
White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Leeds.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/106008/

Version: Accepted Version

Proceedings Paper:
Ruddle, RA, Bernard, J, May, T et al. (2016) Methods and a research agenda for the evaluation of event sequence visualization techniques. In: Proceedings of the IEEE VIS 2016 Workshop on Temporal & Sequential Event Analysis. The Event Event: Temporal & Sequential Event Analysis - An IEEE VIS 2016 Workshop, 24 Oct 2016, Baltimore, Maryland, USA.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.


UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

Methods and a Research Agenda for the Evaluation of Event Sequence Visualization Techniques

Roy A. Ruddle, Jürgen Bernard, Thorsten May, Hendrik Lücke-Tieke, and Jörn Kohlhammer

Abstract— The present paper asks how can visualization help data scientists make sense of event sequences, and makes three main contributions. The first is a research agenda, which we divide into methods for presentation, interaction & computation, and scale-up. Second, we introduce the concept of Event Maps to help with scale-up, and illustrate coarse-, medium- and fine-grained Event Maps with electronic health record (EHR) data for prostate cancer. Third, in an experiment we investigated participants' ability to judge the similarity of event sequences. Contrary to previous research into categorical data, color and shape were better than position for encoding event type. However, even with simple sequences (5 events of 3 types in the target sequence), participants only got 88% correct despite averaging 7.4 seconds to respond. This indicates that simple visualization techniques are not effective.

Index Terms—Visualization, Electronic Health Records, Event Sequences, Research agenda, Evaluation

INTRODUCTION

The broad field of longitudinal data analysis includes situations where users are concerned with the sequence (the order) in which events (categorical data) take place. Such event sequence data are common in domains such as health, where the events may be stages in a patient's care [1] or procedures, diagnoses and prescriptions that are recorded in the patient's EHRs.

The present paper addresses the general question of how can visualization help people make sense of event sequences, and makes three main contributions. First, we outline an agenda of research that is needed to design effective visualizations for event sequences. Second, we introduce the concept of Event Maps to help with scale-up. Third, we describe an exploratory experiment that investigated participants' ability to judge the similarity of event sequences.

1 RELATED WORK

There are two distinct styles for visualizing event sequences, as demonstrated by existing systems. One displays a sequence as a set of symbols (e.g., Lifelines2 [2] and EventFlow [3]). This allows users to identify the events in the sequence just by looking at the visualization, but users need to compare the symbols to judge the similarity of sequences. The other style presents multiple sequences as a network (e.g., DecisionFlow [4] and CAVA [5]), so users have to interact to identify the events in a given sequence but the similarity of sequences is indicated by shared nodes and links.

With both styles, the number of sequences that may be usefully displayed is often far fewer than the number of separate sequences that are in the underlying data. However, providing a suite of interactive filtering and transformation functionality allows users to substantially reduce the number of sequences to be visualized [3, 6]. Alternatively, data mining or machine learning techniques [7] may be used to aggregate a large number of sequences into a smaller number that are suitable for visualization.

- *Roy A. Ruddle is with the University of Leeds. E-mail: r.a.ruddle@leeds.ac.uk.*
- *Jürgen Bernard, Thorsten May, Hendrik Lücke-Tieke, & Jörn Kohlhammer are with the Fraunhofer Institute for Computer Graphics Research. E-mail: {juergen.bernard, thorsten.may,hendrik.luecke-tieke, joern.kohlhammer}@igd.fraunhofer.de.*

Proceedings of the IEEE VIS 2016 Workshop on Temporal & Sequential Event Analysis. Available online at: http://eventevent.github.io

The appearance of an event sequence visualization is dictated by the visual encodings that are used. Although there is a large design space of possibilities, this may be rationalized using the results of experiments that rank encoding channels for different data types. Recent experiments indicate that the position, color or shape channel are the most effective channels for categorical data, whereas position, color or size should be used for ordinal data [8].

Channels such as color or shape offer many different design options. For categorical data, a practical limit is 12 shapes. A similar limit (6 – 12) is suggested for color, but that is reduced to only 3 – 4 for a color- deficient audience [9]. Recent guidance about which shapes and colors to use is provided by [10].

2 RESEARCH AGENDA

Previous research that involved clinicians or health researchers provides several use cases for event sequence visualization. These use cases have aims that include stratifying patients into groups that have similar characteristics [11], understanding the cause of bottlenecks in care provision [1], checking adherence to planned care pathways [3], identifying a patient cohort with certain characteristics [5], and predicting patient outcomes [4].

Previous research has provided fine-grained mappings between EHR systems and user interaction [12]. By contrast, we have identified four fundamental tasks that are orthogonal to the use cases and which visualization is well-placed to support. The tasks are:

a) Simplify sequence (remove or aggregate events to reduce the number that are in a sequence).
b) Find a subsequence (a specific pattern of events in a sequence).
c) Understand longitudinal changes (changes in the pattern from one part of a sequence to another).
d) Compare sequences (similarities and differences between sequences).

In their purest form the first three tasks only need to involve a single sequence although, in practice, real applications will involve multiple sequences. The fourth task concerns between-sequence differences and, therefore, always involves multiple sequences.

Systems such as Lifelines2 [2] and DecisionFlow [4] have been developed from an application perspective, with specific visualization, interaction and computational techniques chosen from a large design space of possibilities. However, there is a considerable gulf between design decisions taken for those systems and existing research into the fundamentals of visualization ([8, 10]). The following sections bridge that gulf by laying out a research agenda for the four fundamental tasks that are described above.

2.1 Presentation

The first things that we need to know is how visual encoding affects users' ability to perform the above task, and when do basic approaches (e.g., a left-justified sequence of symbols) break down in terms of the number of event types, and the number and length of sequences. This will tell us when more sophisticated approaches are needed, and allow unpromising encodings to be ruled out so that others may be studied in greater detail.

The experiment described in §3 makes a first step in that research. Complementary research could use a staircase methodology (e.g., [13]) to determine thresholds at which users' performance becomes unsatisfactory for a given encoding.

Similar methods could be adopted to investigate the 'find a subsequence' and 'understand longitudinal changes' tasks. However, we exclude the 'simplify sequence' task from this part of the research agenda, because it either requires users to interact or computation to be performed (see Section 2.2).

2.2 Interaction and Computation

Systems such as EventFlow [3] and DecisionFlow [4] provide a visual analytics capability, allowing users to interact and perform on-the-fly computations. Evaluations showed that this allowed users to analyze sequences for a few thousand patients and types of event.

Building on research into analysis strategies [6], we subdivide interaction and computation according to whether given functionality changes the presentation of existing data or creates new data (see Table 1). For all four combinations of interaction vs. computation and presentation vs. new data there is functionality that visualization systems could provide to aid users in the analysis of event sequences.

Table 1. Interaction and computation functionality that changes how existing data are presented vs. creates new data.

Change	Interaction	Computation
Presentation	Filter Level of detail	Sort Align
New data	Coalesce Convert	Associate

Users may interactively filter data by selecting certain records, event types and/or times (strategies S1, S2 and S5-8 in [6]). Graphical interfaces make it straightforward to define simple filter criteria (e.g., records that include specific event types), but research is needed to develop interfaces that allow users to easily specify complex criteria (e.g., multiple combinations of events) that would require an excessive number of clicks with today's interfaces. Users may also interactively change the level of detail of events (see S9 in [6]), e.g., by exploiting the hierarchical coding schemes.

Other data presentation functionality includes sorting and alignment, which may be invoked by a user interacting but require non-trivial computation to achieve. Sorting can simplify a visualization to reveal patterns to users [14]. Alignment may be performed manually (e.g., by specifying a common reference time; see S4 in [6]) or computationally with local or global methods from bioinformatics. A research priority is to investigate the effect of existing sorting and alignment methods on the 'find a subsequence', 'longitudinal change' and 'compare sequences' tasks. That research could be performed via experiments similar to those suggested in §2.1 or technical evaluations.

Coalescing repeated events, converting complex events into a single event, and aggregating events over time all: (a) involve the creation of new data, and (b) are likely to be interactive because users need to provide guidance (see S10-13 in [6]). In common with the presentation/interaction functionality (see above), the main need is for research into low-cost interaction methods.

A high-level goal in most of the above use cases is to associate particular sequences of events with certain groups of patients or treatment outcomes (see S3 in [6]). While some expert guidance is essential, substantial computation is necessary to overcome the scale and complexity of the data. For this, an important research area is the design and evaluation of similarity metrics for event sequences, and their validation for each of the four fundamental tasks.

2.3 Scale-up

Current visual analytic systems have only been used to analyze datasets that contain a few thousand patients and event types (see Section 2.2), which falls far short of the quantity of data that is analyzed in some major studies (e.g., 6 million patients in [15]). We propose two approaches that may help to scale-up event sequence visualization and visual analytics techniques for such studies.

The first involves a concept that we term *Event Maps*, and draws inspiration from the *Quality Maps* for visualizing data quality that was proposed by Ward et al. [16]. Event Maps make event types the primary entity in visualizations, so users may see overviews of patterns in event sequence data. We illustrate the concept with a prostate cancer example that is drawn from our work with clinicians from the Universitätsklinikum Hamburg-Eppendorf (UKE) [7, 11].

The UKE data comprises 1940 patients and ten event types. One high-level Event Map simply shows the frequency of each event type, of which five are common, four are rare, and one (H) is in-between (Fig 1a). Another is a histogram showing the number of events for every sequence (Fig 1b).

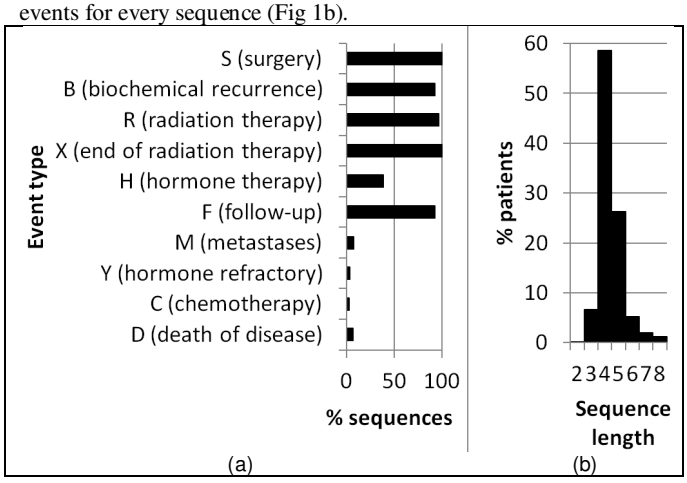

Fig. 1. Event Maps showing: (a) the frequency of each event type, and (b) the number of event types per patient.

Other Event Maps may provide more detail by showing the frequency distribution of events types across sequences (Fig 2a) or of a given event type following another (Fig 2b). This reveals clusters of event types, and indicates that sequences often start with the events B/R/X, with H/F/M in the middle, and C/D at the end. H is unusual because it occurs equally often at the beginning and in the middle (after B and X, respectively).

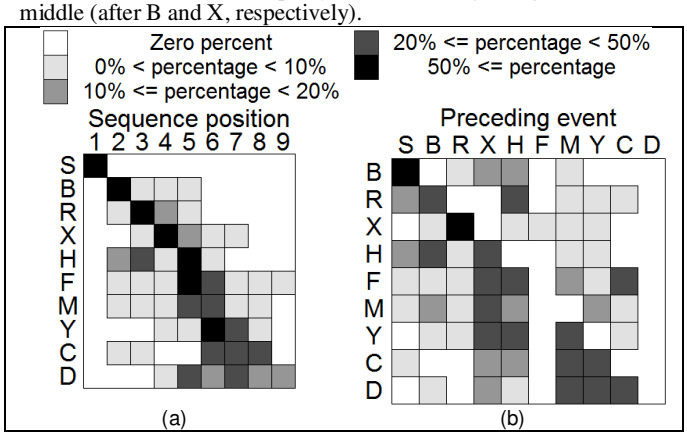

Fig. 2. Event Maps showing the percentage of times each event type: (a) occurs at each position in a sequence, and (b) follows each other event type.

Fine-grained Event Maps help users to understand unusual sequences in the context of those that are common, by showing the percentage of times that each sequence occurs. E.g., sequences may be ordered to group together permutations that contain the same combination of event types (Fig 3).

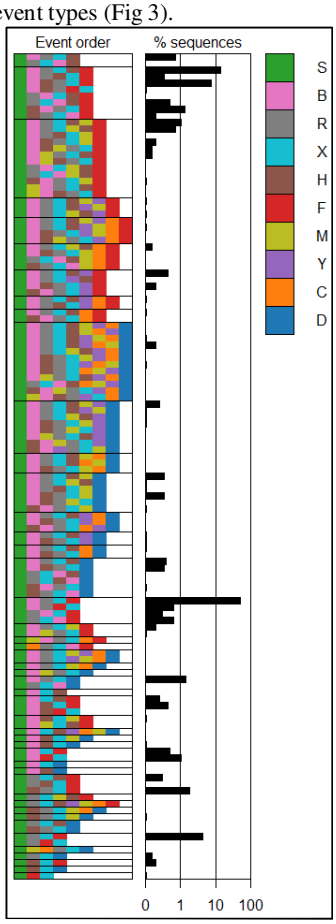

Fig. 3. Fine-grained Event Map showing the percentage of occurrences of each sequence, using horizontal lines to group sequences that contain the same combination of event types.

The second approach is to exploit visualization in the design and optimization of processing pipelines for event sequence data, rather than directly using visualization to gain new insights [17]. This could help users to design models to quantify the similarity of sequences, understand the sensitivity of those models to parameters and assumptions, and validate the models. In turn this would: (a) allow a step-change in the size/complexity of the sequences that it is feasible for users to analyze, and (b) help users to choose a level of detail that balanced the preservation of details against the suppression of noise.

3 Experiment

The experiment involved presenting a series of images to that contained a target sequence and two choices. A participant had to choose which choice was more similar to the target sequence. We used a within participants design with three factors: visual encoding (color vs. shape vs. position), edit type (insert vs. delete vs. substitute vs. insert & delete vs. insert & substitute vs. delete & substitute), and Levenshtein distance between the target and correct choice (2 vs. 3).

3.1 Method

3.1.1 Participants

Thirteen individuals (8 men & 1 woman aged 25 – 45 years, 4 declined to give gender or age). All the participants gave informed consent and were paid for their participation. The study was approved by the Ethics Committee at the first author's institution.

3.1.2 Materials

The experiment was delivered via a web browser using custom-developed Java software. Examples of the trials are shown in Fig 4. The colors and shapes were chosen from a ColorBrewer colorblind-safe palate, and the reordered list of [10], respectively. The instruction on each screen was "Please click on the sequence (A or B) that is most similar to the Target sequence?"

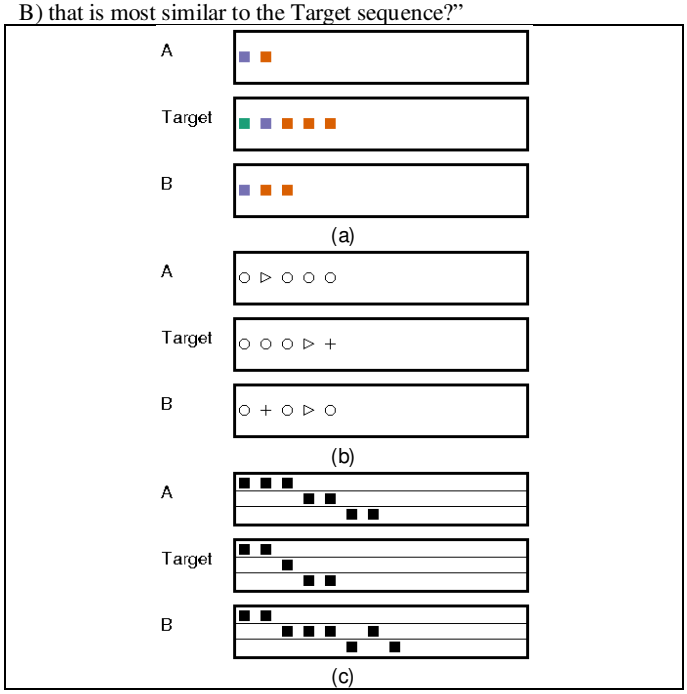

Fig. 4. Examples of the trials: (a) color encoding and delete edit type, (b) shape encoding and substitute edit type, and (c) position encoding and insert edit type.

3.1.3 Procedure

The experiment was divided into two parts: an introduction and the test. In the introduction, a series of slides were used to explain the task, the three types of encoding, and the three basic edit types (insert vs. delete vs. substitute). Then slides were used to present six practice trials to a participant. There was one trial for each of the six edit types (see above), two of which were presented with each encoding. Once the participant had chosen the sequence that was more similar to the target sequence, the correct answer was displayed, together with the edits that needed to be made to change the answer into the target.

The test involved five blocks of trials, with 36 trials in each block (one trial for each combination of encoding, edit type and Levenshtein distance). The trials were presented in a random order. A participant indicated their choice by clicking inside the box that surrounded sequence A or B, which caused the choice and the participant's response time to be recorded, and the screen to be blanked for 1 second before the next trial was displayed. To reduce fatigue, there was a minimum of a 30 seconds pause between blocks.

3.2 Results and Discussion

There was a significant correlation between the time that participants took to complete the whole experiment and the percentage of trials that they got correct ($r(13) = 0.76, p < .01$). Some participants seemed to make fast instinctive judgments whereas others carefully compared the sequences, which we characterize as 'perceptually' and 'cognitively' driven strategies, respectively. Four participants got fewer than 50% of trials correct (chance level) in at least 1 block and were excluded from the analyses of variance (ANOVAs) that follow.

In the ANOVAs the response time data were normalized using a log10 transformation. A † after a p value indicates that the Greenhouse-Geisser correction for sphericity was applied.

To check for learning and fatigue effects, the response time and percentage of correct trials data were analyzed using ANOVAs that treated block as a repeated measure. There were no significant effects, so the data for all blocks were combined. Overall the average was a 7.4 s response time and 88% of trials correct.

A three-way repeated measures ANOVA showed main effects of percentage of correct trials for all of the independent variables (Fig 5): visual encoding ($F(2,16) = 4.44$, $p < .05$), edit type ($F(5, 40) = 2.81$, $p < .05$), and Levenshtein distance ($F(1,8) = 5.21$, $p = .05$).

A three-way repeated measures ANOVA showed a similar pattern of results for response time (Fig 5): visual encoding ($F(2,16) = 3.58$, $p = .05$), edit type ($F(2,17) = 25.74$, $p < .01$†), and Levenshtein distance ($F(1,8) = 10.68$, $p < .05$). There was also a significant edit type × distance interaction ($F(5,40) = 6.68$, $p < .01$).

For percentage correct and response time, position was the worst-performing encoding, which is a notable because other research has ranked it top for categorical data [8]. Delete stood out as the best-performing event type, and participants performed better with the larger Levenshtein distance, probably because that meant the 'wrong' answer only had one event. By contrast, and as expected, participants performed better with the smaller distance for the other five event types. This contrast caused the significant interaction.

Delete vs. insert are opposite types of event, so the results inform us about the effect of sequence length. As length increased participants performed worse, from delete/distance=3 (95% correct) to insert/distance=3 (81% correct). These correspond to the 'wrong' answer having one and nine events, respectively. It also is notable that performance on insert and substitute trials was similar to trials that involved pairs of edit types (e.g., insert & delete).

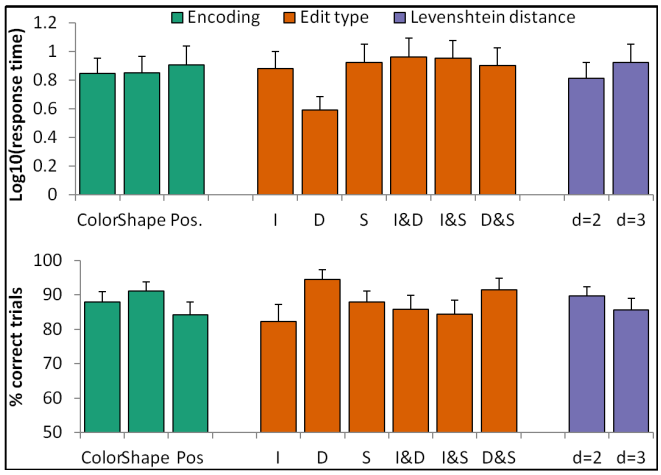

Fig. 5. Mean response time and % of correct trials for the encodings, edit types (I = insert, D = delete, S = substitute) and Levenshtein distances (d=2, d=3). Error bars show standard error of the mean.

4 Conclusion

Within the broad field of temporal data, this paper focuses on event sequences, as are common in domains such as health. We propose a research agenda, and illustrate five types of Event Map that make event types the primary entity in visualizations and are designed to provide a step-change in the scale of data that users can analyze.

We also conducted an exploratory experiment to investigate participants' ability to judge the similarity of event sequences. Color and shape were better than position for encoding event type but, even though participants took a long time to make the judgments (7.4 s), they only got 88% correct. This indicates the need for more effective visualization techniques (see Research Agenda), particularly as the sequences were simpler (5 events of 3 types in the target) than those portrayed in EHR systems (e.g., [2-5]). Our results also raise the question of how correctly, and hence safely, are users judging event sequence similarity with today's systems?

Acknowledgments

This research was supported by the Alexander von Humboldt Foundation.

References

[1] O.A. Johnson, P.S. Hall, and C. Hulme. NETIMIS: Dynamic Simulation of Health Economics Outcomes Using Big Data. *Pharmacoeconomics*, 34:107-114, 2016.

[2] T.D. Wang, C. Plaisant, A.J. Quinn, R. Stanchak, et al. Aligning temporal data by sentinel events: discovering patterns in electronic health records. *Proc. ACM CHI*, pages 457-466, 2008.

[3] M. Monroe, R. Lan, H. Lee, C. Plaisant, et al. Temporal event sequence simplification. *IEEE Trans. Vis. Comput. Graphics*, 19:2227-2236, 2013.

[4] D. Gotz and H. Stavropoulos. DecisionFlow: Visual analytics for high-dimensional temporal event sequence data. *IEEE Trans. Vis. Comput. Graphics*, 20:1783-1792, 2014.

[5] Z. Zhang, D. Gotz, and A. Perer. Iterative cohort analysis and exploration. *Info. Vis.*:1473871614526077, 2014.

[6] F. Du, B. Shneiderman, C. Plaisant, S. Malik, et al. Coping with volume and variety in temporal event sequences: Strategies for sharpening analytic focus. *IEEE Trans. Vis. Comput. Graphics*, in press.

[7] J. Bernard, D. Sessler, A. Bannach, T. May, et al. A visual active learning system for the assessment of patient well-being in prostate cancer research. *Proc. Workshop on Visual Analytics in Healthcare*, pages 1, 2015.

[8] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, et al. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Trans. Vis. Comput. Graphics*, 22:649-658, 2016.

[9] T. Munzner. Visualization Analysis and Design (CRC Press), 2014.

[10] Ç. Demiralp, M.S. Bernstein, and J. Heer. Learning perceptual kernels for visualization design. *IEEE Trans. Vis. Comput. Graphics*, 20:1933-1942, 2014.

[11] J. Bernard, D. Sessler, T. May, T. Schlomm, et al. A visual-interactive system for prostate cancer cohort analysis. *IEEE Comput. Graph. Appl.*, 35:44-55, 2015.

[12] A. Rind, T.D. Wang, A. Wolfgang, S. Miksch, et al. Interactive information visualization to explore and query electronic health records. *Foundations and Trends in Human-Comp. Interaction*, 5:207-298, 2011.

[13] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using weber's law. *IEEE Trans. Vis. Comput. Graphics*, 20:1943-1952, 2014.

[14] M. Behrisch, B. Bach, N.H. Riche, T. Schreck, et al. Matrix reordering methods for table and network visualization. *Computer Graphics Forum*, 35:24, 2016.

[15] A.B. Jensen, P.L. Moseley, T.I. Oprea, S.G. Ellesøe, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications*, 52014.

[16] M. Ward, Z. Xie, D. Yang, and E. Rundensteiner. Quality-aware visual data analysis. *Computational Statistics*, 26:567-584, 2011.

[17] T. von Landesberger, D.W. Fellner, and R.A. Ruddle. Visualization system requirements for data processing pipeline design and optimization. *IEEE Trans. Vis. Comput. Graphics*, in press.