgren, and K. Ng. Dpvis: Visual analytics with hidden markov models for disease progression pathways. *IEEE Transactions on Visualization and Computer Graphics*, 27(9):3685–3700, 2021. doi: 10.1109/TVCG. 2020.2985689 9

[30] B. C. Kwon, J. Verma, and A. Perer. Peekquence: Visual analytics for event sequence data. In *ACM SIGKDD 2016 Workshop on Interactive Data Exploration and Analytics*, vol. 1, 2016. 2

[31] V. L. Lemieux. Provenance: Past, present and future in interdisciplinary and multidisciplinary perspective. In V. L. Lemieux, ed., *Building Trust in Information*, Springer Proceedings in Business and Economics, p. 3–45. Springer International Publishing, 2016. doi: 10.1007/978-3-319-40226 -0_1 2

[32] Y. Liu, T. Althoff, and J. Heer. Paths explored, paths omitted, paths obscured: Decision points & selective reporting in end-to-end data analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, p. 1–14. Association for Computing Machinery, Apr 2020. doi: 10.1145/3313831.3376533 3

[33] K. Madanagopal, E. D. Ragan, and P. Benjamin. Analytic provenance in practice: The role of provenance in real-world visualization and data analysis environments. *IEEE Computer Graphics and Applications*, 39(6):30–45, 2019. 5

[34] A. Mathisen, T. Horak, C. N. Klokmose, K. Grønbæk, and N. Elmqvist. Insideinsights: Integrating data-driven reporting in collaborative visual analytics. *Computer Graphics Forum*, 38(3):649–661, Jun 2019. doi: 10. 1111/cgf.13717 2

[35] S. Mohseni, A. Pachuilo, E. H. Nirjhar, R. Linder, A. Pena, and E. D. Ragan. Analytic provenance datasets: A data repository of human analysis activity and interaction logs. (arXiv:1801.05076), Jan 2018. arXiv:1801.05076 [cs]. 6

[36] P. H. Nguyen, K. Xu, A. Bardill, B. Salman, K. Herd, and B. W. Wong. Sensemap: Supporting browser-based online sensemaking through analytic provenance. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*, p. 91–100, Oct 2016. doi: 10.1109/VAST.2016. 7883515 3

[37] B. F. K. Norambuena and T. Mitra. Narrative maps: An algorithmic approach to represent and extract information narratives. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–33, Jan 2021. doi: 10.1145/3432927 2

[38] B. F. K. Norambuena, T. Mitra, and C. North. Narrative sensemaking: Strategies for narrative maps construction. In *2021 IEEE Visualization Conference (VIS)*, p. 181–185, Oct 2021. doi: 10.1109/VIS49827.2021. 9623296 2

[39] C. North, R. Chang, A. Endert, W. Dou, R. May, B. Pike, and G. Fink. Analytic provenance: process+ interaction+ insight. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pp. 33–36. 2011. 2

[40] ODNI. *National Intelligence Strategy of the United States of America*, p. 17, Jan 2019. 8

[41] H. S. Packer and L. Moreau. *Sentence Templating for Explaining Provenance*, vol. 8628 of *Lecture Notes in Computer Science*, p. 278–280. Springer International Publishing, 2015. doi: 10.1007/978-3-319-16462 -5_33 4

[42] D. Park, M. Suhail, M. Zheng, C. Dunne, E. Ragan, and N. Elmqvist. Storyfacets: A design study on storytelling with visualizations for collaborative data analysis. *Information Visualization*, 21(1):3–16, 2022. doi: 10. 1177/14738716211032653 2

[43] E. S. Patterson and D. D. Woods. Shift changes, updates, and the on-call architecture in space shuttle mission control. *Computer Supported Cooperative Work (CSCW)*, 10(3):317–346, Sep 2001. doi: 10.1023/A: 1012705926828 2

[44] 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, 2015. 2

[45] 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*, CHI '15, p. 2711–2720. Association for Computing Machinery, Apr 2015. doi: 10. 1145/2702123.2702376 1

[46] 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*, CHI '15, p. 2711–2720. Association for Computing Machinery, Apr 2015. doi: 10. 1145/2702123.2702376 2

[47] S. Ram and J. Liu. Understanding the semantics of data provenance to support active conceptual modeling. In *Active Conceptual Modeling of Learning*, p. 17–29. Springer, Berlin, Heidelberg, Nov 2006. doi: 10. 1007/978-3-540-77503-4_3 9

[48] A. Sarvghad, M. Tory, and N. Mahyar. Visualizing dimension coverage to support exploratory analysis. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):21–30, Jan 2017. doi: 10.1109/TVCG.2016. 2598466 3

[49] K. Sasmita and K. M. Swallow. Measuring event segmentation: An investigation into the stability of event boundary agreement across groups. *Behavior Research Methods*, 55(1):428–447, Apr 2022. doi: 10.3758/ s13428-022-01832-5 9

[50] E. Segel and J. Heer. Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1139–1148, Nov 2010. doi: 10.1109/TVCG.2010.179 2

[51] R. Sevastjanova, M. El-Assady, A. J. Bradley, C. Collins, M. Butt, and D. Keim. Visinreport: Complementing visual discourse analytics through personalized insight reports. *IEEE Transactions on Visualization and Computer Graphics*, p. 1–1, 2021. doi: 10.1109/TVCG.2021.3104026 2

[52] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE symposium on visual languages*, pp. 336–343. IEEE, 1996. 8

[53] Y. B. Shrinivasan, D. Gotz, and J. Lu. Connecting the dots in visual analysis. In *2009 IEEE Symposium on Visual Analytics Science and Technology*, p. 123–130. IEEE, Oct 2009. doi: 10.1109/VAST.2009.5333023 2

[54] C. Truong, L. Oudre, and N. Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, Feb 2020. doi: 10.1016/j.sigpro.2019.107299 4

[55] R. Veras and C. Collins. Optimizing hierarchical visualizations with the minimum description length principle. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):631–640, Jan 2017. doi: 10.1109/TVCG. 2016.2598591 2, 9

[56] F. B. Viegas and M. Wattenberg. Communication-minded visualization: A call to action [technical forum]. *IBM Systems Journal*, 45(4):801–812, 2006. doi: 10.1147/sj.454.0801 9

[57] T. D. Wang, C. Plaisant, B. Shneiderman, N. Spring, D. Roseman, G. Marchand, V. Mukherjee, and M. Smith. Temporal summaries: Supporting temporal categorical searching, aggregation and comparison. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1049–1056, Nov 2009. doi: 10.1109/TVCG.2009.187 3, 9

[58] J. D. Wayne, R. Tyagi, G. Reinhardt, D. Rooney, G. Makoul, S. Chopra, and D. A. DaRosa. Simple standardized patient handoff system that increases accuracy and completeness. *Journal of Surgical Education*, 65(6):476–485, Nov 2008. doi: 10.1016/j.jsurg.2008.06.011 2

[59] 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*, 22(1):649–658, Jan 2016. doi: 10.1109/TVCG.2015.2467191 9

[60] K. Xu, A. Ottley, C. Walchshofer, M. Streit, R. Chang, and J. Wenskovitch. Survey on the analysis of user interactions and visualization provenance. *Computer Graphics Forum*, 39(3):757–783, Jun 2020. doi: 10.1111/cgf. 14035 2, 5

[61] J. N. Yan, Z. Gu, and J. M. Rzeszotarski. Tessera: Discretizing data analysis workflows on a task level. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, p. 1–15. ACM, May 2021. doi: 10.1145/3411764.3445728 8, 9

[62] J. M. Zacks, N. K. Speer, K. M. Swallow, T. S. Braver, and J. R. Reynolds. Event perception: A mind-brain perspective. *Psychological Bulletin*, 133:273–293, 2007. doi: 10.1037/0033-2909.133.2.273 3

[63] J. M. Zacks and K. M. Swallow. Event segmentation. *Current Directions in Psychological Science*, 16(2):80–84, Apr 2007. doi: 10.1111/j.1467-8721. 2007.00480.x 3

[64] J. M. Zacks and B. Tversky. Event structure in perception and conception. *Psychological Bulletin*, 127:3–21, 2001. doi: 10.1037/0033-2909.127.1.3 3

[65] J. M. Zacks, B. Tversky, and G. Iyer. Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General*, 130:29–58, 2001. doi: 10.1037/0096-3445.130.1.29 3

[66] J. Zhao, M. Glueck, P. Isenberg, F. Chevalier, and A. Khan. Supporting handoff in asynchronous collaborative sensemaking using knowledge-transfer graphs. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):340–350, Jan 2018. doi: 10.1109/TVCG.2017.2745279 1, 2, 3

[67] J. Zhao, S. Xu, S. Chandrasegaran, C. Bryan, F. Du, A. Mishra, X. Qian, Y. Li, and K.-L. Ma. Chartstory: Automated partitioning, layout, and captioning of charts into comic-style narratives. *IEEE Transactions on Visualization and Computer Graphics*, 29(2):1384–1399, Feb 2023. doi: 10.1109/TVCG.2021.3114211 2, 9