Evaluation of Alternative Glyph Designs for Time Series Data in a Small Multiple Setting

Johannes Fuchs[1] **Fabian Fischer**[1] **Florian Mansmann**[1]
[1]University of Konstanz
fuchs@dbvis.inf.uni-konstanz.de
(fabian.fischer|florian.mansmann)@uni-konstanz.de

Enrico Bertini[2]
[2]NYU Poly
ebertini@poly.edu

Petra Isenberg[3]
[3]INRIA
petra.isenberg@inria.fr

ABSTRACT

We present the results of a controlled experiment to investigate the performance of different temporal glyph designs in a small multiple setting. Analyzing many time series at once is a common yet difficult task in many domains, for example in network monitoring. Several visualization techniques have, thus, been proposed in the literature. Among these, iconic displays or glyphs are an appropriate choice because of their expressiveness and effective use of screen space. Through a controlled experiment, we compare the performance of four glyphs that use different combinations of visual variables to encode two properties of temporal data: a) the position of a data point in time and b) the quantitative value of this data point. Our results show that depending on tasks and data density, the chosen glyphs performed differently. Line Glyphs are generally a good choice for peak and trend detection tasks but radial encodings are more effective for reading values at specific temporal locations. From our qualitative analysis we also contribute implications for designing temporal glyphs for small multiple settings.

Author Keywords

Glyphs; time series; evaluation; small multiples; information visualization.

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: Misc

General Terms

Human Factors

INTRODUCTION

Time series data is the basis for decision making in many different application domains—such as finance, network security, or traffic management—and, thus, constitutes an important area of research for visualization and data analysis. We collaborated, for example, with network security analysts from a large university computer center who need to make decisions based on the amount of daily network traffic for single hosts over time. Detecting trends, spotting peaks, or investigating single points in time from a visual representation are daily analysis tasks of vital importance for our collaborators and analysts in many other domains [18, 20, 26].

For data analysis in such a scenario, glyphs (iconic representations) are an appropriate choice to consider for visually encoding and presenting temporal data. Their advantage lies in their compact way to use screen real estate and the possibility to use them in a small multiple setting. In such a setting, glyphs can enable quick visual comparison of the development of data values over time. However, glyphs come with a trade-off between resolution and increased data density for each time series. They usually do not include axes for reading exact values since they are primarily designed to show multiple attributes in a compact way [36]. A notable example of such a technique is the well-known *sparklines* technique [33].

Yet, due to glyphs' power in presenting multiple time series for comparison, a multitude of designs have been proposed. Different visual variables such as length, color, or position can be used to encode two aspects of temporal data in one glyph: a) the location of a data point in time, and b) the quantitative data value. When confronted with the task of choosing an appropriate glyph design, a visualization designer or practitioner currently has little guidance on which encodings would be most appropriate for which tasks and on which visual features and factors influence people's perception of data encoded in glyphs. While one could follow Cleveland and McGill's ranking of elementary perceptual tasks [10] and try to predict the performance of glyphs based on these results, it is not clear whether their results will hold. Temporal glyphs include dual encodings, are used in specific temporal analysis tasks, and come in many different sizes and densities.

In order to address this lack of guidance on the use of temporal glyphs, we ran a controlled experiment to compare four carefully selected glyphs using two different data densities. These four glyphs were chosen for their use of different combinations of visual variables to encode temporal position and quantitative value of a data point. We evaluated all glyph designs in a small multiple setting as small multiple is the most common usage scenario for temporal glyphs. To our knowledge no other evaluation has been conducted to compare the performance of time series glyphs for small multiple settings based on their data encodings. In particular, we contribute:

- results comparing the task-dependent performance of four glyph designs under two data densities,
- plausible explanations for the observed performance patterns and resulting implications for design,
- a first investigating into the broader issue of how glyphs perform and what factors influence their performance.

RELATED WORK

Time series visualization has a long history going back to at least the 18th century and many different techniques have been developed in the past.

Time Series Visualization Techniques

Willam Playfair [28], for example, used line charts to visualize exports, imports, expenditures or prices and their development over time. Even today these line charts are among the most popular time series visualization techniques and their details are actively discussed in the visualization community, as for example the arc length-based aspect ratio selection [31]. Furthermore, visualization techniques such as stacked graphs (e.g. [37]) aim at making line graphs scalable to analysis tasks involving many time series at once.

Besides line charts, common techniques for visualizing time series are *pixel visualizations* (e.g., Recursive Patterns [4], Circle Segments [5], or Time-Series Bitmaps [22], surveyed in [17]) and *glyph visualizations* (e.g., Sparklines [33] or Tow-Tone Pseudo Coloring [29], surveyed in [35]). Furthermore, properties either inherent or assigned to time have resulted in the development of a number of specialized methods. Periodic patterns can, for example, be visualized with the Concentric Circles Technique [11] or Spirals [8]; likewise several calendar visualizations have been proposed [6, 34] to cope with the irregularities of our Gregorian calendar. Properties assigned to time series often result in multi-dimensional data sets, which can for example be analyzed with axes-based visualizations with radial layouts [32].

Time Series Comparison

Time series comparison is the area most related to our work. Some studies have already been conducted on the evaluation of multiple timeline representations [16] or the comparison of different value ranges for line charts [1]. Alternative techniques for displaying many time series at once are CloudLines [21] or Horizon Graphs [15]. More application driven visualizations, such as systems monitoring (e.g., LiveRAC [25]), project management (a classic: Gantt chart [9]), health (e.g., LifeLines [27]), news (e.g., ThemeRiver [14]) and geographic analysis (e.g., Space-time Cube [19]) make use of various dedicated representation techniques.

Temporal glyphs, the subject of our experiment, are often used in small multiple settings for comparing many different time series at once. Their layout on the plane varies to add additional information like the geographic context on top of a map [13], the ranking in a scatterplot, or a hierarchical data organization [12]. Pearlman and Rheingans use stacked circular glyphs in a graph layout to monitor network traffic over time and visualize the connections [26]. Circular glyphs positioned in a matrix for monitoring the daily traffic of many network devices were also mentioned by Kintzel et al. [18]. These circular representations are similar to the ones used in our experiment. Krasser, however, uses a parallel coordinates plot in combination with glyphs to investigate connections, type of network traffic and the timely sequence [20]. Many glyphs build patterns of different colored stripes over time.

Such a stripe combination is related to one of the designs investigated in our user study.

DESIGN SPACE FOR TEMPORAL GLYPHS

The design space for a basic temporal glyph can be characterized by the visual variables that are used to encode two attributes of temporal data: a) the position of a timepoint on the plane and b) the data value associated with this timepoint. Different visual variables can be used to encode these two attributes. In Table 1 we show some meaningful combinations of visual variables taken from Cleveland and McGill [10] for quantitative data and how they form different glyphs.

Ward [35] describes several categories of glyphs. To narrow down the design space for our experiment we only discuss temporal glyphs with many-to-one mappings where several or all data attributes map to a common type of graphical attribute. This is important in order not to promote certain temporal dimensions and to enable easier intra-record and inter-record comparison, which is fundamental for many tasks involving time series, including the ones chosen for our experiment. While many more different glyph types exist, such as face glyphs, arrows/weathervanes, box glyphs, sticks and trees etc., we focus on two main types of glyphs here: profiles and stars (see [35]). Both types have the advantage that relationships between adjacent data points are easier to see than for other glyphs [35]. While it is theoretically possible to encode temporal position using other visual variables such as length, direction, area, volume, curvature, or shading, no glyph design using these encodings has established itself in practice and is, thus, part of our study.

EXPERIMENT DESIGN

The purpose of our experiment was to compare the performance of different, potentially powerful, temporal glyphs in a small multiple setting. Our three tasks are inspired from our work with network analysts but generalize to other domains in which temporal data has to be compared and analyzed.

Experiment Factors

Our experimental factors were *glyph*, *task*, and *data density*.

Glyphs

We chose the Line Glyph (LIN), Stripe Glyph (STR), Clock Glyph (CLO), and the Star Glyph (STA) for their different characteristics and to assess their performance in a small multiple setting. LIN was chosen as one of the best ranked and most commonly used glyphs in our space and STR for its similar temporal but different value encoding. Glyphs are often designed to encode intuitive pairings of data to visual variables [35] and, thus, we chose two circular designs that take people's potentially intuitive notion of time encoded in a clock-like fashion into account. We chose to test STA for its similar value encoding to LIN and CLO for its similar value encoding to STR.

The Dot Plot was excluded as in our experience the single dots became too small, making it nearly impossible to spot them. The Bar Chart was excluded as well because Cleveland and McGill [10] conjecture that even for values encoded in bar charts the primary elementary task is judging position

Glyph	Temporal Enc.	Data Value Enc. (ranked)	Data Density Issues
Dot Plot	Position CS	Position CS (1)	Small dots difficult to see for small glyphs
Line Glyph	Position CS	Position CS (1)/Direction (3)	May become very dense
Bar Chart	Position CS	Position CS (1)/Length (3)/ Area (4)	May become very dense
Star Glyph	Angle	Length (3)	Small angular differences are hard to distinguish
Stripe Glyph	Position CS	Color Saturation (6)	Color blending for small areas
Clock Glyph	Angle	Color Saturation (6)	Color blending

Table 1. Partial overview of the design space for temporal glyphs. We show combinations of the encodings for quantitative data (cf. Cleveland and McGill's [10]) ranked according to their study results: 1) Position CS, 2) Position NAS, 3) Length/Direction/Angle, 4) Area, 5) Volume/Curvature, 6) Shading/Color Saturation. Other combinations are certainly possible. Position CS = position along a common scale, Position NAS = position along non-aligned scale. Glyph designs written with bold characters are the ones used in our experiment.

along a common scale but that judgements of area and length may also play a role. Therefore, we cannot safely test, which visual variable affects the perception of the data value.

When comparing glyphs visually, the distance between the representations matters. We chose to keep the distance for the different designs identical and, therefore, to have the same uniform small multiple layout. As a consequence it was important to set a fixed aspect ratio for each glyph. To maximize display space for circular glyphs for a fairer comparison we chose a square aspect ratio for each glyph.

For the color encoded glyphs (CLO and STR) we chose a heatmap colorscale, which was motivated by the yellow to red colorscale from ColorBrewer [7]. This scale takes advantage of the fact that the human visual system has maximum sensitivity to luminance changes for the orange-yellow hue [23] and it is also suitable for color blind people.

For each trial, the same type of glyph—but showing different data—was drawn on the screen in a small multiple layout of $8 \times 6 = 48$ glyphs in total (Figure 1). Each glyph was drawn at a resolution of 96×96 pixel.

Tasks
Many different tasks exist that can be performed on time-oriented data [2, 3, 24]. We chose our tasks taking two criteria into account: (1) their ecological validity, i.e. how commonly they are performed in environments where the quick comparison of multiple time series is needed. (2) their heterogeneity in terms of the elementary perceptual tasks, i.e. we picked tasks that involve the comparison of visual variables for encoding data values, investigating different layouts for time and the combination of the two. In terms of ecological validity our tasks were inspired by our work with network security analysts from a large university computer center who had to monitor large amounts of network devices. The ana-

lysts had to be able to efficiently detect anomalous traffic patterns (e.g., peak values in none working hours) to be able to quickly react on the possible threat. Our three tasks were:

Task 1—Peak Detection: Amongst all small multiple glyphs, participants had to select the glyph that contained the highest data value (Figure 1). This task, thus, involved scanning all glyphs for its highest value and comparing across glyphs using length (LIN, STA) or saturation (STR, CLO) judgements.

Task 2—Temporal Location: Among all small multiples, participants were asked to select the glyph with the highest value at a predefined time-point. This time-point was textually shown to the participant in advance (e.g. "3am"). This task, thus, involved first identifying the location of a time-point by making positional (LIN, STR) or angular judgements (STA, CLO) and then comparing the peaks as in Task 1.

Task 3—Trend Detection: Among all small multiples, participants had to select the glyph with the highest value decrease over the whole displayed time period (Figure 2). This task, thus, involved first detecting all decreasing trends and then comparing the first and the last value.

Data Density
In order to test the scalability of each glyph in terms of the number of datapoints it can encode, we tested two data densities. The smaller density consisted of 24 data values (1 for each hour), and the larger of 96 data values (1 for each 15 minutes). The rendered size of the glyphs holding these data points was not varied between each density (Figure 3).

Hypotheses
We previously conducted two exploratory pilot studies with similar glyphs and tasks. From these and the related literature [10, 35] we derive the following hypotheses:
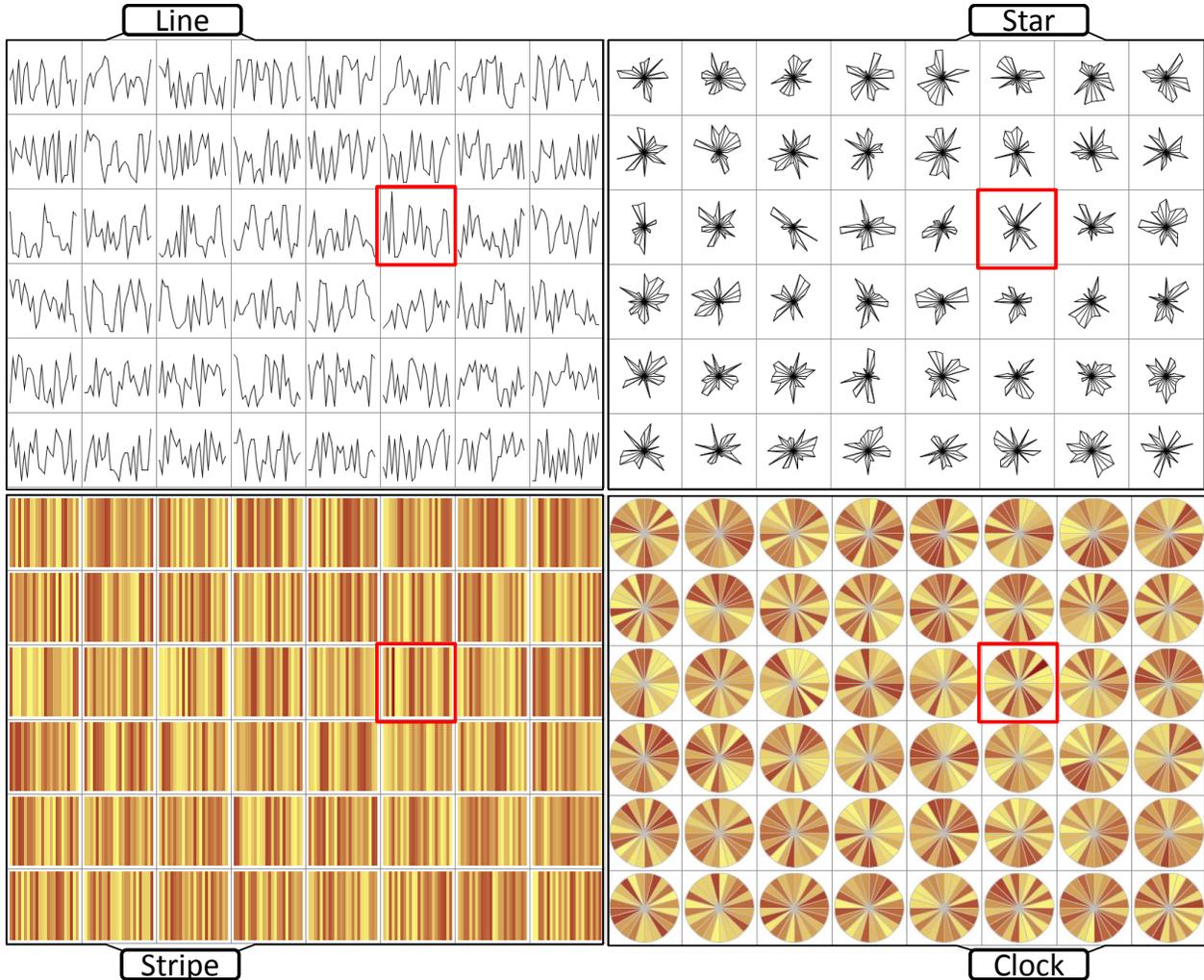
Figure 1. Peak detection: Illustration of the different glyphs with one high data value at a random point in time. For a better understanding the correct glyph is artificially highlighted.
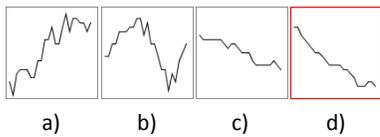

Figure 2. Trend detection: The four glyphs demonstrate different kinds of trends. From left to right: (a) visualizes a positive trend; (b) contains a positive and negative value development but for the whole displayed time interval there is no clear trend visible; (c and d) picture a negative trend over the whole displayed time period with (d) having the higher decrease. The glyph with the highest decrease over the whole displayed time period is artificially highlighted.

Figure 3. Differences between the two datasets for each glyph design.

H1: *For tasks involving primarily a value judgement LIN & STA (position/length encodings) are more accurate and efficient than CLO & STR (color encodings). This effect is strongest for LIN*. This hypothesis is based on Cleveland and McGill's experiments [10] on the perception of position, length, and color. We expect the results to hold for both data densities.

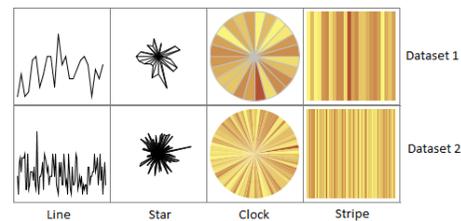**H2:** *For tasks involving primarily a value judgement, CLO & STR (color enc.) are more impacted by higher data density than LIN & STA (position/length enc.).* Color perception may change drastically with varying context colors and size of the object being viewed [30, 36]. We expect color perception to be more impacted than visual acuity on dense line and position encodings.

H3: *When detecting temporal positions, STA & CLO (angular enc.) outperform LIN & STR (position enc.).* Using the familiar clock metaphor, we expect that circular glyphs allow the perception of specific points in time to

be more accurate. This effect is stronger for **CLO** than **STA** as the clock shape is more clearly retained.

H4: *When detecting temporal positions, increasing data density will negatively impact performance with each glyph.*. This is because color judgements are impacted by the size of the object being viewed [30] and angular as well as positional judgements by visual acuity. We expect **CLO & STA** to perform best as they spread out values towards the circumference of the circle giving additional space for perceiving color and position.

H5: *For trend detection, LIN & STA (position and length enc.) are most effective.* In trend detection, two mental sub-tasks have to be integrated by the participant: a) analysis of data development over time (characterized by the slope) and, b) comparison of the first and last data value (trend steepness). We expect the first sub-task to be performed equally well with all glyphs but expect that the comparison of distances between two data values is more difficult with color compared to position/length.

H6: *For trend detection tasks, the participants' performance for each design is not influenced by data density.* For detecting a trend comparing the overall shape rather than single data values is necessary. We expect that increasing the data density will not influence the trend shape and, thus, has no effect on task performance.

Experiment Design

We used a mixed repeated-measures design with the between-subjects variable *task* and the within-subjects independent variables *glyph* and *data density*. The dependent variables were *error*, *time* and *confidence*. Each participant conducted one task with all four glyphs, two densities, and four trial repetitions.

Data

To control the data values and their resulting visual representations, we created synthetic data for the experiment. In total, we created 48 data instances (glyphs) for each repetition, task, and data density. The data was created such that just one glyph represented the correct answer. The glyphs with smaller density held 24, the ones with large density 96 data values. In previous pilot experiments these two values were established as being sufficiently different from one another. Data for each task was created as follows:

Task 1: Each glyph was filled with random noise to a threshold of 80% of its value range according to our experience from pilot studies. For the target glyph a peak value at 100% of the value range was added to the dataset at a random point in time.

Task 2: Each glyph was filled with random noise as in Task 1. A peak value at 100% of the value range was added to the target glyph at a predefined point in time. For the distractor glyphs, peak values of the same value were integrated but at wrong temporal positions.

Task 3: We designed different decreasing trends by varying the values of the first (0–25% of value range) and last data point (75–100% of value range). The target trend decreased 75% of the value range from first to last data value while the distractor glyphs included a decrease of 55%. Along the trend line each data point was varied by zero, one, or two values using a probabilistic function.

Participants

We recruited 24 participants (12 male, 12 female) mainly from the local student population. All participants had normal or corrected-to-normal vision and did not report color blindness. Their age ranged from 19–56 years (median age 24). Each participant had at least finished high school, eight held a Bachelor's, two a Master's degree, and one a Ph. D. The academic background of the participants was quite diverse with no one having a computer science background. 34% of the participants reported to use the computer for more than 30 hours per week and 50% less than 20 hours.

Procedure

The experiment took place in a quiet closed room at our university. In addition to the study participant, the experimenter was the only person present. The participant sat in front of a table at a distance of approx. 50cm from a 24in screen set to a resolution of 1920 × 1200. Participants interacted with the study software using only a mouse.

The experimenter began by explaining the data, the single task, and the design of the different glyphs. The data was presented as financial stock data to provide context.

Only when the participant was familiar with the current glyph design and task, he/she was allowed to proceed. For each glyph and density tested, the participant stepped through four practice trials followed by the four actual study trials. After each trial, the participant entered a confidence score for their answer on a 5-step Likert scale.

The task question was visible on the screen at all times. The presentation order of each glyph was randomized in a Latin square fashion between participants. The glyphs were presented in a 6×8 matrix layout (Figure 1). Each participant saw the same glyphs per trial in different random configurations.

RESULTS

We report on significant results ($p < .05$) from our quantitative analysis (Figure 4) in this section and refer to the qualitative feedback in the discussion section afterwards.

Data Analysis

Task completion time, error rate, and confidence score were recorded for the analysis. We used a repeated-measures ANOVA for the analysis of completion time. Time in our experiment was log-transformed where it did not follow a normal distribution. For the error rate as well as for the confidence score, a non-parametric Friedman's test was used.

Except for the second task we did not observe a strong learning effect between trials. Therefore, we analyzed all four trials for the first and third task, glyph and dataset for each participant. For the second task we analyzed the results of the last three trials. In addition, single answers were marked as outliers when each metric (time, error) was beyond two standard deviations from the mean for a given task and glyph per participant. Outliers were replaced with the closest value
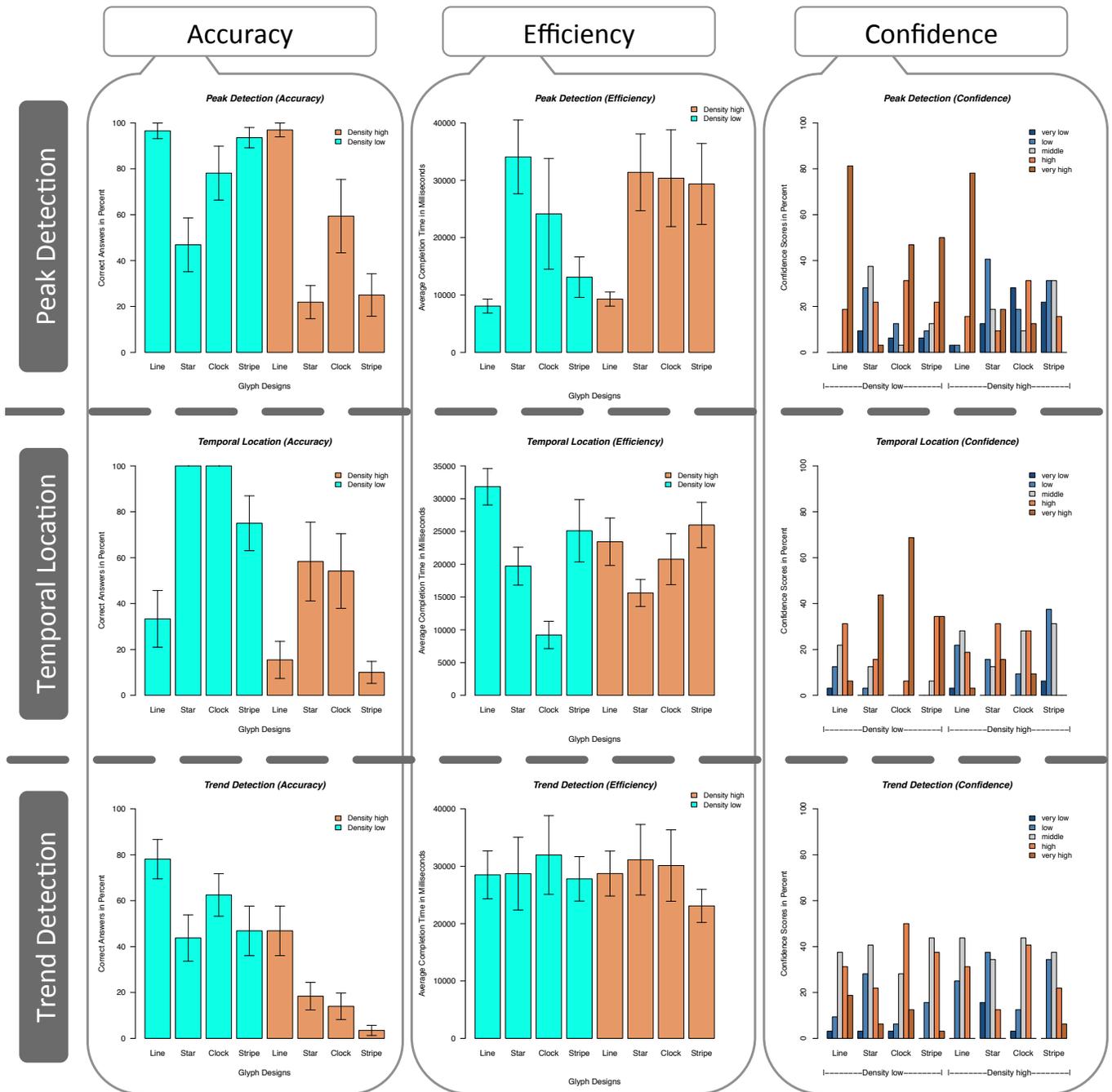
Figure 4. Bar charts with mean and standard deviation showing the results for each task and factor. The x-axis represents the different dependent variables. The y-axis illustrates the different tasks.

two standard deviations from the mean for each participant according to standard procedure. The tasks used in the study differed in their characteristics, so we analyzed the results of each task and dataset independently. Finally, we analyzed the feedback and subjective preference from the post-session interview for a qualitative analysis.

Task 1: Peak Detection

Task 1 consisted of four training repetitions and 2 densities × 4 repetitions with an increasing difficulty for each repeti-

tion block. This setting was used for each glyph design. For the analysis we only considered the more difficult repetition block since the results reveal more interesting insights.

Accuracy

There was a significant effect of *glyph* on *error* for both the low density ($\chi^2(3, N = 32) = 11.62, p < .01$) and the high density condition ($\chi^2(3, N = 32) = 17.59, p < .001$). In the low density condition pair-wise comparisons showed that errors in judgement were significantly worse for STA

(46.9%) than all other designs ($p < .05$). LIN (96.5%) and STR (93.6%) both showed high accuracy with LIN nearly at 100% accuracy. In the high density condition LIN (96.9%) significantly outperformed the other designs by staying at nearly 100% accuracy (all $p < .05$). In addition, CLO (59.4%) performed significantly better than STR (25%) and STA (21.9%) with $p < .01$ in each case. With an increasing data density, STR (from 93.6% to 25%) and STA (from 46.9% to 21.9%) significantly lost accuracy (all $p < .05$).

Efficiency
There was an overall effect of *glyph* on *time* in the low density ($F_{3,21} = 12.1, p < .0001$) and the high density ($F_{3,21} = 11.5, p < .001$) condition. Post-hoc comparisons showed that completion time was significantly higher for STA (34.1 sec.) compared to STR (13.1 sec) and LIN (8 sec.) for the low densities (all $p < .01$). For the higher densities LIN had the fastest completion time (9.3 sec.) compared to the other designs (nearly 30s per repetition on average) ($p < .05$). There was also a significant effect of *glyph* across densities ($F_{3,21} = 4.7, p < .05$). From low to high densities STR (from 13.1 sec. to 29.4 sec.) and CLO (from 24.1 sec. to 30.4 sec.) worsened ($p < .05$), whereas the mean for LIN stayed relatively stable (from 8 sec. to 9.3 sec.).

Confidence
There was an overall effect of *glyph* on *confidence* for both the low density ($\chi^2(3, N = 32) = 15.47, p < .01$) and the high density ($\chi^2(3, N = 32) = 16.28, p < .001$) condition. In the low density condition participants using STA (56.3%) reported a significantly lower confidence score with their answers than for all other designs (all $p < .01$). LIN (96.3%) received the highest confidence with significantly better ratings compared to CLO (80%, $p < .05$) and STA (56%, $p < .001$). In the high density condition LIN (92.5%) is significantly better than the other designs ($p < .001$) and STA (56.3%) better than STR (48.1%) ($p < .05$). From low to high densities STR (from 80% to 48.1%, $p < .05$) and CLO (from 80% to 56.3%, $p < .001$) worsened.

Task 2: Temporal Location
Task 2 consisted of four training repetitions and four real trials for both densities. After the initial training trials we asked participants to detect a different temporal location for the peak value. Therefore, the first real trial was discarded due to the mental recalibration necessary by the participants.

Accuracy
There was a significant effect of *glyph* on *error* for both the low density ($\chi^2(3, N = 32) = 17, p < .001$) and the high density condition ($\chi^2(3, N = 32) = 7.81, p = .05$). In the low density condition pair-wise comparisons showed that errors in judgement were significantly worse for LIN (33.3%) compared to CLO (100%) and STA (100%) (both $p < 0.01$) and STR (75%) compared to CLO (100%) and STA (100%) (both $p < 0.001$). In the high density condition STA (58.3%) significantly outperformed LIN (15.5%) and STR (10%) (both $p < 0.05$). With an increasing data density, STA (from 100% to 58.3%), CLO (from 100% to 54.2%)

and STR (from 75% to 10%) significantly lost accuracy with $p < .05$ in each case.

Efficiency
For the completion time there was only an overall effect of *glyph* on *time* in the low density ($F_{3,21} = 9.1, p < .001$) condition. Post-hoc comparisons showed that CLO (9.2 sec.) significantly outperformed LIN (31.8 sec.) ($p < .01$). There was another significant effect of *glyph* across densities ($F_{3,21} = 5.45, p < .01$). From low to high densities CLO (from 9.2 sec. to 20.8 sec.) deteriorated significantly ($p < .05$).

Confidence
There was an overall effect of *glyph* on *confidence* for both the low density ($\chi^2(3, N = 32) = 13.78, p < .01$) and the high density ($\chi^2(3, N = 32) = 12.12, p < .01$) condition. For the low density condition the results showed a clear picture for the confidence of the participants. The users were significantly more confident when using CLO (73.8%, $p < .05$), and had least confidence with LIN (50%, $p < .05$). For the high density condition the subjects were nearly equally confident using CLO (52.5%) or STA (54.4%), whereas LIN (44.4%, $p < 0.05$) and STR (35%, $p < 0.001$) are ranked worst. From low to high densities STA (from 65.6% to 54.4%, $p < .05$), CLO (from 73.8% to 52.5%, $p < .001$) and STR (from 65.6% to 35%, $p < .001$) worsened.

Task 3: Trend Detection
Task 3 consisted of four training repetitions and four real trials for both densities. For the analysis we discarded the training repetitions and focus only on the real trials.

Accuracy
There was a significant effect of *glyph* on *error* for both the low density ($\chi^2(3, N = 32) = 7.43, p = .05$) and the high density condition ($\chi^2(3, N = 32) = 8.9, p < .05$). In the low density condition pair-wise comparisons showed that errors in judgement were significantly better for LIN (78.1%) compared to STA (43.8%) and STR (46.9%) ($p < .05$). In the high density condition LIN (46.9%) significantly outperformed CLO (14%, $p < .05$) and STR (3.5%, $p < .01$). With an increasing data density, LIN (from 78.1% to 46.9%, $p < .05$), CLO (from 62.5% to 14%, $p < .01$) and STR (from 46.9% to 3.5%, $p < .05$) significantly lost accuracy (all $p < .05$).

Efficiency
For both densities no significant differences can be shown. The participants needed around 30 seconds on average. This was expected to be the maximal amount of time per repetition.

Confidence
There was an overall effect of *glyph* on *confidence* for both the low density ($\chi^2(3, N = 32) = 8.06, p < .05$) and the high density ($\chi^2(3, N = 32) = 7.6, p = .05$) condition. For the low density condition STA (60%) had lower ratings compared to CLO (72.5%, $p < 0.01$) and LIN (70.6%, $p < 0.05$). Same is true for the high density as well with STA (48.8%) being worse compared to CLO (64.4%, $p < 0.01$) and LIN (61.3%, $p < 0.05$). With an increased data density

STA (from 60% to 48.8%, $p < 0.01$) and CLO (from 72.5% to 64.4%, $p < 0.01$) lost significantly confidence.

DISCUSSION

In this section we combine both quantitative and qualitative data collected in our study to explain the varying performance of the different glyph designs according to our hypotheses. An overview of the quantitative results for each task is given in Table 2 where values highlighted in orange signify the best result compared to the other designs.

Task	Measure	LIN	STA	CLO	STR
Peak Detection (value comparison)	accuracy	96%	34%	69%	60%
	efficiency	8s	28.2s	18.6s	16.9s
Peak Detection (time comparison)	accuracy	24%	79%	77%	43%
	efficiency	27.6s	17.7s	15s	25.5s
Trend Detection	accuracy	63%	31%	39%	25%
	efficiency	26.2s	25.5s	27.1s	23.7s

Table 2. Glyph performance for different tasks: This table illustrates the percentage of correct answers (accuracy) and the average time needed (efficiency) for each of the tasks for both densities combined. The orange background signifies the best result compared to the other designs.

Peak Detection

In H1 we conjectured that LIN & STA would outperform CLO & STR due to their position and length encodings for value. The analysis of *error*, however, revealed that nearly no mistakes were made with LIN and only few with STR and that STA had the lowest accuracy followed by CLO. Apparently, the participants had more problems reading value with the circular layouts. This becomes obvious by comparing the most with the least accurate glyph design (i.e., LIN with STA). Both use the same value encoding but differ in the layout of the time dimension. This effect did not change across the two density conditions. STA and STR had a similarly high error rate across densities, CLO deteriorated only slightly, whereas LIN still performed best.

We can, thus, only partially confirm H1. We conclude that polar coordinates must have an effect on *error* for value judgements when the value is encoded with length. The same effect seems not to take place when the value is encoded with color. This can perhaps be explained by the different baselines of the designs. Comparing position/length in a radial design perhaps involves mental rotation to transfer the overall design to a comparable linear layout. This is not true for color encodings, since color does not need an identical baseline.

Another notable effect is the one between CLO and STR: while accuracy was not significantly different for low data density, CLO outperformed STR with high data density. This suggests that CLO is more resilient with respect to data density than STR. We believe this to be due to the fact that the slices in the circular design get more space near the circumference, wheras the slices in the stripe get too small, making the comparison more difficult. This only partially confirms H2: while STR is strongly affected by data density, LIN and CLO are either not affected by data density or affected to a smaller extent (decrease CLO: 18.8%; decrease STR: 68.7%).

The confidence score of the participants for this task was unambiguous with LIN having the highest ratings. In the final interview the participants had to rank the different glyph designs according to their subjective preference. LIN was the most preferred glyph type which matches the performance results of the quantitative analysis.

In the post-session interview, some participants argued that color was better than position/length for data value comparison especially when the distance between the values was very large. Of course, this depends on the color scale used, but seems plausible when the color value is entirely different, which may lead to a preattentive recognition effect. With smaller distances most of the participants commented that they would prefer the position/length encoding. When explaining their performance with STA (i.e. angle/length encoding), participants argued that they had problems comparing lengths with different orientation which further supports our hypothesis that mental rotations may be necessary for comparison and make values harder to compare in these glyphs. Especially in a small multiple setting this is an interesting finding and has to be further tested and considered when arranging glyphs.

Temporal Location

Our results partially support H3. In terms of accuracy both polar designs (CLO and STR) outperformed the linear designs when data density was low. To find an explanation for this result, we looked at the selections made by our participants and discovered an interesting side effect. The data sets corresponding to these wrongly answered questions were enriched with distractors very similar to the correct data instances by showing the same high value but at a different point in time. Participants seemed less likely to select such distractors when using the circular layouts for the time dimension. Participants were significantly more confident and made significantly less mistakes with the polar designs. The participants also reported to like the clock metaphor. Some suggested, however, to visualize only 12 hours at a time for a more intuitive encoding.

When data density was high we observed the same trend, even though only STA showed significant differences with respect to STR and LIN. The good performance of STA can be explained with the combination of the encodings. The length encoding for the data values makes it possible to easily spot the highest value even with lots of datapoints. With the color encodings, participants had problems spotting the peak value. The circular layout performed better than the linear one and worked for estimating the correct point in time.

We saw almost no significant differences between the designs for efficiency (only CLO was better than LIN with low data density and STA better than STR with high data density). Nonetheless, we observed that the overall trend for efficiency did not contradict the trend we found in terms of accuracy.

A significant decrease in performance between the two data densities can only be seen for accuracy. All designs had an increased error rate except for LIN. However, LIN's accuracy had been very low for the low density, thus, a significant de-

crease was nearly not possible. In terms of efficiency only CLO has a higher completion time, whereas, the other designs remained stable. These investigations partially support our hypothesis H4 where we had conjectured that the performance for detecting temporal positions would drop for an increased data density.

Trend Detection
In H5 we had conjectured that LIN & STA would be most effective for this task with the required value judgement as the bottleneck of the two required subtasks. As we expected, in terms of accuracy, the participants performed best using LIN independent from the data density. There was no significant difference between STA, CLO and STR on *error* and no significant results for *time* and, thus, H5 can only be partially confirmed. Independent from the designs, the participants needed around 30 seconds to complete the task.

With an increased data density the accuracy of LIN, CLO and STR dropped significantly. The completion time remained stable with no changes between the two density conditions. Our hypothesis H6 stating that the performance will not change by increasing the data density can, therefore, not be confirmed. Interestingly, participants commented that subjectively the task difficulty was not impacted by higher data density. The qualitative feedback almost matched the quantitative results. Nearly all participants reported to prefer LIN (i.e., position/length encoding) for solving the task.

DESIGN CONSIDERATIONS
With the results gained from the analysis and discussions we derive the following design considerations.

- **To improve value comparison, use a linear layout or switch to color encoding for value:**
  As can be seen in the results for the first and third task, LIN and STA's performance are quite diverse although the value encoding is similar. The polar design has a strong effect on the perception of the position/length encoding.

- **For value encoding, position/length encodings should be preferred to a color encoding:**
  As can be seen in the results gained from Task 1 and 3 where a value comparison was necessary, LIN performs best. Even with an increased data density values could still be compared.

- **Triangular shapes rather may be better than rectangular shapes for color encoding:**
  The slices used in CLO for encoding single data values form a triangular shape because of the circular layout. As can be seen in the results for CLO compared to STR, having more space near the circumference increased participants' performance. Designers could experiment with adding triangular shapes in a linear encoding.

- **Color encodings for higher data densities should be used with caution:**
  The results from task 1 and 3 illustrate, that the performance of the color encoded designs (CLO and STR) depends on the data density. Having a higher data density leads to a decreased performance.

- **Circular layouts rather than linear ones should be preferred for detecting temporal locations:**
  Polar designs are better for detecting specific points in time. This guideline results from the analysis of the second task. Participants performed significantly better using CLO and STA compared to LIN and STR. The clock metaphor increases users' chronological orientation.

- **For time-dependent tasks, sufficient space should be assigned to the designs:**
  Whereas, for solely value comparison tasks the performance of the best design (LIN) is not affected, the accuracy for tasks including temporal information decreases. This is independent from the combination of visual variables used as can be seen for task 2 (STA and CLO) and 3 (LIN). The designs performing best for these tasks are encoded differently but still show the same behavior.

LIMITATIONS
As stated at the beginning, we were inspired by time series data for a daily monitoring task. Especially CLO and STA with their 24 hour clock metaphor profit from this data arrangement. The performance may change with different lengths of time series.

The same is true for the aspect ratio and the size of the single glyphs. The aspect ratio was chosen in order not to greatly disadvantage the circular designs in terms of display space used. However, especially STR would profit from an aspect ratio with more horizontal space. With varying sizes of glyphs, the performance of the designs could change. In our setting we used the minimal space possible to be able to assign one pixel to one data value for the higher data density.

CONCLUSIONS
In this paper, we conducted a controlled experiment with 24 participants to assess the performance of time series visualizations when shrinking their size to glyph representations. In particular, we quantitatively measured accuracy and efficiency, and qualitatively surveyed user confidence and preferences for four glyph types based on three tasks important to our domain experts: peak detection, peak detection at a certain point in time, and trend detection. The four glyphs: Line Glyph, Stripe Glyph, Clock Glyph and Star Glyph were chosen for their varying use of visual variables to encode temporal position and the quantitative value of a data value. The results show that depending on tasks and data density, the chosen glyphs performed differently. We show that the Line Glyph is generally a good choice for peak and trend detection tasks but that radial encodings of time (Star Glyph and Clock Glyph) were more effective when one had to find a particular temporal location. Participants' subjective preferences support these findings. Thus, our study shows that both accuracy and efficiency of tasks such as ours can be boosted when carefully choosing the most appropriate design.

In the future we plan to expand upon this work in two ways: First, we want to test the effect of different small multiple layout techniques for our glyphs (e. g., on a map). Second, it would be interesting to test alternative glyph designs that

cover a larger variety of visual variables for the value encoding in an identical controlled experiment on time series. This would allow us a more general judgement about the applicability of Cleveland and McGill's ranking of visual variables [10] with respect to glyph design. With our current study we complement the research in the field of glyph evaluation by comparing the performance of four temporal glyphs for two peak detection and one trend detection task and provide a first set of design considerations for practitioners.

ACKNOWLEDGMENTS

REFERENCES

1. Aigner, W., Kainz, C., Ma, R., and Miksch, S. Bertin was right: An empirical evaluation of indexing to compare multivariate time-series data using line plots. *Computer Graphics Forum 30*, 1 (2011), 215–228.

2. Aigner, W., Miksch, S., Schumann, H., and Tominski, C. *Visualization of time-oriented data*. Springer-Verlag, 2011.

3. Andrienko, N., and Andrienko, G. *Exploratory analysis of spatial and temporal data*. Springer Berlin,, Germany, 2006.

4. Ankerst, M., Keim, D. A., and Kriegel, H.-P. Recursive pattern: A technique for visualizing very large amounts of data. In *Proc. Visualization (VIS)*, IEEE (1995), 279–286.

5. Ankerst, M., Keim, D. A., and Kriegel, H.-P. Circle segments: A technique for visually exploring large multidimensional data sets. In *Hot Topic Session of Visualization (VIS)*, IEEE (1996).

6. Bederson, B. B., Clamage, A., Czerwinski, M. P., and Robertson, G. G. Datelens: A fisheye calendar interface for pdas. *ACM Trans. Computer-Human Interaction 11*, 1 (2004), 90–119.

7. Brewer, C. A. Colorbrewer—color advice for maps. Accessed online September, 2012, **http://www.colorbrewer.org/**.

8. Carlis, J., and Konstan, J. Interactive visualization of serial periodic data. In *Proc. Symposium on User Interface Software and Technology (UIST)*, ACM (1998), 29–38.

9. Clark, W., Polakov, W., and Trabold, F. *The Gantt chart: A working tool of management*. The Ronald Press Company, 1922.

10. Cleveland, W., and McGill, R. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association* (1984), 531–554.

11. Daassi, C., Dumas, M., Fauvet, M., Nigay, L., and Scholl, P. Visual exploration of temporal object databases. In *Proc. Bases de Données Avancées (BDA)* (2000).

12. Fischer, F., Fuchs, J., and Mansmann, F. ClockMap: Enhancing circular treemaps with temporal glyphs for time-series data. In *Proc. EuroVis Short Papers*, Eurographics (2012), 97–101.

13. Guttorp, P., Sain, S., Wikle, C., Wickham, H., Hofmann, H., Wickham, C., and Cook, D. Glyph-maps for visually exploring temporal patterns in climate data and models. *Environmetrics 23*, 5 (2012), 382–393.

14. Havre, S., Hetzler, B., and Nowell, L. Themeriver: Visualizing theme changes over time. In *Proc. Information Visualization (InfoVis)*, IEEE (2000), 115–123.

15. Heer, J., Kong, N., and Agrawala, M. Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In *Proc. Human Factorsin Computing Systems (CHI)*, ACM (2009), 1303–1312.

16. Javed, W., McDonnel, B., and Elmqvist, N. Graphical perception of multiple time series. *Trans. Visualization and Computer Graphics 16*, 6 (2010), 927–934.

17. Keim, D. A. Designing pixel-oriented visualization techniques: Theory and applications. *Trans. Visualization and Computer Graphics 6*, 1 (2000), 59–78.

18. Kintzel, C., Fuchs, J., and Mansmann, F. Monitoring large ip spaces with clockview. In *Proc. Visualization for Cyber Security (VizSec)*, ACM (2011).

19. Kraak, M. The space-time cube revisited from a geovisualization perspective. In *Proc. International Cartographic Conference (ICC)* (2003), 1988–1996.

20. Krasser, S., Conti, G., Grizzard, J., Gribschaw, J., and Owen, H. Real-time and forensic network data analysis using animated and coordinated visualization. In *Proc. Workshop on Information Assurance and Security*, IEEE (2005), 42–49.

21. Krstajic, M., Bertini, E., and Keim, D. CloudLines: compact display of event episodes in multiple time-series. *Trans. Visualization and Computer Graphics 17*, 12 (2011), 2432–2439.

22. Kumar, N., Lolla, N., Keogh, E., Lonardi, S., and Ratanamahatana, C. Time-series bitmaps: A practical visualization tool for working with large time series databases. In *Proc. Data Mining Conference*, SIAM (2005), 531–535.

23. Levkowitz, H., and Herman, G. Color scales for image data. *Computer Graphics and Applications, IEEE 12*, 1 (1992), 72–80.

24. MacEachren, A. *How maps work*. Guilford Press, 1995.

25. McLachlan, P., Munzner, T., Koutsofios, E., and North, S. LiveRAC: interactive visual exploration of system management time-series data. In *Proc. Human Factors in Computing Systems (CHI)*, ACM (2008), 1483–1492.

26. Pearlman, J., and Rheingans, P. Visualizing network security events using compound glyphs from a service-oriented perspective. In *VIZSEC*. Springer, 2007, 131–146.

27. Plaisant, C., Milash, B., Rose, A., Widoff, S., and Shneiderman, B. Lifelines: visualizing personal histories. In *Proc. Human Factors in Computing Systems*, ACM (1996), 221–227.

28. Playfair, W., and Corry, J. *The commercial and political atlas and statistical breviary*. 1786.

29. Saito, T., Miyamura, H., Yamamoto, M., Saito, H., Hoshiya, Y., and Kaseda, T. Two-tone pseudo coloring: Compact visualization for one-dimensional data. In *Proc. Information Visualization (InfoVis)*, IEEE (2005), 173–180.

30. Stone, M. In Color Perception, Size Matters. *IEEE Computer Graphics and Applications 32*, 2 (Mar./Apr. 2012), 8–13.

31. Talbot, J., Gerth, J., and Hanrahan, P. Arc length-based aspect ratio selection. *Trans. Visualization and Computer Graphics 17*, 12 (2011).

32. Tominski, C., Abello, J., and Schumann, H. Axes-based visualizations with radial layouts. In *Proc. Symposium on Applied Computing*, ACM (2004), 1242–1247.

33. Tufte, E. *Beautiful Evidence*. Graphics Press, 2006.

34. Van Wijk, J. J., and Van Selow, E. R. Cluster and calendar based visualization of time series data. In *Proc. Information Visualization (InfoVis)*, IEEE Computer Society (1999), 4–9.

35. Ward, M. Multivariate data glyphs: Principles and practice. *Handbook of Data Visualization* (2008), 179–198.

36. Ware, C. *Information Visualization: Perception for Design*, 2nd ed. Morgan Kaufmann, 2004.

37. Wattenberg, M., and Kriss, J. Designing for social data analysis. *Trans. Visualization and Computer Graphics 12*, 4 (2006), 549–557.