Coping with Volume and Variety in Temporal Event Sequences: Strategies for Sharpening Analytic Focus

Fan Du, Ben Shneiderman, *Fellow, IEEE*, Catherine Plaisant, Sana Malik, and Adam Perer

Abstract—The growing volume and variety of data presents both opportunities and challenges for visual analytics. Addressing these challenges is needed for big data to provide valuable insights and novel solutions for business, security, social media, and healthcare. In the case of temporal event sequence analytics it is the number of events in the data and variety of temporal sequence patterns that challenges users of visual analytic tools. This paper describes 15 strategies for sharpening analytic focus that analysts can use to reduce the data volume and pattern variety. Four groups of strategies are proposed: (1) extraction strategies, (2) temporal folding, (3) pattern simplification strategies, and (4) iterative strategies. For each strategy, we provide examples of the use and impact of this strategy on volume and/or variety. Examples are selected from 20 case studies gathered from either our own work, the literature, or based on email interviews with individuals who conducted the analyses and developers who observed analysts using the tools. Finally, we discuss how these strategies might be combined and report on the feedback from 10 senior event sequence analysts.

Index Terms—Big data, temporal data, temporal event sequences, workflow, visual analytics, visualization, analytic focus

1 Introduction

The growing volume and variety of data presents both opportunities and challenges for visual analytics [62]. While big data provide valuable insights and novel solutions for business, security, social media, and healthcare, it also presents challenges due to two of its defining characteristics – volume and variety [68]. To obtain cleaner and more structured data, applications have been developed to assist in the data cleaning and wrangling process [4], [28]. Now, researchers are exploring ways to reduce the data volume and variety so as to sharpen the analytic focus. Visual analytics processes are likely to be more rapid when data volume is reduced and patterns are likely to be more discoverable when data variety is trimmed.

The analytic focusing problem is also being addressed in other fields such as knowledge discovery and sensemaking, but this paper emphasizes temporal event sequences, where point and interval events are organized into records. For example, in healthcare, point events may represent doctor visits or tests while interval events may represent a week-long hospitalization or taking a medication for 6 months. A patient record can be represented as a sequence of events, each event being of a particular event category. Descriptive information may be carried in record attributes (e.g., the gender of the patient), and event attributes (e.g., the name of the physician who placed the order).

A growing number of visual analytics and statistical tools have been built for temporal event sequences. These tools often have difficulty in dealing with the growing volume and variety of data:

- *Volume of data:* a dataset may consist of millions of records and hundreds of millions of events, which makes it hard to load and may result in long interactive latency during exploration.
- *Variety of patterns:* a single record may contain thousands or millions of events that fall into thousands of different event categories. Even in smaller datasets the sequential patterns of most records are unique and this variety makes it difficult to generate an overview or to reveal common patterns and anomalies. This definition of pattern variety complements the traditional definition of variety in big data which refers to the variety of data sources (structured, unstructured, semistructured).

While it is useful to have awareness of the data variety, analysts need useful ways of sharpening the analytic focus, leading to useful visualizations of global patterns and anomalies of interest. Just as camera images need to be in focus on objects or faces of interest and telescopes are best when tuned to spectral ranges (visual, ultraviolet, radio, x-ray, etc.), so too analytic tools will be most effective if users can focus their attention. The idea of an analytic pipeline or workflow is well established in mature application domains such as pharmaceutical drug discovery or NASA's remote sensing data analysis, but visual analytics in general and the analysis of temporal event sequences in particular are just beginning to have such workflows.

- *F. Du, B. Shneiderman, and S. Malik are with the Department of Computer Science and the Human-Computer Interaction Lab, University of Maryland. E-mail: {fan, ben, maliks}@cs.umd.edu.*
- *C. Plaisant is with the UMIACS and the Human-Computer Interaction Lab, University of Maryland. E-mail: plaisant@cs.umd.edu.*
- *A. Perer is with the IBM T.J. Watson Research Center. E-mail: adam.perer@us.ibm.com.*

Manuscript received 14 July 2015; revised 19 Feb. 2016; accepted 25 Feb. 2016. Date of publication 9 Mar. 2016; date of current version 3 May 2017. Recommended for acceptance by J. van Wijk.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2016.2539960

While it is useful to provide a comprehensive overview of the data at the start, analysts need ways to sharpen the analytic focus and gain answers to their research questions. Achieving this goal requires data analytic tools to (1) produce effective visualization of global patterns and anomalies of interest, and (2) keep visualizations simple with only necessary information.

This paper describes 15 strategies for sharpening analytic focus that analysts can use to control the data volume and pattern variety. It refines and expands on a set of 10 strategies proposed in a short article [56]. In addition, the paper includes numerous examples of the use of those strategies, selected from 20 case studies, and illustrates how new technologies and user interfaces can support these strategies. After describing the individual strategies and when they were used during the analysis process, we discuss how analysts often iterate over multiple strategies and report on three longer case studies in greater detail. Finally, we summarize the feedback of 10 event sequence analysts who reviewed the list of strategies and propose a basic workflow for applying the strategies.

2 Related Work

This section discusses applications in the related domains of data cleaning and wrangling, knowledge discovery and sensemaking, and temporal event sequence analysis. It also summarizes measures for data volume and pattern variety.

2.1 Data Cleaning and Wrangling

Real-world data are often challenging with respect to volume and variety. To obtain cleaner and more structured data, many applications have been developed to assist in the data cleaning and wrangling process [4], [23], [28], [50], [51]. Particularly, Gschwandtner et al. proposed a taxonomy [26] for dirty time-oriented data and developed a cleaning tool [25] targeting its special characteristics. Wrangling tools have also been built for the manipulation of time-series data such as transforming continuous numeric data into temporal event sequences [10], [19], [32], [43]. While cleaning and wrangling fix essential problems in the data (such as correcting erroneous values or integrating multiple data sources [27]), we assume in this paper that the data are ready for exploration, and discuss the next step of sharpening the analytic focus by reducing the data volume and pattern variety.

2.2 Data Focusing

Data focusing has been described in the field of knowledge discovery and sensemaking, where focusing techniques are employed to reduce the data volume by extracting subsets of data [52]. While statistical and machine learning algorithms are designed to handle huge data volumes, they may achieve even better results on subsets than on the entire data [5]. Instance and feature selection are two important tasks for data focusing in knowledge discovery and sensemaking, targeted at reducing the number of tuples and attributes, respectively. A unified instance selection framework has been proposed and evaluated, which creates a focused sample of the entire data by selecting representative prototypes from groups of similar tuples [35], [52], [53]. The goal of feature selection is to choose a subset of features to optimize the feature space according to certain criteria. Motoda and Liu summarized four groups of major feature selection algorithms [44]. These data focusing techniques are related to some of our extraction strategies. However, besides records and features (event categories), event sequences also contain patterns of ordered temporal events, which need to be handled particularly in order to reduce data variety. This paper emphasizes temporal event sequence analysis as opposed to multidimensional data and provides a more diverse set of analytic focusing strategies.

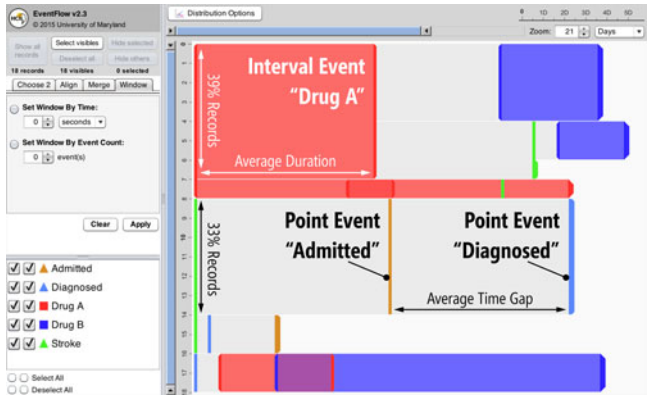

Fig. 1. The interface of EventFlow [41]. In the aggregated overview, the horizontal axis represents time and the vertical axis represents the number of records. Each row shows the event sequence of a record while multiple sequences with similar temporal patterns are visually combined and encoded as color bars.

2.3 Applications of Analytic Focusing Strategies

Analytic focusing strategies have been implemented in many temporal event sequence analysis tools. For example, EventFlow [41] visualizes a large number of event sequences by visually combining records with similar temporal patterns into an aggregated overview (Fig. 1). The aggregation is more effective when the number of unique complete sequences is controlled. EventFlow includes features that allow flexible filter-based and substitution-based strategies for data simplification and analytic focusing [42]. DecisionFlow [24] produces simplified visualizations by aggregating event episodes based on user queries. This goal-driven analytic focusing strategy enables it to handle large numbers of event sequences and event categories. Frequence [47] detects and visualizes frequent temporal patterns at different levels of detail to reduce the amount of data that need to be focused on. Outflow [63] presents a hierarchical clustering approach to aggregate similar patient records to reduce pattern variety. Scribe Radar [64] targets large-scale user activity data and employs extraction and temporal windowing strategies to produce focused visualizations based on analytical goals. Examples of the use of analytic focusing strategies can also be found in other temporal event sequence analysis tools [6], [14], [31], [60].

Interactions play an essential role in supporting these analytic focusing activities, which helps analysts specify the strategies to apply and fine tune the parameters after inspecting the effects. For example, in the light of the generic interaction frameworks [7], [65], "filter" or "query" interactions can be used for extracting records of

interest [24], "reconfigure" interactions can support splitting long streams of event sequences into periodic units and rearranging them to reveal cyclic patterns [42], [47], and "abstract" interactions can aggregate detailed patterns or event categories into a high-level overview [47], [64]. Showing results in a timely manner is critical for maintaining users' performance [36] but becomes difficult as the data volume grows. To address this issue, we introduce two iterative strategies to guide users starting from small and scaling up iteratively.

2.4 Measuring Volume and Variety

We separate the measures of volume and variety. Users of relational databases measure data volume by their size, usually in bytes, number of rows, or number of columns [3]. For temporal event sequences, the relational table approach is a poor fit, so visual analytic tools define their own data storage formats, invalidating traditional metrics. Since records and events are the basic elements of event sequences, a reasonable method to measure the volume of temporal event sequence data is by number of records and events.

Users of visualization tools measure the visual complexity by the number of distinct elements displayed on the screen [17]. For example, the number of dots in a scatterplot, the number of bars in a bar chart, and the number of nodes or links in a node-link diagram. EventFlow researchers measured the visual complexity of temporal event sequences by the number and average height of the aggregated chunks on the display [42]. These visual complexity metrics reflect the variety of the underlying data but are specific to the application. In this paper we will use a more general tool-independent metric for the variety of temporal event sequence data: the number of unique complete sequences, which indicates the number of patterns in the temporal event sequences and decreases as the data is simplified. We believe that meaningful insights are more likely to be found if data analysts can reduce the variety of patterns based on their analytical goals.

3 A Taxonomy of Analytic Focusing Strategies

This taxonomy of analytic focusing strategies for temporal event sequences is based on (1) the authors' combined experience in developing and evaluating multiple temporal event sequence analysis tools and dozens of case studies with real data and users, (2) the information contained in case studies gathered from the literature, and (3) email interviews with eight individuals who conducted such case studies and developers who observed analysts using the tools. Many of the case studies mentioned as examples below used EventFlow [41], but several other visual analytic tools (e.g., [46], [47], [48], [58], [61], [64]) are also presented. For each strategy, we include multiple examples taken from case studies. We describe the rationale for using the strategy and when possible indicate the resulting reduction in number of records or events (data volume) and unique complete sequences (pattern variety). In the older case studies, the reduction numbers could not always be collected because the original data were not available anymore, partners had moved on, or the functionality used to apply the strategy had changed and the operations could not be reproduced.

3.1 Extraction Strategies

3.1.1 S1: Goal-Driven Record Extracting

The experience from case studies strongly indicates that the question at hand often requires only a fraction of the records found in large datasets. For example, in the **ASTHMA** case study [40], the US Army Pharmacovigilance Center had 15 million patient histories. When the analytical goal was to determine how medications for asthma had been prescribed in the past six years, they extracted a set of 182,000 records. For Washington Hospital Center, among their over 1 million patients only 3,600 had been administered the radiology **CONTRAST** whose effect was being studied [61]. In a study of **EPILEPSY** at the Rennes University Hospital, out of 1.1 million epileptic patients only 4,800 had the inclusion criteria for the study. In these case studies, less than 1 percent of the total records were extracted.

Traditional query and extraction tools (e.g., i2b2 [2] or BTRIS [1] in the world of medicine) are needed before the focused analysis can begin. More advanced techniques are also possible, such as refining the extraction by sequential pattern or integrating statistics with analytics to filter the records by relevant features. In a Frequence [47] case study on **FOURSQUARE** check-in data, the analysts were interested in users who lived in New York City and had an active Twitter account. After extracting records by users' geographical and Twitter activity attributes, the number of records (users) was reduced by 91 percent (from over 200,000 to 17,739). Then, to investigate the check-in pattern of "Professional→Food", the analysts further refined the extraction and retained records that contained this pattern.

CareFlow [46] was integrated with patient similarity analytics [59] to focus on only clinically relevant patients. In the **HEART** case study, the database originally contained 50,625 patients. Using standard extraction techniques, researchers were able to determine that 4,644 met the criteria for heart failure. However, the complexity of these patients, who often suffered from many comorbidities, was tamed by filtering the dataset to show only patients with clinical proximity, determined by the devised similarity measures from analytics [59]. CareFlow could then be productively used to discover the common temporal patterns for the most clinically similar patients. For instance, in Fig. 2, limiting the view to the 300 most similar patients provides a comprehensible overview of the treatment progression, with a flow graph of only 47 nodes and 111 edges.

Based on our observations and interviews, this strategy is most often applied at the start of an analysis. Analysts continue using it during the analysis to eliminate errors and outliers or to focus on groups of records of interest. For example, when the analytical goal is to construct cohorts (e.g., [30], [66]), record extraction occurs at the very end, possibly followed by loading the extracted data in other tools (e.g., [38], [67]) for further exploration.

3.1.2 S2: Goal-Driven Event Category Extracting

Some studies require a large portion of the records in a database, but only a small fraction of the events in each record.

Fig. 2. Rather than visualizing all patients, CareFlow [46] employs a Goal-Driven Record Extraction strategy (S1) to focus users on only the most relevant 300 patients by applying similarity analytics, which results in a comprehensible flow graph of only 47 nodes and 111 edges.

For example, the **EPILEPSY** study only extracted the "seizure" events and the ten preceding administrations of anti-epileptic drugs. In the **ASTHMA** study [40], only event categories that represented asthma related medications were retained. This reduced the number of events by 34 percent (from 1,607 to 1,054) and the number of unique complete sequences by 33 percent (from 98 to 66). In another study related to **PROSTATE CANCER** radiation treatment [45], the analytical goal was finding what durations and intensity of radiation produce the fewest bone fractures yet still curtail the cancer. The analysts removed events such as eye, dental exams, and even the procedure's details to trim the dataset and greatly focus the analysis.

We observed that analysts often start working with even fewer event categories than those extracted and progressively add the remaining categories as needed to refine the analysis. In the **ASTHMA** case study [40], researchers started with each category of medications one by one and then looked at two at a time. Even three drugs at once turned out to introduce too much pattern variety and required applying other strategies to simplify the display. In another case study called **DEVIATION** [15], the goal of the National Children's Hospital analysts was to see if a required set of actions were taken in the correct order and what were the deviations. The data volume was small so events of all categories could be seen at once. However, the sequences were so diverse that no clear patterns for the deviations could be seen at first. The analysts started over with only two event categories (the first two event categories of the sequence performed in correct order), resulting in an immediate 98 percent reduction in the number of unique complete sequences and allowing them to easily spot a widespread protocol deviation. They then included other event categories one by one to find more complex deviations.

Based on our observations and interviews, this strategy is often used at the beginning of an analysis, to help analysts get started with a simpler view. Tools have been developed to guide the choice of a small number of meaningful event categories to look at. For example, Choose2 [39] suggests pairs of event categories based on metrics such as "maximize record coverage" (Fig. 3). This strategy continues to be useful during the analysis, with some users exploring a few event categories at a time but in a systematic fashion.

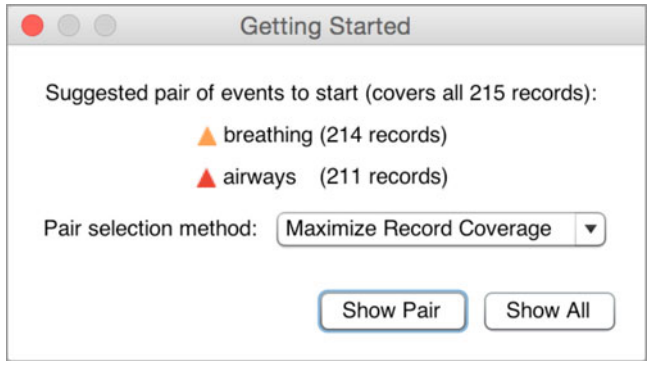

Fig. 3. The interface of Choose2 [39]. It automatically selects a pair of event categories to help users get started.

3.1.3 S3: Identifying Features Linked to Outcome

The holy grail of event sequence analysis remains understanding what sequences of events lead to a better outcome than others, leading to hypotheses about causation. Outcomes are often defined by inclusion of certain events in a record. For example, the ultimate outcome of the treatment of patients in the emergency department can be whether they were discharged alive or discharged dead, and the outcome of an online shopping clickstream is whether the person purchased something or not [37]. Identifying features (e.g., event categories, patterns) whose occurrences are correlated to an outcome is a useful step in the analysis of records with different outcomes.

In the **HEART** case study using CareFlow [46], the outcome was defined as negative if a patient was re-admitted into a hospital (or died) within 12 months of the initial diagnosis of heart disease. Of the 300 similar patients shown in Fig. 2, 76 patients had negative outcomes and 224 had positive outcomes. In this visualization, outcome is encoded visually, with positive outcomes as green and negative outcomes as red. As almost 75 percent of the patients managed to stay healthy, most of the event sequences are green. However, users can change the outcome measure interactively to test new hypotheses, or filter to only patients with a certain outcome. For instance, if users wish to examine only the paths that led to mortality, the display would update to show the progression of only the 10 patients who died.

This strategy is typically used at the beginning of an analysis and guides the extraction of records, event categories, and patterns. When the data exhibit extraordinary complexity, analysts might simplify the data before using this strategy to ensure meaningful features. Tools have been designed for comparing temporal event sequences and identifying features linked to outcomes [24], [38], [47], [67].

3.1.4 S4: Aligning

A common question in event sequence analysis is "what happened before and after a particular event of interest? (e.g., emergency room visit)", often in the context of a cause-and-effect investigation. A useful strategy is to align all records by the occurrence of a selected alignment event [60] (e.g., the first, Nth, or last occurrence). A side effect of the alignment is that records without the alignment event are removed from the visualization.

This alignment strategy has been commonly used. For example, in the **LIVER** case study [34], the analyst wanted to understand how patients developed liver disease after starting to use total parenteral nutrition (TPN) and records were aligned by the first TPN event so that the liver disease events after the alignment point could be reviewed. In the **TWITTER** case study [64], the analysts aligned the user logs by the first use of the feature of interest to analyze what other events preceded and followed the use of that feature. In the **BASKETBALL** case study [42], the analysts aligned the play-by-play sequence records by offense-to-defense transitions to investigate how well the University of Maryland team performed during these transitions.

Analysts might use the aligning strategy at any time during the analysis, but aligning is often followed by a temporal windowing (S5) to focus on events just before, just after, or just around the alignment.

3.1.5 S5: Temporal Windowing

In many cases only a relatively small window of time matters for a given analytical goal. The window selection might be arbitrary (e.g., only the most recent two weeks of data or the first 10 events) or goal driven (a period around alignment events). The window size could be defined by a time duration or number of events. In the case study of **PROSTATE CANCER** treatment [45], the window was set to be 10 years from each patient' diagnosis of prostate cancer. In the **TWITTER** case study [64], to understand users' behavior at their first website visit, the analysts examined only 10 events from the beginning of each session. Additional examples of the use of this strategy are given in the detailed case studies of Section 4.

Temporal windowing is typically applied at the beginning of an analysis for specifying the time span of interest or complying with the data scope imposed by an Institutional Review Board (IRB) review. It often follows aligning (S4) to extract events around the alignment point. When the records being studied are very long, temporal windowing dramatically reduces both data volume and pattern variety.

3.1.6 S6: Selecting Milestone Events

In many situations the large volume of data comes from streams of repetitious events. A typical strategy we observed is to keep only the events corresponding to "milestones". For example in social media log analysis, such as the use of Twitter, individual records typically include 100 s or 1,000 s of tweeting events. Sharpening the analytic focus might require thoughtful selection of milestone events such as the 1st, 10th, 100th, and possibly 1,000th tweets in each person's record. This dramatically reduces the clutter of tweets and allows analysts to study the timing of those milestones in respect to other events. For example, relationship to retweets, mentions, replies, etc. becomes clearer. Similarly, analysts might choose to retain only the dates of the 1st, 10th, 100th, and 1,000th followers. In the medical domain, we observed analysts only keeping the first diagnosis of diabetes instead of keeping all occurrences (the diagnosis has to be recorded at every doctor visit even if the disease never goes away). Analysts sometimes picked the third or fifth, just to be sure the first one was not a coding error.

In the **LOG** case study (described in Section 4.3), the analysts found that events of some categories repeatedly occurred within a short period of time, and selected milestone events to simplify the records. For every 50 events of the same category, only the first one was selected (i.e., the 1st, 51st, 101st, etc). Looking at a month of activity of one person, this strategy reduced the number of events by 98 percent (from 12,012 to 290) and the number of unique complete sequences by 30 percent (from 27 to 19) while still representing the varying amount of activities in each day. The most common daily pattern could be immediately identified.

This strategy has been used after visual review of the raw data, i.e., at the start of the analysis. The non-milestone events were then hidden from view, or removed entirely from the data when data volume was an issue.

3.1.7 S7: Random Sampling of Records

If the previous strategies fail to sufficiently reduce the data volume, random sampling of records – such as extracting every 10th record – may become a reasonable strategy. In the **ASTHMA** study [40], from a total number of 182,000 patients, the US Army Pharmacovigilance Center selected a random sample of 100 asthma patients under age 65 with a new LABA (Long-Acting Beta Agonists) prescription. With random sampling of records, analysts are likely to want to extract as many records as possible or to balance groups of cohort, so tools that iteratively alter the size of the selected set would be useful. There is often some benefit in getting a rough indication of the prevalence of the patterns being sought [22].

Random sampling was only used as a last resort (e.g., when the data didn't even load). Two other potential strategies, random sampling of events or event categories within records, were not observed in the case studies and do not seem useful (even though selecting milestone events in a stream could be considered as a form of "goal oriented sampling" of events).

3.2 Folding Strategy

3.2.1 S8: Temporal Folding

Some datasets have records that are long streams which may be more successfully analyzed by folding (or splitting) each record into yearly, monthly, or weekly units. In the radiology **CONTRAST** case study [61], each patient record was split into segments centered around each administration of contrast material. In the **FOURSQUARE** case study [47], to break down the long streams of users' check-in events, the analyst folded each stream with a 6-hour sliding window, yielding short sequences of events that occur exactly within the window. In a study of interpersonal **VIOLENCE** [49] conducted at Yale University, the 90-day record of each of the 141 participants consisted of detailed events such as drug and alcohol use as well as incidents of arguments, physical violence, sexual abuse, etc. The pattern variety was overwhelming until the long streams were broken into weekly records, thereby revealing weekend conflicts and drug use. Further breaking the records by day showed yet different patterns.

Interactive or automatic folding may detect cyclic phenomena and reduce the variety of patterns to visualize.

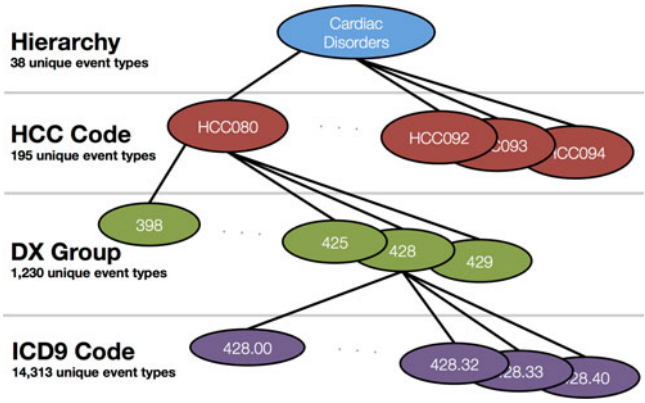

Fig. 4. An illustration of grouping using pre-existing hierarchies.

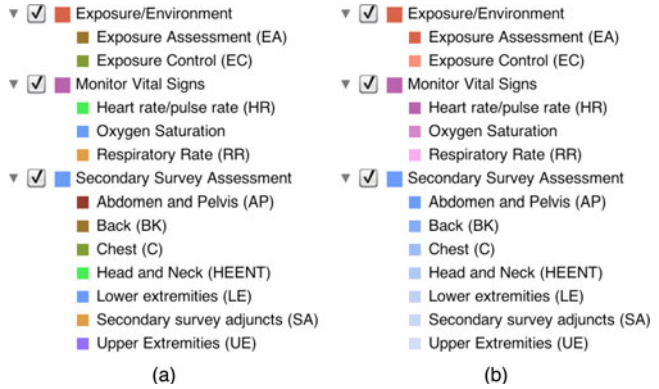

Fig. 5. Illustrations of color encoding before (a) and after (b) visual grouping in the WORKFLOW case study.

Folding alone does not address data volume issues as the number of folded records increases – although the number of events remains constant. However, once the temporal folding is done, pattern variety may be reduced, and record extraction may become useful again (e.g., keeping only days with violent events). In the case studies, analysts usually started by inspecting a sample of the data in visualization tools to decide how to fold, e.g., that it was better to use 4 am as the end of the day instead of midnight in the **VIOLENCE** case [49]. None of the tools used in the case studies included built-in folding features so it was done in the source database and then the analysis restarted.

3.3 Pattern Simplification Strategies

3.3.1 S9: Grouping Event Categories

With the explosion of the number of event categories, aggregation becomes necessary [18]. For example, seeing global patterns is impossible when there are over 400 types of lung cancer and over 200 types of bone cancer. Replacing all lung cancers with a single event category and all bone cancers with a single event category reduces the pattern variety. While the number of events remains the same, the simplification sharpens the analytic focus and allows the analysts to determine that lung cancers often spread to bones, but bone cancers rarely spread to the lungs.

In an EventFlow case study with the Children's National Hospital (**WORKFLOW**) the analysts reduced the pattern variety by aggregating 61 emergency department procedures into 18 meaningful groups – using their domain knowledge. This strategy reduced the number of unique complete sequences by 39 percent (from 449 to 274). Scribe Radar [64] supports a 6-level hierarchy for event categories, from the high-level "client" events (e.g., web, iPhone, and Android) to the low-level "action" events (e.g., click and hover). It used page-level events (e.g., home, profile, search) in the **TWITTER** case study for investigating how people begin spending time on Twitter, and drilled into section-level events (e.g., the follower and following sections in the profile page) to analyze where people performed the "follow" action.

Sometimes, events can be grouped using existing ontologies. Since most medical events (e.g., diagnoses and medications) are organized in a standard ICD9 code hierarchy, case studies using Frequence [47] and Care Pathway Explorer [48] took advantage of them to mine frequent patterns of medical events at different levels of detail. For example, Fig. 4 shows a set of events under cardiac disorders in the diagnosis hierarchy, which contains four different levels. The first level is the Hierarchy Name, which is the highest level in the Hierarchical Condition Categories (HCC) used in Medicare Risk Adjustment provided by Centers for Medicare and Medicaid Services (CMS). This level has 38 distinct event categories. The second level is the more detailed Hierarchical Condition Categories, which contains 195 different codes. The third level contains 1,230 unique Diagnosis (DX) group names (the first three digits of the ICD9 code). The fourth-level contains 14,313 different codes of the International Classification of Diagnosis ninth edition (ICD9). Similar hierarchies exist for other data types, such as medications, procedures, and labs. All levels in these hierarchies are a many-to-one mapping to the higher levels.

While grouping event categories is often used to preprocess the data before analysis (e.g., to aggregate drugs names by classes [40], [46]), dynamic aggregation (i.e., undoing and redoing the grouping) may be needed to provide variable levels of detail during the analysis. For example, in the **WORKFLOW** case study, the analysts began by grouping a large number of event categories into high-level ones. Then, other strategies were applied and observations were made. To verify the findings, analysts needed to undo several groupings to see a detailed display with low-level event categories. This requires dynamic undo capabilities that go beyond the standard rolling back to a past state.

Besides grouping event categories by modifying the underlying data, we also observed analysts mapping similar event categories with similar colors, helping them to reduce the number of distinct colors on the display and perceive event categories with similar colors as a group (Gestalt principle of similarity [29]). This visual grouping approach does not reduce data volume or number of unique sequences, but seemed particularly useful when analysts were trying to generate a more focused visualization including low-level event categories. Fig. 5 shows examples of color encoding before and after visual grouping in **WORKFLOW**.

3.3.2 S10: Coalescing Repeating Point Events into One

When dealing with patient histories, a major simplification is to convert multiple point events, such as 25 normal blood pressure readings over 12 months into a simpler more meaningful single interval event that shows normal blood

pressure during the 12 months. In the **HYPERTENSION** case study, researchers analyzed a cohort of 1,294 patients with hypertension enrolled in a chronic disease management program at the Vanderbilt University Medical Center [58]. Researchers were interested in understanding the trends of blood pressure in these patients, as keeping blood pressure in control is critical for reducing morbidity and mortality. In total, there were 192,864 blood pressure measurements for all 1,294 patients. However, to model control, all consecutive instances of in-control and out-of-control assessments were merged into intervals. This reduced the number of blood pressure measurement events by 97 percent (from 192,864 to 6,537), yielding only about five intervals per patient.

In most cases, analysts apply this strategy in visualization tools so as to inspect the effect and decide on the level of simplification needed. However, it may also be applied in the source database to reduce the data volume, especially when the number of repeating events is very large.

3.3.3 S11: Coalescing Repeating Interval Events into One

The pharmacovigilance **ASTHMA** project raised this issue for patients who received repeated prescriptions for the same medication [40]. Patients often refilled the prescription early, which appeared as an overlap of two intervals, or delayed their refill, which appeared as a gap between the intervals. Analysts simplified the patterns by merging intervals with overlaps of less than 15 days or gaps of less than 10 days resulting in long intervals indicating the "drug episode". For a subset of 100 asthma patients, applying this strategy on the LABA events (prescriptions of Long-Acting Beta Agonists) reduced the number of LABA events by 19 percent (from 355 to 288), and the number of unique LABA patterns by 45 percent (from 31 to 17). This strategy is another version of S10 for interval events. One additional occasion of applying this strategy is after an S10, for simplifying the interval events produced by S10.

3.3.4 S12: Converting Hidden Complex Events into One

In many application domains, some high level complex events such as a heart attack or surgery may consist of 20-100 events that all happened within a time period (e.g., certain blood tests or imaging over multiple days). These detail events may not be relevant to the analysis, so all of them can be identified and replaced by a single event. Data mining approaches such as knowledge-based temporal abstraction [55] or frequent pattern mining [9] might also be able to identify such complex events, leading to suggested simplifications. EventFlow [41] allows users to search for a specific event pattern – including temporal constraints and the absence of events – and replace it with a shorter pattern or even a single event (Fig. 6).

In the **HYPERLIPIDEMIA** case study of Care Pathway Explorer [48], a cohort of 14,036 patients with hyperlipidemia was analyzed. The patients had a total of 70,379 diagnosis events and 97,189 medication events during their first year after diagnosis. At this level, few insights were found so the analysts decided to focus on specific subcohorts with pre-conditions. After filtering to patients with a hypertension pre-condition, the patient count was reduced to 2,800 patients, featuring 14,979 hyperlipidemia-related diagnoses and 24,898 medication events. Then, frequent sequence mining analytics were run to reduce this event set into only 28 frequent event sequences that could be visualized easily using a Sankey diagram. Unveiling these hidden event sequences led to interpretable insights, such as finding that patients with higher low-density lipoprotein cholesterol levels tend to use higher levels of pain medication.

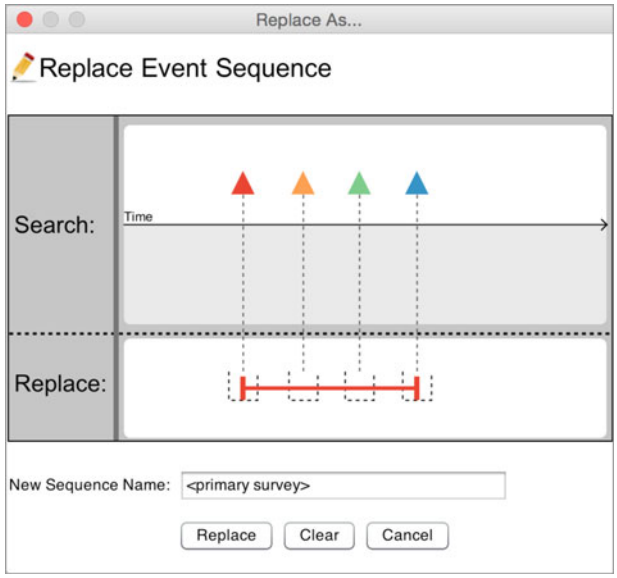

Fig. 6. The Search & Replace interface of EventFlow [41]. Users can search for a pattern and replace it with another pattern. In this figure, a pattern of four point events is converted into an interval event.

Besides converting events in the actual data, we also saw analysts assign rainbow colors to sequences that form a process. For example, in the **DEVIATION** case study [15], the analysts assigned red, orange, green, blue, and purple to the five procedures that should be performed in order to follow the protocol (Fig. 9b). The rainbow colors helped the analysts easily detect correct sequences and deviations.

This strategy seems to be more useful after other extraction and simplification strategies have been used. Inspecting the data in visualization tools helps analysts search for those complex events and spot variations.

3.3.5 S13: Bucketing by Time Period

When the number of events per record and the variety of patterns is so high that most sequences become unique, a powerful strategy is to bucket events by a fixed time period (such as one minute when dealing with computer usage or weeks when dealing with long term medication use). Events occurring during the time period are replaced by one event – or just a few. Analysts can define rules to summarize each time period, e.g., using the most common event category, a combination of the most common and least common event category, or dynamic characteristics of the data such as the occurrence of spikes, increases or drops, etc. This strategy creates events at regular intervals and produces patterns which can be further simplified using the other analytic focusing strategies. Reviewing sample raw data helps

TABLE 1
Examples of 20 Case Studies using Multiple Analytic Focusing Strategies Gathered from the Literature, or Based on Email Interviews with Individuals Who Conducted the Analyses and Developers Who Observed Analysts Using the Tools

	ASTHMA [40]	BASKETBALL [42]	CONTRAST [61]	DEVIATION [15]	DRUG [13]	EHR COHORTS [20]	EPILEPSY [41]	FOURSQUARE [47]	HEART [46]	HYPERLIPIDEMIA [48]	HYPERTENSION [58]	LIVER [34]	LOG [41]	MEMuRY [33]	MOTION [12]	PROSTATE CANCER [45]	PROSTATE CANCER 2 [11]	TWITTER [64]	VIOLENCE [49]	WORKFLOW [41]
S1: Goal-Driven Record Extracting	•	•	•	•	•	•	•	•	•	•	•	•	•	•			•	•	•	•
S2: Goal-Driven Event Category Extracting	•	•		•	•		•	•	•	•	•	•		•	•	•	•	•		•
S3: Identifying Features Linked to Outcome								•	•	•	•	•				•				
S4: Aligning	•	•	•	•	•		•	•	•	•		•			•	•	•	•	•	•
S5: Temporal Windowing					•	•	•	•	•	•	•	•	•		•	•	•	•		•
S6: Selecting Milestone Events								•	•	•										
S7: Random Sampling of Records	•				•															
S8: Temporal Folding		•	•					•	•					•				•		
S9: Grouping Event Categories	•			•	•		•	•	•	•	•	•		•	•	•	•	•	•	•
S10: Coalescing Repeating Point Events into One				•		•		•	•	•	•		•	•						•
S11: Coalescing Repeating Interval Events into One	•				•							•	•							
S12: Converting Hidden Complex Events into One	•			•	•			•		•										
S13: Bucketing by Time Period												•								
S14: Analyzing Small Subset then Larger One	•				•															
S15: Partitioning					•							•			•			•		

A bullet point indicates a strategy was used in a case study.

define the bucketing strategy. We have seen this strategy only executed using external scripts. As of today, it has not been implemented in the tools we surveyed.

3.4 Iterative Strategies

Extraction strategies can be executed in traditional database systems, then the extracted results loaded in the interactive visual analytic tools. However, when volume and variety remained a problem after the initial extraction, most analysts started by looking at a small subset of data or chose a partitioning strategy.

3.4.1 S14: Analyzing A Small Subset then A Larger One

It is common practice to conduct visual inspections of even small subsets of data to facilitate the data cleaning and wrangling process. Our interviews and case studies indicate that analysts also find the visual inspection of a subset of data to be effective for guiding the choice of extraction strategies (e.g., identifying interesting event categories) and for devising meaningful simplification strategies (e.g., merging short intervals into longer ones). Since visual inspection of the transformation results allows analysts to refine their work and further sharpen analytic focus, the data transformations need to be defined and tested within the interactive visual analytic tool. Analysts can define and test the extraction, temporal folding, and pattern simplification strategies on a subset of data. When they are satisfied with the results, they may apply the strategies to larger subsets or the entire data. In the case of EventFlow [41], most of the strategies have been implemented and can be saved, allowing them to be reapplied on a larger dataset at load time before the visualization is generated. Those data transformations could also be executed in the source database as well, or in a separate analytic focusing tool. Progressive Visual Analytics techniques [21], [57] may provide complementary solutions.

3.4.2 S15: Partitioning

We also observed analysts use a "divide and conquer" strategy by partitioning a large dataset into disjoint subsets that could be analyzed independently without losing any information. One common practice is to partition by event categories. For example, in a dataset of patients' drug prescriptions, disjoint subsets can be created for patients who have only taken certain drugs, such as only Drug A, only B, A and B. This partition method is particularly effective for data where event categories are unevenly distributed (i.e., each record contains only some of the event categories). Another approach is to create the subsets based on record attributes. For example, a large set of customer histories might be partitioned by the state of residence, or age groups. After the partition, each disjoint subset contains a complete (i.e., all records satisfying the condition are included) but a small portion of the entire data, which preserves the temporal patterns among events of different categories. Sometimes the differences of patterns in each partition are significant, so the analysis by partitions may produce clearer insights than trying to analyze the full dataset.

4 Examples of Combining Multiple Strategies

Our case study review shows that most case studies applied multiple strategies, as summarized in Table 1, which includes 20 case studies using 12 different visualization tools [11], [12], [20], [33], [41], [46], [47], [48], [58], [60], [61], [64]. To illustrate how analysts combine and iterate over multiple strategies, we describe three case studies in detail.

4.1 DEVIATION Case Study

In the **DEVIATION** case study [15], researchers at the National Children's Hospital were studying how well the recommended trauma treatment protocol (ATLS [8]) had been followed. ATLS specifies two order-sensitive surveys. Specifically, the primary survey consists of four injury evaluations: Airway (*A*), Breathing (*B*), Circulation (*C*), and Disability (*D*), and is followed by the secondary survey for identifying other injuries. In this case study, the hospital wanted to find out what percentage of the patients were treated following the correct protocol and what the

deviations were. The dataset was small (215 records, 1,991 events) but the variety of patterns was surprisingly large – which was unexpected since the protocol should be followed consistently (Fig. 9a). The analysts could not readily discern the common deviations and had to apply a series of focusing strategies to reach their goal.

First, the analysts chose only two event categories (Airway and Breathing) (**S2: category extract**), resulting in a 98 percent reduction in the number of unique complete sequences (from 208 to only 4), which allowed them to immediately spot a common problem: 14 percent of the patients were treated in a $B \rightarrow A$ order, while the airway evaluation was missing for 2 percent of the patients. The analysts then added Circulation events, which fell in two different categories: "central pulse" and "distal pulse". Since the two event categories of pulse checks were both acceptable circulation evaluations, the analysts combined them into a single one named "any pulse" (**S9: category grouping**). This made clear that 17 percent of the patients received two consecutive pulse checks (once for each type of checks), which was an acceptable practice. To remove the noise, the analysts replaced any two consecutive "any pulse"events with a single one, at the time of the earlier pulse check (**S10: n points to one**).

Finally, the analysts added the event categories for disability evaluation and secondary survey. This produced an overview of all events related to the protocol (Fig. 9b). The analysts found that the most prevalent pattern was $A \rightarrow B \rightarrow C \rightarrow D$ followed by a secondary survey. This indicated that 48 percent of the patients had been correctly treated following the protocol. To focus on the deviations, the analysts removed the records that were in the correct order (**S1: record extract**), which made it easier to review the 28 deviations. They spot that the most common deviation was to perform the disability evaluation during the secondary survey, which occurred in 21 percent of the cases. Compared with the original data, the successive analytic focusing strategies had reduced the number of events by 73 percent (from 1,991 to 542), and the number of unique complete patterns by 87 percent (from 208 to 28).

4.2 DRUG Case Study

In the **DRUG** case study [13], analysts from the business and pharmacy schools aimed to describe the patterns of hypertension drug prescriptions, for example, when patients start taking multiple drugs, switch from one drug to another, or interrupt their prescriptions (i.e., significant gaps, indicating poor adherence to the drug regimen). The drug dataset contained 790,638 patients with 9,435,650 prescriptions of 5 drugs, extracted using a combination of record extraction (**S1: record extract**) and temporal windowing with a two year window (**S5: windowing**). Each prescription is an interval event of 30 or 60 days, and each drug is an event category. Aggregation of similar drugs into larger drug classes had already been done in the database prior to extraction (**S9: category grouping**). The entire dataset failed to load into the visualization tool, so the analysts started by looking at a sample of a few thousand of records (**S14: small then large**). This led them to decide to partition the dataset (**S15: partitioning**) by separating out the patients who only ever took 1 drug. This reduced the number of events by 36 percent (6,052,157 remaining for the patients who had used at least 2 drugs during their treatment). The analysts continued the process and partitioned the data into 31 disjoint subsets that could be analyzed separately (i.e., each permutation for patients who took 1 drug, 2 drugs, then 3, 4 and the last subset for patients who used all 5 of the drugs).

The partitioning allowed the largest subset (713,971 events, i.e., less than 8 percent of the original data) to fit in the visualization tool for visual inspection. The analysts started with a medium size subset consisting of 37,613 patients who took only 2 drugs (Fig. 10a). They merged the repeating short intervals of the same event category into long ones, by removing small gaps and overlaps in the prescriptions (**S11: n intervals to one**). This strategy reduced the number of events by 74 percent (from 588,847 to 151,274), and the number of unique complete sequences by 84 percent (from 26,906 to 4,417). To focus on overlaps and interruptions among prescriptions, the analysts searched and replaced the overlap and interruption patterns with new events (**S12: pattern to one**). Since new events were created, this strategy increased the number of events by 57 percent (from 151,274 to 237,078), and the number of unique complete sequences by 51 percent (from 4,417 to 6,659). Finally, the analysts grouped the event categories of single drugs into a new category (**S9: category grouping**), representing the periods when patients were taking only one drug of any kind. The number of events stayed the same but the number of unique complete sequences was reduced by 42 percent (from 6,659 to 3,871).

In total, for this 2-drug subset the above strategies reduced the number of events by 60 percent (from 588,847 to 237,078) and the number of unique complete sequences by 86 percent (from 26,906 to 3,871). A focused display was produced and helped the analysts to answer their research questions (Fig. 10b). The analysts reapplied these strategies to simplify the other 30 subsets and combined their findings.

The partitioning strategy requires analysts to keep in mind the relative size of each subset. We observed them resizing the displays of different subsets to match the relative number of patients. For example, in Fig. 7 the analysts were analyzing three subsets, looking at the patterns of two drugs (alone or combined) for a total of 166,478 patient records. They repeated this display layout for all possible pairs of drugs, compared their relative number of records, and then moved on to patients using 3, 4 or 5 drugs. A useful addition would be visual tools for representing the relative size of partitions.

4.3 LOG Case Study

LOG is a recent case study in which security researchers were interested in developing new methods for detecting insider threats. The dataset [16] contained approximately 180 million events from the monitoring of computer usage, consisting of six categories (e.g., login, email, web browsing, etc.) with an average of 33 thousand events per user. A black-box anomaly detection algorithm [54] was used to compute a suspiciousness score per person per day. Unfortunately, the algorithm did not provide any explanations for the suspiciousness scores, so EventFlow was used to help the researchers understand what might be anomalous in the temporal event patterns of the person-days

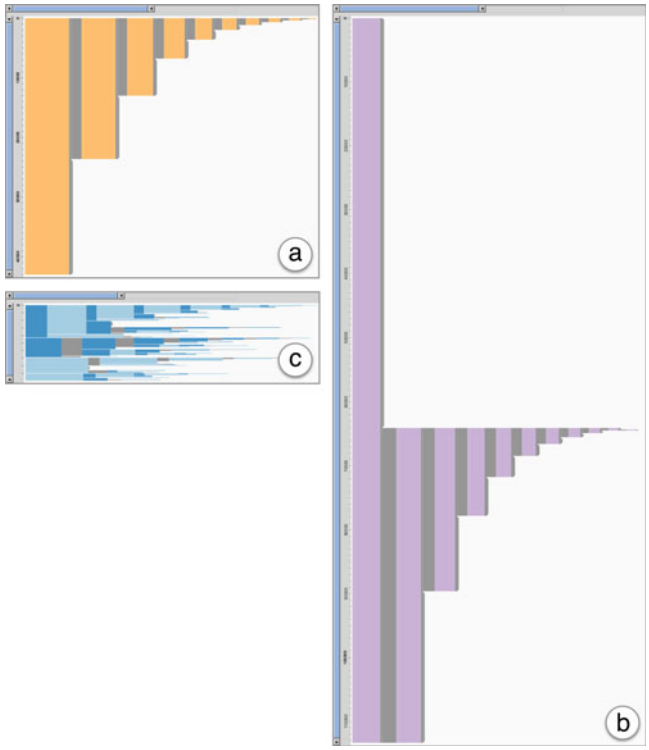

Fig. 7. The displays of the disjoint subsets of patients who took the yellow drug only (a), the purple drug only (b), and both drugs (c) (with light blue representing the use of a single drug, dark blue representing the concurrent use of the two drugs, and gray representing interruptions). The analysts resized the displays by the number of records in each subset to keep track of their relative sizes.

determined to be suspicious. The records of users with low scores during the whole period were removed (**S1: record extract**). Temporal windowing was used to keep only one month worth of data around the suspicious person-days (**S5: windowing**), and temporal folding was applied to cut the month-long streams into smaller day-long records (**S8: folding**). These three strategies together reduced the data volume by 99 percent, keeping only 1,311,337 events. As the researchers needed to review one user at a time, they partitioned the remaining data and created a subset for each suspicious user, yielding 56 disjoint subsets of 23,416 events on average (**S15: partitioning**). Each subset was then loaded into the visualization tool for visual inspection.

Since the computer usage events of the same category often repeatedly occurred within a short period of time and created visual clutter, the researchers coalesced repeating events (**S10: n points to one**) using EventFlow's "Search & Replace" operation. Repeating point events adjacent to each other within 10 minutes were converted into an interval, starting at the time of the first point event and ending at the time of the last one. For a medium size subset of 12,012 events, the number of unique complete sequences remained the same, but the visual clutter was eliminated due to a 96 percent reduction in the number of events (from 12,012 to 462).

To further simplify the data, the researchers then looked for patterns representative of suspicious days (**S3: link to outcome**) using a separate High Volume Hypothesis Testing tool (CoCo [38]), which supports comparing two groups of temporal event sequences. For each subset, CoCo identified event categories that occurred significantly more or less

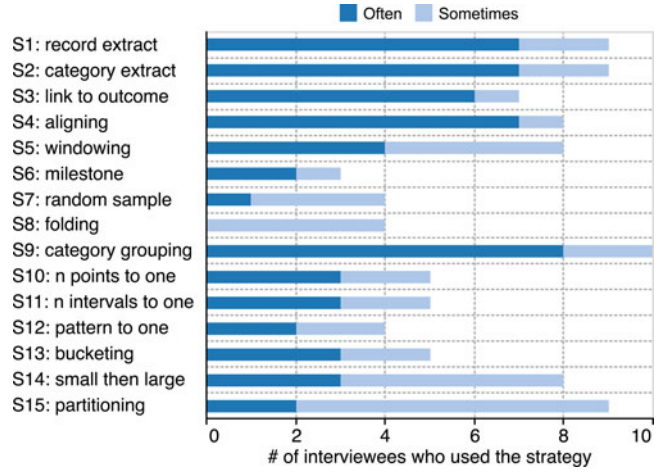

Fig. 8. The number of interviewees who used each strategy before. Data collected from our interviews with 10 event sequence analysts. Dark blue represents frequent use while light blue indicates occasional use.

prevalently in high scored days than low scored days. Thus, analysts inspected a display that used only these differentiating event categories (**S2: category extract**). For the above medium size subset, this strategy further reduced the number of events by 92 percent (from 462 to 24), and the number of unique complete sequences by 74 percent (from 27 to 7). Comparisons in temporal patterns between days with high and low scores were made based on the simplified visualization. As we continue this case study, we hope that sets of series of strategies can be combined into a semi-automated workflow.

5 Follow-Up Interviews with Analysts

To validate the proposed list of strategies, we interviewed senior event sequence analysts. An invitation was broadcasted to a mailing list of people who had shown interest in EventFlow. Ten responded to participate and provided feedback on the focusing strategies (4 by video conferencing, 6 by email). On average, these participants had 5 years of experience in event sequence analysis on a wide range of datasets including electronic health records, computer logs, public transportation logs, and students' academic records. Case studies from four of the participants had been reviewed earlier and had guided our development of the strategies. The six others were new. We described the 15 strategies and asked the participants which ones they had used before and how often. For those strategies they had never used, we asked them to provide possible reasons (e.g., not useful for their data). Then, we asked if they thought that having the list of strategies would be useful for their future studies, or for helping novice analysts to get started. Finally, the participants gave general comments in an open-ended conversion.

5.1 Use of the Strategies

Fig. 8 shows the use of the strategies reported by the interviewees. Overall, each strategy had been used by at least three interviewees in their studies. Particularly, S1 (record extract), S2 (category extract), S9 (category grouping), and S15 (partitioning) were the most popular strategies and have been used by almost all interviewees, while S6

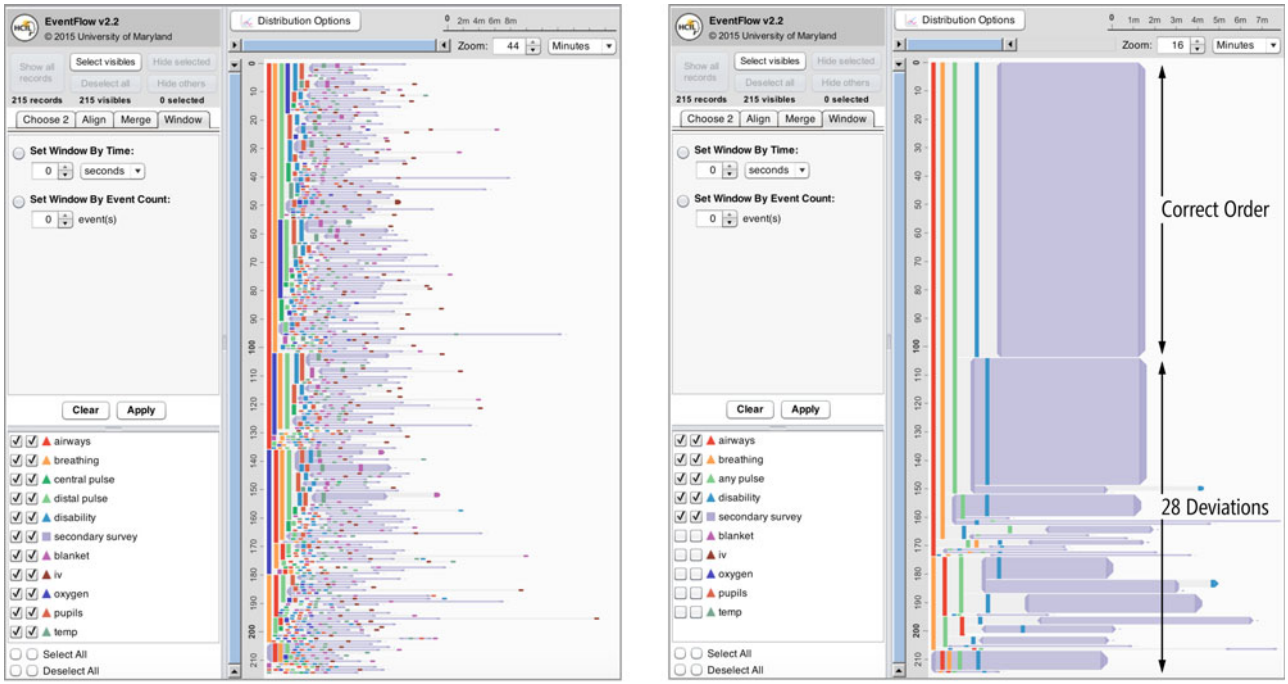

Fig. 9. The displays of the DEVIATION case study using EventFlow [41], before (a) and after (b) applying analytic focusing strategies. From the simplified display (b), analysts could see that approximately half of the records contained events in the correct order (shown at the top), and the common deviations were clearly visible below.

(milestone), S7 (random sample), S8 (folding), and S12 (pattern to one) were used by less than half of them.

The most common reason for not using a strategy is *"it is not useful for my data."* For example, an interviewee commented on S6 (milestone): *"If you have many events you may want to use this [strategy], but our data is very sparse."* Another one who analyzed students' academic records commented on S7 (random sample) *"our data is small and we want to see the whole population"* and on S10 (n points to one) *"the repetitions have meanings in our data and we want to keep them."*

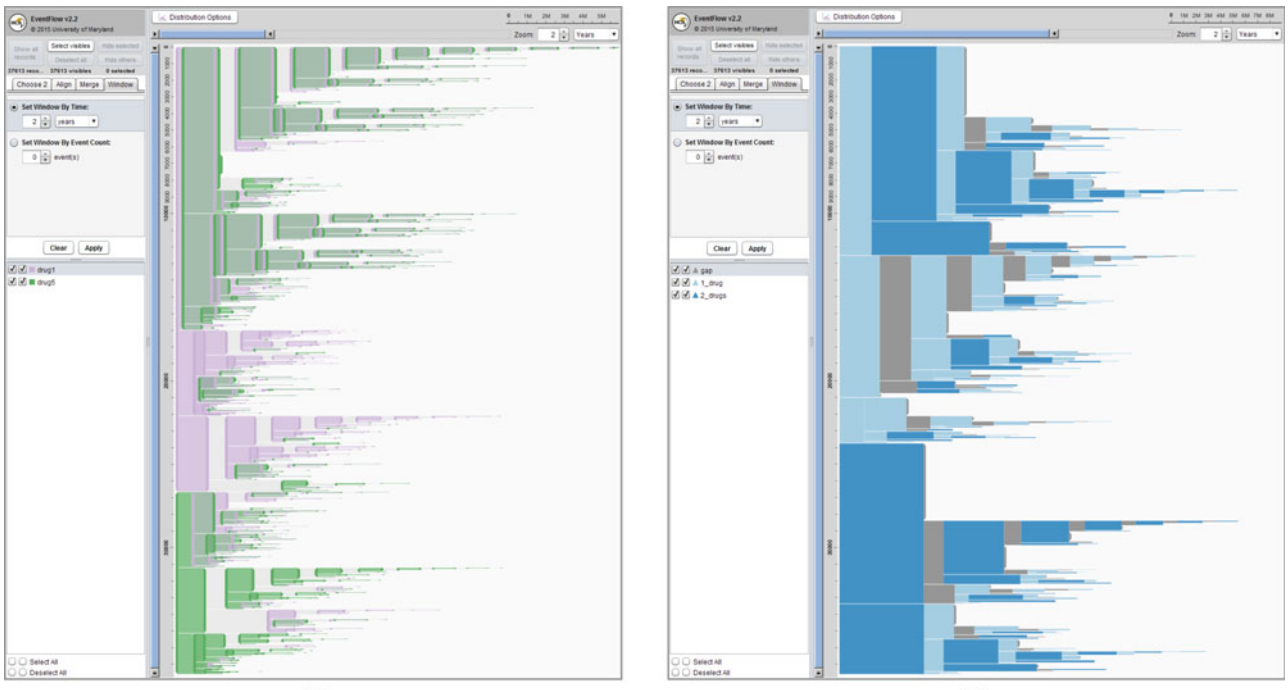

Fig. 10. The displays of the DRUG case study using EventFlow [41], before (a) and after (b) applying analytic focusing strategies. In (a), green and purple bars represent the prescriptions of two drugs. In (b), bars in light blue represent the use of a single drug, in dark blue represent the concurrent use of the two drugs, and in gray represent interruptions.

Besides, the interviewees also mentioned they had never thought about some of the strategies and would try them. Particularly, S8 (folding) and S12 (pattern to one) were noted by three interviewees as inspiring. One who worked on healthcare explained S12 could help him *"aggregate the medical processes and make the data cleaner."* *"I have not thought about it. This inspires me!"* he stated.

All interviewees thought the list of strategies would be useful for their future studies. One mentioned *"it helps me think about different ways of analyzing the data"* and another stated *"it reminds me of something I have used before so I don't have to start from scratch [when analyzing a new dataset]."* As for helping novice analysts get started, some interviewees thought the strategies were useful: *"They (novices) may not know which [strategies] are the best but can start by checking them one by one ... they can get more ideas after that."* Also, one suggested: *"[Providing] examples of how other analysts use the strategies would be helpful."* Some interviewees expressed caution: *"Using the strategies requires basic analytic skills and domain knowledge ... they [novices] may get overwhelmed easily."*

5.2 Missing Strategies

In the open-ended conversations, interviewees mentioned possible new strategies that are not included in the list of observed strategies. One interviewee analyzed drug prescription patterns where many events were logged using the same timestamps as the patients' visits. He requested a strategy for dealing with the complexity caused by the concurrency (e.g., aggregating the concurrent events into an event set). Another whose data was sparse with some events distributed far from the others asked for a strategy for deciding when to remove the timestamps and only keeping the sequential information. Besides, grouping similar event sequences and only showing a few representatives in each group was suggested for creating thumbnails of complex datasets. As the field of event sequence analysis expands and new strategies emerge, our list of strategies can be expanded to guide new users.

5.3 Undo Support

Another topic discussed by several interviewees was the need for undo support. One interviewee stated: *"Users should feel safe that every step they take can be consistently and cleanly backtracked without fear of corrupting the visualization or, worse, the underlying data."* Some others emphasized *"you don't have to restart when you make a mistake"* and one suggested *"it (undo) would enable going back and forth to see the effect [of a strategy]."* Despite its importance, designing an undo model and its interface for the analytic focusing strategies described in this paper remains a challenge and will be our future work.

6 Analytic Focusing Workflow

In the 20 case studies we examined, the workflow for reaching insight varied greatly, depending on the data, the goal of the analysis, and the experience of the analyst. Still all analyses applied one or more extraction strategies and one or more simplification strategies, while folding and iterative strategies were valuable strategies to consider. With the belief that complex workflows are often derived from an idealized one, we propose a basic workflow to guide analysts. Beginning at the raw data, extraction strategies are applied to extract relevant records, event categories, or events (**S1-7**), thereby significantly reducing the data volume. When dealing with activity streams, the event sequences could be extremely long and would be more meaningful if analysts apply the folding strategy (**S8**) to replace a long sequence with many shorter ones so that cyclic patterns are easier to recognize. Pattern simplification strategies (**S9-13**) are used after extraction and folding. Data analysts, who are driven by analytical goals, need to visually inspect the data during the workflow to pick the strategies, review the results and tune the parameters. Useful insights are more visible in the simplified data.

However, this linear workflow does not guarantee useful insights. Analysts usually develop customized workflows based on their data, analytical goals, and intermediate findings. Besides, when large datasets inhibit loading and carrying out exploration, starting from a small sample (**S14**) or a partition (**S15**) enables exploratory operations to determine the best workflow. Analysts can use the captured workflow as a macro to reapply the operations on the full dataset or other partitions.

Event sequence analysis is still a new domain compared to the analysis of other data types, so future prototypes and case studies will likely lead to a refined workflow.

7 Conclusion

The paper described a set of 15 strategies for sharpening analytic focus that analysts can use to reduce the data volume and pattern variety including (1) extraction strategies, (2) temporal folding, (3) pattern simplification strategies, and (4) iterative strategies. For each strategy, we provided numerous examples of use and of the impact on volume and/or variety. The strategies are based on 20 case studies conducted with domain experts using real data, and refined based on interviews which allowed us to provide details not available in the original research papers describing the prototypes. The case studies reveal the richness and diversity of application domains that can benefit from visual analytic tools for temporal event sequences and demonstrate the growing community of users who put them to work.

Acknowledgments

The authors thank the analysts and developers who shared their experience and case studies with them, and the reviewers for their valuable feedback. They appreciate the partial support for this research from the Center for Health-related Informatics and Bioimaging (CHIB) at the University of Maryland, Oracle, and Leidos.

References

[1] (2008). Biomedical Translational Research Information System [Online]. Available: http://www.btris.nih.gov/
[2] (2004). Informatics for Integrating Biology and the Bedside [Online]. Available: https://www.i2b2.org/
[3] (2000). Limits in SQLite [Online]. Available: https://www.sqlite.org/limits.html
[4] (2010). Open Refine [Online]. Available: http://openrefine.org/
[5] D. W. Aha and R. L. Bankert, "A comparative evaluation of sequential feature selection algorithms," in *Learning from Data*. New York, NY, USA: Springer, 1996, pp. 199–206.

[6] W. Aigner, S. Miksch, H. Schumann, and C. Tominski, *Visualization of Time-Oriented Data*. New York, NY, USA: Springer, 2011.

[7] R. Amar, J. Eagan, and J. Stasko, "Low-level components of analytic activity in information visualization," in *Proc. IEEE Symp. Inform. Vis.*, Oct. 2005, pp. 111–117.

[8] R. H. Alexander and H. J. Proctor, "Advanced trauma life support course for physicians," *Resource Document*, pp. 2:317–2:318, 1993.

[9] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu, "Sequential pattern mining using a bitmap representation," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 429–435.

[10] I. Batal, L. Sacchi, R. Bellazzi, and M. Hauskrecht, "A temporal abstraction framework for classifying clinical temporal data," in *Proc. AMIA Annu. Symp.*, 2009, pp. 29–33.

[11] J. Bernard, D. Sessler, T. May, T. Schlomm, D. Pehrke, and J. Kohlhammer, "A visual-interactive system for prostate cancer stratifications," in *Proc. IEEE Vis Workshop Vis. Electron. Health Rec. Data*, 2014.

[12] J. Bernard, N. Wilhelm, B. Kruger, T. May, T. Schreck, and J. Kohlhammer, "MotionExplorer: Exploratory search in human motion capture data based on hierarchical aggregation," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2257–2266, Dec. 2013.

[13] M. V. Bjarnadóttir, S. Malik, E. Onukwugha, T. Gooden, and C. Plaisant, "Understanding adherence and prescription patterns using large-scale claims data," *PharmacoEconomics*, vol. 34, no. 2, pp. 169–179, 2015.

[14] M. Burch, F. Beck, and S. Diehl, "Timeline trees: Visualizing sequences of transactions in information hierarchies," in *Proc. Working Conf. Adv. Vis. Interfaces*, 2008, pp. 75–82.

[15] E. Carter, R. Burd, M. Monroe, C. Plaisant, and B. Shneiderman, "Using EventFlow to analyze task performance during trauma resuscitation," in *Proc. Workshop Interactive Syst. Healthcare*, 2013.

[16] Defense Advanced Research Projects Agency. Anomaly detection at multiple scales (ADAMS) broad agency announcement DARPA-BAA-11-04.

[17] S. G. Eick and A. F. Karr, "Visual scalability," *J. Comput. Graphical Statist.*, vol. 11, no. 1, pp. 22–43, 2002.

[18] N. Elmqvist and J.-D. Fekete, "Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines," *IEEE Trans. Vis. Comput. Graphics*, vol. 16, no. 3, pp. 439–454, May 2010.

[19] P. Federico, S. Hoffmann, A. Rind, W. Aigner, and S. Miksch, "Qualizon graphs: Space-efficient time-series visualization with qualitative abstractions," in *Proc. Working Conf. Adv. Vis. Interfaces*, 2014, pp. 273–280.

[20] P. Federico, J. Unger, A. Amor-Amors, L. Sacchi, D. Klimov, and S. Miksch, "Gnaeus: Utilizing clinical guidelines for knowledge-assisted visualization of EHR cohorts," in *Proc. EuroVis Workshop Vis. Anal.*, 2015, pp. 79–83.

[21] J.-D. Fekete, "ProgressiVis: A toolkit for steerable progressive analytics and visualization," in *Proc. 1st Workshop Data Syst. Interactive Anal.*, 2015, p. 5.

[22] D. Fisher, I. Popov, S. Drucker, and M. schraefel, "Trust me, I'm partially right: Incremental visualization lets analysts explore large datasets faster," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2012, pp. 1673–1682.

[23] H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C.-A. Saita, "Declarative data cleaning: Language, model, and algorithms," in *Proc. 27th Int. Conf. Very Large Data Bases*, 2001, pp. 371–380.

[24] D. Gotz and H. Stavropoulos, "DecisionFlow: Visual analytics for high-dimensional temporal event sequence data," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 12, pp. 1783–1792, Dec. 2014.

[25] T. Gschwandtner, W. Aigner, S. Miksch, J. Gärtner, S. Kriglstein, M. Pohl, and N. Suchy, "TimeCleanser: A visual analytics approach for data cleansing of time-oriented data," in *Proc. 14th Int. Conf. Knowl. Technol. Data-Driven Bus.*, 2014, pp. 18:1–18:8.

[26] T. Gschwandtner, J. Gärtner, W. Aigner, and S. Miksch, "A taxonomy of dirty time-oriented data," in *Proc. Multidisciplinary Research Practice Inform. Syst.*, pp. 58–72. New York: Springer, 2012.

[27] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono, "Research directions in data wrangling: Visualizations and transformations for usable and credible data," *Inform. Vis.*, vol. 10, no. 4, pp. 271–288, 2011.

[28] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. "Wrangler: Interactive visual specification of data transformation scripts," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2011, pp. 3363–3372.

[29] K. Koflka, *Principles of Gestalt Psychology*. New York, NY, USA: Har, 1935.

[30] J. Krause, A. Perer, and H. Stavropoulos, "Supporting iterative cohort construction with visual temporal queries," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 91–100, Jan. 2016.

[31] M. Krstajic, E. Bertini, and D. Keim, "CloudLines: Compact display of event episodes in multiple time-series," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 12, pp. 2432–2439, Dec. 2011.

[32] T. Lammarsch, W. Aigner, A. Bertone, M. Bögl, T. Gschwandtner, S. Miksch, and A. Rind, "Interactive visual transformation for symbolic representation of time-oriented data," in *Proc. Human-Computer Interaction Knowl. Discovery Complex, Unstructured, Big Data*, 2013, pp. 400–419.

[33] T. Lammarsch, W. Aigner, A. Bertone, S. Miksch, and A. Rind. "Mind the time: Unleashing temporal aspects in pattern discovery," *Comput. Graphics*, vol. 38, pp. 38–50, 2014.

[34] G. Lipori, "The use of recursive partitioning via temporal visual toolsets to efficiently isolate patient cohorts," in *Proc. Human-Comput. Interaction Lab Symp.*, 2013.

[35] H. Liu and H. Motoda, "On issues of instance selection," *Data Mining Knowl. Discovery*, vol. 6, no. 2, pp. 115–130, 2002.

[36] Z. Liu and J. Heer, "The effects of interactive latency on exploratory visual analysis," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 2122–2131, Dec. 2014.

[37] S. Malik and E. Koh, "High-volume hypothesis testing for large-scale web log analysis," in *Proc. Extended Abstracts Human Factors Comput. Syst.*, 2016.

[38] S. Malik, B. Shneiderman, F. Du, C. Plaisant, and M. V. Bjarnadóttir, "High-volume hypothesis testing: Systematic exploration of event sequence comparisons," *ACM Trans. Interactive Intell. Syst.*, vol. 6, no. 1, pp. 9:1–9:23, 2016.

[39] M. L. Mauriello, B. Shneiderman, F. Du, S. Malik, and C. Plaisant, "Simplifying overviews of temporal event sequences," in *Proc. Extended Abstracts Human Factors Comput. Syst.*, 2016.

[40] T. E. Meyer, M. Monroe, C. Plaisant, R. Lan, K. Wongsuphasawat, T. S. Coster, S. Gold, J. Millstein, and B. Shneiderman, "Visualizing patterns of drug prescriptions with EventFlow: A pilot study of Asthma medications in the military health system," in *Proc. Workshop on Visual Analytics in Healthcare*, pp. 55–58, 2013.

[41] M. Monroe, "Interactive event sequence query and transformation," Ph.D. dissertation, Univ. Maryland, College Park, MD, USA, 2014.

[42] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman, "Temporal event sequence simplification," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2227–2236, Dec. 2013.

[43] R. Moskovitch and Y. Shahar, "Medical temporal-knowledge discovery via temporal abstraction." in *Proc. AMIA Annu. Symp. Proc.*, 2009, pp. 452–456.

[44] H. Motoda and H. Liu, "Feature selection, extraction and construction," *Commun. Inst. Inform. Comput. Machinery*, vol. 5, pp. 67–72, 2002.

[45] E. Onukwugha, Y. Kwok, C. Yong, C. Mullins, B. Seal, and A. Hussain, "Variation in the length of radiation therapy among men diagnosed with incident metastatic prostate cancer," in *Int'l J. Radiation Oncology, Bio., Phys.*, vol. 87, no. 2, p. S350, 2013.

[46] A. Perer and D. Gotz, "Data-driven exploration of care plans for patients," in *Proc. Extended Abstracts SIGCHI Conf. Human Factors Comput. Syst.*, 2013, pp. 439–444.

[47] A. Perer and F. Wang, "Frequence: Interactive mining and visualization of temporal frequent event sequences," in *Proc. 19th Int. Conf. Intell. User Interfaces*, 2014, pp. 153–162.

[48] A. Perer, F. Wang, and J. Hu, "Mining and exploring care pathways from electronic medical records with visual analytics," *J. Biomed. Inform.*, vol. 56, pp. 369–378, 2015.

[49] S. Powsner and T. Sullivan, "Data preparation for EventFlow: A case study of over 12000 daily reports from women in abusive relationships," in *Proc. Human-Comput. Interaction Lab Symp.*, 2014.

[50] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, Dec. 2000.

[51] V. Raman and J. M. Hellerstein, "Potter's wheel: An interactive data cleaning system," in *Proc. 27th Int. Conf. Very Large Data Bases*, 2001, pp. 381–390.

[52] T. Reinartz. *Focusing Solutions for Data Mining: Analytical Studies and Experimental Results in Real-World Domains*. New York, NY, USA: Springer-Verlag, 1999.

[53] T. Reinartz, "A unifying view on instance selection," *Data Mining Knowl. Discovery*, vol. 6, no. 2, pp. 191–210, 2002.
[54] T. E. Senator, H. G. Goldberg, A. Memory, W. T. Young, B. Rees, R. Pierce, D. Huang, M. Reardon, D. A. Bader, E. Chow, I. Essa, J. Jones, V. Bettadapura, D. H. Chau, O. Green, O. Kaya, A. Zakrzewska, E. Briscoe, R. I. L. Mappus, R. McColl, L. Weiss, T. G. Dietterich, A. Fern, W.-K. Wong, S. Das, A. Emmott, J. Irvine, J.-Y. Lee, D. Koutra, C. Faloutsos, D. Corkill, L. Friedland, A. Gentzel, and D. Jensen, "Detecting insider threats in a real corporate database of computer usage activity," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 1393–1401.
[55] Y. Shahar, "A framework for knowledge-based temporal abstraction," *Artif. Intell.*, vol. 90, no. 12, pp. 79–133, 1997.
[56] B. Shneiderman and C. Plaisant, "Sharpening analytic focus to cope with big data volume and variety," *IEEE Comput. Graphics Appl.*, vol. 35, no. 3, pp. 10–14, May-Jun. 2015.
[57] C. Stolper, A. Perer, and D. Gotz, "Progressive visual analytics: User-driven visual exploration of in-progress analytics," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 12, pp. 1653–1662, Dec. 2014.
[58] J. Sun, C. D. McNaughton, P. Zhang, A. Perer, A. Gkoulalas-Divanis, J. C. Denny, J. Kirby, T. Lasko, A. Saip, and B. A. Malin, "Predicting changes in hypertension control using electronic health records from a chronic disease management program," *J. Am. Med. Inform. Assoc.*, vol. 21, no. 2, pp. 337–344, 2014.
[59] J. Sun, F. Wang, J. Hu, and S. Edabollahi, "Supervised patient similarity measure of heterogeneous patient records," *SIGKDD Explor. Newslett.*, vol. 14, no. 1, pp. 16–24, 2012.
[60] T. D. Wang, C. Plaisant, A. J. Quinn, R. Stanchak, S. Murphy, and B. Shneiderman, "Aligning temporal data by sentinel events: Discovering patterns in electronic health records," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2008, pp. 457–466.
[61] T. D. Wang, C. Plaisant, B. Shneiderman, N. Spring, D. Roseman, G. Marchand, V. Mukherjee, and M. Smith, "Temporal summaries: Supporting temporal categorical searching, aggregation and comparison," *IEEE Trans. Vis. Comput. Graphics*, vol. 15, no. 6, pp. 1049–1056, Apr. 2009.
[62] P. C. Wong and J. Thomas, "Visual analytics," *IEEE Comput. Graphics Appl.*, vol. 24, no. 5, pp. 20–21, Sep. 2004.
[63] K. Wongsuphasawat and D. Gotz, "Exploring flow, factors, and outcomes of temporal event sequences with the Outflow visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 12, pp. 2659–2668, Dec. 2012.
[64] K. Wongsuphasawat and J. Lin, "Using visualizations to monitor changes and harvest insights from a global-scale logging infrastructure at Twitter," in *Proc. IEEE Conf. Vis. Anal. Sci. Technol.*, 2014, pp. 113–122.
[65] J. S. Yi, Y. ah Kang, J. Stasko, and J. Jacko, "Toward a deeper understanding of the role of interaction in information visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 13, no. 6, pp. 1224–1231, Nov. 2007.
[66] E. Zgraggen, S. M. Drucker, D. Fisher, and R. DeLine, "(s|qu)eries: Visual regular expressions for querying and exploring event sequences," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2015, pp. 2683–2692.
[67] J. Zhao, Z. Liu, M. Dontcheva, A. Hertzmann, and A. Wilson, "MatrixWave: Visual comparison of event sequence data," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2015, pp. 259–268.
[68] P. Zikopoulos, C. Eaton, and Others. *Understanding big data: Analytics for enterprise class Hadoop and streaming data*. McGraw-Hill Osborne Media, 2011.

Fan Du received the bachelor's degree from Zhejiang University in 2013. He is currently working toward the PhD degree in computer science at the University of Maryland and is a member of the Human-Computer Interaction Lab. His research focuses on data visualization and human-computer interaction, especially on analyzing electronic health records, and user activity logs.

Ben Shneiderman is a distinguished university professor in the Department of Computer Science and was the founding director from 1983 to 2000 of the Human-Computer Interaction Laboratory at the University of Maryland. He received the ACM SIGCHI Lifetime Achievement Award in 2001. His research interests include human-computer interaction, information visualization, and user interface design. He is a fellow of the ACM, AAAS, and the IEEE and a member of the National Academy of Engineering.

Catherine Plaisant received the doctorat d'Ingénieur degree in France in 1982. She is a senior research scientist at the Human-Computer Interaction Laboratory of the University of Maryland Institute for Advanced Computer Studies. She enjoys most working with multidisciplinary teams on designing and evaluating new interface technologies that are usable and useful. She has written more than 100 refereed technical publications on the subjects of information visualization, evaluation methods, electronic health record interfaces, digital libraries, online help, and so on. She coauthored with Ben Shneiderman the fifth edition of *Designing the User Interface*.

Sana Malik received the BS degree from the University of Delaware in 2011 and the MS degree from the University of Maryland in 2014, both in computer science. She is currently working toward the PhD degree in computer science and is a member of the Human-Computer Interaction Lab at the University of Maryland, College Park. Her research interests are information visualization, temporal event sequences, and data analytics.

Adam Perer received the BS degree in computer science from Case Western Reserve University, in 2002, and the MS and PhD degrees in computer science from the University of Maryland, College Park, in 2005 and 2008, respectively. He is currently a research scientist in the Healthcare Analytics Research Group at the IBM T.J. Watson Research Center in New York. He was previously a research scientist in the IBM Research's Visual Communication Lab in Cambridge and IBM Research's Social Technologies Group in Haifa. His research focuses on making sense of big data through visualization and interactive analytics. He is a coauthor of more than 20 conference and journal papers in the fields of information visualization, visual analytics, and human-computer interaction. He serves on the organizing and program committees of leading visualization conferences, including InfoVis and EuroVis.