This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3444594

Investigating Professional Analyst Strategies in Immersive Space to Think

Kylie Davidson, Lee Lisle, Ibrahim A. Tahmid, Kirsten Whitley, Chris North, and Doug A. Bowman

Abstract—Existing research on sensemaking in immersive analytics systems primarily focuses on understanding how users complete analysis within these systems with quantitative and qualitative datasets. However, these user studies mainly concentrate on understanding analysis styles and methodologies from a predominantly novice user study population. While this approach provides excellent initial insights into what users may do within IA systems, it fails to address how professionals may utilize an immersive analytic system for analysis tasks. In our work, we build upon an existing immersive analytics concept - “Immersive Space to Think” to understand how professional user populations differ from novice users in immersive analytic system usage. We conducted a user study with 11 professional intelligence analysts who completed three analysis sessions each. Using our results from this study, we provide deep analysis into how professional users complete sensemaking within immersive analytic systems, compare our findings to previously published findings with a novice user population, and provide insights into how to develop better IA systems to support the professional analyst’s strategies within these systems.

Index Terms—Human-Computer Interaction, Immersive Analytics, Virtual Reality, Information Visualization, Sensemaking

1 Introduction

In the world of big data, there is a growing need to provide better ways of filtering and sorting the large amounts of data presented to us. In addition, providing analysis spaces for understanding the presented data will be critical to parse through the data to find patterns and develop key insights. With the advent of new lower-cost immersive technologies, there are new opportunities for using these technologies to build an analysis space of the future for completing sensemaking tasks. Immersive Analytics (IA) is an emerging research field combining previous research on visual analytic systems and data visualization with new immersive technologies such as virtual/augmented reality interfaces. These immersive technologies provide richer sensory information, including better depth cues, increased spatial orientation understanding, better peripheral awareness, and increased engagement compared to traditional desktop displays [1], [2], [3]. Additionally, they build on the ideas of distributed cognition [4], offering an immersive way to externalize cognition into the environment around you.

With this growing research area, there has been a large body of work dedicated to understanding how users work within these spaces with both quantitative data using 3D visualizations [5], [6], [7] and non-quantitative data such as text-based analysis [8], [9]. Many of these non-quantitative data-based studies have focused on understanding how these immersive analytic systems can support the sensemaking process [10] including single-session analysis [8], [9], multi-session analysis [11] and trying to understand how strategies used and structures made within these systems in these systems change as users become more familiar with their usage [12]. These studies have provided many insights into how IA systems can benefit the sensemaking task.

However, there is a large gap in our understanding of professional users within IA systems. Some research has been dedicated to using professional users within immersive analytic prototypes to understand how professional economists explored economic data [7]. However, to our knowledge, a study has not yet been done on how professional intelligence analysts utilize an IA system during an analysis task. Additionally, we are unaware of work that aims to compare professional and novice analysts within these systems. This paper seeks to fill that gap by providing a deep data analysis on professional intelligence analysts’ usage in an IA system - Immersive Space to Think, as well as providing insights into how these professional intelligence analysts differed or were similar to novice user populations studies before.

The key contributions of this work are as follows:

- A deep understanding of professional intelligence analysts’ multi-session sensemaking process in immersive space
- A detailed comparison of novice vs. professional analysts in immersive sensemaking
- A validation of previous findings presented in IA sensemaking systems

2 Related Work

2.1 Intelligence Analysis and Sensemaking

Some examples of professions that focus on sensemaking are journalism [13] and intelligence analysis [10]. In Intelligence analysis, analysts parse large amounts of structured and unstructured data to evaluate the information for potential threats. Pirolli and Card [10] defined the sensemaking

- *Kirsten Whitley is with the US Department of Defense*
- *Kylie Davidson, Lee Lisle, Ibrahim A. Tahmid, Chris North, and Doug A. Bowman are with the Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, 24060.*
  E-mails: kyliedavidson@vt.edu, llisle@vt.edu, iatahmid@vt.edu, north@cs.vt.edu, dbowman@vt.edu,

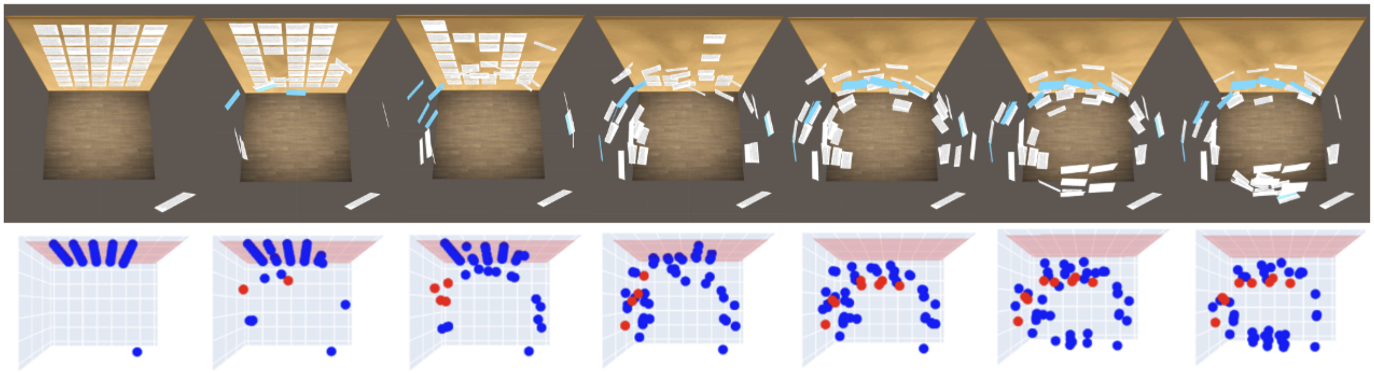


Fig. 1: High-Level View of an Immersive Space to Think Analysis. This example is of Participant Five's analysis over time. From left to right: start, middle of session 1, end of session 1, middle of session 2, end of session 2, middle of session 3, end of session 3. Top Row: Top-down orientation into the immersive space. Bottom Row: top-down view of 3D scatter plots used for data analysis in this work. The blue dots represent the documents provided to the users for analysis, and the red dots represent notes/labels created by the user during analysis.

process as having one main loop - the *Sensemaking Process* or *Knowledge Generation Process* with two sub loops - *foraging loop* and *sensemaking loop*. During the foraging loop, information is gathered, sources are read, and relationships are drawn between the collected information. During the sensemaking loop, schemas are formed, hypotheses are developed, and a case is built on relevant findings. The end goal of the sensemaking process is **presentation**, where the end story is told about the underlying analysis and its conclusions. In the case of intelligence analysis, this presentation could be the published intelligence report, or in journalism, the end product can be a news article. The sensemaking process is bidirectional, and as one completes the analysis, one can move up and down through the different stages of the process as needed while conducting this analysis.

The sensemaking process is both time and effort-intensive [10]. Additionally, the further into the process, the more effort the analysis requires. Simple sensemaking tasks such as deciding what to cook for dinner or which product to buy may not require much time/effort. However, when we look at the scale of intelligence analysis, these tasks may take weeks, months, or even years to complete. Previous work on sensemaking in immersive analytic systems using an intelligence analysis task has used novice user populations [9], [11]. This study gave the research community insights into how IA can support multi-session sensemaking tasks and the overall organizational structures and strategies used during analysis. However, evaluating sensemaking with solely a novice user population could be inadequate to understand how these systems would be used for real complex sensemaking tasks such as intelligence analysis. For example, professionally trained intelligence analysts have critical and analytic thinking skills which could lead to different overall strategies for sensemaking or different ways in which they organize the information in the environment around them. This work aims to build on previous work looking at multi-session analysis within IA systems and compare and contrast how professional analysts differ from novice analysts within IA sensemaking.

Since sensemaking is a cognitively demanding task, many systems have been developed to support analysts while completing their analysis. Visual Analytics (VA) aims to facilitate analytic reasoning or sensemaking using interactive visual interfaces [14]. VA systems aim to assist the user in exploring data, identifying regions of interest, and synthesizing large amounts of data in a timely manner [15]. Overall, VA systems should support analysts understanding, reasoning, and decision-making during sensemaking. There are many ways in which a VA system can help in the sensemaking process, with Endert et al. [16] emphasizing that most approaches target the foraging loop or sensemaking loop support. Many VA system approaches use semantic interaction or human-in-the-loop machine learning to provide assistance [17], [18], [19]. At the same time, another approach for VA is to provide a large physical "space to think" for completing an analysis task using multiple monitors [3]. In these large spaces, the space around the user can be used to distribute cognition into different regions of the space and later accessed as a form of externalized memory during the sensemaking process. As technology evolves, new immersive technologies such as virtual and augmented reality provide new opportunities for VA in immersive spaces - coined Immersive Analytics.

2.2 Immersive Analytics

Immersive Analytics (IA) is built on the ideas of VA and is defined as "the science of analytic reasoning facilitated by immersive human-computer interfaces" [20]. Ideally, these systems should support many stages of the sensemaking process, such as data exploration and schemas development through data visualization and embodied interactions. IA systems provided new opportunities that VA systems are unable to support through providing better depth cues, expansive 360 degree 6 degrees of freedom space for interaction, limited distractions, and better spatial understanding [1], [2], [3]. These immersive technologies allow the user to "step through the glass" [21] of traditional display desktop technologies, allowing users to be inside the data. Some applications areas of IA include situated analytics/spatial data analysis [6], [22], [23], [24], immersive visualization [5], [25], and collaboration [26], [27], [28].

Another IA system approach builds off of the Space to Think concept by Andrews et al. [3]. In this approach to IA, a large-tracked area is provided to a user during a sensemaking or analysis task. There is a lot of research surrounding this application area, including understanding how the users utilize the space [8], [9], understanding the deeper structures made within the spaces while completing an analysis task [11], [12], [24], [29], [30], trying to determine how much space is beneficial for analysis [31], which technology is better for analysis in these spaces [32], how we can merge traditional 2D and immersive 3D systems through cross-virtuality [33], cognitive load analysis of these systems [34], and how we can support organization with clustering techniques [35].

The research on this application area is growing, but there is still limited research on understanding how professional users may utilize these systems for analysis. Batch et al. [7] conducted a user study with professional economist/data scientist users, providing critical insights into system usage in both an “in-the-wild study” leading to system feature changes and a mixed methods study. However, to our knowledge, there is no published work on how professional intelligence analysts utilize an IA system for sensemaking or comparison study to understand how professional and novice analysts differ in their organizations and strategies. This work aims to fill this gap.

3 Immersive Space to Think Approach

Immersive Space to Think (IST) is a concept for immersive analytic systems to provide a large-tracked space while completing a sensemaking task involving reading, making connections, and organizing multiple text-based documents [9], [11]. The IST concept builds on distributed cognition, allowing users to offload cognition from their analysis into the immersive space around them through the organizations and structures created during their analysis. In our version of IST, we achieve the immersive space through a VR head-worn display where users interact in the system with hand-held controllers. In addition, we have a tracked keyboard on a wheeled table that can be used for text input during the analysis (see figure 2). For our research, we are focusing on a text-based sensemaking task. In this version of IST, all of the documents used for analysis begin in a randomized order on a virtual bulletin board. Then using the tracked controllers, the users can pull documents off the virtual bulletin board and organize them into the environment around them. Beyond the ability to organize the documents in the 3D immersive space, the version of IST used in this study provided some additional sensemaking features.

Label Making - IST supports creating text labels to use within the space. **Note Taking** - The system also supports notes, which the user can create to make annotations within the space. The note feature also lets users offload information into the immersive environment. **Searching** - Using the text input system, IST supports searching for relevant keywords/phrases. **Highlighting** - Users can select text within a document and highlight it yellow for emphasis. **Copy To/Paste From Clipboard** - To better support the user in offloading information into notes or the search bar, IST supports the ability to copy text from a document and paste it elsewhere within the IST system.

As this study aimed to evaluate the differences between Novice users and Professionally Trained analysts, the features used in this IST study were the same as in the recent paper of Davidson et al. [11].

4 User Study

4.1 Goals and Research Questions

A professional intelligence analyst uses analytic thinking skills to work on complex sensemaking tasks daily. These professionals are trained to complete these complex tasks and develop a unique set of critical thinking skills to achieve their goals. To fully understand how the IST concept can support the sensemaking task, we believe that evaluating professional intelligence analyst sensemaking within this system can provide us with new insights on how these tools would be used “in-the-wild”. This user study aimed to understand how professionally trained intelligence analysts complete sensemaking within an IA system and gather feedback from those analysts on future design ideas for IA prototypes. This study also aims to understand how professionally trained analysts and novice users differ in IA system usage, as the type of strategies and organizational tools used by these analysts may differ from those identified in previous work.

To compare these two populations, we rely on the data and user study design as seen in Davidson et al.’s previous work [11].

Our research questions are as follows:

4.1.1 Research Questions

1) **RQ1 - Spatial Structures**
   a) What spatial structures do professional analysts form in IST?
   b) How do these spatial structures change over multiple sessions?
   c) How do these structures compare to novice users?
2) **RQ2 - Transformation Path**
   a) How do the spatial structures formed map to the 1-dimensional written reports?
   b) How do these transformation paths compare to novice user transformation paths?
3) **RQ3 - Physical Navigation**
   a) How do professional analysts navigate the space while completing the sensemaking task?
   b) How do professional analysts compare to novice users in physical navigation while completing a sensemaking task?
4) **RQ4 - Strategies**
   a) What strategies do the professional analysts use, and how does this correlate to the quality of the written reports?

b) How do these strategies change over the analysis sessions?
c) How do these strategies compare to those of novice users?

4.2 Apparatus

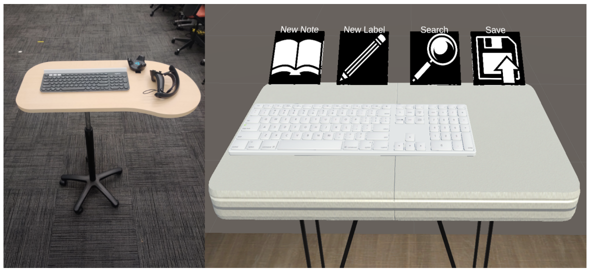


Fig. 2: Left: Text Entry System. A HTC Vive Tracker on the wheeled table, a wireless keyboard, and space for the controller to be set down while typing. The keyboard was replaced with a 2019 Mac Book Pro using Google Docs during report writing. Right: VR Headset view of the tracked keyboard and table. There are four menu buttons attached to the keyboard. These buttons are New Note, New Label, Search, and Save.

A desktop computer was used to run the IST system using an HTC VIVE Pro [1] head-worn display with a wireless adapter to allow for a tether free walking in the space. The participants held one Valve Index wireless controller in their dominant hand for interacting with the system. Participants used a wheeled keyboard table with a wireless keyboard for text input. Small amounts of foam were cut out of the VIVE Pro headset to provide better support for text entry so the participants could see their fingers on the physical keyboard. A 2019 MacBook Pro with Chrome Remote Desktop connected to the desktop PC was used for the report writing. An example of the wheeled table used for text entry and report writing can be seen in figure 2. A Steam VR 2.0 Lighthouse tracking system was used for a tracked area of 3x3 meters. In the tutorial phase of the study, the participants were trained on the size of the tracked space and were also shown that the virtual floor represented the boundary of the tracked area. In addition, the experimenter watched to ensure that the participant did not get too close to the physical room boundaries. The amount of tracked space in this study differed from the prior work by Davidson et al. [11], where the participants had 4 x 8 meters. This was due to the constraint of recruiting professional analysts at a location convenient to them. We highlight this as a limitation and apply normalization when necessary in our analysis to draw comparisons between the two groups.

4.3 Experimental Task

This experiment aimed to understand how professional analysts complete sensemaking in an immersive analytic prototype. Additionally, we aimed to evaluate how professional and novice users differ on structures and strategies used to inform future design ideas for immersive analytics systems. For this comparison, we utilized the same experimental task as seen in [11]. This task asked the participants to analyze a set of text-based documents, of which our participants would have no prior knowledge, to write a report of their findings at the end.

The task provided to the participants was as follows:

"Today is April 27th, 2003. You are an intelligence analyst working for the federal government. It is believed that terrorists are planning an imminent attack on the United States. Other analysts have gathered a set of potentially relevant documents containing information about potential suspects. These documents have been loaded into the Immersive Space to Think system for your analysis. Your goal is to analyze the information and develop a specific hypothesis about any potential planned terrorist attack(s) against the US. Your hypothesis should identify **who, what, when, and where**.

Over the course of 3 sessions, you will develop a report for the Office of Homeland Security. To do this effectively, you must prepare a defensible and persuasive report that describes exactly what your conclusions are based on the documents provided. During session one, you will focus on analyzing the documents and preparing a specific hypothesis of your findings (who, what, when, and where). During session two, you will continue your analysis and prepare a written outline of your findings. During session three, you will finalize your analysis and write a 1-2 page report of your findings. Your report must state what action or actions the terrorist(s) are planning, where they will occur, and when they will occur, using evidence and citations from the documents provided. Our hope is that your report can be used to thwart the terrorist(s) threat."

Participants were notified that the milestones for both the outline and report were designated as goals to be fulfilled by the conclusion of sessions two and three. They were also informed that they could progress ahead or revisit the checkpoints as necessary to support their sensemaking process. Our analysis identified participants who proactively completed the checkpoints ahead of schedule and returned to checkpoints as needed to support their process.

4.4 Dataset

Again, since this study aimed to understand similarities/differences of sensemaking between professionals and novices in our prototype IST, the dataset used in this task is the same as seen in Davidson et al. This allowed us to compare both the strategies used for completing their sensemaking task and the overall spatial structures that were formed as the task was a controlled variable between these two separately studied [11].

This dataset, Sign of the Crescent, is a fictional intelligence analysis dataset containing 40 text-based documents. Each document is about a paragraph in length. The dataset contains many features, including names/aliases, addresses, phone numbers, bank account information, places of business, passport numbers, etc. Additionally, this dataset contains relevant (23) and distractor (17) documents that add complexity to the analysis task.

[1] https://www.vive.com/eu/product/vive-pro/

The dataset contains three major plotlines related to a set of coordination documents. Plotlines boil down to the locations where the events related to the plotline are happening (Atlanta, Boston, or New York). Plotline 1 contains five documents, Plotline 2 contains seven documents, Plotline 3 contains four documents, and there are seven coordination documents. Along with these features, the documents purported to come from three different reporting organizations: FBI, CIA, and Sanctioned Intercepts. There were 24 FBI documents, 11 CIA documents, and 5 Sanctioned Intercept documents.

4.5 Participants

Due to the challenges of scheduling professional intelligence analysts for a user study, we had a relatively small pool with N equal to 11. The analysts were US Department of Defense employees and had an average of 9.72 years of professional intelligence analysis experience, with a standard deviation of 5.12 years. To protect the identities of the participants, age was not collected. However, we view years of analysis experience as a good proxy for age, with the age range representative of the US working-class population. We had six male and five female participants with the following VR/AR experience levels: Two participants had no previous experience, six participants had tried VR/AR once or twice, two participants had tried VR/AR 3-10 times, and one participant had tried VR/AR more than ten times. All participants had normal or corrected vision (glasses or contacts). The participants were recruited using email and word-of-mouth recruiting. The Virginia Tech institutional review board approved this study to ensure participant protection.

4.6 Measures

Before the study began, participants were guided in the consent process and informed that they could withdraw their consent at any time, in which case, any data collected from them would be deleted. After completing the consent process, the participants were asked to complete a **pre-study questionnaire**, which collected background data such as gender, analysis experience, VR/AR experience, general sensemaking strategies, etc.

Log Files - Collected during each session, and containing the participant's head movement, controller movement, keyboard movement, as well as any system-level interactions with the system (i.e., new notes, new label, grabbing document, hovering over a document, highlighting, etc.). All movement data was reported at about ten times/second.

Save Files - Generated once a minute during the session and contains the location of each IST artifact (documents, notes, and labels). We can create minute-by-minute snapshots of the immersive space over the entire sensemaking task using the save files.

Screen Recordings - Taken from the participants' first-person point of view during their session.

Interviews - After each session, interviews were conducted on session-specific questions, and after the study, the participant took part in a post-task interview.

Outlines/Reports - Generated by the participants during the experimental task. The Outlines were generated during session two, and reports were generated during session three within a Google Doc using the laptop placed on the tracked table.

In addition to the data collected for this study, we graded the reports for correctness and quality. For correctness, we evaluated the reports based on the ground trust of the dataset using a rubric created to look at the *Who, What, When, and Where* questions. The rubrics were developed with non-participant professional intelligence analysts. The rubric was designed to produce subjective quality scores, which measured conciseness, persuasiveness, clarity, completeness, relevance, and bias. These rubrics are the same ones used in Davidson et al.'s. previous work [11] and a copy of the report grading rubrics can be found in the appendix.

4.7 Procedure

This study follows the same procedure as seen in Davidson et al. [11]. The study was broken down into three sessions, which are detailed below.

4.7.1 Session 1

Pre-session - Before participant arrival, all equipment used during the study was sanitized to help prevent the spread of virus. Upon arrival, the participants were asked to review the consent document and ask if they had any questions. Upon receiving the consent of the participants, the study began with the background questionnaire.

Tutorial - Before entering the headset, the participants were trained on the Valve index controller and HTC Vive Pro adjustment mechanisms. After this, the participant would don the headset for the tutorial on how to use the IST system. The tutorial began walking around the immersive space to understand the physical boundaries of the tracked area. Then, using a tutorial dataset, the participants were trained to manipulate the documents, use the keyboard/table, and all the system features of IST. After learning all of the features, the participants were given 5 minutes to practice using the features, walking around in the headset, and using the keyboard/table.

After completing the tutorial, the participants were given a quick break where they could take the headset off. During this time, they were provided a printout version of the experimental task, as seen in section 4.3, and the participants were encouraged to ask questions if clarification was needed before beginning the study. Once the participant was ready, they would don the headset and begin the task.

Session - Participants were instructed to start their analysis and work towards an initial hypothesis. Participants were given about 40 minutes for this session. Participants were encouraged to ask questions on the task or the system if they needed help.

Interview - Session 1 concluded with a 5-minute post-session interview to gather information on the strategies and organizational structures used during the analysis task thus far.

4.7.2 Session 2

Onboarding - Before beginning the study portion of the session, the participants were re-familiarized with how to use the equipment and all the adjustments on the headset.

Session - The participants were given 65 minutes to continue their analysis and develop an outline of their findings. If, at the 50-minute mark, the participant had not started their outline, they were instructed to do so.

Interview - Session 2 concluded with a 5-minute post-session interview aimed at collecting information on the strategies and organizational structures used. Additionally, we wanted to understand if the structures/strategies used changed from the previous session.

4.7.3 Session 3

Session - The participants were given 60 minutes to finish the analysis and develop their 1-2 page report of their findings.

Interview - During this 15-minute interview, the participants answered questions about the task, including overall strategy, overall spatial layout, and how the layouts changed over time.

During each session, the participants were given time warnings at the halfway point and 15, 5, and 1 minute remaining. This was to help the participants use their time efficiently and keep track of their progress.

5 Results and Discussion

To evaluate RQ1, we looked at the overall structures formed by the participants during their analysis and how those structures changed across their analysis process. For RQ2, we looked at mapping the 3D layouts and the 1D reports written during the analysis using the document reference (citation) order. For RQ3, we focused on the participant's movement in the physical space during the analysis task. Lastly, for RQ4, we looked at the system-level interactions and overall analysis strategies to understand the relationships between actions and report scores.

5.1 RQ1: Spatial Layouts and Change Over Time

5.1.1 Spatial Layouts

To understand the overall spatial structure formed by the analysts, we use a bird's eye view into the analysis space to understand the high-level organizational pattern used for document layout. In previous work, four high-level organizational patterns have been detected within immersive spaces. Using the updated definitions of these organizations proposed by Davidson et al. [12], these layouts are *Semi-Cylindrical, Cylindrical, Environmental, and Planar*. Semi-cylindrical is defined as having documents curved around the user but not on all sides as seen in figure 3 left, whereas cylindrical layouts have document positions in the full 360 degrees around the user as seen in figure 3 middle. Environmental Layouts use the physical features of the space, such as the floor/bulletin board for organizing the documents onto a plane as seen in figure 3 right, and planar layouts use planes within the space that are not curved around the user while also not matching a physical feature of the space as seen in environmental.

To categorize the overall spatial organization of their analysis space, we look at the final layout created by the participants. This study found four semi-cylindrical layouts, six cylindrical layouts, and one environmental layout.

Organizational Features Digging deeper into the overall structures formed, we can also look at what additional features the participants used to create organization within the spaces. Previous work by Davidson et al. examined five other deeper organizational elements within the final layouts of the participants, including relevance vs distractor, plotline encoding, reporting agency sorting, timeline formation, and trash piles. We examine these features within the space below and two new features not yet found within the final layouts - network graphs and scratch space.

5.1.2 Relevant vs. Distractor Documents

One feature of the space evaluated in the past was the distance of relevant or distractor documents to the center of the tracked area. In the analysis of the professional users, we found that distractor documents had an average distance of 2.256 m (SD = 0.6898 m) to the center of the tracked space, with relevant documents having an average of 2.290 m (SD = 0.6842 m) to the center of the tracked space. At first glance, it seems concerning that the relevant documents are further from the center of the space. However, when looking at the overall organization of these spaces, the participants' schemas (document locations) are not often centered within the tracked space. For this reason, we chose to calculate the center point of the space based on where the participant spent the most time (dwelled) during their analysis. Then, using this new center point, we calculated the relevant and distractor document distances to that point. In this new calculation, we found the distractor documents were, on average, 1.428 m (SD = 0.3283 m) from the center, and relevant documents were 1.388 m (SD = 0.3047) from the center. Using a t-test, we did not find that these were statistically significant (p = 0.1825), but we suggest that future work verify these results with large participant pools.

5.1.3 Reporting Agency Sorting

Another feature we found within the deeper organization strategies was sorting by reporting agency according to the task dataset 4.4. In this study, we had four participants who organized the data into distinct categories based on the reporting agency of the documents. Figure 4, right has an example of the reporting agency sorting where the FBI reporting documents are organized on the bulletin board, the CIA documents are organized on the right-hand side, and the Sanctioned Intercepts documents on the side of the space opposite from the bulletin board.

5.1.4 Timelines

In our analysis of the spatial layout created by the participants, we identified that all participants (11/11) utilized a timeline column during their analysis of the documented set. Figure 4 Left shows an example of two long-column timelines.

5.1.5 Network Graphs

A new feature we discovered in analyzing these spatial layouts was a network graph generated by P10, as seen in figure 4. In their analysis, they used the ceiling of the analysis space and the label feature to create a social network

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3444594

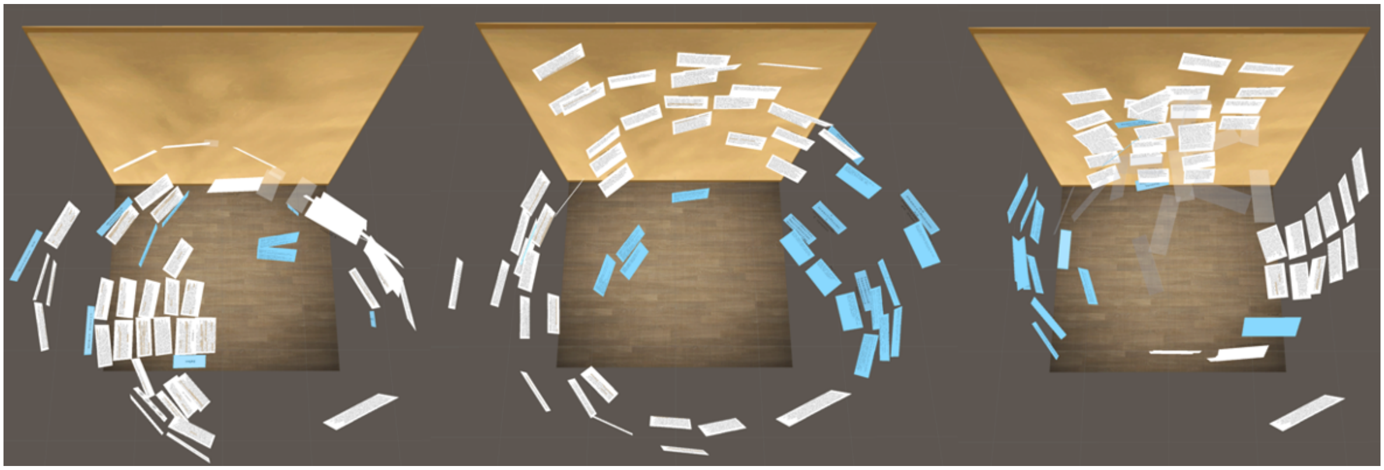


Fig. 3: Top-Down View at High-Level Organizational Layout of the analysis spaces. Left: Semi-Cylindrical Layout, Middle: Cylindrical Layout, Right: Environmental Layout. The white documents are those provided for analysis, and the blue documents indicate a note or label created during the analysis. In the bottom right-hand corner is a copy of the task as detailed in section 4.3.

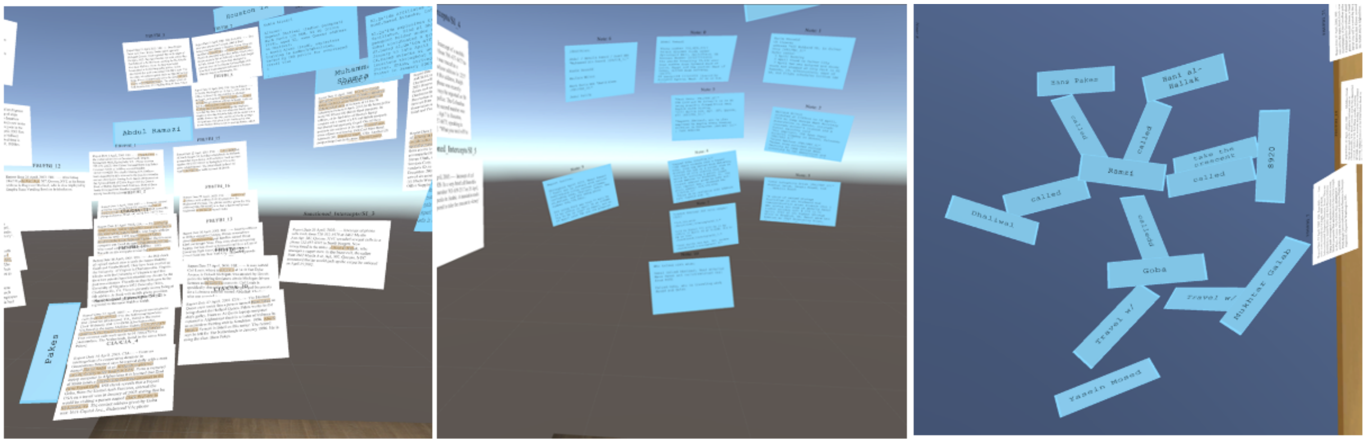


Fig. 4: Deeper Organizational Features found in the analysis spaces. Left: Column/Timeline organization with additional labeling of people of interest within the columns, Middle: Scratch Space for the analyst's comments, Right: Network Graph created with the label feature.

map. For example, the participant had a "Person 1" label with another label, "Called", then another label, "Person 2," allowing them to build a graph of Person 1 → Called → Person 2 →, etc., making the information into a more coherent narrative and glanceable. While we only had one user create a social network graph this way, we want to highlight the use of this deeper organization created by the participant and suggest building support for generating graphs like these in the future.

5.1.6 Scratch Space

Lastly, in our analysis of the deeper organizations used within the space, we also identified the use of scratch space with three of the participants. An example of scratch space can be seen in figure 4 middle. In scratch space, a designated section of the analysis space was used to organize the analyst's notes/comments, separate from the provided text documents. This approach to managing their notes created a division between the ground truth of the dataset (i.e., the provided text documents) and the analyst's thoughts/comments (i.e., interpretation of the ground truth provided).

5.1.7 Change Over Time

To understand how IST supports different stages of the sensemaking task, we wanted to understand the following:

1) How do the spatial layouts change across the analysis task?
2) How does document interaction change across the analysis task?
3) How do these trends compare to those found in a previous study using novice analysts?

To understand how IST supports the overall sensemaking task, we look at how the spatial layouts created by the participants change across the analysis sessions. In doing so, we can look at the evolution of the schemas formed during the sensemaking process. An example of how P5's spatial structures evolved across the analysis can be seen in figure 1. In this stepped evolution, we can see a semi-cylindrical layout begin to form by the end of session 1, all of the documents being removed from the bulletin board by the end of session 2, and some columns/timelines forming on the back side of the tracked space into the end of session 3. In addition to visually inspecting the spaces and how they changed across the sensemaking task, we wanted to evaluate how the users grouped information over the sessions and if those groups became more distinct (i.e., more organized). For this analysis, we leverage clustering algorithms as seen in previous work to understand how the overall organization changed over time [12]. For our analysis, we ran OPTICS (Ordering Points to Identify the Clustering Structure) clustering [36] on the final save of each session created by the participants. We chose to use the OPTICS clustering technique because it performed well at clustering the columns and clusters formed by the participants when we compared it to K-Means [37], Hierarchical Clustering Techniques [38], and DBSCAN [39]. In addition, OPTICS is a density-based clustering technique that also detects noise (documents too far away from others to be considered part of a cluster) within the algorithm. In our OPTICS analysis, we used min samples=2, metric=l1 with the remainder of the parameters set to the defaults within [40]). An example of the clustering produced by OPTICS can be seen in figure 5

To evaluate the clustering over time, we utilized two different clustering evaluation methods - Silhouette Scores [41] and Calinski Harabasz (CH) Scores [42]. Silhouette scores range from -1.0 to 1.0 and indicate how close a point or document is to its own cluster (cohesion) compared to the other clusters (separation). CH is also known as the variance ratio criterion and is an indicator for the sum of between-cluster variance and within-cluster variance for each cluster. With CH, the higher the score, the better the overall clustering, and for silhouette scores, the closer to 1.0 indicates a very *strong* clustering. Overall, we expected to see that the silhouette and CH scores would increase across the sensemaking task.

The average silhouette scores were 0.32, 0.40, and 0.42 for sessions 1, 2, and 3, respectively. We performed an Analysis of Variance (ANOVA) test with a random effect of participant to account for individual differences between participants, using Tukey's HSD method for post-hoc comparisons which adjusts for multiple comparisons and controls the overall Type I error rate. We found a significant difference in silhouette scores (F(2,20) = 4.8452, (p = 0.0192)). The post-hoc tests revealed a statistically significant difference between session 1 and 3 (p = 0.0214) with session 3 higher than session 1, and a trend towards significance between sessions 1 and 2 with session 2 higher than session 1 (p = 0.0690). There was no difference between sessions 2 and 3 (p = 0.8403). The average CH Scores were 43.03, 54.55, and 56.49 for sessions 1, 2, and 3, respectively. We found a significant difference in CH scores (F(2, 20) = 5.025 (p = 0.0171)). The post-hoc tests revealed a statistically significant difference between session 1 and 3 (p = 0.0215) with 3 higher than session 1, and a trend towards significance between session 1 and 2 (p = 0.0522) with session 2 higher than session 1. There was no difference between sessions 2 and 3 (p = 0.9060). Combining the overall change over time graphics as seen in figure 1 and the clustering analysis, we can see that these spaces are 1) constantly changing, 2) becoming more organized during the analysis, and 3) most interestingly, the spaces (spatial structures) are refined up to the final stages of sensemaking where the participant is preparing the presentation of their findings in the written report.

Next, we were interested in understanding the overall trends of the document movement across the sensemaking sessions. Based on previous analysis, we expected to see a decrease in the distance a document moved across the session and the number of documents moved in each session. The analysis of document movement trends over each session can be seen in figure 6a and 6b. Overall, we saw a decrease in the number of documents moved per session from sessions 1-3 and 2-3. We also saw that the distances those documents moved decreased from sessions 1-3. Using an Each Pair Students T-Test with Bonferroni's correction, we saw no statistical differences between distance or count moved from sessions 1-2.

5.1.8 Novice vs. Professional Analysts

Finally, to address RQ1C, we look at how the professional analyst populations' overall trends compare to the novice analysts' identified trends in previous work. First, in high-level spatial structure analysis, we see that the analysts' population structures match those identified in previous work. The analysts formed semi-cylindrical, cylindrical, and environmental layouts within the immersive space. In addition to seeing trends between the populations with the high-level structures used for document organization, we also found some trends within the deeper organizational features of the immersive spaces. One feature we saw every analyst use was timelines/columns, which was also seen with 6/8 novice users.

A feature we saw with only one novice, sorting by reporting agency, was identified in 4/11 analyst sensemaking layouts. This work also uncovered two new organizational features that were not identified with the novice population: Network Graphs (1/11) and Scratch Space (3/11). While these deeper organizational features are not as popular as the column/timeline pattern, we believe they are essential to support future IST systems designs because they give the analyst more control over the organization or location of details within their space.

Additionally, our analysis of the professional analyst population's spatial structures identified similar patterns for how the spaces evolved through incremental refinement. Overall, most of the documents are moved in analysis early on in the sensemaking process (sessions 1 and 2) with a decrease in the count but not zero movement into the later stages (session 3). In addition, we see a decrease in the distances that these documents move from session 1 to 3 with less overall distance as the analyst moves further into the sensemaking process.

In summary, we found that the analyst population utilized the same high-level structures previously identified within IST analysis. This confirms that IST systems may benefit from previously proposed assistive high-level organizational features. Additionally, the analyst population utilized the same deeper organization features previously identified in IST analysis, and some new organization features not

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3444594

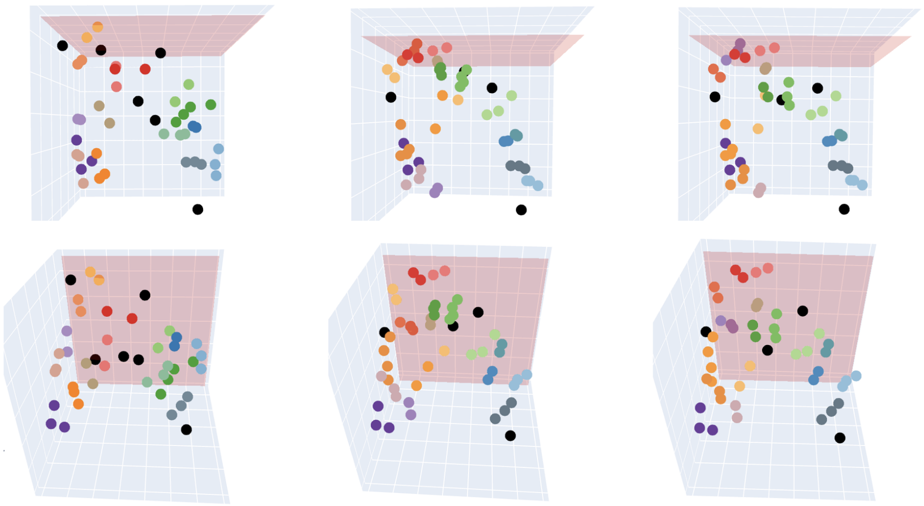
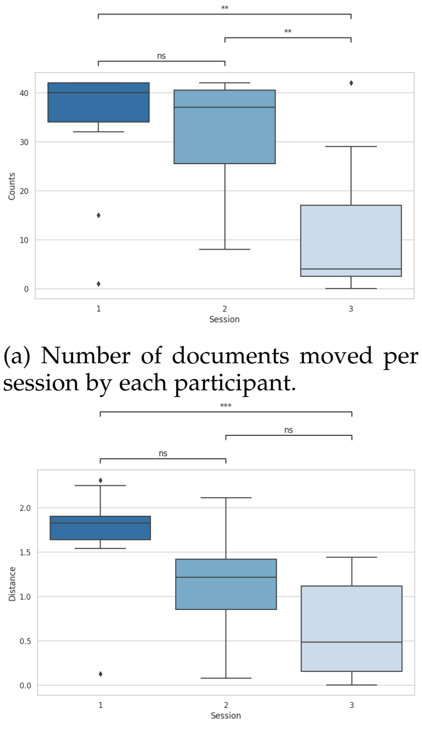


Fig. 5: From left to right, End of Session One, End of Session Two, and End of Session Three of P4's analysis space. The top row is a top-down view into the space, and the bottom row is a rotated view looking towards the bulletin board. The colors in this image indicate the clustering applied to each document by the OPTICS algorithm. The black documents were "noise" detected by the OPTICS algorithm.



(a) Number of documents moved per session by each participant.

(b) Average distances the participant moved a document during each session.

Fig. 6: The total number of documents moved was significantly different from session 1 to 3 (p = 0.004) and again from session 2 to 3 (p = 0.007) There are also significant differences in the average distance a document moved from sessions 1 to 3 (p = 0.0007) using an Each Pair Students T-Test with Bonferroni's correction.

yet identified. Using these findings, we can verify many of the previous results from IST systems with professional users and provide new insights into the features that can be developed to better support the overall sensemaking process in systems like IST.

5.2 RQ2:Mapping of Spatial layouts to Report Structure

Along with an in-depth analysis of the spatial structures and organization of the analysis space, we were also interested in understanding how the analysis space mapped to the reports written by the participants. The objective of this analysis was to evaluate the potential of future assistive report-generating features (e.g., automatically generating a report outline based on the spatial layout of documents in IST). In the instructions given to the participants, they were asked to include references to reports when citing information from them. Then, using these references, we generated a reference order of the reports (e.g., CIA_8, FBI_2,....SI_4). Then, using this ordered list of references, we located the document within the immersive space in the 3D scatter plots and drew a line through the immersive document layout, as we can see in figure 7. This line represents the documents directly referenced within the written report, which was then colored based on which paragraph the document was cited in (starting with red and ending in pink).

We drew this line on each final layout and then calculated the average total distance along the reference path of 11.25m (sd = 5.0974) for each participant. This total distance represents the distance between the first point to the second point, the third point, etc. Additionally, we looked at the distance along the reference path normalized by the total pairwise distance between all documents within the space (spread), which had an average of 0.00823m (sd = 0.0045). This value gives us a better understanding of the total distance along the path relative to the total spread of documents within their spatial layout.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3444594

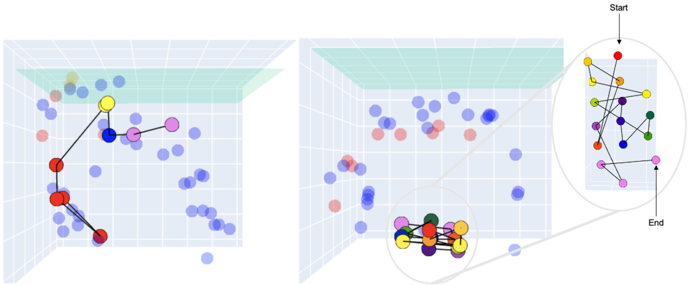


Fig. 7: Left: P3 spatial layout with a reference line drawn through it. The green mesh indicates the bulletin board. The red, opaque documents are cited in paragraph one, the yellow documents in paragraph 2, the blue in paragraph 3, and the pink in paragraph 4 of the report. Right: P5 reference order of moving relatively top-down through the columns within their analysis space for citing documents. The zoomed portion is oriented as if the participant is standing in the center of the analysis space, staring at documents. Most Red, Orange, and Yellow documents are on the top of the columns, with Pinks, and Purples and Blues toward the bottom of the columns. In this figure, the documents are cited in paragraph order Red, Orange, Yellow, Green, Blue, Indigo, Violet, Pink (ROYGBIV + Pink).

The next feature of the reference lines that we evaluated was the shape of the pathway through the space. In previous work, three pathways were presented: left → right, right → left, and no clear path, meaning there were many intersections along the line with no clear shape to the line. For clarity, we will rename these to clockwise (left → right) and counterclockwise (right → left). An excellent example of a clockwise reference pattern can be seen in figure 7 left. The red documents are all cited within the first paragraph, and then we can follow the reference order in a clockwise pattern through to the pink documents, which are cited in the last paragraph together. In our analysis of the professional analysts, we found five roughly clockwise patterns, three roughly counterclockwise patterns, one new pattern, which we call top-down, and two participants with no distinct pattern. In the top-down pattern, the participant referenced documents at the top of columns created in their analysis first, then they worked their way down the columns toward the bottom. Figure 7 right shows an example of the top-down layouts.

5.2.1 *Novice vs. Professional Analysts*

In looking at the document reference paths in the analysis space, the novice reference paths were longer than those of the professional analysts. However, this was due to the total tracked analysis space of the novices being a 4x8 meter tracked space compared to the 3x3 meter tracked space provided for professional analysts. When we look at the total normalized values of the reference line, we do not see any statistical significance between the professional and novice user populations. However, when we look at the reference line's overall shape, we see more professionals with a roughly identifiable transformation path than the novices. In the shape analysis, we found that 9/11 of the professional analysts had a clockwise, counter-clockwise, or top-to-bottom transformation path compared to the novices, where only 4/8 had a roughly clockwise or counter-clockwise path. This could be due to the professional analysts having a better organization overall than the novice users or a better ability to cluster relevant information for the report writing.

5.3 RQ3: Physical Navigation

One of the study's goals was to understand how the analysts navigate the immersive space (i.e., walking) while they complete the sensemaking task. To evaluate the participants' physical navigation, we tracked their movement during their analysis to understand the space usage while they were completing their sensemaking. In our files, we collected the headset position x, y, z, yaw, pitch, and roll ten times per second while the participant was in the headset. We took the average headset position, once per second, to use in the following analysis.

5.3.1 *Professional Analysts Physical Navigation*

Across the entire analysis task, we found that the participants navigated 434.32 m (sd = 138.64) of the space. With breakdowns of 131.04 m (sd = 36.38 m) in session 1, 201.78 m (sd = 75.92 m) in session 2, and 101.49 m (50.04 m) in session 3. To account for variation in the amount of time spent in the headset, we also present the amount navigated normalized by number of seconds spent in the headset: total average normalized 0.11 (sd = 0.050), session 1 normalized 0.05 (sd = 0.02), session two normalized 0.04 (sd = 0.015), and session three normalized 0.03 (sd = 0.009). Additionally, we visualize this data in figure 8.

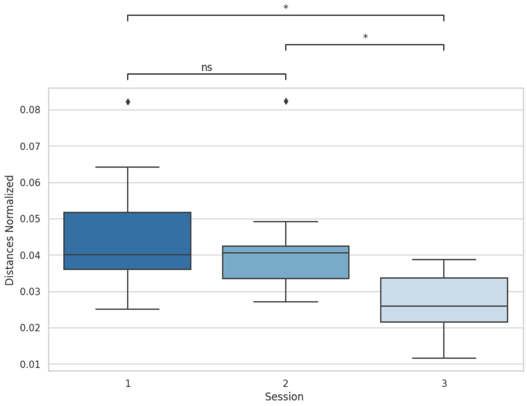


Fig. 8: Movement data normalized by time in the headset. We can see a decrease in movement over time with significant differences between sessions 1 and 3 (p = 0.02) and 2 and 3 (p = 0.03) using an Each Pair Students T-Test with Bonferroni's correction.

When we look at the data normalized by the amount of time spent in the headset, we can see a downward trend in navigation as the participants work into the later stages of the sensemaking process. We see no statistically significant difference in the amount navigated between sessions one and two. However, there was a difference in the amount navigated between sessions 2 to 3 ($p < 0.05$) and sessions 1 to 3 ($p < 0.05$) using Each pairs Student's T-Tests for pairwise comparisons with Bonferroni's correction.

Considering the task provided, it makes sense that the participants spent less time navigating the space during session three, which focused on report writing. However, when we look at the average navigation (pre-normalization), we can see 101.4930 m on average were navigated by our participants during that session, showing that, even in the late stages of the sensemaking process, using physical navigation to access externalized memory to be used by our participants.

In addition to looking across the participants for overall trends in movement, we also looked deeply at each participant's movement trajectory as seen in figure 9 and their dwelled areas as seen in figure 10. One trend we found when we looked deeper at the individual participants was two groups of participants when it came to navigation within the immersive space. We had participants who navigated more of the space - **Movers**, and then we had participants who were primarily stationary during their analysis - **Stationary**. This is supported by the large standard deviation we found in the above data. In figure 9, we can see a mover on the top row compared to a stationary participant on the bottom row. Overall, we had six movers and five stationary participants. In our analysis, we looked at both the shape of the movement trajectory and the amount of space navigated by the participant.

In addition to looking at the overall trajectory of the movement, we also looked at where the participants spent time during their analysis, which we define as dwell. When we look at the overall trends of dwellings, we see a decrease in the number of intense dwelled locations across the sessions, indicated by a darker green color in the heatmaps. For instance, with P6 in figure 10, we can see the number of locations with intense green (more time spent in that location) decrease across the session to one primary dwell location during the final session. While this trend is more pronounced in the mover group of participants, we also saw their trend across the stationary participants.

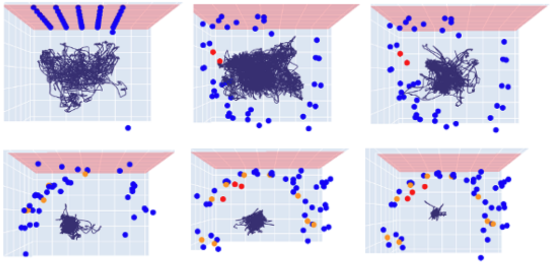


Fig. 9: From left to right, session 1, session 2, and session 3 for P6's (top) and P11's (bottom) analysis. The red plane represents the bulletin board, the blue points indicate documents, the red colored points are notes created by the user, and the yellow points are labels created by the user.

5.3.2 Novice vs. Professional Analysts

Overall, the patterns we saw with physical navigation were similar to those of our novice users.

In our analysis of the professional analyst movement in IST, we saw a couple of trends. First, we saw decreased movement as the participants moved further in the sensemaking process. Additionally, we saw two trends in movement style in the space - **Movers vs Stationary** users. Lastly, we saw trends in the decreased number of intense dwell spots over time, as seen in our analysis of time spent in the different regions of the analysis space.

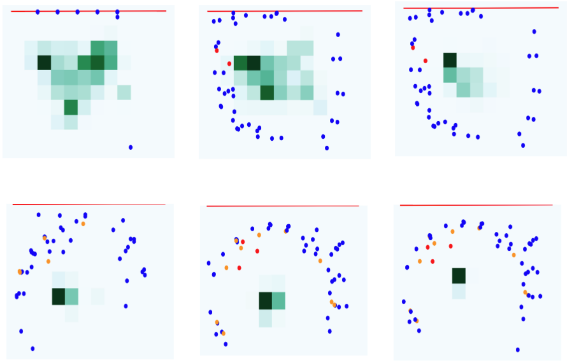


Fig. 10: From left to right, session 1, session 2, and session 3 for P6's (top) and P11's (bottom) analysis. The red plane represents the bulletin board, the blue points indicate documents, the red colored points are notes created by the user, and the yellow points are labels created by the user. The heat maps show the time spent (dwell) within a one sq ft area of the tracked space.

When we compare these findings to those reported in previous work [11], we see many similarities between novice and professional analysts. When looking at the normalized movement data, we see a decrease in movement from sessions 2 to 3, which is supported in both the analyst and novice populations. However, with the professional analysts, we also saw a decrease in movement from sessions 1 to 3, which we did not see in our novice population.

Another similarity we found was with the patterns in the trajectory of movement between the analysts and the novices. In both cases, there was a distinction between the movers who navigated most of the space while conducting their analysis and the stationary who tended to stay in one spot. We also found similar trends in the dwell analysis that looked at time spent within square foot regions of the space. Overall, we saw a decrease in the number of dwelled locations as the participants progressed further into the sensemaking process, with the caveat that this trend is more evident within the **mover** participant group than with the stationary group.

Overall, the analysis of movement data between the professional and novice analysts confirmed the findings by Davidson et al. [11]. Overall, the participants tend to decrease movement as they advance into the later stages of the sensemaking process. There are two types of analysis styles when it comes to movement within the IST system, and we see a decrease in the number of dwelled areas, more obviously in the movers; however, we do see similar trends in the more stationary participants.

5.4 RQ4: Strategies and Report Scores

Our last research question, RQ4, aimed to understand if specific spatial layouts, strategies, or interactions within

IST correlate to better quality sensemaking. We needed an indicator for sensemaking quality for our analysis, so we relied on the reports written by the analysts during session three of their analysis. We evaluated these reports for two criteria - Correctness and Quality using the rubrics described in section 4.6.

Correctness - To grade the overall correctness of the reports, the experimenter evaluated each report in a randomized order. A point was awarded for each correct piece of information provided in the report based on the ground truth of the dataset.

Quality - To grade the reports for overall quality, we worked with a set of professional intelligence analysts. Before grading the reports, the experimenter met with the analysts to onboard them to the grading process and provide them with detailed instruction sheets including a document that specified the order to grade the reports (randomized for each grader). The graders were then provided a PDF of each of the reports written by the participants, and the grading rubric via a Google Form. To ensure the graders were provided the same materials as those who graded the novice participants' reports, the analysts who graded these reports were also provided with *Getting Started* materials that included the task the participants were given 4.3, the solution to the analysis, and the documents used during the analysis, which they were able to reference if needed during the grading process. The analysts were given a month to find flexibility in their work lives to complete their grading and asked to complete their grading within 1-week after starting. In total, we had three professional analysts complete the quality report grading.

5.4.1 High-Level Strategies

First, we take a high-level look at the overall strategies that the professional analysts utilized for their sensemaking within IST. For these high-level categorizations, we build on the previous work of Davidson et al., and Kang et al. [12], [43] In this study, we identified three high-level strategies for the analysts' sensemaking. "Build from Detail" (S1 in Davidson et al.), where the participants read all documents first, then sort the documents into groupings, and refine those groupings as needed. Overall, we have two participants who employ this sensemaking strategy. Next, we have "Hit the Keywords" (S3 in Davidson et al.), where the participants skim documents for keywords, execute searches on those keywords, and then group and refine based on those queries. We had 6 participants employ this sensemaking strategy. Next, we had "Overview, Filter, and Details" (S4 in Davidson et al.), where the participants get an overview of the content of the documents first, filter or sort the documents into rough groupings (normally based on reporting agency information) and then revisited relevant documents for details and deeper understanding. We had three participants employ this method.

In our analysis of these strategies, we attempted to correlate the strategies and overall sensemaking performance. We did not find any statistically significant correlation between the overall sensemaking strategy used and the report's correctness or quality scores scores.

5.4.2 Note Taking

Another item we were interested in was the use of notes during the analysis process. In our comparison of the two populations, we found that the total notes created by analysts (mean = 5.09, SD = 4.13) were more than that of the novice analysts (mean = 2.63, SD = 1.41) using a t-test (p=0.0446). However, the total number of notes does not indicate the quality or content of the notes.

Therefore, to dig a bit deeper into this, we were interested in how we can evaluate the content of the notes to understand if there are more profound differences in the types of notes used during this analysis. To evaluate the quality of the content of the notes, we looked at the total number of notations (all marks made in the notes), the total number of words (words, symbols, and abbreviations of the notes), and lastly, the total number of content words (nouns, verbs, adjectives, and adverbs) identified by Siegel [44]. For this analysis, we used the following steps: 1) get all notes (analyst and novice), 2) drop any empty notes, 3) count the number of notes for each participant, 4) combine all of the notes' content into a text file, 5) create a cleaned version of note content, 5) count the total number of characters in note content, 6) split the note content into words, 8) count the words, 9) apply the Natural Language ToolKit's (NLTK) part-of-speech tagging to each word in the notes [45], 10) count the total number of parts of speech, and 11) normalize each part of speech by total word count.

This analysis found two interesting differences between the novice and professional analysts' notes. The first difference we detected was that the professional analysts (mean = 130.21, SD=63.73) used more nouns than the novices (mean = 72, SD= 93.10) (p=0.0406) and the professionals (mean = 11.77, SD = 7.78) utilized more digits than the novices (mean = 2.73, SD = 4.38) (p = 0.0045). This finding was quite interesting since we did not find any difference in the notes' total number of characters or words. Indicating the analyst notes contained more pertinent content relating to people, places, things, phone numbers, addresses, and bank account information than the novices.

5.4.3 Overall Highest vs Lowest Scorers interactions

Taking our analysis further, we wanted to understand trends/differences between the overall highest and lowest scorers. For this analysis, we combined the data from the professional intelligence analyst and the novice users. Then, we took the five highest (4 Analyst, 1 Novice) and five lowest (1 Analyst, 4 Novice) scores from the combined population. Using the combined data, we created ridge line plots looking at the interaction frequency of each log file interaction (e.g., highlights, searches, grabs, etc.) between the two groups. We normalized each interaction between 0 and 1 for comparison purposes, and the resulting ridgeline plot can be found in figure 11.

One interesting difference we found was that the highest scorers tended to use the copy-to-clipboard features earlier in the sensemaking process than the lower scorers. Another interesting difference between the highest and lowest scorers was using notes. The highest scorers tended to create new notes more frequently than the lower score during sensemaking. Interestingly, however, lower scorers tended

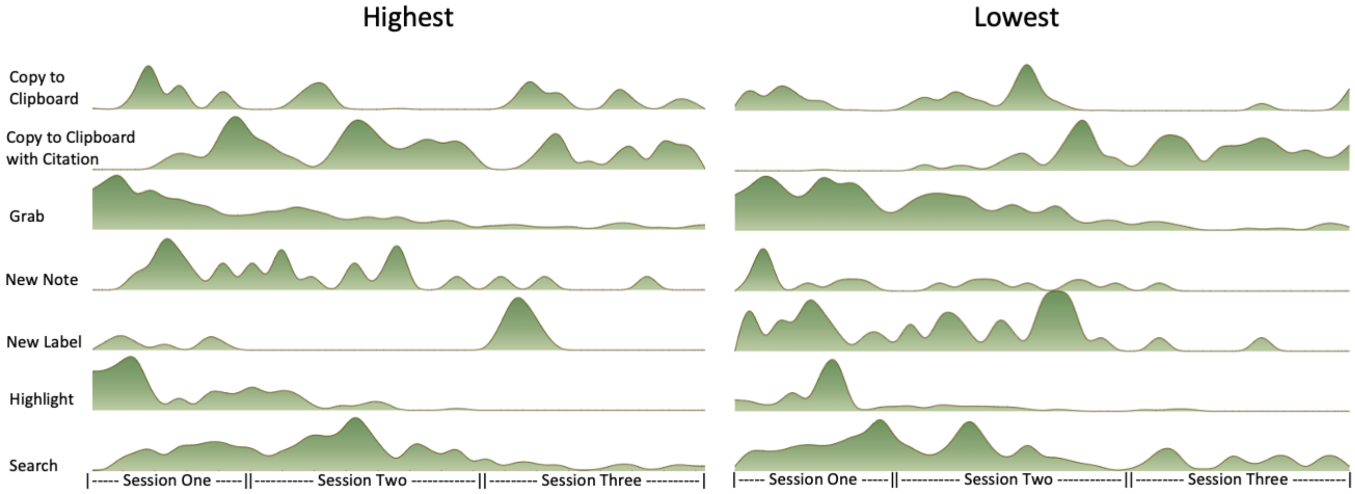


Fig. 11: Left - Highest scorers, Right - Lowest Scorers. The x-axis represents time within each of the sessions from the start of the analysis to the end of the analysis. The y-axis is the frequency of each interaction normalized between 0 and 1.

to produce more labels than the higher scorers. Another interesting difference we detected between our highest and lowest scorers was the frequency of searches during the third analysis session. The lower scorers tended to rely on the search feature more during this last analysis session, which could indicate that they relied on the software to locate critical information they planned to use in the report writing versus relying on their externalized memory to access that information.

5.4.4 System Log Interactions

Another feature of the analysis we were interested in was whether specific interactions with the system during the analysis (e.g., highlighting, searching, note-taking, etc.) correlated with higher or lower performance on the sensemaking task. The idea is to learn from what the professional analysts did during their sensemaking to guide what leads to effective sensemaking within systems like IST. In this analysis, we found two statistically significant correlations. First, we found that the total number of notes positively correlated (0.5197) with the report's correctness (p=0.0259). Additionally, we found that the number of notes created during session one positively correlated (0.5640) with report correctness (p=0.0188).

5.4.5 Novice vs Professional Analysts

We found similar trends when comparing the professionals to the novice analyst populations studied before. When we look at the high-level strategies, we identify very similar high-level strategies to those found with the novice users. In our deeper analysis of feature usage, we found that the professional analysts produced more notes with more relevant content during their sensemaking than the novice users.

In addition to the high-level strategies, we looked deeper at the system-level interactions through the log file interactions. In this analysis, we found differences in the quality of the notes between the two populations and correlations between note creation and the correctness score. Lastly, we presented some trends in interaction frequency between the highest and lowest scorers across the two populations. Including search frequency late into the sensemaking process, note creation, label making, and copying to the clipboard.

5.5 Key Takeaways

At a high level, our analysis identified many similarities between the novice and professional user populations. There are several possible reasons for these similarities. One interpretation is that the sensemaking affordances and features of IA systems, such as IST, enable novice users to perform similarly to professionals in sensemaking tasks. Perhaps the ability to externalize the analysis process by organizing documents and annotations in 3D space helps novices think more systematically and, therefore, perform more like a trained analyst. Another perspective is that when professional users engage in an analysis task using an unfamiliar system like IST, they become novices within that system despite their professional sensemaking skills. This speculation is based on the familiarity of the tool to the participants. For instance, novice users have no expectations about how intelligence analysis tools should function, while professionals work with specific tools daily, expecting them to support their analysis process optimally. Therefore, when presented with IST, professional users are asked to perform an analysis task unfamiliarly, without their familiar tools.

Similarly, another interpretation regarding our results could be rooted in the analyst's ability to apply their expertise/domain knowledge during this task. In intelligence analysis, analysts often focus on specific areas, building domain expertise, similar to a student concentrating research efforts on a particular topic (e.g., Immersive Analytics) during their dissertation. However, in our study, professional analysts complete a fictional analysis task where their specific domain expertise cannot be applied. This suggests that we removed some domain-specific elements from the professional analysts' analyses, leading them to perform more similarly to novices within the system. While all three interpretations likely have some merit, since the overall correctness scores were low, we suggest that the second and third interpretations are most likely. In other words, we speculate that IST does improve the analysis performance of novices [8] but also reduces the analysis performance of professionals, at least initially.

In our study, we observed similarities between the professional and novice users. However, we also noticed some intriguing differences between these populations. One possible interpretation of these differences is that, while the professional users may have experienced some reduction in familiarity and domain-specific knowledge, these professionals are skilled in sensemaking and the analysis process, which could have contributed to identifying these interesting and novel differences, such as the analyst who built a network graph or use of the utilization scratch space to keep *analysts comments* separate from the ground truth of the dataset.

Overall, we believe our results suggest that novice user populations can provide our research community with novel insights into immersive analytic sensemaking in traditional user studies. However, to gather these insights during an inherently challenging task such as text-based sensemaking, it is essential to identify highly motivated novices, i.e., those interested in the analysis task. This can be facilitated by providing user-study payment [11] or recruiting participants interested in the experimental task subject [9], [32]. Additionally, we believe that our results could suggest a need for more genuine longitudinal user studies with professional users, which could allow us to see more significant differences between user study populations due to the longitudinal nature, allowing the professionals to get over the unfamiliarity of using the system during analysis. However, there are many logistical difficulties in running a genuinely longitudinal study. Lastly, in a similar vein, we believe that in other non-traditional user studies (e.g., expert feedback sessions or in-the-wild case studies), it is essential to gather feedback from professional users to gain insights into how a prototype can better support them in their daily tasks.

6 Limitations

We recruited 11 professional analysts with an average of 9.72 years of analysis experience. However, this is still a small sample size regarding power in statistical analysis. Due to this, we report all of our findings with caution that future work should be done to help validate the results.

Additionally, to recruit professional intelligence analysts in a convenient way, we set up the experiment in a room accessible to the analysts on the campus where they worked. Unfortunately, this room only had about 3x3 meters of open space for tracking, which led to a varied amount of tracked space between the professionals and novice populations. Where necessary in this paper, normalization has been applied to draw comparisons between the groupings. We do not believe this has affected the results of this study but still present it as a possible limitation.

Another limitation of this work was surrounding the text-entry system used to follow the user-study design of Davidson et al.'s prior work [11]. A slightly inaccurate keyboard model was placed on the VR environment table, to help participants locate the keyboard on the tracked table. This may have caused challenges for participants when entering text within the system. Future work in immersive analytics could utilize AR or AR pass-through portals for text entry, as seen in Giovannelli et al.'s work [46].

7 Conclusions and Future Work

This paper aimed to fill a gap in the existing literature by providing an in-depth evaluation of professional analyst sensemaking in an immersive analytic system. Our analysis found that the high-level organizational structures used by the analysts were similar to those discovered using novice user populations. Additionally, we found that the professional analysts utilized all of the previously discovered deeper organizational patterns presented in Davidson et al.'s previous work [11], adding two new features: Network Graphs and Scratch Space. Additionally, we found that the overall movement patterns of the professional analysts matched those of the novice users. Lastly, we identified new trends between the highest and lowest scores on the reports produced at the end of the sensemaking task.

Using these findings, we speculate on why we identified many similarities within the user study populations and suggest that for future work in more traditional user studies, identifying motivated novice users can be sufficient in our community to gather key insights into user interactions within IA systems. We also suggest features that can be designed to support sensemaking in immersive analytic systems better. For instance, linking features can be added to allow users to externalize their cognition in new ways to help better the ability to create graphs. Additionally, when we look at the mapping of the 3D spatial layouts and the 1D reports of professional analysts, we have identified more distinct mapping patterns that show promise in creating assistance for report writing in this late-stage part of sensemaking. Future work could be done to examine how to transform the 3D spatial structure into a report outline using the information-rich data of document placement and generative large-language models. Additionally, more work can be done to examine how additional organizational features assist in the sensemaking process. Lastly, to help unify the two sensemaking subloops, work on semantic interaction could be a fruitful next step in immersive analytic research.

Acknowledgments

This work was supported in part by NSF I/UCRC CNS-1822080 via the NSF Center for Space, High-performance, and Resilient Computing (SHREC).

References

[1] D. A. Bowman and R. P. McMahan, "Virtual reality: How much immersion is enough?" *Computer*, vol. 40, no. 7, pp. 36–43, 2007.

[2] K. Marriott, F. Schreiber, T. Dwyer, K. Klein, N. H. Riche, T. Itoh, W. Stuerzlinger, and B. H. Thomas, *Immersive Analytics*. Springer, 2018, vol. 11190.

[3] C. Andrews, A. Endert, and C. North, "Space to think: large high-resolution displays for sensemaking," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2010, pp. 55–64.

[4] J. Hollan, E. Hutchins, and D. Kirsh, "Distributed cognition: Toward a new foundation for human-computer interaction research," *ACM Trans. Comput.-Hum. Interact.*, vol. 7, no. 2, p. 174–196, Jun. 2000. [Online]. Available: https://doi.org/10.1145/353485.353487

[5] M. Cordeil, A. Cunningham, T. Dwyer, B. H. Thomas, and K. Marriott, "Imaxes: Immersive axes as embodied affordances for interactive multivariate data visualisation," in *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. ACM, 2017, pp. 71–83.

[6] Y. Yang, T. Dwyer, K. Marriott, B. Jenny, and S. Goodwin, "Tilt map: Interactive transitions between choropleth map, prism map and bar chart in immersive environments," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 12, pp. 4507–4519, 2021.

[7] A. Batch, A. Cunningham, M. Cordeil, N. Elmqvist, T. Dwyer, B. H. Thomas, and K. Marriott, "There is no spoon: Evaluating performance, space use, and presence with expert domain users in immersive analytics," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 536–546, 2019.

[8] L. Lisle, X. Chen, J. E. Gitre, C. North, and D. A. Bowman, "Evaluating the benefits of the immersive space to think," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 2020, pp. 331–337.

[9] L. Lisle, K. Davidson, E. J. Gitre, C. North, and D. A. Bowman, "Sensemaking strategies with immersive space to think," in *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, 2021, pp. 529–537.

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

[11] K. Davidson, L. Lisle, K. Whitley, D. A. Bowman, and C. North, "Exploring the evolution of sensemaking strategies in immersive space to think," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–15, 2022.

[12] K. Davidson, L. Lisle, I. A. Tahmid, K. Whitley, C. North, and D. A. Bowman, "Uncovering best practices in immersive space to think," in *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2023, pp. 1094–1103.

[13] S. Cohen, J. T. Hamilton, and F. Turner, "Computational journalism," *Communications of the ACM*, vol. 54, no. 10, pp. 66–71, 2011.

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

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

[16] A. Endert, W. Ribarsky, C. Turkay, B. W. Wong, I. Nabney, I. D. Blanco, and F. Rossi, "The state of the art in integrating machine learning into visual analytics," in *Computer Graphics Forum*, vol. 36, no. 8. Wiley Online Library, 2017, pp. 458–486.

[17] L. Bradel, C. North, L. House, and S. Leman, "Multi-model semantic interaction for text analytics," in *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2014, pp. 163–172.

[18] A. Endert, P. Fiaux, and C. North, "Semantic interaction for sensemaking: inferring analytical reasoning for model steering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2879–2888, 2012.

[19] D. Sacha, M. Kraus, D. A. Keim, and M. Chen, "Vis4ml: An ontology for visual analytics assisted machine learning," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 385–395, 2018.

[20] T. Chandler, M. Cordeil, T. Czauderna, T. Dwyer, J. Glowacki, C. Goncu, M. Klapperstueck, K. Klein, K. Marriott, F. Schreiber *et al.*, "Immersive analytics," in *2015 Big Data Visual Analytics (BDVA)*. IEEE, 2015, pp. 1–8.

[21] M. Slater and S. Wilbur, "A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments," *Presence: Teleoperators and Virtual Environments*, vol. 6, no. 6, pp. 603–616, 12 1997. [Online]. Available: https://doi.org/10.1162/pres.1997.6.6.603

[22] N. ElSayed, B. Thomas, K. Marriott, J. Piantadosi, and R. Smith, "Situated analytics," in *2015 Big Data Visual Analytics (BDVA)*. IEEE, 2015, pp. 1–8.

[23] A. Rowden, S. Aslan, E. Krokos, K. Whitley, and A. Varshney, "Waverider: Immersive visualization of indoor signal propagation," in *Proceedings of the 2022 ACM Symposium on Spatial User Interaction*, 2022, pp. 1–12.

[24] K. A. Satriadi, B. Ens, M. Cordeil, T. Czauderna, and B. Jenny, "Maps around me: 3d multiview layouts in immersive spaces," *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. ISS, nov 2020. [Online]. Available: https://doi.org/10.1145/3427329

[25] C. Donalek, S. G. Djorgovski, A. Cioc, A. Wang, J. Zhang, E. Lawler, S. Yeh, A. Mahabal, M. Graham, A. Drake *et al.*, "Immersive and collaborative data visualization using virtual reality platforms," in *2014 IEEE International Conference on Big Data (Big Data)*. IEEE, 2014, pp. 609–614.

[26] M. Cordeil, T. Dwyer, K. Klein, B. Laha, K. Marriott, and B. H. Thomas, "Immersive collaborative analysis of network connectivity: Cave-style or head-mounted display?" *IEEE transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 441–450, 2016.

[27] B. Lee, X. Hu, M. Cordeil, A. Prouzeau, B. Jenny, and T. Dwyer, "Shared surfaces and spaces: Collaborative data visualisation in a co-located immersive environment," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2020.

[28] B. Ens, S. Goodwin, A. Prouzeau, F. Anderson, F. Y. Wang, S. Gratzl, Z. Lucarelli, B. Moyle, J. Smiley, and T. Dwyer, "Uplift: A tangible and immersive tabletop system for casual collaborative visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1193–1203, 2020.

[29] W. Luo, A. Lehmann, H. Widengren, and R. Dachselt, "Where should we put it? layout and placement strategies of documents in augmented reality for collaborative sensemaking," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3491102.3501946

[30] W. Luo, A. Lehmann, Y. Yang, and R. Dachselt, "Investigating document layout and placement strategies for collaborative sensemaking in augmented reality," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–7.

[31] L. Lisle, K. Davidson, L. Pavanatto, I. A. Tahmid, C. North, and D. A. Bowman, "Spaces to think: A comparison of small, large, and immersive displays for the sensemaking process," in *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2023, pp. 1084–1093.

[32] L. Lisle, K. Davidson, E. J. K. Gitre, C. North, and D. A. Bowman, "Different realities: a comparison of augmented and virtual reality for the sensemaking process," *Frontiers in Virtual Reality*, vol. 4, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frvir.2023.1177855

[33] M. R. Seraji, P. Piray, V. Zahednejad, and W. Stuerzlinger, "Analyzing user behaviour patterns in a cross-virtuality immersive analytics system," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 5, pp. 2613–2623, 2024.

[34] A. Galati, R. Schoppa, and A. Lu, "Exploring the sensemaking process through interactions and fnirs in immersive visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 5, pp. 2714–2724, 2021.

[35] I. A. Tahmid, L. Lisle, K. Davidson, C. North, and D. A. Bowman, "Evaluating the benefits of explicit and semi-automated clusters for immersive sensemaking," in *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2022, pp. 479–488.

[36] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: Ordering points to identify the clustering structure," in *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '99. New York, NY, USA: Association for Computing Machinery, 1999, p. 49–60. [Online]. Available: https://doi.org/10.1145/304182.304187

[37] J. MacQueen, "Classification and analysis of multivariate observations," in *5th Berkeley Symp. Math. Statist. Probability*. University of California Los Angeles LA USA, 1967, pp. 281–297.

[38] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," *arXiv preprint arXiv:1109.2378*, 2011.

[39] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[41] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[42] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[43] Y.-a. Kang, C. Gorg, and J. Stasko, "Evaluating visual analytics systems for investigative analysis: Deriving design principles from a case study," in *2009 IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2009, pp. 139–146.

[44] J. Siegel, "Did you take "good" notes?: On methods for evaluating student notetaking performance," *Journal of English for Academic Purposes*, vol. 35, pp. 85–92, 2018.

[45] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[46] A. Giovannelli, L. Lisle, and D. A. Bowman, "Exploring the impact of visual information on intermittent typing in virtual reality," in *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2022, pp. 8–17.

Kylie Davidson is a Ph.D. student at Virginia Tech. She is a member of the Center for Human-Computer Interaction (HCI) and the Sanghani Center on campus. She focuses on HCI research as it is applied to virtual/augmented reality for analytics. She is a student member of the IEEE Computer Society.

Lee Lisle is a Postdoctoral Associate at Virginia Tech National Security Institute's Hume Center. He focuses on human-computer interaction research as it applies to mixed reality, specializing in 3D user interfaces and immersive analytics of non-quantitative data.

Ibrahim A. Tahmid is a Ph.D. student at Virginia Tech. He focuses on human-computer interaction research, specializing in enhancing immersive analytics by leveraging rich sensory data to generate automated suggestions during sensemaking tasks. He is a student member of the IEEE Computer Society.

Kirsten Whitley works in the Department of Defense, researching visual analytics for intelligence and cyber analysts, recently concentrating on immersive analytics and its opportunities for government use cases. Her successes stem from combining cognitive task analyses of analysis expertise with emerging technology. She has also supported the visual analytics research community, sponsoring Visualization for Cyber Security (VizSec) in its early years, serving as VizSec 2014 general chair, and introducing cyber use cases to the annual VAST challenge competition (2010-2012). She earned her Ph.D. in computer science at Vanderbilt University.

Chris North is a professor of computer science at Virginia Tech. He is the associate director of the Sanghani Center. His research seeks to create effective methods for human-in-the-loop analytics of big data. His work falls in areas of visual analytics, information visualization, human-computer interaction, large high-resolution display and interaction techniques, and visualization evaluation methods, with applied work in intelligence analysis, cyber security, bioinformatics, and GIS.

Doug A. Bowman received a Ph.D. degree in Computer Science from Georgia Tech. He is a Professor in the Department of Computer Science and Director of the Center for Human-Computer Interaction at Virginia Tech. He is a member of the IEEE Computer Society