Eurographics Conference on Visualization (EuroVis) 2023
D. Archambault, R. Bujack, and T. Schreck
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 42 (2023), Number 3

A Comparative Evaluation of Visual Summarization Techniques for Event Sequences

Kazi Tasnim Zinat[1], Jinhua Yang[1], Arjun Gandhi[1], Nistha Mitra[1] & Zhicheng Liu[1]

[1] University of Maryland College Park, Maryland, United States

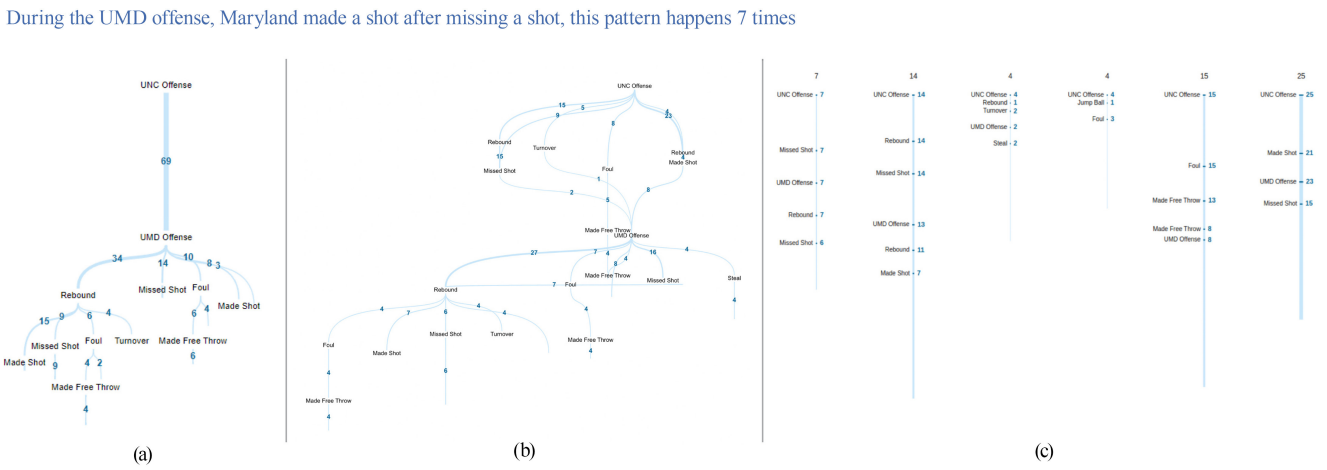


Figure 1: *In our study, we generated visual summaries of the same dataset using three different techniques: (a) CoreFlow [LKD*17], (b) SentenTree [HWS17], and (c) Sequence Synopsis [CXR18]. These visual summaries were shown one at a time, and the participants were asked to rate how closely the visualization depicted given insights about the dataset and provide a justification for their ratings.*

Abstract

Real-world event sequences are often complex and heterogeneous, making it difficult to create meaningful visualizations using simple data aggregation and visual encoding techniques. Consequently, visualization researchers have developed numerous visual summarization techniques to generate concise overviews of sequential data. These techniques vary widely in terms of summary structures and contents, and currently there is a knowledge gap in understanding the effectiveness of these techniques. In this work, we present the design and results of an insight-based crowdsourcing experiment evaluating three existing visual summarization techniques: CoreFlow, SentenTree, and Sequence Synopsis. We compare the visual summaries generated by these techniques across three tasks, on six datasets, at six levels of granularity. We analyze the effects of these variables on summary quality as rated by participants and completion time of the experiment tasks. Our analysis shows that Sequence Synopsis produces the highest-quality visual summaries for all three tasks, but understanding Sequence Synopsis results also takes the longest time. We also find that the participants evaluate visual summary quality based on two aspects: content and interpretability. We discuss the implications of our findings on developing and evaluating new visual summarization techniques.

CCS Concepts

• ***Human-centered computing*** → ***Visualization design and evaluation methods;*** *Empirical studies in visualization;*

1. Introduction

In many application domains, discrete events are recorded for specific entities, and ordered temporally to form sequences. For exam-

ple, healthcare providers keep records of lab results or treatment events for each patient; businesses collect clickstreams to increase conversion; software developers log user behavior to identify potential usability issues. These datasets are often complex and heterogeneous: few sequences are identical to each other, and there is usually high variability between sequences in terms of the number and type of events and their orders. Visualizations based on simple visual encoding and aggregation are therefore inadequate. Extensive research has thus focused on techniques that combine computational methods with visual interfaces [WSSM12; MWP*12; GCGC13; GXZ*18; GJG*19; MSM*21]. In particular, a number of techniques try to generate *visual summaries* of event sequences [LWD*17; LKD*17; CXR18; CPYQ18; MLL*13; PW14] by showing only important events and salient structures that serve as overviews. Figure 1 shows exemplary visual summaries of a basketball match dataset consisting of 69 sequences and 465 events, generated by three techniques.

Despite advances in novel visual summarization techniques, we have little understanding of their effectiveness. To date, there have been no empirical studies comparing these techniques through controlled experiments. The lack of systematic evaluation is problematic: researchers have no established baselines and methods to measure and innovate new techniques; practitioners have no guidance on choosing a suitable technique for their data and analytic needs.

In this paper, we present the design and results of an insight-based crowdsourcing experiment evaluating three existing visual summarization techniques: CoreFlow, SentenTree, and Sequence Synopsis. We chose these techniques based on considerations such as applicability across different domains, diversity of summary structures, and adjustable summary granularity. Our focus is on the underlying algorithms, not interactive systems. The algorithms enable *automated* generation of visual summaries, which serve as overviews in visualization [Shn96; KAF*08].

In the experiment, the participants evaluate how closely visual summaries generated by the techniques at different granularity levels match known ground truths for different datasets. We analyze the participants' ratings, the time spent on evaluating the summaries, as well as their justifications for the ratings. We find that (1) visual summaries generated by Sequence Synopsis receive the highest ratings, but they also require more time to understand; (2) two factors influence the perceived quality of a visual summary: content and interpretability. We discuss the implications of our findings on developing and evaluating new summarization techniques.

This paper makes the following contributions:

- **Experiment Design.** To the best of our knowledge, this is the first controlled experiment to compare event sequence visual summarization techniques. We identify dataset, task, granularity as independent variables and measure technique effectiveness through user rating, completion time, and text responses.
- **Result Analysis.** Our analysis of the experiment data deepens the understanding on factors influencing summary effectiveness, criteria for assessing summary of quality, and trade-offs in event sequence visual summarization.
- **System Implementation.** We re-implemented three existing automated event sequence summary techniques as well as three different approaches to visualize summary structures. We plan to open source these implementations.

2. Background and Related Work

2.1. Visualization Techniques for Event Sequence Data

Early event sequence visualization tools focused on displaying each individual records, [Kar94; PMR*96; PMS*03; HOB94; WPQ*08; FKSS06]. and they can only handle a small number of sequences. Later tools aggregate events across sequences to generate tree structures or directed-acyclic graph (DAG) structures [WGP*11; WG12; MWP*12; MLL*13; GCGC13; PG13; GS14]. Figure 2 shows these aggregation approaches. A tree structure can then be visualized as an icicle plot [WGP*11] or a sunburst chart [SZ00], and a graph structure can be visualized using a node-link diagram, a Sankey diagram [WG12; GCGC13; GS14], or concatenated adjacency matrices [ZLD*15]. Besides aggregation, these tools also support additional functionalities, such as querying and filtering to simplify visual overviews of complex sequences [MWP*12; VJC09]. Many of these functionalities require human knowledge to interactively generate a meaningful visualization.

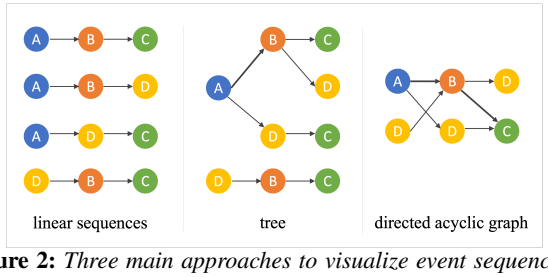


Figure 2: *Three main approaches to visualize event sequences as linear sequences, a tree with a virtual root, and a directed acyclic graph. Each circle represents an event, order goes from left to right.*

2.2. Visual Summarization: Mining and Visualizing Patterns

It is time consuming to interactively aggregate and simplify events, and this method is often not scalable enough for existing datasets [LWD*17; MLL*13; LDDH16]. Automated techniques have been thus developed to extract important events and patterns from a dataset, and display the extracted results in a concise visual summary. Most automated techniques mine frequent patterns as part of the summarization process, and the structures of the extracted patterns are also in the form of linear sequences, tree, or directed acyclic graphs. For example, Frequence [PW14], Patterns and Sequences [LWD*17], Chronodes [PCK*18], and Peekquence [KVP16] extract linear sequential patterns. Sequence Synopsis [CXR18] performs sequence clustering first, and then mines frequent sequential patterns based on an information-theoretic approach to minimize description length [Grü07]. Instead of extracting linear patterns, CoreFlow [LKD*17] mines branching patterns using a recursive rank-divide-trim approach. Despite the misleading term "tree" in its name, SentenTree [HWS17] uses a breadth first approach to find graph-like patterns in tweets. This technique can be applied to event sequence data in general. The extracted patterns can then be visualized according to their structures. Sequence Synopsis [CXR18] and Patterns and Sequences [LWD*17] display the sequential patterns as linear ordered event sets. CoreFlow [LKD*17] visualizes the extracting branching pattern using an icicle plot. SentenTree [HWS17] uses a node-link diagram to show

the extracted DAG pattern. Linear patterns can also be merged and visualized as a Sankey diagram, as shown in Frequence [PW14].

2.3. Segment, Align, and Cluster Event Sequences

Besides mining-based approaches, researchers have also developed visual analytic techniques to segment, align, and cluster sequences. To segment sequences, EventThread [GXZ*18] and EventThread2 [GJG*19] uses the tensor analysis with an unsupervised stage analysis algorithm to find progression states in event sequences. STBins [QBW*20] renders segment similarity based on temporal binning with Jaccard coefficient-based segment similarity measure. DPvis [KAS*20] encodes event sequences via hidden Markov model to identify disease progression pathways. To align sequences by key events, Sequence Braiding [BZS*20] computes pairwise alignment of input sequences, and orders them through a constrained intersection reduction algorithm. To cluster sequences, Wei et al. [WSSM12] clusters sequences using a self-organizing map. Cadence [GZW*20] supports dynamic hierarchical aggregation of high dimensional event sequence data. Sequen-C [MSM*21] tries to combine both sequence clustering and alignment to support multi-level visual analytics. All these techniques still try to show all the events and sequences with additional visual structures such as alignment and clusters. In this paper, our focus is on the evaluation of mining-based techniques that have a data-reduction component, where the analytical results contain much fewer events and sequences than the original dataset, and the visualizations are best considered as *visual summaries* of event sequences.

2.4. Interactive Visual Analytics

Most event sequence visual analytics works provide interfaces to perform interactive analysis. For example, EventPad [CvWvW18], MAQUI [LLMB19], DecisionFlow [GS14], and s|queries [ZDF*15] allow users to query events and patterns. In Eloquence [VN18]- users can interactively add local constraints while patterns are mined using PrefixSpan [PHM*04]. Segmentifier [DM19] proposes a high-level analysis model with support for data wrangling and downstream data analysis. Progressive Insights [SPG14] provide intermediate results by modifying BFS SPAM mining to prioritize interesting patterns and prune uninteresting ones. Whether automatic summarization techniques are used without human input or in a mixed-initiative approach, their effectiveness plays an important role in determining the analytic outcomes. Our goal in this paper is to evaluate the effectiveness of automated summarization techniques for providing such overviews.

3. Motivation and Approach

We surveyed 14 research papers that proposed novel visual summarization techniques over the past 8 years [CXR18; CPYQ18; GXZ*18; HWS17; KVP16; LKD*17; LWD*17; PW14; PCK*18; PWH15; GJG*19; GWP14; RPP*17; WZT*16]. One paper [LKD*17] explored comparison with other techniques, but only showed sample visualizations for illustrative purposes. 13 out of 14 evaluated the proposed techniques through qualitative case studies with domain experts. While case studies demonstrate the ecological validity of the work, they do not provide an objective account on how the techniques compare to each other.

In this paper, we use the term *techniques* to refer to the underlying *algorithms*, not the interactive systems, in the original papers. The algorithms determine both the summary *content* (i.e., what events and relations are included) and summary *structure* (i.e., linear sequences, tree, DAG). We focus on evaluating algorithms based on the following reasons: 1) the algorithms are usually presented as primary contributions in the original papers, 2) the case studies in these papers show that users rely on the algorithmically generated visual summaries as overviews to understand their datasets, and 3) the different user interface and interaction designs in the original papers introduce confounding factors, making it hard to understand technique effectiveness.

To date, there has been no established methods or metrics to evaluate event sequence summarization techniques. Existing studies on visualization effectiveness usually focus on graphical perception [HB10; KH18], where the methods and results cannot be directly applied to our problem. First, these studies exclusively focus on the choice of encoding methods (i.e., how data attributes are represented using visual properties of marks). In our case, however, significant data reduction is performed before visual encoding to make the visualizations readable. Merely focusing on encoding overlooks the importance of data reduction. Second, most of the prior studies focus on low-level tasks such as looking up and comparing data values. Understanding how well users perform these tasks does not shed light on the more important questions on the quality of visual summaries. For example, while we care if users can read and understand the visual summary presented to them, we are also interested in assessing the ease or difficulty to detect the presence of patterns in the visual summary.

Given the lack of established methods, we have considered different experimental settings. First, we may simply evaluate the algorithms by objectively checking if the generated visual summaries contain pre-formulated ground truth associated with a dataset. However, this approach overlooks the importance of graphical perception: people's ability to identify the ground truth can be influenced by both the summary structure and content. An in-lab controlled study can address this problem. However, it is important to include multiple datasets from different domains in the study since dataset domain and properties can influence technique effectiveness [LKD*17]. Furthermore, we want to compare at least three techniques, and include multiple levels of summary granularity. An in-lab design is not likely to scale well for these factors.

Based on these considerations, we propose to adapt the insight-based methodology [SND05; Nor06] in a crowdsourcing experiment. We focus on how well the visual summaries generated by different techniques depict pre-formulated insights about existing datasets, as judged by the participants. An alternative approach where users explore the generated visual summaries to reach insights would be more ecologically valid, but such open-ended designs are difficult to control and monitor on a crowdsourcing platform. We argue it is a reasonable proxy to test if a visual summary depicts given insights: it would be harder to reach insights if the summary presents less relevant information or is hard to interpret.

4. Insight-Based Crowdsourcing Experiment Design

4.1. Visual Summarization Techniques

Dozens of visual summarization techniques for event sequences are available [GGJ*21]. We choose the techniques to be included in this study based on the following criteria. First, the techniques should be domain agnostic and can be applied to datasets from different problem domains. Second, we want to include three types of summary structures: linear sequences, tree, and directed acyclic graph (DAG). Third, the techniques should generate summaries consisting of much fewer events and sequences compared to the input dataset; simple clustering methods, for example, do not satisfy this requirement. Fourth, we focus on automated summarization techniques in this study, so the generation of visual summaries should require minimal human input. Finally, we would like to compare the visual summaries at different levels of granularity, so the techniques should support controlling summary granularity through appropriate parameters.

Based on these criteria, we choose CoreFlow [LKD*17], SentenTree [HWS17], and Sequence Synopsis [CXR18]. These techniques mine frequent events and patterns and visualize them as linear sequences, a tree, and a directed acyclic graph, respectively. SentenTree [HWS17] is the only technique we know that produces a DAG summary. Although it was originally developed for text data, it can be directly extended to event sequences. They all have granularity parameters that can be tuned, and can be applied to event sequences from different domains.

4.1.1. Overview of Algorithmic Approaches

CoreFlow [LKD*17] recursively applies a rank-divide-trim approach. Events are initially ranked using a pre-defined metric, such as the frequency of occurrence and average index (the mean value of index positions) across sequences. The top-ranked event is added to the summary; and the sequences are partitioned into two groups based on whether they contain the top-ranked event. Finally, the sequences containing the top-ranked event are trimmed. These three operations are recursively applied to resulting sequence groups until either all sequences have been processed or a predefined minimum support threshold is reached. (i.e., a threshold below which the mining algorithm will stop).

SentenTree [HWS17] also uses a rank and divide approach. But instead of pruning the sub-sequences up till the first occurrence of the top ranked event, SentenTree also extracts frequent patterns above the minimum support in these sub-sequences. Given the same minimum support, SentenTree usually mines more events and patterns than CoreFlow.

Sequence Synopsis [CXR18] uses the minimum description length [Grü07] principle to cluster sequences and identify a representative sequential pattern per cluster. The algorithm performs iterative merging to find clusters and associated patterns, while optimizing for the number of generated patterns, and the edits required to obtain the original dataset from the patterns.

4.1.2. Visualization Design

Our goal is to compare the summarizing algorithms. However, the visual representations and styles for the generated summaries in the original works vary greatly. To eliminate the potential confounding effects, we did not follow the original visualization designs. Instead, we chose a minimalistic and consistent design for the visualizations (Figure 1). Each event is represented by a node with a label, events connected by links form a pattern. The vertical position represents the order of events, going from top to bottom. The width of a link is proportional to the number of sequences. To facilitate reading, we place numeric labels representing the number of sequences next to each node in Sequence Synopsis, and on top of each link in CoreFlow and SetenTree.

We use the tidy tree layout [RT81] for CoreFlow and the Sugiyama layout [STT81] for SentenTree. The tidy tree layout tries to achieve symmetry and compactness in node positioning. The Sugiyama layout layers nodes and optimizes their placement and ordering to reduce edge crossings. For Sequence Synopsis, we set the patterns in an equidistant layout. Vertical position of the events encode the average index position across sequences they appear in. The start and end nodes of mining are hidden to reduce visual clutter. In SentenTree, if a node has two or more predecessors, then the average position is set after the predecessor with the highest average index position. This design decision guarantees a node will always appear after all its predecessors. Finally, the same color scheme is used across techniques for consistency.

4.1.3. Implementation

Among these three techniques, only SentenTree [HWS17] has been open-sourced. The available implementation assumes tweets as input, and cannot readily handle generic event sequence data. The visualization implementation does not conform to the design guidelines discussed above either. Therefore, we re-implemented the summarization algorithms The supplemental materials contain descriptions of our implementation and how we verify its correctness. To avoid computational latency being a confounding factor, we precompute the summaries, render visualizations based on the results, and save the visualization as images to be used in the study.

4.1.4. Granularity

Each of these techniques supports parameterized tuning of summary granularity. CoreFlow and SentenTree use minimum support to determine the percentage of sequences an event must appear in to be mined and included in the pattern. Sequence Synopsis uses two parameters α and λ. α balances the trade-off between minimizing information loss and reducing visual clutter. λ controls the total number of patterns. We select to control λ as the granularity parameter. In order to find the right balance between visual clutter and missing information, while maintaining a wide range of variation, we experimented with varying degrees of granularity. We decided upon using six granularity levels for each technique, with the minimum support value ranging from 5% to 30% with increments of 5% for CoreFlow and SentenTree, and value of λ ranging from 90% to 15% with decrements of 15% for Sequence Synopsis.

4.2. Datasets, Analytical Tasks and Insights

We searched for event sequence datasets by reviewing published papers at visualization and mining conferences and journals, and examining public dataset repositories [DG17; PDF*16]. We collected a candidate set of 15 potential event sequence datasets from

Dataset	Description	# Unique Events	Sequence Length			# Sequences	Total Events
			Min	Max	Median		
Pediatric Trauma Unit (Trauma) [CBM*13; MMPS13]	Order of trauma response events for children	11	5	11	9	215	1,991
Emergency Department (Emergency) [Pla17]	Patients' movement through the hospital after being brought to emergency	6	3	16	4.5	100	451
UMD vs UNC Basketball Match (Basketball) [Mon13]	A play-by-play event log of a basketball match	13	4	13	6	69	465
VAST Challenge (VAST) [WCJ*17]	Published as a challenge in 2017 VAST. The dataset is about the movement of cars in a nature preserve. Each sequence records the locations a car passed by during its trip	6	2	49	8	1,000	9,443
Issue Workflow [AAS*19]	Workflows related to bug fixes on an Apache software project	16	2	21	11	45	177
Career [MSD*16; GJC*22; GXZ*18; GJG*19]	Career path milestone events for university professors over 23 years	10	11	32	17	40	767

Table 1: *Dataset description and event details*

Dataset	Domain	Task	Example Insight
Trauma	Medical	Anomaly Detection	Only about half out of 215 patients went through the process in the right order: airway→ breathing → pulse → gcs → secondary_survey
Emergency	Medical	Common Pattern Identification	Out of the 100 people, about a third (37 people) are discharged alive after going to the emergency room
Basketball	Sports	Common Pattern Identification	During the UMD offense, Maryland made a shot after missing a shot, this pattern happens 7 times
VAST	Transport	Clustering	Approximately 17 cars (17%) are "pass-throughs"
Workflow	Technical	Clustering	For 6 issues that were created, nothing happened afterwards.
Career	Academic	Anomaly Detection	8 persons do not have any publications or attend any conferences after they became a professor

Table 2: *Domain and assigned task information of the datasets along with example insight*

various domains. To reduce bias, we exclude datasets used in any of the three original papers. We identified the following information:

- Application Domain: Example domains include but are not limited to healthcare, software, sports, and human activities.
- Data Size: We record statistics about the size of the dataset in terms of the number of unique events, number of total events, number of sequences, and max/min/median sequence length.
- Insights: There is no established benchmark ground truth readily available for our intended evaluation task. However, many datasets have associated insights that can be used as ground truth. These insights are usually included in the case study section of corresponding publication; sometimes supplemental videos also present these insights in detail. For each dataset, we curated a set of insights both from the paper and the video transcript (if available). Table 2 shows example insights.
- Analytical Tasks: After curating the insights for each dataset, we identify the representative high-level analytical tasks that these insights support. We identified three tasks: Anomaly Detection, Common Pattern Identification, and Clustering.

Based on these dimensions, we select six datasets to be used in the study. The goal is to include diverse application domains and varying data size, and to have the same number of datasets for each of the three analytical tasks. As we prioritized datasets from different domains with insights from previous studies, we have limited control over different size parameters such as number of unique events and sequence length. We use the associated insights in the corresponding publications and supplemental videos for Emergency [Pla17], Trauma [CBM*13; MMPS13], Career [MSD*16; GJC*22; GXZ*18; GJG*19] and Basketball [Mon13] dataset. For the VAST [WCJ*17] and Workflow [AAS*19] datasets, two of the authors independently analyzed these using EventFlow [MLL*13] for the Clustering task, and calibrated the findings. We then picked one primary analytical task for each dataset, and identified three insights that can support the task. Table 1 contains a brief description of the selected datasets and complexity. Table 2 has the domain information and task for each dataset with an example insight.

4.3. Study Design

We used an insight-based approach to design our experiment, showing participants one visual summary at a time from a dataset and asking them to rate how accurately the summary matches each of the three insights associated with that dataset. With three techniques, six granularity levels, three analytical tasks, and six datasets (two per task), we generated $3 \times 6 \times 3 \times 2 = 108$ unique visual summaries. With three (3) insights for each dataset, we have a total of $108 \times 3 = 324$ unique combinations. It is not possible for a single participant to experience all these combinations in a reasonable amount of time, since we have to explain the meaning

of each dataset to them. We considered various plausible combinations. Finally, we decided that it was important to expose every participant to all three techniques, without overwhelming them with the unfamiliar events of multiple datasets, while also avoiding learning effect. Therefore, we choose a mixed approach using a within-participants design for technique and insight, and a between-participants design for dataset, task, and granularity. That is, each participant would experience 9 combinations in total: all three (3) techniques and all three (3) insights associated with a dataset, and only one (1) granularity level and one (1) dataset, which corresponds to one (1) task. We implemented the study design in JavaScript on Qualtrics to guarantee equal participant distribution across granularity levels and datasets, and to ensure participants accurately encounter the combinations mentioned above.

4.4. Study Procedure

We developed a three phase evaluation process to ensure the participants have adequate visual and data literacy. The participants first complete a tutorial and a pre-screening test. They proceed to the main experiment if they satisfy the prescreening criterion. Details of participant selection criteria is mentioned in 4.5.

4.4.1. Tutorial

We generated visual summaries using the three techniques on check-in information of 25 individuals randomly sampled from the New York City Foursquare check-in dataset [YZZY15]. The 20-month check-in data was divided into weeks. This dataset is not part of the main experiment. In the tutorial, the participants click through a series of displays, which incrementally highlights different parts of the visual summaries to describe what the nodes and links signifies. The tutorial in particular focuses on the interpretation of branches in CoreFlow and SentenTree visualizations, and the number of sequences and events in Sequence Synoposis visualizations. The tutorial is included in the supplemental materials.

4.4.2. Pre-Screening Test

We sampled the first 1000 events of a baseball game dataset (it was not a part of the main experiment) and created six questions to test the participant's literacy in terms of different low-level tasks. If a participant answers at least 50% of the questions correctly, they are invited to the main experiment. The supplemental materials include the questions from the pre-screening test.

4.4.3. Main Experiment

In the main experiment, a brief description of the dataset and the problem domain is initially presented to the participant. For example, the following text explains the emergency department dataset:

"*The data is a sample of 100 patients, showing how a patient moves through the hospital over time. All the possible events found in the data are: arrival at the hospital (Arrival), going to the emergency room (Emergency), to the ICU (ICU), to normal floor room (Floor) and discharged alive (Discharge-Alive) or not (Die).*"

In each of the following pages, the participant sees an insight and one visual summary generated by one of the three techniques. To eliminate any order effects, we use the loop and merge feature on Qualtrics to randomize the order of techniques. The participant rates each technique on a 7 point Likert scale, evaluating its ability to accurately match the insights for a given dataset. we conducted a preliminary study with ten participants to verify the soundness of our study method and estimate the average completion time.

As described in Section 4.3, each participant rates a total of nine (3 Insights $\times$ 3 techniques = 9) visual summaries. They also provide a text justification for the ratings assigned. In the pilot study, we asked for a justification for each of the nine ratings, but the feedback from the participants indicated that this was too time consuming and the responses were very similar for the same technique. We thus only collect three justifications, one for each technique, from each participant. In addition to the ratings and text justifications, we also recorded the time they spent on each visual summary.

4.5. Participants

We recruited participants from multiple sources. We started recruiting participants with $\geq$ 90% HIT Approval Rate on Mechanical Turk but encountered issues with the quality of responses; some participants did not pass the pre-screening test or did not understand the visual summaries- which was evident based on the text justification. Only 29.2% of responses were high quality. Therefore, we switched to Prolific which has better filtering options. We recruited US residents who were at least 18 years old, had an approval rate of $\geq$ 95%, and used a desktop. We limited educational status to those who attended technical/community college or had an undergraduate degree. We also recruited participants from the student body at our university. All of them had at least a Bachelor's degree. We finalized 180 participants, yielding 1620 observations (180 Participants $\times$ 3 Techniques $\times$ 3 Insights). We made sure each condition has the same number of observations. All participants were compensated above minimum wage. Table 3 displays the acceptance rate for each platform.

Source	Total participant	Accepted	Acceptance Rate
MTurk	24	7	29.2%
Prolific	150	135	90.0%
Students	61	55	90.2%
Total	**235**	**197**	**83.82%**
Further Filtered		17	
Finally Accepted		**180**	

Table 3: *Participant details*

5. Analysis of Likert Scale Ratings of Visual Summaries

We first analyze the effects of various independent factors on summary quality, as measured using the Likert scale ratings assigned by the participants. We perform exploratory analysis by creating visualizations of aggregated ratings, and build linear mixed-effects models to assess the effects of the independent variables. Mixed effects models are appropriate for our multi-level study design involving repeated measures, and have been used in similar empirical studies [LH14; KH18]. We model *technique*, *granularity*, and *task* as fixed effects and *participant*, *dataset*, and *insight* as random effects to extend the findings beyond the participants and datasets used in the study. Insights are nested under datasets in the models since each dataset has a unique set of insights.

We test the statistical significance of the fixed effects using likelihood-ratio tests [Win13]: we build a full model (with the fixed effect in question) and a reduced model (without the fixed effect in

Technique1	Technique2	Rating	Time
CoreFlow	SentenTree	0.16	0.09
CoreFlow	SequenceSynopsis	0.33	0.36
SentenTree	SequenceSynopsis	0.24	0.33

Table 4: *Cohen's d values for effect size estimation*

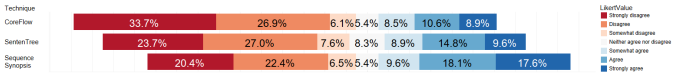
question), and compare these models to obtain p-values. The degrees of freedom for the comparison model is the difference in estimated parameters between the full and reduced model. The linear mixed-effects models are implemented using the R package lme4 [BMBW14]. Below we report the results.

5.1. Technique Influences Visual Summary Quality

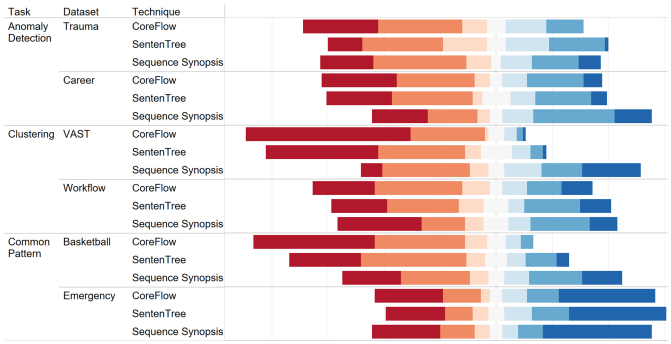
Figure 3 shows the distribution of ratings for each technique, as well as a more detailed breakdown of the ratings by task and dataset. Overall, Sequence Synposis (mean rating 3.86) outperforms SentenTree (mean 3.35) and CoreFlow (mean 2.95).

This observed pattern in the figure is confirmed in the statistical analysis. We find a strong main effect of technique on the Likert scale ratings based on likelihood-ratio tests on random intercept models: $\chi^2(2, N = 1620) = 74.14$, $p < 0.001$. For CoreFlow, the estimated intercept is (3.03 ± 0.43 std. error). The estimated SentenTree intercept is 0.39 higher (3.42 ± 0.10 std. error). The estimate for Sequence Synopsis is the highest at (3.94 ± 0.10 std. error). The random intercept models assume that the effects of technique are the same for all the participants and all the datasets. Following the recommendations by Barr et al. [BLST13], we also build random slope models, the effect remains significant: $\chi^2(2, N = 1620) = 46.30$, $p < 0.001$ where we assume the effects of technique vary for each participant, and $\chi^2(2, N = 1620) = 11.786$, $p < 0.01$, where we assume the effects of technique vary for each dataset.

We also built models for pairwise technique comparison , tested the significance and calculated the effect size in table 4. All the paired models show statistically significant effects of technique, implying that Sequence Synopsis significantly performs better than SentenTree, and SentenTree significantly outperforms CoreFlow.



(a) *Likert scale ratings distribution across Technique*



(b) *Likert scale ratings distribution across Task, Dataset and Technique*

Figure 3: *(a) the percentages of Likert scale ratings for each technique, out of the three techniques, Sequence Synopsis has the highest mean, followed by SentenTree and CoreFlow; (b) a detailed break-down of the ratings by task and dataset.*

Variables	Technique	Task	Granularity
Likert Rating	✓	✗	✓(SentenTree)
Completion Time	✓	✗	✓

Table 5: *Analysis summary of statistical significance*

5.2. Granularity Matters for SentenTree

Based on likelihood ratio tests on random intercept models, we can not find statistically significant effect of granularity in the overall Likert scale ratings. We also analyze the within-group effect of granularity for all three techniques. The effect is not statistically significant for CoreFlow and Sequence Synopsis. However, granularity has a significant effect on the ratings for SentenTree in random intercept models: $\chi^2(1, N = 540) = 6.54$, $p < 0.05$. The effect remains significant for random slope model: $\chi^2(1, N = 540) = 4.48$, $p < 0.05$. Higher levels of abstraction in visual summaries produced by SentenTree lead to lower ratings. The rating decreases by 0.15 for every 5% increase in granularity.

We find significant interaction effects between technique and granularity: $\chi^2(2, N = 1620) = 16.354$, $p < 0.001$. The effects of technique vary by granularity: for every 5% increase in granularity, the rating drops by 0.006 for CoreFlow and by 0.15 for SentenTree, but rises by 0.09 for Sequence Synposis.

5.3. Task Has No Significant Effect on Rating

We do not find significant main effects for tasks on the Likert scale rating. We also model the interaction effects between *task* and *technique*, and between *task* and *granularity*. The technique-task interaction has a weak significance: $\chi^2(4, N = 1620) = 8.92$, $p < 0.1$, but the granularity-task interaction is not significant. For CoreFlow, the estimated intercept for the Anomaly Detection task is 3.15, with the intercept for the Clustering task being 0.56 lower at (2.59 ± 0.60 std. error) and for the Common Pattern task is 0.12 higher at (3.27 ± 0.60 std. error). All three techniques have the highest rating for Common Pattern task, with Sequence Synopsis performing best (intercept estimate 4.02 ± 0.26 std. error).

6. Analysis of Completion Time

Our analysis of the completion time – the time participants spent on evaluating each visual summary – is similar to our analysis of Likert scale ratings. We use log-transformed completion time as the response variable to build linear mixed effects models and conduct likelihood ratio tests as mentioned in Section 5. Table 5 shows our statistical analysis summary.

6.1. Technique Influences Completion Time

To illustrate the effect sizes, we compute the bootstrapped means and 95% confidence intervals for log-transformed completion time by sampling individuals with replacement (Figure 4). Overall, participants spend more time evaluating visual summaries generated by Sequence Synposis compared to SentenTree and CoreFlow.

The statistical findings corroborate the observed pattern. Based on likelihood-ratio tests on random intercept models, we find a significant main effect of technique on log completion time: $\chi^2(2, N = 1620) = 98.92$, $p < 0.001$. For CoreFlow, the estimated intercept is (3.43 ± 0.14 std. error). The estimated SentenTree intercept is 0.1 higher (3.53 ± 0.04 std. error). The estimate for sequence synopsis

is the highest (3.82 ± 0.04 std. error). We also build random slope models, the effects remain significant: $\chi^2(2, N = 1620) = 52.66$, $p < 0.001$, where we assume the effects of technique vary for each participant, and $\chi^2(2, N = 1620) = 24.71$, $p < 0.001$, where we assume the effects of technique vary for each dataset.

Additionally, we developed models for pairwise technique comparison, tested the significance and calculated effect sizes (Table 4). All the paired models show statistical significance.

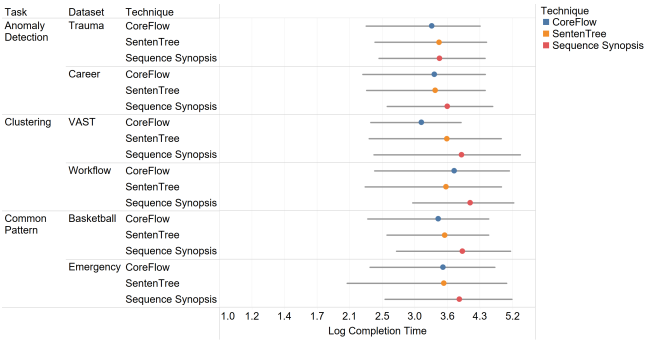


Figure 4: *Bootstrapped means and 95% confidence intervals for log completion times across datasets and techniques.*

6.2. Task Has No Significant Effect on Completion Time

We find no main effects of task on completion time, nor significant interaction effects between task and granularity. The task-technique interaction, however, is statistically significant: $\chi^2(2, N = 1620) = 15.77$, $p < 0.01$. The projected intercept for the Anomaly Detection task in CoreFlow is 3.43, with the intercepts for the Clustering and Common Pattern tasks being 0.25 (3.68 ± 0.18 std. error) and 0.21 (3.64 ± 0.18 std. error) higher, respectively.

The common pattern identification task has the greatest estimate for CoreFlow (estimate 3.44). The clustering task take the longest time for SentenTree (estimate 3.56) and Sequence Synopsis (estimate 3.97). All tasks involving rating CoreFlow are finished by participants in the shortest amount of time.

6.3. Granularity Affects Completion Time

We discover that the level of granularity has a statistically significant impact on the log-transformed completion time based on likelihood ratio tests on random intercept models: $\chi^2(1, N = 1620) = 13.25$, $p < 0.001$. The intercept estimate for granularity is 0.08 lower, indicating that with every 5% increase in granularity, the log-transformed completion time drops by 0.08.

We also analyze the within-group effect of granularity for all techniques. The effect is not statistically significant for Sequence Synopsis, but is significant for CoreFlow ($\chi^2(1, N = 540) = 8.04$, $p < 0.01$), and SentenTree ($\chi^2(1, N = 540) = 14.58$, $p < 0.001$).

7. Analysis of Qualitative Data

In addition to the Likert scale ratings, we also asked the participants to provide text responses to justify their ratings. These qualitative responses can help us better understand the criteria and reasoning used by the participants to assess the quality of visual summaries.

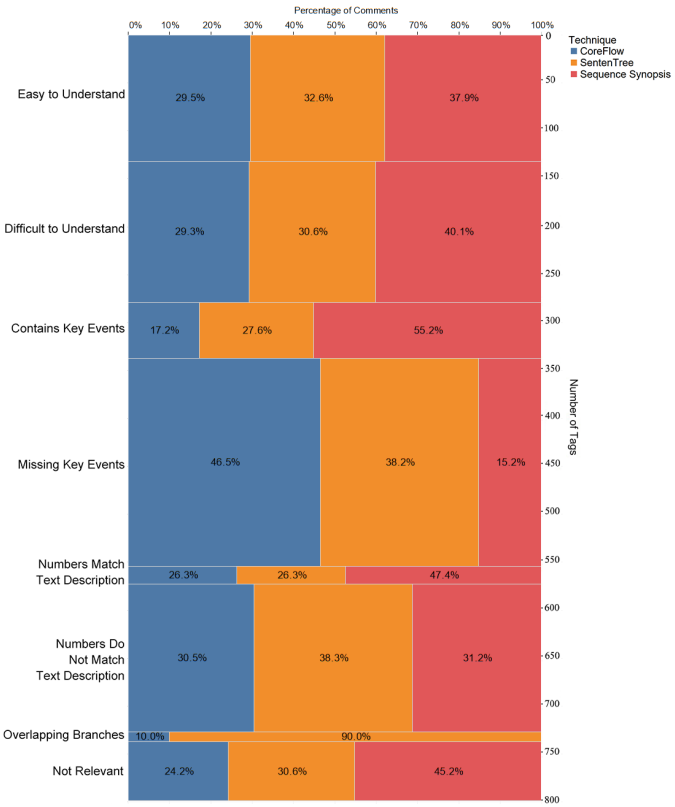


Figure 5: *Mosaic plot showing distribution of techniques for different tag categories. The bar height represents the number of comments for each tag category, and the bar width represents the percentage of comments contributed by a technique within each tag category. Missing Key Events is the largest tag category. Most of the comments with this tag belongs to CoreFlow (46.5%)*

We performed open coding on the responses from 180 participants. Most comments touched upon two aspects of visual summary quality: *content* (i.e., how closely the events and patterns included in the visualization match those mentioned in the insight), and *interpretability* (i.e., how easy it is to read the visualization). We identified four tags related to content: *Contains Key Events*, *Missing Key Events*, *Numbers Match Text Description*, and *Numbers Do Not Match Text Description*. The first two are concerned with whether important events are included in the visualization, and the remaining two focus on whether the quantitative information such as the number of sequences is consistent with the given insight. We also identified three tags related to interpretability: *Easy to Understand*, *Difficult to Understand*, and *Overlapping Branches*, the first two are self-explanatory and the last one focuses on link crossing and cluttered views. Finally, we created a tag *Not Relevant* to cover responses that are not intelligible or provide irrelevant information on the quality of visual summaries.

We then manually labeled each response with these identified tags. A response can mention multiple aspects of the summary quality, hence assigned multiple tags. In total, we assigned 799 tags. Figure 5 shows the number of comments associated with each tag, and the percentage distribution of each technique under each tag. The tag categories with the most comments are *Missing Key Events*

(217), *Numbers Do Not Match Text Description* (154), *Difficult to Understand* (147), and *Easy to Understand* (132).

7.1. Content

Many responses indicate that the visual summaries lack complete and consistent information compared to the given insights- this observation applies to all three techniques. One participant remarks, "*Some of the activities in the facts were not shown in the image. The activities that were shown had numerical discrepancies to the fact*". This response falls under the *Missing Key Events* and *Numbers Do Not Match Text Description* tag. Similar comments were found across all three techniques. On the other hand, some visualizations do conform to the insight: "*The numbers are all there, as far as my understanding is. That's why I chose strongly agree for each option.*" The participants rate the image favorably when *Numbers Match Text Description* and the image *Contains Key Events*.

Sequence Synopsis outperforms the other techniques in terms of including key events, with only 15.2% of Missing Key Events tags associated with it, compared to 38.2% for SentenTree and 46.5% for CoreFlow. Furthermore, 55.2% of *Contains Key Events tags* are associated with Sequence Synopsis, compared to 27.6% for SentenTree and 17.2% for CoreFlow. CoreFlow only mines the most frequent event and trims subsequences preceding it, leading to the omission of less frequent events above the minimum support threshold. SentenTree iteratively extends the most common summary sequence, making it prone to missing events. In contrast, Sequence Synopsis uses a clustering and merging strategy that allows even less frequent events to form a pattern, preventing significant information loss during merging with more frequent events.

Sequence Synopsis also performs best in terms of numeric information accuracy. One interesting observation is techniques have almost equal share in the *Numbers Do Not Match Text Description* tag. It might be due to how the algorithms group the sequences into patterns, or the fact that a pattern described in the insight is distributed across multiple branches or sequences.

7.2. Interpretability

Sequence Synopsis contributed the largest share of comments for both *Easy to Understand* and *Difficult to Understand*, followed by SentenTree, and CoreFlow. The interpretability of the summaries varies by dataset and granularity. For the Basketball dataset at a granularity of 0.15, four out of five participants found Sequence Synopsis *Difficult to Understand*, and none found it *Easy to Understand*. On the contrary, for the Career dataset at a granularity of 0.3, three out of five participants found Sequence Synopsis *Easy to Understand*, and none found it *Difficult to Understand*.

Participants had mixed reactions to the branching patterns in CoreFlow and SentenTree visualizations. Some preferred CoreFlow's simplicity: "*This image is easy to understand and made the questions easy to answer as well because the path the numbers take is quite simple*", while others found the single graph structure in SentenTree easier to comprehend than Sequence Synopsis: " *This image appears to be the same as the previous image, except with less information trees shown. Since there is only focus on one information tree in this image, I would say it is slightly easier to understand*". One participant also noted the ease of following the branches to identify anomalies: " *This image was easy to read what the proper order of assessment was, what happened when there were deviations. You could also determine the assessment outline after the deviations from the norm occurred.*".

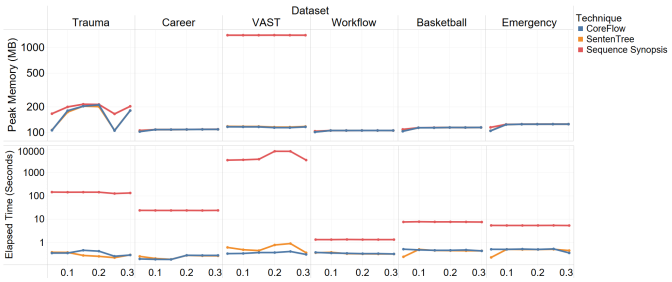


Figure 6: *Time elapsed and memory consumption across datasets at different granularities (finer to coarser, from left to right). The three techniques have similar memory consumption, except for the VAST dataset. Sequence Synopsis has the longest execution time.*

However, for some participants, the branching patterns were unfavorable: "*This image is not easy to follow because I'm not sure what one of the bifurcations means since it is not labeled. It also seems to be very confusing as to what were the sequence of events.*" In finer granularity SentenTree representations, *Overlapping Branches* are a prominent issue , which led to confusion and difficulty of understanding: "*The image is difficult to follow because there is a part of the graphic where pathways overlap and it is hard to attribute which data goes to which step*".

The participants also have mixed reactions towards the individual linear sequence representation in Sequence Synopsis visualizations. Some participants prefer the distinction: "*It's easy to understand because it's step by step*". Another comments: "*The images are easier to understand, because it delineates different threads of event . . .*". On the other hand, some individuals find it difficult to consolidate information across sequences: "*It's tough to aggregate this information from the different vertical bars easily.*"

The patterns related to content can explain the Likert scale ratings. Sequence Synopsis may have received higher ratings because it includes more key events and accurate numeric information than the other algorithms. The missing or inconsistent information may have contributed to CoreFlow and SentenTree obtaining a lower rating. We did not observe any strong correlation between accuracy and granularity, which is consistent with our finding in section 5 that granularity has no impact on Likert scale ratings.

In comparison with content, the relationship between interpretability and the ratings is less clear. The technique with the highest Likert score ratings, Sequence Synopsis, accounts for the largest share (40.1%) of comments with the *Difficult to Understand* tag. The difficulty in understanding is in line with the fact that the participants spent more time evaluating the Sequence Synopsis graphics. However, the largest share (37.9%) of comments with the tag *Easy to Understand* also belong to Sequence Synopsis. Further investigation is needed to understand how Sequence Synopsis' interpretability changes according to different datasets and granularity.

8. Discussion

Possible Explanations of Rating Results. Our analysis in section 5 shows that Sequence Synopsis performs the best in terms of sum-

mary quality for all three tasks. Sequence Synopsis uses a clustering and merging based information-theoretic mining approach with minimum description length principle. This enables the technique to penalize information loss while also taking visual clutter reduction and number of summary sequences into account. These strategies help to produce accurate summary results. Both CoreFlow and SentenTree use frequent pattern based mining techniques, where the next most frequent event is added to the summary sequence in a greedy algorithmic approach. There is no mechanism to account for information loss due to exclusion of less frequent events. SentenTree has an option to regulate the overall number of events in the visualization, but does not effectively control complexity or visual clutter in the branches, leading to lower Likert scale ratings.

Trade-offs between Ratings and Reading/Computation Time. The task completion time is inversely correlated with technique ratings: while Sequence Synopsis performs best in terms of rating, it also requires the longest time for the participants to understand its visualization results. CoreFlow, on the other hand, obtains the lowest ratings, due to its omission of key events, but produces simpler visual summaries that require the least comprehension time among the three techniques. The need to strike a balance between summary complexity and accuracy likely applies generally to all visual summarization techniques for event sequences.

Real-world applications must also consider computational time and memory cost. Figure 6 displays the peak memory consumption and computational time of the techniques to mine the datasets at different levels of granularity (finer to coarser from left to right). Note that all the axis scales are logarithmic. Except for the Trauma and VAST datasets, the three techniques exhibit comparable maximum memory usage. Sequence Synopsis for the VAST dataset uses up to 1.4 GB of memory at its maximum, significantly higher than the other techniques having maximum memory usage around 200 MB. In terms of elapsed mining time, CoreFlow and SentenTree complete mining for each dataset under a minute, whereas Sequence Synopsis takes significant longer time, in particular, about one hour for the VAST dataset. If we take efficiency into account, Sequence Synopsis requires more computation time and memory resource in addition to human time. These factors are important when selecting techniques to use in practical situations.

Visual Summarization Techniques Need Improvement in General. There is no silver bullet for creating perfect visual summaries of event sequence data, as demonstrated by our multifaceted analysis of human evaluation as well as computational resources. Sequence Synopsis, the technique with highest average rating, has a score of only 3.86 on a 7 point Likert scale. There is still ample room for improvement in all the facets, including more accurate information content in the visual summaries, less computational resources, and enhanced interpretability.

9. Limitations and Future Work

Potential Difference from Original Implementation. As described in 4.1.3, we evaluated the techniques mentioned using our own implementations. Although we follow the algorithm descriptions in the paper, our implementations might not accurately reflect the efficiency of the original code.

Evaluating Interactive Systems. Our current approach focuses on summary content and structure, which can be further refined and explored interactively to reach insights. Future research can shed light on how the algorithms and interactive exploration together influence insight generation.

Inclusion of Larger Datasets. The largest dataset in our experiment has 1000 sequences. Real world datasets are often much larger in size and contains more event types. This could potentially limit the generalizability of the findings to larger datasets. Despite this limitation, the experiment still provides valuable insights and serves as a starting point for further investigation.

Interpreting and Controlling Granularity. SentenTree and CoreFlow use minimum support parameter to control granularity, where lower values indicate finer granularity and more events included in the visualization. Sequence Synopsis, on the other hand, regulates the number of summary sequences, rather than individual event frequency. The granularity variable in our experiments provides only an approximation and requires careful interpretation. Future research should explore ways to control granularity based on both pattern number and individual event frequency.

Assessing Factors Influencing Technique Effectiveness. Our study evaluates the effectivenss of three techniques holistically. Each technique has multiple components working in conjunction. In particular, data reduction method (frequent pattern mining or information-theoretic clustering) and summary structure (linear sequences, tree, or graph) seem to play important roles. Further study is required to assess the effects of these individual components and their interaction on the outcome of visual summarization.

Guidelines for Selecting Visual Summarization Techniques. Our analysis can offer some crude guidance on choosing a technique for a given dataset and task. For instance, Sequence Synopsis is the best option when the data contains numerous important but sporadically occurring events. CoreFlow may be the best option if the user is looking for a quick summary with the most common events. SentenTree may be the best if the user is interested in the way events unfold in a branch-and-merge structure. However, our current analysis is insufficient to offer detailed and quantified technique recommendations. Further work is necessary to model the effects of dataset characteristics on technique effectiveness.

10. Conclusion

In this work, we present the experiment design and results comparing the effectiveness of three different visual summarization techniques for event sequence data. To the best of our knowledge, our study is the first of its kind to offer a comprehensive comparison of such techniques. The insight-based method can potentially inform future work that evaluates the effectiveness of new or existing techniques. Our quantitative and qualitative analysis results also provide insights on the techniques' performance, the trade-offs involved in event sequence visual summarization, and rooms of improvement for new summarization techniques.

Acknowledgments. We would like to thank Catherine Plaisant and Ben Shneiderman for providing access to the datasets and EventFlow. We also thank Hannah Bako and Yishan Ding for their assistance with the study design, as well as Esmotara Rima for helping with the front-end implementation.

References

[AAS*19] ALLARD, TONY, ALVINO, PAUL, SHING, LESLIE, et al. "A dataset to facilitate automated workflow analysis". *PLOS ONE* 14.2 (Feb. 2019), 1–22. DOI: 10.1371/journal.pone.0211486. URL: https://doi.org/10.1371/journal.pone.0211486 5.

[BLST13] BARR, DALE J., LEVY, ROGER, SCHEEPERS, CHRISTOPH, and TILY, HARRY J. "Random effects structure for confirmatory hypothesis testing: Keep it maximal". *Journal of Memory and Language* 68.3 (2013), 255–278. ISSN: 0749-596X. DOI: https://doi.org/10.1016/j.jml.2012.11.001. URL: https://www.sciencedirect.com/science/article/pii/S0749596X12001180 7.

[BMBW14] BATES, DOUGLAS, MÄCHLER, MARTIN, BOLKER, BEN, and WALKER, STEVE. "Fitting linear mixed-effects models using lme4". *arXiv preprint arXiv:1406.5823* (2014) 7.

[BZS*20] BARTOLOMEO, SARA DI, ZHANG, YIXUAN, SHENG, FANGFANG, et al. "Sequence Braiding: Visual Overviews of Temporal Event Sequences and Attributes". *IEEE Transactions on Visualization and Computer Graphics* (2020). DOI: 10.1109/tvcg.2020.3030442 3.

[CBM*13] CARTER, ELIZABETH A., BURD, RANDALL S., MONROE, MEGAN, et al. *Using EventFlow to Analyze Task Performance During Trauma Resuscitation*. 2013. URL: http://www.cs.umd.edu/hcil/trs/2013-19/2013-19.pdf 5.

[CPYQ18] CHEN, YUANZHE, PURI, ABISHEK, YUAN, LINPING, and QU, HUAMIN. "StageMap: Extracting and Summarizing Progression Stages in Event Sequences". *2018 IEEE International Conference on Big Data (Big Data)*. 2018, 975–981. DOI: 10.1109/BigData.2018.8622571 2, 3.

[CvWvW18] CAPPERS, BRAM C.M., van WIJK, JARKE J., and van WIJK, JARKE J. "Exploring Multivariate Event Sequences Using Rules, Aggregations, and Selections". *IEEE Transactions on Visualization and Computer Graphics* (2018). DOI: 10.1109/tvcg.2017.2745278 3.

[CXR18] CHEN, YUANZHE, XU, PANPAN, and REN, LIU. "Sequence Synopsis: Optimize Visual Summary of Temporal Event Data". *IEEE Transactions on Visualization and Computer Graphics* (2018). DOI: 10.1109/tvcg.2017.2745083 1–4.

[DG17] DUA, DHEERU and GRAFF, CASEY. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml 4.

[DM19] DEXTRAS-ROMAGNINO, KIMBERLY and MUNZNER, TAMARA. "Segmentifier: Interactive Refinement of Clickstream Data". *Computer Graphics Forum* (2019). DOI: 10.1111/cgf.13715 3.

[FKSS06] FAILS, JERRY ALAN, KARLSON, AMY, SHAHAMAT, LAYLA, and SHNEIDERMAN, BEN. "A visual interface for multivariate temporal data: Finding patterns of events across multiple histories". (2006), 167–174 2.

[GCGC13] GOTZ, DAVID, CAO, NAN, GOLDBRAICH, ESTHER, and CARMELI, BOAZ. "GapFlow : Visualizing Gaps in Care for Medical Treatment Plans". 2013 2.

[GGJ*21] GUO, YI, GUO, SHUNAN, JIN, ZHUOCHEN, et al. "Survey on visual analysis of event sequence data". *IEEE Transactions on Visualization and Computer Graphics* 28.12 (2021), 5091–5112 4.

[GJC*22] GUO, SHUNAN, JIN, ZHUOCHEN, CHEN, QING, et al. "Interpretable Anomaly Detection in Event Sequences via Sequence Matching and Visual Comparison". *IEEE Transactions on Visualization and Computer Graphics* 28.12 (2022), 4531–4545. DOI: 10.1109/TVCG.2021.3093585 5.

[GJG*19] GUO, SHUNAN, JIN, ZHUOCHEN, GOTZ, DAVID, et al. "Visual Progression Analysis of Event Sequence Data". *IEEE Transactions on Visualization and Computer Graphics* (2019). DOI: 10.1109/tvcg.2018.2864885 2, 3, 5.

[Grü07] GRÜNWALD, PETER. *The Minimum Description Length Principle*. Jan. 2007. ISBN: 9780262256292. DOI: 10.7551/mitpress/4643.001.0001 2, 4.

[GS14] GOTZ, DAVID and STAVROPOULOS, HARRY. "DecisionFlow: Visual Analytics for High-Dimensional Temporal Event Sequence Data". *IEEE Transactions on Visualization and Computer Graphics* (2014). DOI: 10.1109/tvcg.2014.2346682 2, 3.

[GWP14] GOTZ, DAVID, WANG, FEI, and PERER, ADAM. "A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data". *Journal of Biomedical Informatics* (2014). DOI: 10.1016/j.jbi.2014.01.007 3.

[GXZ*18] GUO, SHUNAN, XU, KE, ZHAO, RONGWEN, et al. "EventThread: Visual Summarization and Stage Analysis of Event Sequence Data". *IEEE Transactions on Visualization and Computer Graphics* (2018). DOI: 10.1109/tvcg.2017.2745320 2, 3, 5.

[GZW*20] GOTZ, DAVID, ZHANG, JONATHAN, WANG, WENYUAN, et al. "Visual Analysis of High-Dimensional Event Sequence Data via Dynamic Hierarchical Aggregation". *IEEE Transactions on Visualization and Computer Graphics* (2020). DOI: 10.1109/tvcg.2019.2934661 3.

[HB10] HEER, JEFFREY and BOSTOCK, MICHAEL. "Crowdsourcing graphical perception: using mechanical turk to assess visualization design". *CHI* (2010). DOI: 10.1145/1753326.1753357 3.

[HOB94] HARRISON, BEVERLY L., OWEN, RUSSELL, and BAECKER, RONALD M. "Timelines: An Interactive System for the Collection and Visualization of Temporal Data". GI '94 (1994), 141–148. ISSN: 0713-5424. URL: http://graphicsinterface.org/wp-content/uploads/gi1994-17.pdf 2.

[HWS17] HU, MENGDIE, WONGSUPHASAWAT, KRIST, and STASKO, JOHN T. "Visualizing Social Media Content with SentenTree". *IEEE Transactions on Visualization and Computer Graphics* (2017). DOI: 10.1109/tvcg.2016.2598590 1–4.

[KAF*08] KEIM, DANIEL A., ANDRIENKO, GENNADY, FEKETE, JEAN-DANIEL, et al. "Visual Analytics: Definition, Process, and Challenges". *Information Visualization* (2008). DOI: 10.1007/978-3-540-70956-5_7 2.

[Kar94] KARAM, GERALD M. "Visualization using timelines". *ISSTA '94* (1994). DOI: 10.1145/186258.187157 2.

[KAS*20] KWON, BUM CHUL, ANAND, VIBHA, SEVERSON, KRISTEN A., et al. "DPVis: Visual Analytics With Hidden Markov Models for Disease Progression Pathways". *IEEE Transactions on Visualization and Computer Graphics* (2020). DOI: 10.1109/tvcg.2020.2985689 3.

[KH18] KIM, YOUNGHOON and HEER, JEFFREY. "Assessing Effects of Task and Data Distribution on the Effectiveness of Visual Encodings". *Computer Graphics Forum* (2018). DOI: 10.1111/cgf.13409 3, 6.

[KVP16] KWON, BUM CHUL, VERMA, JANU, and PERER, ADAM. "Peekquence: Visual Analytics for Event Sequence Data". *ACM SIGKDD Workshop on Interactive Data Exploration and Analytics (IDEA)*. 2016 2, 3.

[LDDH16] LIU, ZHICHENG, DEV, HIMEL, DONTCHEVA, MIRA, and HOFFMAN, MATTHEW. "Mining, pruning and visualizing frequent patterns for temporal event sequence analysis". *Proceedings of the IEEE VIS 2016 Workshop on Temporal & Sequential Event Analysis*. 2016, 2–4 2.

[LH14] LIU, ZHICHENG and HEER, JEFFREY. "The Effects of Interactive Latency on Exploratory Visual Analysis". *IEEE Transactions on Visualization and Computer Graphics* (2014). DOI: 10.1109/tvcg.2014.2346452 6.

[LKD*17] LIU, ZHICHENG, KERR, BERNARD, DONTCHEVA, MIRA, et al. "CoreFlow: Extracting and Visualizing Branching Patterns from Event Sequences". *Computer Graphics Forum* (2017). DOI: 10.1111/cgf.13208 1–4.

[LLMB19] LAW, PO-MING, LIU, ZHICHENG, MALIK, SANA, and BASOLE, RAHUL C. "MAQUI: Interweaving Queries and Pattern Mining for Recursive Event Sequence Exploration". *IEEE Transactions on Visualization and Computer Graphics* (2019). DOI: 10.1109/tvcg.2018.2864886 3.

[LWD*17] LIU, ZHICHENG, WANG, YANG, DONTCHEVA, MIRA, et al. "Patterns and Sequences: Interactive Exploration of Clickstreams to Understand Common Visitor Paths". *IEEE Transactions on Visualization and Computer Graphics* (2017). DOI: 10.1109/tvcg.2016.2598797 2, 3.

[MLL*13] MONROE, MEGAN, LAN, RONGJIAN, LEE, HANSEUNG, et al. "Temporal Event Sequence Simplification". *IEEE Transactions on Visualization and Computer Graphics* (2013). DOI: 10.1109/tvcg.2013.200 2, 5.

[MMPS13] MONROE, MEGAN, MALIK, SANA, PLAISANT, CATHERINE, and SHNEIDERMAN, BEN. *Introduction to EventFlow for Temporal Event Sequence Analysis*. 2013. URL: https://www.youtube.com/watch?v=FHgcJDnW8q8 5.

[Mon13] MONROE, MEGAN. *Basketball Play-By-Play Analysis Using EventFlow*. 2013. URL: https://vimeo.com/66965934?embedded=true&source=vimeo_logo&owner=5782051 5.

[MSD*16] MAURIELLO, MATTHEW LOUIS, SHNEIDERMAN, BEN, DU, FAN, et al. "Simplifying overviews of temporal event sequences". *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2016, 2217–2224 5.

[MSM*21] MAGALLANES, JESSICA, STONE, TONY, MORRIS, PAUL D, et al. "Sequen-c: A multilevel overview of temporal event sequences". *IEEE Transactions on Visualization and Computer Graphics* 28.1 (2021), 901–911 2, 3.

[MWP*12] MONROE, MEGAN, WONGSUPHASAWAT, KRIST, PLAISANT, CATHERINE, et al. "Exploring Point and Interval Event Patterns: Display Methods and Interactive Visual Query". (May 2012) 2.

[Nor06] NORTH, CHRIS. "Toward measuring visualization insight". *IEEE Computer Graphics and Applications* (2006). DOI: 10.1109/mcg.2006.70 3.

[PCK*18] POLACK, PETER J., CHEN, SHANG-TSE, KAHNG, MINSUK, et al. "Chronodes: Interactive Multifocus Exploration of Event Sequences". *Ksii Transactions on Internet and Information Systems* (2018). DOI: 10.1145/3152888 2, 3.

[PDF*16] PERER, ADAM, DRUCKER, STEVEN, FISHER, DANIEL, et al. *An IEEE VIS 2016 Workshop The Event Event: Temporal Sequential Event Analysis*. 2016. URL: https://eventevent.github.io/ 4.

[PG13] PERER, ADAM and GOTZ, DAVID. "Data-driven exploration of care plans for patients". *CHI Extended Abstracts* (2013). DOI: 10.1145/2468356.2468434 2.

[PHM*04] PEI, JIAN, HAN, JIAWEI, MORTAZAVI-ASL, B., et al. "Mining sequential patterns by pattern-growth: the PrefixSpan approach". *IEEE Transactions on Knowledge and Data Engineering* (2004). DOI: 10.1109/tkde.2004.77 3.

[Pla17] PLAISANT, CATHERINE. *Eventflow demo (short for talks) - Visual analytics for temporal event data*. 2017. URL: https://youtu.be/fLe1GawokXc?t=87 5.

[PMR*96] PLAISANT, CATHERINE, MILASH, BRETT, ROSE, ANNE, et al. "LifeLines: visualizing personal histories". (1996), 221–227 2.

[PMS*03] PLAISANT, CATHERINE, MUSHLIN, RICHARD, SNYDER, AARON, et al. "LifeLines: Using Visualization to Enhance Navigation and Analysis of Patient Records". (2003), 308–312 2.

[PW14] PERER, ADAM and WANG, FEI. "Frequence: Interactive mining and visualization of temporal frequent event sequences". *Proceedings of the 19th international conference on Intelligent User Interfaces*. 2014, 153–162 2, 3.

[PWH15] PERER, ADAM, WANG, FEI, and HU, JIANYING. "Mining and exploring care pathways from electronic medical records with visual analytics". *Journal of Biomedical Informatics* (2015). DOI: 10.1016/j.jbi.2015.06.020 3.

[QBW*20] QI, JI, BLOEMEN, VINCENT, WANG, SHIHAN, et al. "STBins: Visual Tracking and Comparison of Multiple Data Sequences Using Temporal Binning". *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), 1054–1063. DOI: 10.1109/TVCG.2019.2934289 3.

[RPP*17] ROBINSON, ANTHONY C., PEUQUET, DONNA J., PEZANOWSKI, SCOTT, et al. "Design and evaluation of a geovisual analytics system for uncovering patterns in spatio-temporal event data". *Cartography and Geographic Information Science* 44.3 (2017), 216–228. DOI: 10.1080/15230406.2016.1139467. eprint: https://doi.org/10.1080/15230406.2016.1139467. URL: https://doi.org/10.1080/15230406.2016.1139467 3.

[RT81] REINGOLD, EDWARD M. and TILFORD, JACK. "Tidier Drawings of Trees". *IEEE Transactions on Software Engineering* SE-7 (1981), 223–228 4.

[Shn96] SHNEIDERMAN, BEN. "The eyes have it: a task by data type taxonomy for information visualizations". *Proceedings 1996 IEEE Symposium on Visual Languages* (1996). DOI: 10.1109/vl.1996.545307 2.

[SND05] SARAIYA, PURVI, NORTH, CHRIS, and DUCA, KAREN. "An insight-based methodology for evaluating bioinformatics visualizations". *IEEE Transactions on Visualization and Computer Graphics* (2005). DOI: 10.1109/tvcg.2005.53 3.

[SPG14] STOLPER, CHARLES D., PERER, ADAM, and GOTZ, DAVID. "Progressive Visual Analytics: User-Driven Visual Exploration of In-Progress Analytics". *IEEE Transactions on Visualization and Computer Graphics* (2014). DOI: 10.1109/tvcg.2014.2346574 3.

[STT81] SUGIYAMA, KOZO, TAGAWA, SHOJIRO, and TODA, MITSUHIKO. "Methods for Visual Understanding of Hierarchical System Structures". *IEEE Transactions on Systems, Man, and Cybernetics* 11.2 (1981), 109–125. DOI: 10.1109/TSMC.1981.4308636 4.

[SZ00] STASKO, JOHN T. and ZHANG, EUGENE. "Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations". *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings* (2000). DOI: 10.1109/infvis.2000.885091 2.

[VJC09] VROTSOU, KATERINA, JOHANSSON, JIMMY, and COOPER, MATTHEW. "ActiviTree: Interactive Visual Exploration of Sequences in Event-Based Data Using Graph Similarity". *IEEE Transactions on Visualization and Computer Graphics* (2009). DOI: 10.1109/tvcg.2009.117 2.

[VN18] VROTSOU, KATERINA and NORDMAN, AIDA. "Exploratory visual sequence mining based on pattern-growth". *IEEE transactions on visualization and computer graphics* 25.8 (2018), 2597–2610 3.

[WCJ*17] WHITING, MARK A., COOK, KRIS, JORDAN CROUSER, R., et al. "VAST Challenge 2017: Mystery at the Wildlife Preserve". *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 2017, 173–178. DOI: 10.1109/VAST.2017.8585503 5.

[WG12] WONGSUPHASAWAT, KRIST and GOTZ, DAVID. "Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization". *IEEE Transactions on Visualization and Computer Graphics* (2012). DOI: 10.1109/tvcg.2012.225 2.

[WGP*11] WONGSUPHASAWAT, KRIST, GUERRA GÓMEZ, JOHN ALEXIS, PLAISANT, CATHERINE, et al. "LifeFlow: visualizing an overview of event sequences". *Proceedings of the SIGCHI conference on human factors in computing systems*. 2011, 1747–1756 2.

[Win13] WINTER, BODO. "Linear models and linear mixed effects models in R with linguistic applications". *arXiv preprint arXiv:1308.5499* (2013) 6.

[WPQ*08] WANG, TAOWEI DAVID, PLAISANT, CATHERINE, QUINN, ALEXANDER J., et al. "Aligning temporal data by sentinel events: discovering patterns in electronic health records". *CHI* (2008). DOI: 10.1145/1357054.1357129 2.

[WSSM12] WEI, JISHANG, SHEN, ZEQIAN, SUNDARESAN, NEEL, and MA, KWAN-LIU. "Visual cluster exploration of web clickstream data". *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2012). DOI: 10.1109/vast.2012.6400494 2, 3.

[WZT*16] WANG, GANG, ZHANG, XINYI, TANG, SHILIANG, et al. "Unsupervised clickstream clustering for user behavior analysis". *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2016, 225–236 3.

[YZZY15] YANG, DINGQI, ZHANG, DAQING, ZHENG, VINCENT W., and YU, ZHIYONG. "Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs". *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45.1 (2015), 129–142. DOI: 10.1109/TSMC.2014.2327053 6.

[ZDF*15] ZGRAGGEN, EMANUEL, DRUCKER, STEVEN, FISHER, DANYEL, et al. "(s|qu)eries: Visual Regular Expressions for Querying and Exploring Event Sequences". Proceedings of CHI 2015. ACM - Association for Computing Machinery, Apr. 2015. URL: https://www.microsoft.com/en-us/research/publication/squeries-visual-regular-expressions-for-querying-and-exploring-event-sequences/ 3.

[ZLD*15] ZHAO, JIAN, LIU, ZHICHENG, DONTCHEVA, MIRA, et al. "Matrixwave: Visual comparison of event sequence data". *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2015, 259–268 2.