Visual Encodings of Temporal Uncertainty: A Comparative User Study

Theresia Gschwandtner, Markus Bögl, Paolo Federico, and Silvia Miksch

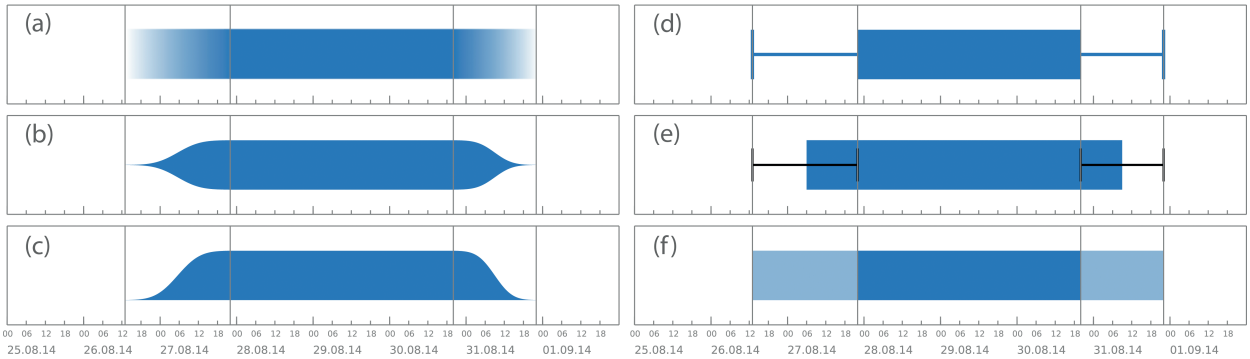

Fig. 1: Six different visual encodings of start/end uncertainty of temporal intervals used in the user study: (a) gradient plot, (b) violin plot, (c) accumulated probability plot, (d) error bars, (e) centered error bars, and (f) ambiguation. We designed encodings (a)–(c) to encode statistical uncertainty and encodings (d)–(f) to encode bounded uncertainty. All encodings were used to estimate earliest start, latest start, earliest end, and latest end, as well as minimum, maximum, and average interval duration. Moreover, encodings (a)–(c) were used to estimate the probability that the interval has already started/ended at a marked position in time.

Abstract—A number of studies have investigated different ways of visualizing uncertainty. However, in the temporal dimension, it is still an open question how to best represent uncertainty, since the special characteristics of time require special visual encodings and may provoke different interpretations. Thus, we have conducted a comprehensive study comparing alternative visual encodings of intervals with uncertain start and end times: gradient plots, violin plots, accumulated probability plots, error bars, centered error bars, and ambiguation. Our results reveal significant differences in error rates and completion time for these different visualization types and different tasks. We recommend using ambiguation – using a lighter color value to represent uncertain regions – or error bars for judging durations and temporal bounds, and gradient plots – using fading color or transparency – for judging probability values.

Index Terms—Uncertainty, temporal intervals, visualization.

1 Introduction

Many real world data sets contain some amount of uncertain information. Finding visualization techniques that communicate these uncertainties instead of neglecting them is, therefore, imperative to provide the user with correct information. In recent years, there has been a lot of effort to design such visualization techniques which incorporate uncertainty into the representation; in particular in the temporal domain, there are several approaches to visually communicate temporal indeterminacy [11, 5, 16, 3, 9]. Information that is temporally indeterminate can be characterized as 'do not know when information', or more precisely, as 'do not know exactly when information'. Examples of this are inexact knowledge (e.g., 'time when the earth was formed'), future planning data (e.g., 'it will take 2-3 weeks'), or imprecise event times (e.g., 'one or two days ago').

If there is no complete or exact information about time specifications or if time primitives are converted from one granularity to another, uncertainties are introduced and have to be dealt with. Indeterminacy might be introduced by explicit specification (e.g., earliest beginning and latest beginning of an interval) or is implicitly present in the case of multiple granularities. For instance, the statement 'Activity A started on June 14, 2009 and ended on June 17, 2009' can be modeled by the beginning instant 'June 14, 2009' and the end instant 'June 17, 2009' both at the granularity of days. If we look at this interval from a granularity of hours, the interval might begin and end at any point in time between 0 a.m. and 12 p.m. of the specified day [2].

1.1 Types of Temporal Uncertainty

There are many ways to model time in information systems and time is modeled differently for different applications depending on the particular problems. Aigner et al. [2] outline in detail major design aspects and features which are particularly important when modeling time with respect to visualizations. They distinguish between three different time primitives: (1) instants, i.e., single points in time, (2) intervals, i.e., durations between two instants, as well as (3) spans, i.e., durations that are not anchored in time.

Olston and Mackinlay [19] describe two different types of uncertainty: *statistical uncertainty* and *bounded uncertainty*. In case of *statistical uncertainty*, the probability of candidates within an uncertain interval follows a statistical model, for instance, a normal distribution. In case of *bounded uncertainty*, on the other hand, no assumptions can be made about the probability distribution of possible values inside the interval, i.e., all candidates are equally likely. Be it that they actually are equally likely or that there is just not enough information available to make statistical assumptions.

In addition, there may be dependencies between uncertain aspects of time primitives. For points in time there can be only one uncertain aspect, which is the exact point in time when this event actually takes

• *Theresia Gschwandtner, Markus Bögl, Paolo Federico, and Silvia Miksch are with Vienna University of Technology. E-mail: {last name}@ifs.tuwien.ac.at.*

Manuscript received 31 Mar. 2015; accepted 1 Aug. 2015; date of publication xx Aug. 2015; date of current version 25 Oct. 2015. For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

place. For intervals the start time and the end time can be uncertain. There is a *dependency* between these two uncertainties when, in case the interval starts early, it is more likely that it also ends early. For instance, when a work package of a project starts earlier than expected, it is more likely that it also ends earlier than expected. However, this is not always the case. For instance, if the uncertainty of start and end of an interval stems from a coarse granularity.

This led us to categorize eight different types of temporal uncertainty:

- Bounded uncertainty of an instant (e.g., we know something took place at May 14, 1988 but we do not have any information about at what exact time it took place)
- Statistical uncertainty of an instant (e.g., radiocarbon dating the time of death of a plant gives you an estimated age together with an error and confidence interval)
- Bounded uncertainty of start and end of an interval (e.g., uncertainties about start and end time stem from a coarse granularity)
- Statistical uncertainty of start and end of an interval (e.g., phenological seasons: an early start of winter does not imply an early end of winter, but there is statistical probability for the time of start and end)
- Bounded uncertainty of start and end of an interval with dependency (e.g., an early start of a project plan makes an early end more likely; it is known when this plan has to start and when it has to end, but no statistical assumptions can be made about start and end time)
- Statistical uncertainty start and end of an interval with dependency (e.g., start and end time of studies of an individual; an early start makes an early end more likely and there are statistical assumptions on start and end age)
- Bounded uncertainty of a span (e.g., I do not know if it took my friend 5 or 10 hours to find the perfect present one day before my birthday)
- Statistical uncertainty of a span (e.g., the duration of pregnancy)

These different types of temporal uncertainty require different types of visualizations. The visualization of statistical uncertainty of an instant could be similar to statistical uncertainty of any numerical value. Corell and Gleicher [6] investigated different encodings to visually represent this type of uncertainty. Also the bounded uncertainty of an instant may be visualized similar to the methods proposed in [19]. However, other types of temporal uncertainty require more specific types of visualizations and the suitability of alternative visual encodings of these has not been sufficiently evaluated yet. Thus, we have investigated different types of visual encodings to represent statistical and bounded uncertainty of start and end of an interval (without dependency between start and end time).

2 Visualization of Uncertainty

When it comes to a comprehensive categorization of different methods to visually encode uncertainty, most existing work focuses on geospatial data or spatio-temporal data (e.g., [20, 27]). According to these findings, uncertainty can be mapped to visual attributes, such as color, transparency, line width, texture, and sharpness or focus; other visualization methods include for instance, side-by-side displays of competing results, side-by-side displays of data values and uncertainty values (adjacent maps), animation, additional transparent layers, additional symbols, glyphs (most commonly error bars), contouring, confidence intervals. There are also non-visual methods, like mapping uncertainty to sound. However, the representation of complex temporal uncertainties is quite often accomplished by glyphs.

2.1 Temporal Uncertainty

In 1765 Joseph Priestley created a graphical representation of the life spans of famous historical persons [22]. He introduced the use of a horizontal line to represent an interval of time – which is a common method nowadays. Moreover, he also considered the visual representation of temporal uncertainties using dots. He even considered different levels of uncertainty by using solid lines, different amounts of dots before and after lines, as well as dots below lines.

Program Evolution and Review Technique (PERT) charts have been developed by the US Navy in 1950 to facilitate the planning and management of complex projects. PERT charts graphically depict the single tasks of a project together with their temporal inter-task dependencies by using boxes to represent tasks and arrows to depict predecessors and successors of each task. Scheduling information and buffer time of each tasks are given as numerical values within the boxes. These charts are especially useful when the exact duration or start and end times of tasks are hard to define. Often PERT Charts are used to identify critical time constraints (the critical path) within the project, i.e., if there is no buffer time between the single tasks of the critical path; as a consequence, the delay of one of these tasks would always lead to a delay of the whole project. However, PERT charts represent the crucial information rather in textual form than using visual means.

Sets Of Possible Occurrences (SOPO) diagrams were first described in a landmark paper by Jean-Francois Rit in 1986 [25]. Both the x- and y-axis represent time. The x-axis encodes the starting time of a process, while the y-axis encodes the finishing time. A diagram is created by the limiting parameters of earliest and the latest start of the process, the earliest and the latest end of the process, as well as the minimum and the maximum duration of the process. The resulting area represents all possible intervals that fit the temporal constraints of the given process. This diagram has been used in the medical application SOPOView [16], which provided visualization of therapy plans. The conducted evaluation shows that most of the physicians found the SOPO-diagram complex and confusing. They argued it would take too much time to get familiar with the visualization, which they could not spare in their busy job. Furthermore, it was remarked that a lot of the test users could not identify the benefits of the visualization.

The TimeAnnotationGlyph (part of the AsbruView project [11]) is focused on the communication of the exact temporal dimensions of clinical treatment plans. This is accomplished by using time-glyphs. A time-glyph consists of two bars; the first bar indicates the maximum duration of the plan. Beneath this bar the earliest start, latest start, earliest end, and latest end are indicated by little arrows pointing towards these time points. On top of the maximum duration bar two diamonds represent the latest start and the earliest end. The minimum duration of the plan is represented by a second bar on top of these two diamonds.

Another method to visualize temporal uncertainty using metaphors is Paint-Strips [5]. They are composed of a painted horizontal line (represents definite time). However, at the end of the line there can be a paint roller, attached to a weight. This roller can move freely until it hits a wall. This paint roller represents the earliest and latest finishing time of the process. Additionally, several paint rollers can be attached to the same weight, implying dependencies between processes. This technique is easy to understand. However, as a consequence of the basic design it is limited in its expressiveness and not suitable to represent complex constraints.

PlanningLines [3] is a visualization method to communicate temporal uncertainties of processes and was designed to provide a more efficient alternative to PERT-charts. The temporal representation of processes was derived from the visualization of LifeLines [21]. A PlanningLines glyph allows for depicting complex temporal uncertainties, like earliest start, latest start, earliest end, and latest end, as well as minimum and maximum duration. It consists of two encapsulated bars representing the minimum and maximum duration, which are bounded by two caps that represent the start and end intervals.

Decision charts [9] are a graphical representation for depicting future decisions and potential alternative outcomes along with their probabilities over time. In contrast to the other approaches which use an ordered time domain, decision charts use a branching time model.

From these examples we can see that the visualization of complex temporal uncertainties is quite often accomplished by specialized glyphs. More general efforts have been made in investigating different types of visual encodings to represent uncertainty of data values in other domains.

2.2 Uncertainty of Data Values

In 2009 Sanyal et al. [26] conducted a study to examine the effectiveness of four commonly used uncertainty visualization techniques – traditional error bars – somehow equivalent to our centered error bars (see Section 2.3) but for line charts and surfaces, size of marks, color of marks, and color of lines or surface areas. The visualizations were based on either 1D or 2D data sets (which result in two- and three-dimensional visualizations). The study was conducted with 27 participants who had to solve four assignments which were divided equally into search and counting tasks. These were devised to simulate an exploratory navigation of the data set. Search tasks were aimed at identifying locations of high or low uncertainty from within an area. For counting tasks, the users had to count the different data features or uncertainty features in a visualization. The results showed some interesting observations. The results for using error bars to indicate uncertainty stood out most. Users performed very poorly in the questions where error bars were used, regardless of the data dimension (even though it took the most time to solve the tasks). They argued that a possible reason could be the high density of the data sets which made it hard to identify the single error bars. Another interesting result is that there was a significant difference in user performance between searching for locations of high uncertainty and searching for locations of low uncertainty. For 1D data, using the size of marks to indicate uncertainty had the best results for data values with high uncertainty. Mark color and surface color were the most successful techniques to search for low uncertainty. For 2D data, surface coloring performed best overall, however, results for counting data features were not optimal. In these cases (i.e., counting tasks) the other techniques were more successful, but they were all outperformed by surface coloring in the other disciplines. It should be noted that there is no clear winner among the four visualization technique, just best choices for a specific task. The results showed lower error rates for 1D data than for 2D data, even though the latter tasks took longer time to be solved.

Corell and Gleicher [6] presented a user study on the effectiveness of four different visual encodings of statistical uncertainty of data values, i.e., bar charts with error bars (again similar to our centered error bars), modified box plots, gradient plots, and violin plots. Their results revealed some problems with error bars: (1) a 'within-the-bar bias' which means that representing the mean value by a bar provides a false metaphor of containment and participants thought that values within this bar are more likely than values outside this bar, and (2) a 'binary interpretation', which means that values are either within the margins of the error bars, or they are not, which leads to difficulties in judging the correct probability.

In 2012 MacEachren et al. [15] conducted two connected studies, of which one targeted intuitiveness of abstract visual variables and iconic symbols in uncertainty visualizations. The second study focused on performance while map reading tasks had to be solved using the most intuitive abstract and iconic representations of uncertainty. It was limited to the use of symbols to represent different items of discrete data linked with uncertainty. The types of uncertainty used were accuracy, precision, and trustworthiness. These were mapped to three data types, namely space, time, and attribute, resulting in nine different conditions for uncertainty. The symbols that were used can be split into two groups: abstract and iconic. The first group consisted of symbols which vary only in one visual variable. The eleven abstract visual variables have been designed in the works of Bertin [4], Morrison [17] and MacEachren [14, 13]. The second group consisted of symbols which have an icon, a graphical metaphor for their represented value, for instance, the picture of a clock to signify time. Some of their results reflect that iconic symbols of uncertainty can be more intuitive and more accurately judged when aggregated while abstract visual encodings can lead to quicker judgments.

While two of these studies report on problems with error bars, although for different reasons, it is hard to nominate a definite winner to visualize uncertainty. However, we took these studies as an inspiration to find a set of encodings that might be useful to represent uncertainties of start and end times of intervals.

2.3 Visual Encodings of Uncertain Intervals

We formulated our design goals partly in accordance with [6] and modified them for the representation of intervals:

1. The visual encoding should clearly present the effect size – the certain part of the interval should be clearly represented in combination with the uncertain start and end time.

2. The visual encoding should be compatible with the well-known representation of temporal intervals as horizontal bars on a temporal axis.

3. Encodings of statistical uncertainty should explicitly map the underlying probability distribution to a continuous visual variable.

4. Encodings of bounded uncertainty should not provoke the interpretation of varying probabilities (i.e., no changes in the visual variable).

Besides the studies mentioned in the previous section, a number of different ways of encoding uncertain information can be found in the literature. Bertin [4] defined these seven visual variables for visual representations: location, size, color hue, color value, grain, orientation, and shape. Morrison [17] added color saturation and arrangement, and MacEachren [14, 13] suggested three more variables specifically for the visualization of uncertainty: clarity (fuzziness), resolution (of boundaries and images), and transparency. However, some of these visual encodings are quite similar, thus we grouped them into (1) focus, contour crispness, fill clarity, fog, and resolution [13], fuzziness, clarity, and blur [13, 15], (2) fog obscuring uncertain items [13], (3) color value [7, 15], (4) texture [7], (5) location [15], (6) orientation [13], (7) transparency [8] (MacEachren evaluated transparency as being 'acceptable' [15]), (8) color hue [7, 13] (MacEachren refuted its suitability [15]), (9) color saturation [13] (MacEachren refuted its suitability [15]), (10) shape [13] (MacEachren refuted its suitability [15]).

However, not all of these visual variables are applicable when representing uncertainty of start and end times. One of our design goals was to preserve the well known representation of intervals as horizontal bars (compare, for instance, [22, 21, 5, 3]). Thus, we came up with a set of six different visual encodings for uncertain start and end of temporal intervals that go along with representing the certain part of the interval (i.e., the time span at which the event takes place for sure) by a horizontal bar along a time axis. Three of our encodings we designed for the representation of bounded uncertainty (1–3) and the other three we designed for the representation of statistical uncertainty (4–6). In Figure 1 we present these different types:

(a) **Gradient plots**: using a color gradient form a solid color (e.g., blue) to white (in accordance to [6])

(b) **Violin plots**: encoding the probability that a point in time falls into the interval by the height of the bar (in accordance to [6])

(c) **Accumulated probability**: similar to violin plots but representing the uncertainty by accumulating the probability distribution of the start and end times

(d) **Error bars**: as used in [19]

(e) **Centered error bars**: error bars that are centered around the most likely start and end times

(f) **Ambiguation**: as used in [19]

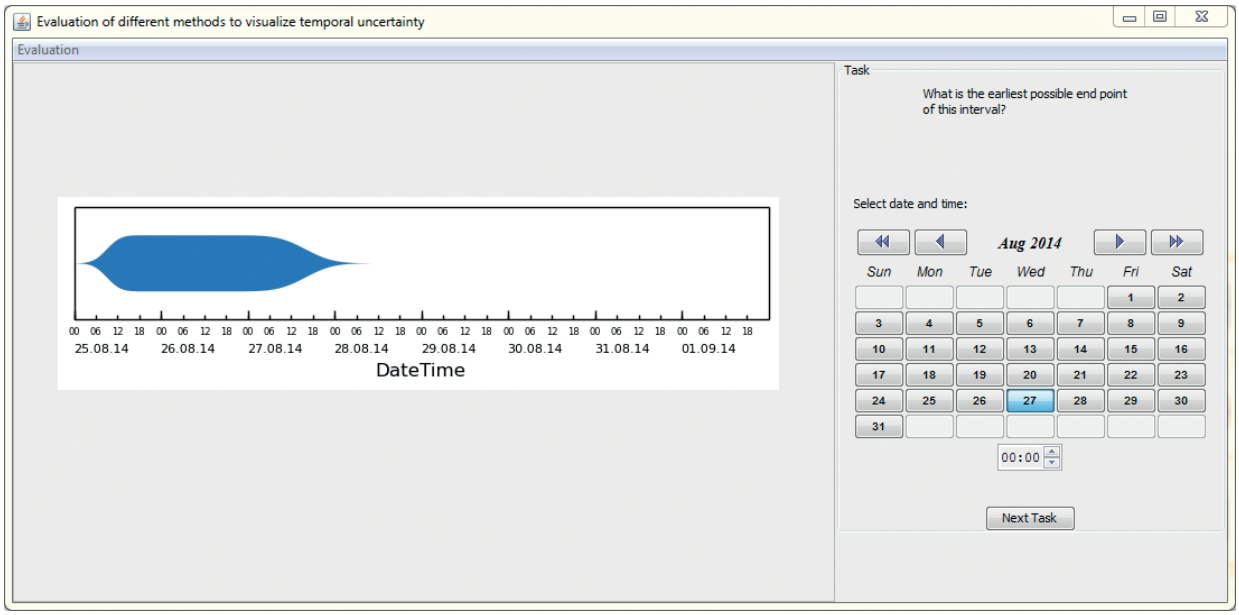

Fig. 2: We used EvalBench [1] for the evaluation session. This is a screenshot of one question of Task 2c.

Although Olston and Mackinlay [19] used error bars to represent statistical uncertainty, we assumed that our study participants would interpret them as representing bounded uncertainty, since they do not explicitly encode the underlying probability function. In Figure 1 the start and end of uncertain regions are marked by lines. These lines were not given in the pictures presented in the user study.

3 Study Design

In the following subsections we give a detailed outline of our study design, including the design goals, data and visualization preparation, our hypotheses, and how we conducted the user study session.

3.1 Data

The figures we presented to our subjects are based on generated artificial data that was then used to generate the visual representations. In a first step we generated the data using the R environment for statistical computing [24]. We generated a day and a month in the year 2014 using the *runif* function to draw random deviates from a uniform distribution. Applying the same method, we added a random number of hours to the date to generate a time stamp for the earliest start and repeated adding hours for the latest start, earliest end, and latest end. We limited the different intervals between the respective time points by specifying the minimum and maximum number of hours of each interval. Based on these time stamps we calculated the mean start time (i.e., *earliest start + (latest start - earliest start) / 2*) and the mean end time accordingly. We chose the different parameters in a way that we could provide a fixed time scale for all visualizations but still generate most possible combinations of uncertain start and end regions and interval durations. We did this to cover a wide range of possible occurrences of uncertainty in real life. We used six labels for each of the six types visual encodings and assigned the labels to the random data. For the questions asking about the probability that the interval started or ended at a specific time point (Task 4a and 4b), we generated time points in the interval between earliest start and latest start as well as earliest end and latest end. Therefore, we used the *rnorm* to draw random time points from a normal distribution and calculated the respective probability values. The resulting data frame containing all necessary data was exported as a comma separated file and used to generate the visual representations. For generating the figures out of the random data, we used *Python* [23] and *matplotlib* [10]. For the figures using a gradient, violin, or accumulated probability transition between the earliest and latest start/end, we computed a normal cumulative distribution function (normal cdf) with $\mu = 0, \sigma = 1$ and used these values to weight the alpha level, width, or height respectively. Therefore, we used the range of $\mu \pm 3\sigma$ to have a smooth transition from earliest and latest start/end with the μ at the mean start/end time point. We decided to not embed our questions within a use case or domain in order to not foster any domain-specific presumptions or interpretations.

3.2 Hypotheses

We formulated the following five hypotheses:

- H1 Participants naturally interpret gradient plots, violin plots, and accumulated probability plots to represent statistical uncertainty while they interpret error bars, centered error bars, and ambiguation to represent bounded uncertainty.
- H2 Error bars, centered error bars, and ambiguation are better suited (in terms of error and time) to represent earliest start, latest start, earliest end, and latest end than gradient plots, violin plots, and accumulated probability plots.
- H3 Error bars, centered error bars, and ambiguation are better suited (in terms of error and time) to judge minimum and maximum duration of an interval.
- H4 Gradient plots, violin plots, and accumulated probability plots are better suited (in terms of error and time) than ambiguation and error bars to judge the average duration of an interval.
- H5 Gradient plots, violin plots, and accumulated probability plots are equally suited (in terms of error and time) when judging the probability that a marked point in time falls into the represented interval.

3.3 Participants and Evaluation Session

We had 73 participants which we recruited from our bachelor students in computer science (14 participants were female). All students were taking our course in information design and visualization which implies a certain knowledge about visual representations. To conduct the evaluation session we used EvalBench [1] (see Figure 2), a software library designed especially for evaluating visualizations and we conducted a pilot testing session with two participants. We decided for a within-subject study, so each participant had to solve the following tasks for all types of visual encodings:

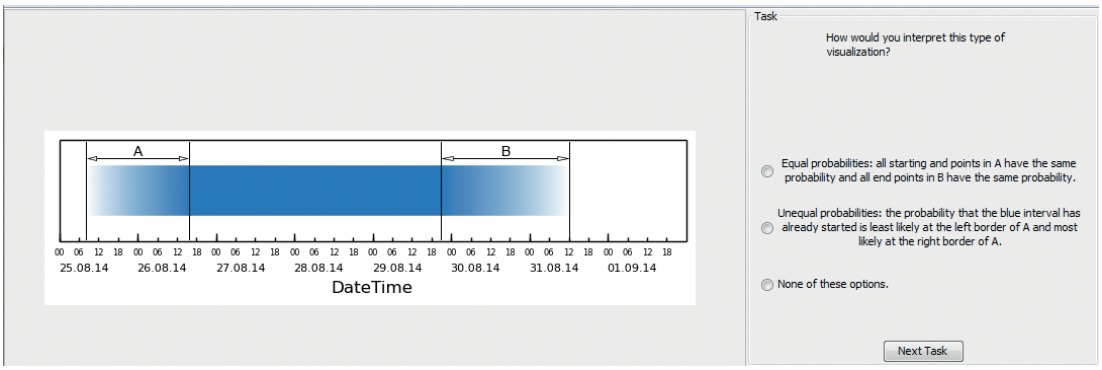

Fig. 3: This is a screenshot of one question of Task 1.

Task 1 – Interpretation: We presented one picture of each type with marked start and end intervals (labeled by 'A' and 'B'; see Figure 3). For each type the participants answered the question

1 How would you interpret this type of visualization?

by choosing one of three possible answers: (•) Equal probabilities: all starting points in A have the same probability and all end points in B have the same probability, (•) Unequal probabilities: the probability that the interval has already started is least likely at the left border of A and most likely at the right border of A, and (•) None of these options. This task was preceded by a detailed explanation of what we mean with this two different types of uncertainty.

Task 2 – Start and End Time: We presented one picture of each visual encoding type for each of these subtasks:

2a What is the earliest possible start point of this interval?

2b What is the latest possible start point of this interval?

2c What is the earliest possible end point of this interval?

2d What is the latest possible end point of this interval?

Participants were asked to answer by selecting a date and time by means of a date picker tool (see Figure 2).

Task 3 – Interval Duration: We presented one picture of each visual encoding type for each of these subtasks:

3a What is the maximum duration of the interval (in hours)?

3b What is the minimum duration of the interval (in hours)?

3c What is the average duration of the interval (in hours)?

Participants were asked to answer by freely picking a numeric value (number of hours).

Task 4 – Probabilities: We presented three pictures of each of the types gradient plot, accumulated probability plot, and violin plot for each subtask and marked different positions in uncertain regions with a red line (see Figure 4), or, in case of Task 4c, with two red lines (one in the uncertain start region and one in the uncertain end region). Again, this task session was preceded by an explanation about the underlying probability distribution and its representation.

4a What is the statistical probability (in %) that the interval has already started at the marked point in time (red line)?

4b What is the statistical probability (in %) that the interval has already ended at the marked point in time (red line)?

4c Which of the two marked points in time (red lines) are more likely to fall into the interval?

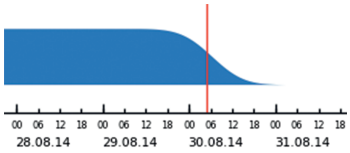

Fig. 4: Task 4b: judging the probability that the interval has already ended at the marked point in time.

Task 4a and Task 4b were answered by picking a numeric value between 0 and 100. To answer Task 4c, participants could choose one of three options: (•) the earlier point in time is more likely contained in the interval, (•) the later point in time is more likely contained in the interval, (•) both points in time are equally likely contained in the interval.

Task 5 – Preferences: For the final task we presented one picture of each type and asked the participants:

5 What are your personal preferences regarding this type of visualization?

on a five point Likert scale from 'I don't like it' to 'I like it'.

For these tasks (except for Task 1 and Task 5) we measured the completion time and error, i.e., depending on the question type we measured right or wrong, or error magnitude (the absolute distance from the expected value). In order to balance our results with respect to fatigue and learning effects, questions within one task were presented to the participants in random order.

4 Results

We had 73 submissions from which we had to exclude one because the logs of completion time and error rate were empty. Thus, we had a total of 72 submissions. Given the varying numbers of pictures we presented for each type for different task we had different sample sizes for different tasks (see Table 1) with a minimum of n = 72.

4.1 Analysis Approach

In order to analyze our results we used the R environment for statistical computing [24]. We conducted one way analysis of variances (ANOVA) for each task and subtask. Since our data does not always completely follow a normal distribution (which is a prerequisite for ANOVA), we backed up our results with the non-parametric Kruskal-Wallis test [12], which led to similar but even more significant results (see Table 2 and 3). ANOVA and Kruskal-Wallis tests give initial information if there is a significant difference in the means and variances (of error and time) of the different groups, i.e., the different visual encodings. To understand which of the groups differ significantly, we conducted Games-Howell post-hoc tests or Nemenyi post-hoc tests [18] respectively. Moreover, we visually investigated boxplot

Table 1: Sample Sizes for Different Tasks.

task	# encoding types	# pictures per type	# answers per type	# answers after exclusion*
Task 1	6	1	72	72
Task 2a	6	1	72	67
Task 2b	6	1	72	58
Task 2c	6	1	72	64
Task 2d	6	1	72	66
Task 3	6	1	72	72
Task 4a	3	3	216	216
Task 4b	3	3	216	191
Task 4c	3	3	216	216
Task 5	6	1	72	72

**we describe the exclusion of answers in Section 4.2*

representations of the results as well as mean and median values in order to form our conclusions.

However, before actually running these tests, we manually inspected the results. Among a large majority of reasonable results, we also noticed some peculiarities.

4.2 Problems and Observations

In this section we provide a fair discussion of the problems we encountered, the data cleansing we had to conduct at different tasks, and also of the mistakes we made. In particular, when inspecting the results we noticed different problems for different tasks:

Task 2: Confusion of similar questions: When inspecting the answers for Task 2 it was obvious that the participants quite frequently confused earliest start, latest start, earliest end, and latest end. This is not surprising, since they had to go through a lot of questions and may have gotten less careful over time. To deal with this problem, we prepared a second data set from the original set of answers for comparison. For this second answer set we excluded answers that could not be considered as a valid answer to the question for these two reasons: the given answer, i.e., the point in time, is much too far away from the time we were looking for or the given answer points directly to one of the other start and end points. Our exact algorithm was: (1) exclude points in time that are not visible in the picture (outside the scale), (2) exclude answers that differ more than 18 hours from the point in time we were looking for, i.e., outside a 36 hour interval around the point we were looking for (considering our overall time frame of 192 hours for the whole picture, these 36 hours seemed to be a rather conservative approach), and (3) exclude answers that point to an other start or end time with a precision of 6 hours or less (6 hours are equivalent to one tick on the time scale of the provided pictures). This led to the exclusion of 103 answers from 1728 answers (i.e., 17 earliest start, 39 latest start, 31 earliest end, and 16 latest end answers). Leaving a minimum of 67 answers for each visual encoding for Task 2a, 58 answers for Task 2b, 64 answers for Task 2c, and 66 answers for Task 2d. However, we based our analysis on the original answer set and used this second answer set only for comparison.

Table 2: P-Values of ANOVA and Kruskal-Wallis Tests for **Errors**.

task	ANOVA p-value	Kruskal-Wallis p-value
earliest start	0.892	< 0.001 ***
latest start	0.130	< 0.001 ***
earliest end	0.008 **	< 0.001 ***
latest end	0.274	< 0.001 ***
together	**0.097**	$<$ **0.001** ***
minimum duration	< 0.001 ***	< 0.001 ***
average duration	0.188	< 0.001 ***
maximum duration	0.047 *	< 0.001 ***
together	**0.002** **	$<$ **0.001** ***
probability (deleted)	< 0.001 ***	< 0.001 ***
probability (all)	$<$ **0.001** ***	$<$ **0.001** ***

*Significance levels: p-value<0.001: ***, p-value<0.01: **, p-value<0.05: **

Table 3: P-Values of ANOVA and Kruskal-Wallis Tests for **Time**.

task	ANOVA p-value	Kruskal-Wallis p-value
earliest start	0.005 **	< 0.001 ***
latest start	0.046 *	< 0.001 ***
earliest end	0.217	< 0.001 ***
latest end	0.995	0.029 *
together	**0.931**	**0.139**
minimum duration	< 0.001 ***	< 0.001 ***
average duration	0.383	0.211
maximum duration	< 0.001 ***	< 0.001 ***
together	**0.004** **	$<$ **0.001** ***
probability (deleted)	0.048 *	0.006 **
probability (all)	**0.043** *	**0.003** **

*Significance levels: p-value<0.001: ***, p-value<0.01: **, p-value<0.05: **

Task 4b: Badly phrased question: Inspecting the answers to Task 4b, i.e., judging the probability to which the interval has already ended a marked position in time (in the uncertain end region), we again noticed some obvious misunderstandings. In particular, there were many answers that gave us the estimated probability to which the marked point in time falls into the interval (i.e., indicated by the height of accumulated probability plots and violin plots, and by the amount of transparency of the gradient plot). This is correct when asking for the probability to which the interval has already started at the marked position. However, in case of the probability to which the interval has already ended at the marked position, this answer is the inverted probability of what we were looking for (i.e., 100%-given answer). Thus, we decided to still use the original set of given answers for analysis, but also prepare two more sets of answers for further investigation: one where we excluded the inverted answers and one where we re-inverted these answers. Our exact algorithm to identify these inverted answers was: answers that deviate from the correct probability by more than 50% and the deviation of their inverted values from the correct value is less than 15%. This led to the exclusion/inversion of 62 answers out of 648 answers, which left a minimum of 191 answers for each visualization type when excluding them. However, these three data sets of different degrees of preparation led to only minor changes in outcome.

Task 4c: Unequal conditions: We prepared pictures of gradient plots, violin plots, and accumulated probability plots (three different pictures for each type) where two points in time were marked – one in the uncertain start region and one in the uncertain end region – and asked the participants which point is more likely to fall into the interval or if they are both equally likely. However, the results showed very bad scoring for gradient plots which does not go along with the results of the other tasks. We believe we made a mistake in presenting two out of three gradient plots with equal probabilities of the two marked positions to fall into the interval (in contrast to much more easily distinguishable probability settings for the other plots). We decided to not present these results, because we believe they only reflect our mistake in not preparing pictures of balanced distinguishability for the three visual encodings. On the other hand, we already tested the effectiveness of these visual encodings in communicating probabilities in Task 4a and Task 4b.

4.3 Outcome

In this section we evaluate our results and reflect on our hypotheses.

H1 – Interpretation of visual encodings: Hypothesis 1 is **partly confirmed**. There are very clear results for most visual encodings and these match our assumptions, that accumulated probability plots, violin plots, and gradient plots are interpreted as statistical probability and the others as bounded probability (see Figure 5). However, centered error bars present an outlier here. Our results do not show a clear winner: 47% expected this visual encoding to represent bounded uncertainty and 36% expected it to represent statistical uncertainty. Moreover, 17% of the participants thought it represents neither of them. This is by far the highest percentage value of the category 'other' among all visual encoding types, which we believe also represents the participants confusion with this type of visualization.

Another notable finding is that error bars as well as centered error bars are usually quite often used to represent statistical uncertainty [19], but we assumed that participants would consider them to represent bounded uncertainty because they do not explicitly encode the probability distribution. This assumption was backed up by the study results, which show that 89%

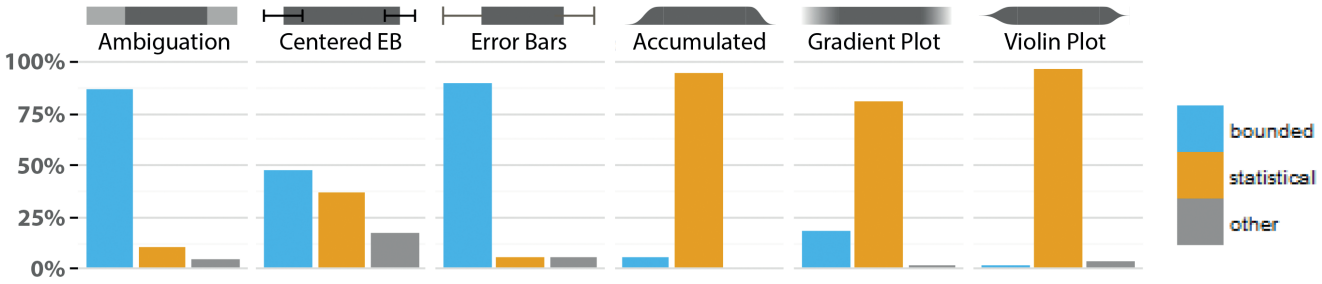

Fig. 5: **Ad-hoc interpretation** of the different visualization types. Centered error bars led to confusion: 47% expected this them to represent bounded uncertainty, 36% expected them to represent statistical uncertainty, and 17% of thought they represent either of which.

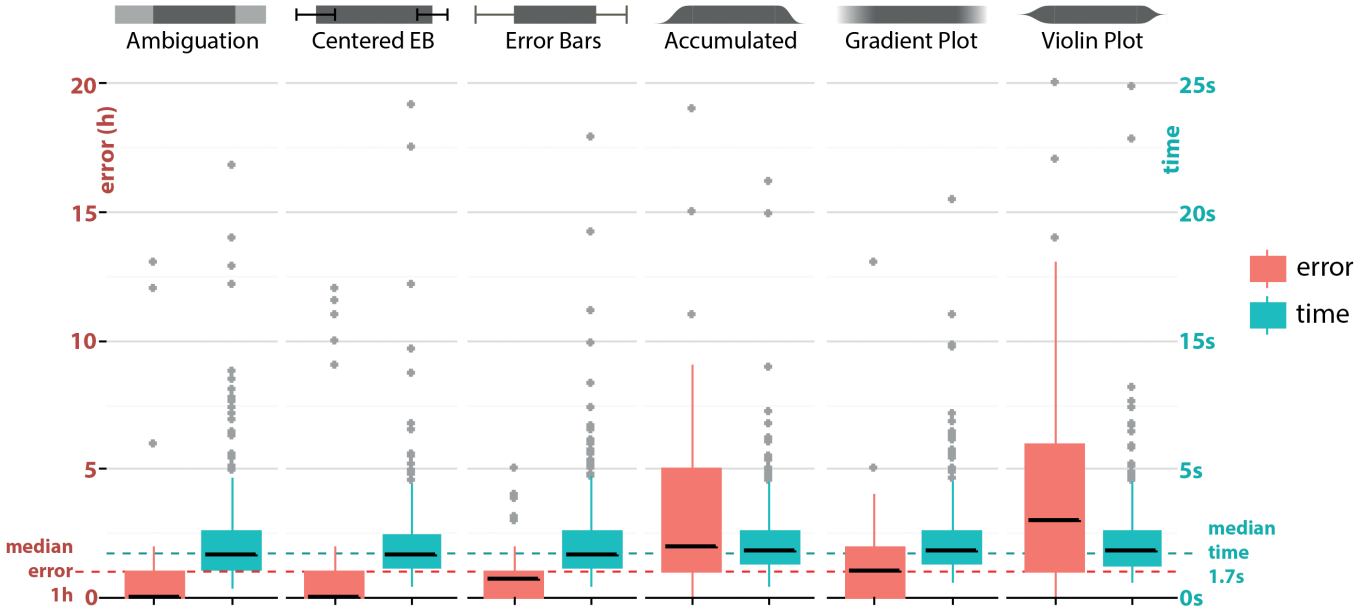

Fig. 6: Time and error rates for judging **earliest start, latest start, earliest end, and latest end**. Ambiguation, centered error bars, and error bars lead to significantly ($p<0.001$) better error rates than accumulated probability plots, gradient plots, and violin plots. Gradient plots, however, score significantly ($p<0.001$) better than accumulated probability plots and violin plots.

of the participants expected error bars to represent bounded uncertainty, and besides the general confusion with centered error bars there are still more votes for bounded uncertainty than for statistical uncertainty.

H2 – Effectiveness in representing start and end time: Hypothesis 2 can be **confirmed**. We assumed that ambiguation, error bars, and centered error bars are better suited to represent start and end times.
Error: In Table 2 we see that there is a big difference between p-values of the ANOVA test and p-values of the Kruskal-Wallis test because this answer set did not meet the preconditions of the ANOVA test. Thus, we used the Kruskal-Wallis test in combination with the Nemenyi post-hoc test. The Nemenyi post-hoc test showed that, as expected, ambiguation, error bars, and centered error bars lead to significantly ($p<0.001$) better error rates than accumulated probability plots, violin plots, and gradient plots. When looking at the comparison answer set (from which we excluded obviously misguided answers), we find very similar results with no difference in significance levels.
Time: Considering the completion time there are significant differences for the single subtasks, but not when looking at all four cases together, so these differences are distributed among the different visualization types and cancel each other out.
Other observations: While there are no significant differences between ambiguation, error bars, and centered error bars, gradient plots score significantly ($p<0.001$) better than the two other encodings of statistical uncertainty (i.e., accumulated probability plots and violin plots). The results show that violin plots and accumulated probability plots have a much higher error rate than all other visual encodings, especially when assessing latest start and earliest end. Participants estimated the latest start too early and the earliest end too late. Moreover, even in the cleansed comparison data set we found many outliers for centered error bars, most of them pointing to either the beginning or the end of the bar that represents the actual interval. Thus, we assume that the visual encodings of centered error bars led to confusion.

H3 – Effectiveness in representing minimum and maximum duration: Again, hypothesis 3 can be **confirmed only partly**. We assumed that error bars, centered error bars, and ambiguation are better suited to represent minimum and maximum duration of intervals. Since the answer data set did not meet all constraints of the ANOVA test, we again used the Kruskal-Wallis test in combination with the Nemenyi post-hoc test.
Error: Accumulated probability and violin plots have significantly ($p<0.001$) worse error rates than all the other visual encodings for judging the minimum and maximum duration. Again, gradient plots score better than expected and do not significantly differ from ambiguation, error bars, and centered error bars (see Figure 7). The mean error of gradient plots even makes the second place after error bars and the median error of gradient plots is second best after centered error bars.
Time: There are also significant differences in completion time. Error bars score significantly ($p<0.001$) better than accumulated probability plots, ambiguation, and centered error bars. In addition, gradient plots and violin plots score significantly ($p<0.001$) better than accumulated probability plots.

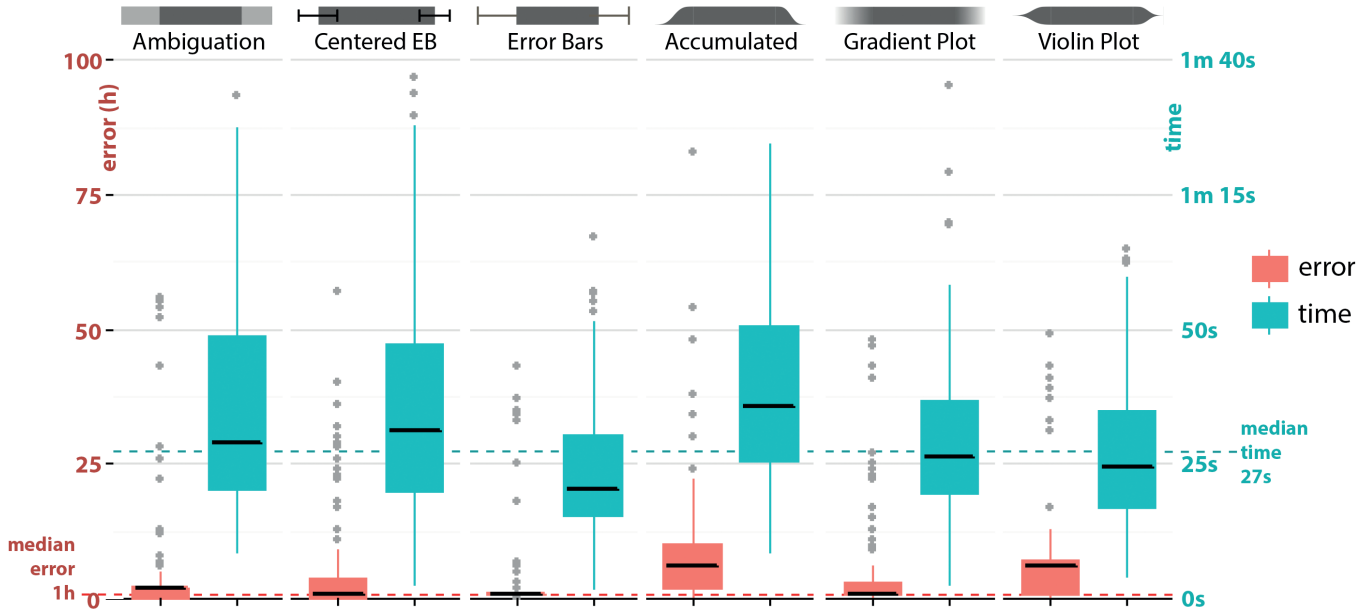

Fig. 7: Time and error rates for judging the **minimum and maximum duration** of intervals. Accumulated probability plots and violin plots have significantly ($p<0.001$) worse error rates than all other visual encodings.

H4 – Effectiveness in representing average duration: We thought that the visual encodings intended to represent statistical uncertainty are better suited to judge the average duration of an interval than the visual encodings for bounded uncertainty, since they give visual hints of the probability of start and end times. However, hypothesis 4 must be **rejected** since they score slightly worse in terms of median errors (see Figure 8). In addition, centered error bars lead to significantly better error rates than gradient plots ($p<0.001$), accumulated probability plots ($p<0.01$), and violin plots ($p<0.05$). This seems legitimate, given that the mean start time lies in the middle between earliest start and latest start and the same is true for the mean end point for all types of visual encodings. Thus, the interval bar of centered error bars encodes exactly the average duration of the interval.
Time: There are no significant differences in terms of completion time.
Other observations: Considering the combined results of error rates for minimum, maximum, and average duration, violin plots and accumulated probability plots score significantly ($p<0.001$) worse than all other visual encodings, and gradient plots score only significantly ($p<0.05$) worse than centered error bars and error bars.

H5 – Effectiveness in representing probability of start and end time: We assumed that gradient plots, violin plots, and accumulated probability plots are equally suited to represent the probability of which a marked point in time falls into the represented interval.
Error: Considering the original answer set of Task 4 (no cleansing), gradient plots have a significantly ($p<0.001$) lower error rate than violin plots (see Figure 9). Also accumulated probability plots lead to better error rates than violin plots (but not significantly: $p=0.06$). When deleting or correcting the inverted answers – both measures lead to very similar results – this picture becomes more clear: gradient plots have a significantly lower error rate than violin plots ($p<0.001$) and accumulated probability plots ($p<0.01$). And again, accumulated probability plots lead to slightly better error rates than violin plots.
Time: Considering completion time, there are no significant differences, but interestingly, gradient plots lead to the longest median completion times.
From this results, we draw the conclusion that hypothesis 5 must also be **rejected**, since gradient plots are significantly better suited to represent the probability of which a marked point in time falls into the represented interval – at least in terms of error rates.

5 Limitations and Further Work

Our study had some limitations that could be interesting aspects of further work. For instance, we did not compare the visual encodings explicitly showing the probability to which a point in time falls into the interval (i.e., gradient plots, accumulated probability plots, and violin plots) to those visual encodings which do not (i.e., ambiguation, error bars, and centered error bars) in their effectiveness for judging these probabilities. It would be important to understand if the explicit visual mapping of these probabilities actually provide a significant benefit. Moreover, we had to exclude Task 4c (i.e., which of two marked positions is more likely to fall into the interval) because of mistakes in the task design. It would be interesting to find out if this analysis leads to additional insights. Another aspect which should be considered is that we have based our visualizations of probabilities on normal probability distributions. It is not clear if the same visual encodings are equally suited to communicate other types of probability distributions. Also, the visualizations tested in this study were lim-

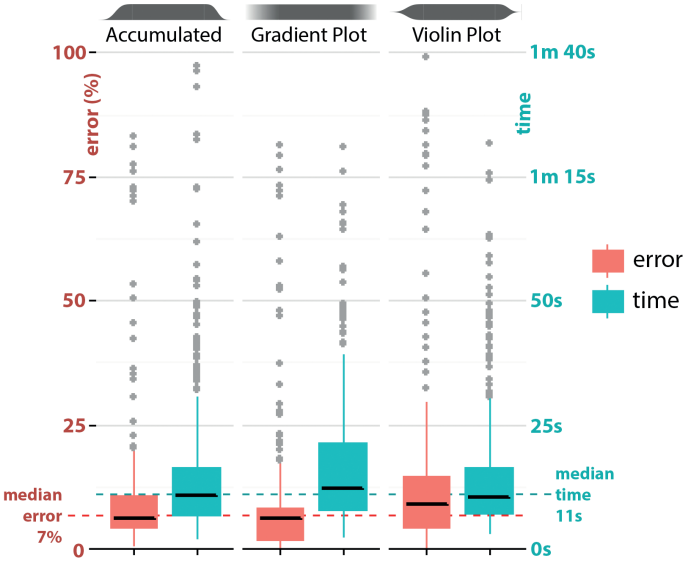

Fig. 9: Time and error rates for judging **probability values**. Gradient plots lead to significantly better error rates than violin plots ($p<0.001$) and accumulated probability plots ($p<0.01$).

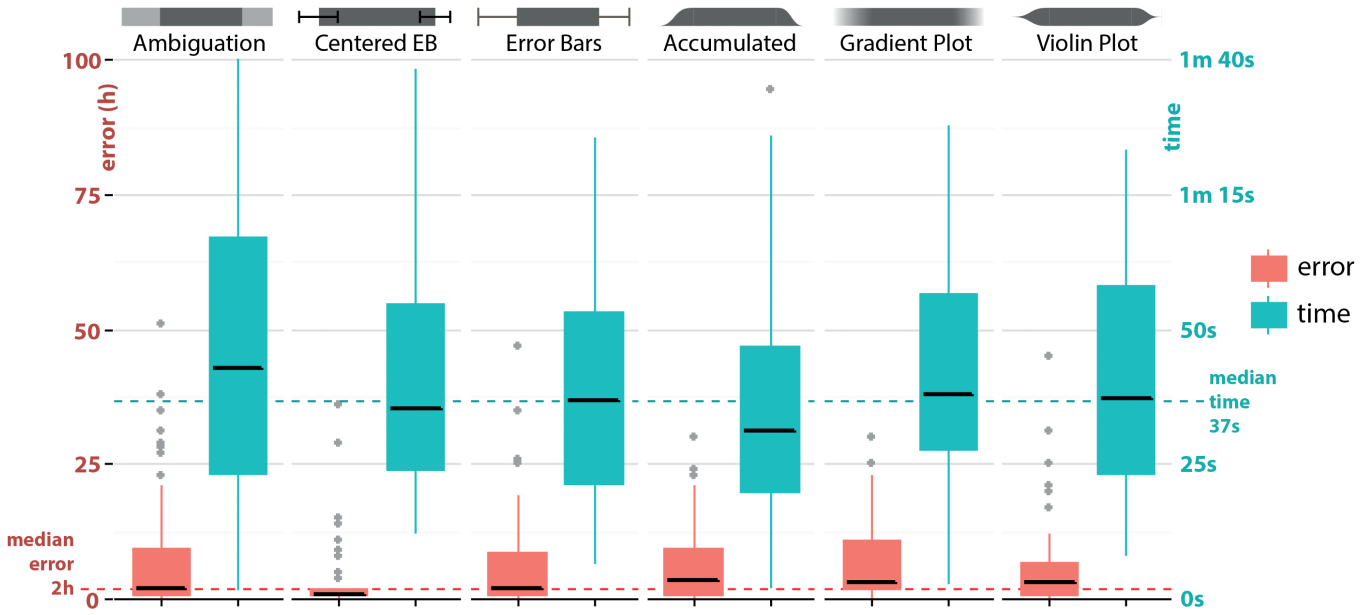

Fig. 8: Time and error rates for judging **average duration**. Centered error bars lead to significantly better error rates than accumulated probability plots ($p<0.01$), gradient plots ($p<0.001$), and violin plots ($p<0.05$).

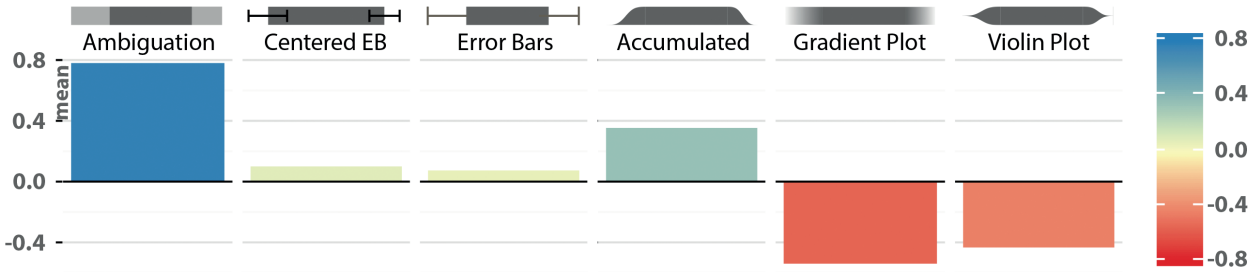

Fig. 10: **Preferences** about the different visualization types. We asked our study participants to express their personal preferences regarding the different visualization techniques on a five point Likert scale from -2 ('I don't like it') to 2 ('I like it').

ited to very simplistic, non-interactive representations of one interval at a time. More complex settings and interactions are relevant to be addressed in further work.

In general there is still a need to investigate other types of temporal uncertainties, for instance, different types of visual encodings to represent (bounded and statistical) uncertainties with dependencies (see Section 1.1). There are some approaches to represent these dependencies (e.g., [11, 3]), however, no comprehensive comparative study of these approaches has been conducted.

6 Conclusion

We have conducted a comprehensive user study comparing six different ways of encoding temporal uncertainty (i.e., uncertain start and end times of intervals) in terms of error and completion time. Our results show that centered error bars led to confusion on how to interpret them even though we gave an explanation marking earliest start, latest start, earliest end, and latest end in the beginning of the evaluation session. Other studies have already reported on problems with error bars [26, 6], however for different reasons. As usual, the best suited visual encoding depends on the task. For representing earliest start, latest start, earliest end, and latest end, as well as minimum, maximum, and average duration of intervals, error bars and ambiguation scored best. However, if also a statistical probability distribution should be represented, we recommend gradient plots, since they not only score best in representing these probabilities (in terms of error rates) but are also significantly superior to accumulated probability plots and violin plots regarding the other tasks. Interestingly, participants liked gradient plots the least when reporting on their preferences (see Figure 10).

Acknowledgments

We wish to thank Peter Filzmoser and Margit Pohl for their counseling regarding the statistical analysis. Moreover, the research leading to these results has received funding from the Centre for Visual Analytics Science and Technology CVAST, funded by the Austrian Federal Ministry of Science, Research, and Economy in the exceptional Laura Bassi Centres of Excellence initiative (#822746).

REFERENCES

[1] W. Aigner, S. Hoffmann, and A. Rind. Evalbench: A software library for visualization evaluation. *Computer Graphics Forum*, 32(3):41–50, 2013.

[2] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of time-oriented data*. Human-Computer Interaction. Springer-Verlag, 1st edition, 2011.

[3] W. Aigner, S. Miksch, B. Thurnher, and S. Biffl. PlanningLines: Novel glyphs for representing temporal uncertainties and their evaluation. In *Proceedings of the 9th International Conference on Information Visualisation (IV 2005)*, pages 457–463. IEEE Computer Society Press, 2005.

[4] J. Bertin. *Semiology of graphics: diagrams, networks, maps (translated by William J. Berg)*. University of Wisconsin Press, 1967/1983.

[5] L. Chittaro and C. Combi. Visual definition of temporal clinical abstractions: A user interface based on novel metaphors. In S. Quaglini, P. Barahona, and S. Andreassen, editors, *Proceedings of the 8th Conference on Artificial Intelligence in Medicine in Europe (AIME 2001), Lecture Notes in Computer Science*, pages 227–230. Springer-Verlag, 2001.

[6] M. Correll and M. Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2142–2151, 2014.

[7] T. J. Davis and C. Keller. Modelling and visualizing multiple spatial uncertainties. *Computers & Geosciences – Special issue on exploratory cartographic visualization*, 23(4):397–408, 1997. Exploratory Cartograpic Visualisation.

[8] I. Drecki. Visualisation of uncertainty in geographical data. In W. Shi, P. Fisher, and M. F. Goodchild, editors, *Spatial Data Quality, Chapter 10*, pages 140—159. Taylor & Francis, London, UK, 2002.

[9] R. L. Harris. *Information Graphics: A Comprehensive Illustrated Reference*. Oxford University Press, Inc., 1999.

[10] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[11] R. Kosara and S. Miksch. Metaphors of movement: A visualization and user interface for time-oriented, skeletal plans. *Artificial Intelligence in Medicine (AIIM), Special Issue: Information Visualization in Medicine*, 22(2):111–131, 2001.

[12] W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.

[13] A. M. MacEachren. Visualizing uncertain information. *Cartographic Perspectives*, (13):10–19, 1992.

[14] A. M. MacEachren. *How maps work: representation, visualization, and design*. Guilford Press, 1995.

[15] A. M. MacEachren, R. E. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan. Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2496–2505, 2012.

[16] P. Messner. Time Shapes – A visualization for temporal uncertainty in planning. MSc thesis in Business Informatics, Vienna University of Technology, Institute of Software Technology and Interactive Systems, April 2000.

[17] J. L. Morrison. A theoretical framework for cartographic generalization with the emphasis on the process of symbolization. *International Yearbook of Cartography*, 14:115–127, 1974.

[18] P. Nemenyi. Distribution-free multiple comparisons. Master's thesis, Princeton University, supervised by J.W. Tukey, 1963.

[19] C. Olston and J. D. Mackinlay. Visualizing data with bounded uncertainty. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02)*, pages 37–41. IEEE Computer Society Press, 2002.

[20] A. T. Pang, C. M. Wittenbrink, and S. K. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, 1997.

[21] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, and B. Shneiderman. Lifelines: Using visualization to enhance navigation and analysis of patient records. In *Proceedings of the AMIA Symposium (AMIA 1998)*, pages 76–80. Hanley & Belfus, Inc., 1998.

[22] J. Priestley. *A Chart of Biography*. Johnson, J., 1765.

[23] Python Software Foundation. *Python Language Reference, version 2.7*, 2015.

[24] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[25] J.-F. Rit. Propagating temporal constraints for scheduling. In *Proceedings of the 5th National Conference on Artificial Intelligence (AAAI-86)*, pages 383–388. AAAI Press, 1986.

[26] J. Sanyal, S. Zhang, G. Bhattacharya, P. Amburn, and R. Moorhead. A user study to compare four uncertainty visualization methods for 1D and 2D datasets. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1209–1218, 2009.

[27] H. Senaratne and L. Gerharz. An assessment and categorisation of quantitative uncertainty visualisation methods for geospatial data. In *Proceedings of the 14th AGILE International Conference on Geographic Information Science*, 2011.