Comparing Similarity Perception in Time Series Visualizations

Anna Gogolou, Theophanis Tsandilas, Themis Palpanas, Anastasia Bezerianos

Comparing Similarity Perception in Time Series Visualizations

Anna Gogolou, Theophanis Tsandilas, Themis Palpanas, and Anastasia Bezerianos

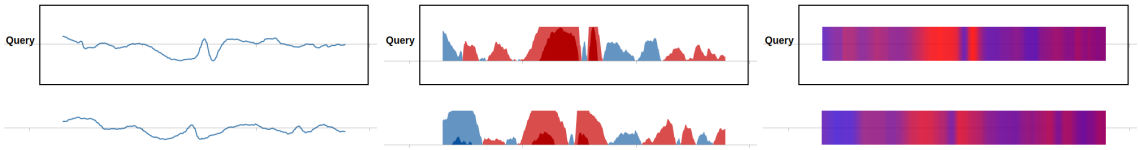


Fig. 1. Three time series visualizations compared in order to understand if we perceive similarity differently with each one (Line Chart left, Horizon Graph middle, Colorfield right). This example shows a query and one of the four possible answers participants had to choose from using each of the three visualizations. The answer here comes from an automatic similarity search algorithm (DTW).

Abstract— A common challenge faced by many domain experts working with time series data is how to identify and compare similar patterns. This operation is fundamental in high-level tasks, such as detecting recurring phenomena or creating clusters of similar temporal sequences. While automatic measures exist to compute time series similarity, human intervention is often required to visually inspect these automatically generated results. The visualization literature has examined similarity perception and its relation to automatic similarity measures for line charts, but has not yet considered if alternative visual representations, such as horizon graphs and colorfields, alter this perception. Motivated by how neuroscientists evaluate epileptiform patterns, we conducted two experiments that study how these three visualization techniques affect similarity perception in EEG signals. We seek to understand if the time series results returned from automatic similarity measures are perceived in a similar manner, irrespective of the visualization technique; and if what people perceive as similar with each visualization aligns with different automatic measures and their similarity constraints. Our findings indicate that horizon graphs align with similarity measures that allow local variations in temporal position or speed (i.e., dynamic time warping) more than the two other techniques. On the other hand, horizon graphs do not align with measures that are insensitive to amplitude and y-offset scaling (i.e., measures based on z-normalization), but the inverse seems to be the case for line charts and colorfields. Overall, our work indicates that the choice of visualization affects what temporal patterns we consider as similar, i.e., the notion of similarity in time series is not visualization independent.

Index Terms—Time series, similarity perception, automatic similarity search, line charts, horizon graphs, colorfields, evaluation.

1 Introduction

Time series are temporal sequences of data points that derive from measurements and recordings of a range of natural processes or human activities. A city's temperature per hour, a person's blood oxygen saturation per day, and an electroencephalography (EEG) signal are all examples of time series data. Large time series collections are becoming increasingly commonplace [46], and their analysis involves a diverse range of tasks, such as searching for pattern templates or anomalies, identifying reoccurring waveforms, or classifying time series subsequences into clusters of similar patterns, all of which involve the notion of similarity between time series. Data-mining research has developed a wide range of techniques to automate such tasks [23]. In many situations however, automated techniques fail to produce satisfactory results, thus experts rely on visual analytic tools to perform their tasks. For example, in EEG data, comparing time series to identify *epileptiform discharges* is difficult [35]. These temporal patterns take a variety of different forms that are very specific to individual patients, while very similar patterns appear in normal background activity. Although several techniques claim to automatically detect such patterns [32], medical experts still visually inspect the EEG data of their patients. This process is especially time consuming, as experts need to visually scan a large number of temporal signals recorded from multiple EEG sensors, find, and compare these patterns.

In such scenarios, the use of visualization techniques that accurately and effectively communicate similar patterns between time series becomes important. Times series are commonly represented as *line charts*, but a considerable amount of work in Information Visualization has examined alternative visual encodings, such as *horizon graphs* [29, 34, 47, 50, 53] and *colorfields* [2, 15, 45, 53, 59]. This literature has focused on elementary visual tasks that require estimation, e.g., estimation of averages, or point comparison and discrimination tasks. Visual pattern matching is a more complex task that requires the simultaneous comparison of a large number of features and likely incorporates many of these previously mentioned tasks. Thus, previous results say very little about how people access the similarity of two or more time series when using different time-series visualizations.

In this paper, we examine how line- and color-encoding techniques affect what time series we perceive as similar. Specifically, we present the results of two laboratory experiments that compare three representative techniques: (1) line charts, (2) horizon graphs, and (3) color fields. In addition to task performance, we assess the reliability (or subjectivity) of participants' answers and examine whether the above techniques penalize or favor similarity *invariances* [6, 16, 21] that are often required by certain application domains. For example, two patterns might be considered as similar, irrespective of their amplitudes (amplitude invariance) or their stretching along the time dimension (time-scale invariance). We want to understand whether the three visualizations exaggerate or deemphasize such deformations. To this end, we assess the perception of similarity between time series, with respect to representative similarity distance measures that are well known to be invariant to certain properties of a time series [6]. Our first experiment investigates *local-scale* (or *warping*) invariance by contrasting similarity perception with *euclidean distance (ED)* and *dynamic time warping (DTW)*. Our second experiment in turn investigates *amplitude* and *offset* invariance by contrasting similarity perception with and without *z-normalization*.

- *Anna Gogolou is with Inria, Univ. Paris-Sud, and Univ. Paris-Saclay, France. E-mail: anna.gogolou@inria.fr.*
- *Theophanis Tsandilas is with Inria, Univ. Paris-Sud & CNRS, and Univ. Paris-Saclay, France. E-mail: fanis@lri.fr.*
- *Themis Palpanas is with Univ. Paris-Descartes, France. E-mail: themis@mi.parisdescartes.fr.*
- *Anastasia Bezerianos is with Univ. Paris-Sud & CNRS, Inria, and Univ. Paris-Saclay, France. E-mail: anastasia.bezerianos@lri.fr.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

In contrast to previous studies, that used human sketched [21, 41] or artificially generated [16] query patterns, the queries in our experiments are extracted from annotated EEG data and express real patterns of interest. A major challenge is how to derive patterns that are representative of real data and tasks, but that also highlight the differences of the tested similarity measures. We address this challenge by selecting query patterns for which the different distance similarity measures produce clearly distinct answers. This enables us to assess whether similarity perception with each visual encoding technique is invariant to warping as well as to amplitude and offset deformations in the signals.

To summarize, this work is the first to investigate how humans perceive similarity between time series with both line- and color-encoding visualization techniques. Our results answer two major questions: (1) how easy or difficult it is to visually identify similar patterns with different visualization techniques; and (2) whether similarity perception with these techniques is invariant to representative signal deformations.

2 Related Work

We now discuss previous work on time series visualization, search, and perception, in particular with respect to their similarity.

2.1 Time Series Visualization

Since the first line charts were used by Lambert and by Playfair in the 18th century [63], several visualizations were introduced for time series (see [1, 43] for an overview of time oriented visualizations). The goal of these techniques vary, for example some communicate the periodical nature of data (e.g., [8, 66]), others aggregate multiple time series trough clustering (e.g., [64]), and yet others focus on examining how to interactively explore and compare a set of time series (e.g., [67, 68]).

One aspect that has received considerable attention is the scalability of time series visualizations. One of the oldest visualization approaches is to present *line charts* in small multiples [63] or sparklines [42]. More recent approaches extend the line chart representation itself. For example, the two-tone pseudo coloring and *horizon graphs* [50, 53] split the vertical range of values in a line chart into a few vertical bands, that are then colored and superimposed. This representation saves vertical space, while maintaining the overall line shape. Others address scalability using color-based representations, often referred to as heatmap or *colorfields*. Instead of using position to encode the range of values over time (as is done in line charts), these visualizations use vertical color strips, whose color saturation or brightness encodes value. This approach is seen in many systems [2, 15, 45, 53] and scales well as multiple such sequences of small height can be stacked together [37,59]. As remarked by Javed et al. [34], in order to represent multiple time series, the above representations split the space (mainly vertically) and attempt to optimize the vertical footprint of each individual time series.

Alternatively, multiple visualizations can occupy the same space [34]. Multiple line charts, often of different colors, can be superimposed or can be replaced by variations of area charts that attempt to optimize space (e.g., stacked [11] or braided [34] graphs). The majority of these space sharing techniques do not scale well for a large number of time series due to clutter. Moreover, the stacked variations, which do not have a common baseline, could complicate comparison tasks such as determining similarity. We focused on techniques that split space, as our motivating scenario for similarity search (see Sec. 3) revealed that it is important to be able to see a large number of time series together.

2.2 Studies on Time Series Perception

A number of perception studies have compared different time series visualizations under a variety of tasks, in particular visualizations that use positional or color encodings.

Correll et al. [15] investigated the efficiency of representations using either position (line charts) or color (colorfields) when estimating averages. They found that people are better at estimating high-level statistical overview tasks, such as averages, when using colorfields. Albers et al. [2] compared eight different time series visualisations that used both positional and color encodings (among other variations). They found that positional visualizations were more efficient for tasks requiring point comparisons (e.g., maxima), whereas color once again performed better for summary comparisons (e.g., ranges, averages).

Fuchs et al. [24] studied glyphs, presented in small multiples. Position/length and color were among the variations used for the different glyph designs. They did not test averaging tasks, but they found that for peak and trend detection tasks, line glyphs worked best.

For positional encodings, Heer et al. [29] compared line charts with variations of horizon graphs for a value comparison and estimation task. They focused mainly on the effects of chart size and layering and found that horizon graphs performed better than line charts for small chart sizes. Later, Perin et al. [47] improved the efficiency of horizon graphs by allowing an interactive adjustment of the band baseline. As discussed, Javed et al. [34] compared visualizations which split or share the same space, under peak, trend, and discrimination tasks. They found that while shared-space (superimposed) techniques worked well for small numbers of time series, split-space ones worked better for large numbers, and that horizon graphs were faster than line charts for discrimination tasks but slower for peak and trend detection.

Similarity search likely involves both point comparisons, such as finding maxima, and overview comparisons, such as comparing the overall shape of time lines. It is thus unclear if position or color-based visualizations are best suited for similarity tasks. In this work, we focus on three visualization techniques that rely on position (line charts), color (colorfields), or both (horizon graphs). These techniques can also scale well to multiple time series when presented as small multiples.

2.3 Time Series Similarity

Analysts often define a subsequence of interest as *query* and use automated tools to search for similar patterns. We discuss data-mining research on similarity search algorithms and then visualization research on how to specify similarly queries and evaluate results.

Similarity Algorithms. Data-mining research has proposed a plethora of algorithms (distance measures) that assess the distance between two time series. Ding et al. [18] group them in four categories. The simplest type are *Lock-step* measures, such as the Euclidean Distance (ED) [22], that perform point-by-point value comparison between two time series. ED can be combined with data normalization, often called *z-normalization* [26], which considers as similar patterns that may vary in amplitude and y-offset. Another commonly used group are *Elastic* measures, that allow horizontal "stretching" and/or "compression" of a time series when searching for similar ones. For example, Dynamic Time Warping (DTW) [7] accounts for similar sequences that vary in speed or are shifted temporally (temporal warping). Other categories are less common and include more specialized measures that are *threshold-based*, e.g., TQuEST [4], or *pattern-based*, e.g., SpADe [14].

To evaluate similarity measures, Ding et al. [18] performed a nearest neighbor classification (1NN) by using distances of nine different similarity algorithms and then compared their classification accuracy with respect to pre-labeled classes [13]. Based on their analysis, they concluded that there is no superior measure, as their classification accuracy depends on the dataset and its domain. Among their findings is that, on small datasets, DTW can be significantly more accurate than ED, but, as the size of the dataset increases their accuracies converge. In our work, we focus on DTW, ED and its variations because: (i) they are the most commonly used measures in the visualization and data-mining literature; (ii) they are efficient [16, 18]; and (iii) they are appropriate for our motivating domain (see Sec. 3).

Interactive Querying. There has been a growing interest in interactive exploration and querying of time series. Early examples express queries through visual filtering. For example, TimeSearcher [30] allows users to specify their queries through "time box" selections (rectangle regions). In Querylines [52], users create line segments to define the filters for their queries. Later approaches focus on algorithmic similarity, for example through the automatic detection of specific "motifs", simple shapes such as spikes or sinks that users can combine to form queries [27]. Others [44] examine how to automatically extract a grammar to express time series approximately and simplify the search of matches to a sketched query, or they have focused on algorithmic

performance and scalability of similarity search [69, 70]. Recently, Qetch [41] presented a sketch-based querying system and a similarity algorithm that is scale independent. With few exceptions [41], these approaches have not been evaluated through user studies.

Another approach is to use similarity algorithms developed by the data-mining community. Buono et al. [10] enable users to interactively select part of an existing time series to form a query, that is then matched to possible results using ED. Others define query patterns through sketching [16, 31, 41, 54, 65]. Most sketch-based systems use ED [10, 31], but more recent work [16, 41, 54] has considered additional measures. All these approaches rely on line chart visual representations. While we do not study querying in this paper, this line of work motivates our research, as we want to understand how people assess similarity in the results of their queries.

Studies on Similarity Perception. Few studies have investigate subjective user evaluation of similarity results. TimeSketch [21] proposed a crowdsource procedure where crowdworkers ranked time series w.r.t. to their similarity to a small set of sketched queries. The goal was to produce a human-generated ranking and then compare it to the ranking of similarity algorithms. They found DTW to be the closest to human ranking, with ED performing worse or similarly, and SpADe performing badly for small queries. This procedure helps derive human-driven similarity measures and provides insights about how close they are to algorithmic measures, but it is unclear how it can apply to non-sketched queries. Mannino and Abouzied [41] compared their own matching algorithm with ED and DTW by using again simplified query patterns sketched by hand. Their studies showed that the results of their matching algorithm were ranked higher that those of DTW (and ED), but focused on a small set of sketched queries rather than a large set of real time series patterns as is our case. Correll and Gleicher [16] in turn examined whether similarity perception is *invariant* [6] to signal deformations. In particular, they examined how humans rated the similarity between a simplified pattern (the query) and a target that was the original query transformed in different ways. Their results indicated that most transformations did not decrease similarity and that no single algorithm could match human judgements. This work again used line chart visualizations, while we consider similarity across different visual representations. We explain finer differences to this study in Sec. 4.

3 Motivation

Our motivation stems from a real problem presented to us by a team of neuroscientists, experts in the analysis of EEG recordings for the diagnosis of epileptic events. Our experimental task is inspired by the user interfaces that such experts use to visually analyze EEG data. The pool of our experimental data was also provided directly by them.

In two 1h sessions we met with two and three neuroscientists respectively from the MEG/EEG Center of the ICM Brain and Spine Institute[1]. They are looking for tools to improve the detection of "epileptiform discharges". These are abnormal patterns that have been linked to various cognitive disruptions and reoccurrences of epileptic seizures [60]. They are often not isolated cases but may appear as periodic patterns [36], whose periodicity, may vary significantly from one patient to another.

Epileptiform discharges are events which are characterized by a spike of 20-70 milliseconds (ms) usually followed by a sharp wave lasting 70-200 ms [17, 56, 57]. As opposed to epileptic seizures that produce large disturbances in the EEG signal of a patient, epileptiform discharges are especially hard to detect. Although data-mining research has developed algorithms to automatically detect their patterns [32], according to our experts, such algorithms result in many false positives and are not useful in practice. Main reasons for this problem is that epileptiform discharges take a range of different forms and often resemble normal background activity due to regular artifacts such as pulses of the heart, the eyes, or the muscles [35]. In addition, their patterns vary greatly across patients so machine-learning approaches cannot help.

For these reasons, medical experts do not trust automated techniques and still visually scan the data to identify abnormal events, using tools such as the one depicted in Figure 2. This can be a very tedious and

[1]The ICM Brain and Spine Institute, `https://icm-institute.org/en/`

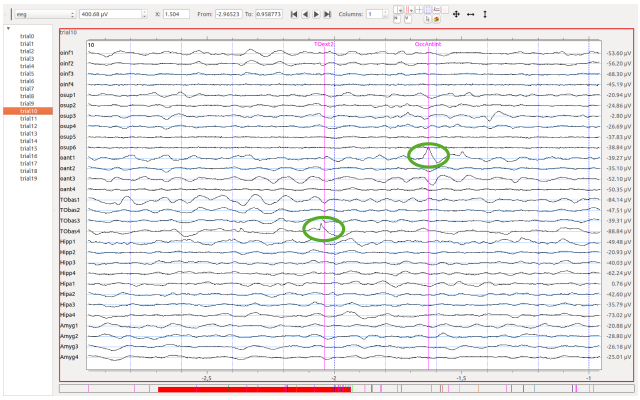


Fig. 2. The Muse tool[1] used by neuroscientists to visualize measurements from 295 electrodes and sensors placed on patients. Here the neuroscientist has restricted their view to 6 groups of sensors (30 in total) from one recording trial (trial 10). Purple lines indicate manual annotations of epileptiform discharges that neuroscientists have detected on different sensors. The particular discharges are highlighted in a green oval for illustration purposes only (these highlights are not part of the tool). The scroll bar in the bottom indicates what time frame of the series is currently visible, and is augmented with indications of where manual annotations exist (small colored line segments).

complex task. Experts need to visually inspect around 300 sensors and several thousand data points per sensor (see Sec. 4.2). And even when they find candidate events, they often need to consult additional resources (e.g., 3D representations of the location of the electrodes placed on the scalp) to make their decisions and annotate their data.

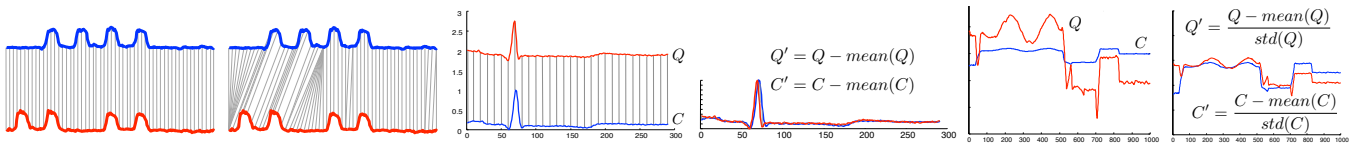
In an attempt to aid our users, we tried to understand if it would be possible for them to first manually identify a small number of epileptiform discharges and use them as patterns to automatically detect similar subsequences. The experts could then visually verify whether they are similar and decide if they are also potential discharges. To this end, we requested information about what types of variations or deformations in the patterns could indicate similar signals.

The experts were able to verbally describe roughly the signal they were looking for. They explained that the duration of spikes and waves can vary and are not consistent even for a single patient, thus stressed or compressed signals are of interest (invariant to time-warping). When asked, they also explained that the height of the pattern can vary across patients (invariant to amplitude). But they could not say to what extent the amplitude of the spikes and discharges is important, i.e., to what extent signals could be considered similar if they differed in amplitude. In some cases we got the response that a spike can be too small (i.e., in some cases amplitude may play a role) but this can only be determined by looking at the background noise - the parts of the signal before and after the spike. Or that to interpret a spike they needed access to views from other sensors. The importance of context in detecting such discharges is well documented [17, 56, 57]. These are all very subtle properties that need to be evaluated case by case, and in context, stressing further the need for human intervention.

As our experts explained, identifying these types of discharges requires a lot of experience, and some of their decisions remain subjective. Past work has shown that agreement even between different experts can be particularly low [35]. While this task relies on extensive experience and involves substantial domain knowledge, it still raises an interesting question. Do visualizations actually help viewers understand what temporal patterns are similar, or are there aspects of the invariances of interest that are not communicated well? We set out to investigate if different types of visualizations communicate or de-emphasize invariances in a similar way, or if visualizations need to be chosen appropriately.

4 Goals and Research Strategy

Given that users like neuroscientists rely on visualization tools to take decisions, understanding how a visualization may affect what time series are perceived as similar is important. The similarity criteria used by experts can be complex and highly uncertain, and the extent to which



(a) Euclidean Distance (ED) (b) DTW (c) Z-Normalization: y-offset shifting (d) Z-Normalization: amplitude scaling

Fig. 3. Overview of how the algorithms we used perform matching for similarity: (a) Euclidean Distance computes the $L2$ distance between all corresponding points of two time series of equal length. (b) DTW allows the matching of points between two time series, even if these points are not aligned on the time axis. (c-d) Z-normalization transforms a time series into a new series of the same length that has zero mean and standard deviation (std) one. It enables similarity search independent of y-offset and amplitude scaling. (Images courtesy of E. Keogh)

signal deformations satisfy such criteria often depends on thresholds that may vary from case to case. Thus, we are especially interested in knowing which visual encodings are sensitive to deformations of a time series signal and which of them are "invariant" to those deformations. Such knowledge can help us design tools that better match the invariances required by different application domains. It can also help us support users' tasks by proposing alternative visualizations, as different visualizations may emphasize (or deemphasize) the perception of different deformations in the signal.

4.1 Experimental Approach

As discussed in Sec. 2, previous work has studied deformation invariances from an algorithmic perspective. Batista et al. [6] enumerate several types of invariance: temporal warping, uniform scaling, amplitude and offset, phase, trend, complexity, etc. Correll and Gleicher [16] consider these types of invariances to design a sketch-based query system that is flexible enough to accommodate algorithms with different invariance characteristics. They then present the results of an experiment that investigates how sensitive or invariant similarity perception is with respect to different deformations when using line charts.

While inspired by this research, our goal is different. We are interested in how *different visualizations* affect similarity perception, thus we treat the visualization techniques as our primary experimental factor. Although we also seek to understand how different techniques support invariances, the way we control for invariances is different. In particular, our approach is based on the observation that signal deformations emerge naturally in real data, taking complex forms that cannot be easily reproduced with artificially created patterns. Thus, as opposed to Correll and Gleicher [16], we do not directly control signal deformations as experimental factors. In Correll and Gleicher's experiment, the patterns of interest take elementary forms (upward and downward lines, sine waves, Perlin noise, etc.) and are transformed uniformly along the time dimension. This approach allows for stricter control and simplifies the experimental design but does not capture the way people compare patterns in real data. For example, when determining if two time series are similar, a user may have to assess temporal stretches or vertical shifts that occur in small portions of the signal in combination with other deformations. In such scenarios, the perception of similarity is likely to rely on a mix of very subtle signal characteristics.

Given these considerations, we decided to use real data to generate our experimental tasks, based on the application domain and scenario that we described in the previous section. We also decided to concentrate on the invariances that are most relevant to these data.

4.2 Dataset

We used a real dataset provided to us by our collaborating neuroscientists (see Sec. 3). The dataset contains measurements from 295 electrodes and sensors placed on patients' scalps: Among them 151 signals come from Magneto-Encephalo-Graphy (MEG), 33 from Electro-Encephalo-Graphy (EEG), and 39 from intracranial Electro-Encephalo-Graphy (iEEG) sensors. Measurements last six seconds and are captured at a sampling rate of 1250 Hz. All our data comes from 154 such recordings of the same patient, that each contains 295 long time series - 1 per sensor, of 7500 data-points each ($\sim$ 341 million data points in total). We used this dataset to generate experimental trials.

To understand similarity, we need to compare time series with interesting temporal patterns. How to determine interesting patterns is a difficult problem. Synthetic patterns can lead to artificially looking results, while randomly selecting ones from a real data set may result in empty or noisy patterns. Eichmann and Zgraggen [21] addressed this problem by collecting sketched patterns by non-expert people. However, this approach is only appropriate for simplified human-created patterns that may capture the intricacies of real patterns in the data.

Our dataset allows for a better solution. Neuroscientists have manually annotated this dataset by adding markers at time points that correspond to potential interictal epileptiform discharges. Thus the dataset already contains real patterns of interest. We used the area around these annotated events as *potential queries* for our similarity search algorithms. The dataset contains a total of 205 annotations.

4.3 Controlling for Invariances

When considering time series to compare against the potential queries, we focus on ones that contain deformations that are important to our experts. They indicated (see Sec. 3) that patterns that are invariant to - i.e., allow for variations in (i) *time warping* and (ii) *amplitude* and *offset* are of interest. Time-warping invariance is important since EEG signals often vary in transient or rhythmic activity [40], e.g., they may include slow delta waves with frequencies lower than 4 Hz, as well as fast beta waves with frequencies greater than 13 Hz. Amplitude and offset invariance is important because experts are often interested in clustering spikes based on their shape independently of their vertical height or shift [51]. Other invariances, such as noise and trend, are usually unwanted. Medical experts preprocess their data by applying filters that remove noise or long additive trends in the signals. Finally, global invariances such as uniform scaling are less interesting, as they can be supported by global-scaling tools that are independent of visualization.

As we do not treat invariances as experimental factors, we do not directly vary their levels. However, we control them by using similarity algorithms that are well known to support them (see Figure 3). For time-warping invariance, we use Dynamic Time Warping (DTW) [7]. For amplitude and offset invariance, we use z-normalization [26]. Both algorithms are well established and widely used in the data-mining literature [6]. We do not consider Hough Transform [16], as it combines invariances of both DTW and z-normalization. We contrast the results of the above algorithms with the results of the simple Euclidean Distance (ED) by asking participants to choose between them. We note that in the experiment participants see the *original* time series and their values (not the deformed versions used by the similarity algorithms).

This approach shares similarities with that of Eichmann and Zgraggen [21], who compared how people rank the results of multiple algorithms that measure similarity. For many queries, however, similarity algorithms may return identical or similar results. To deal with this constraint, we developed an automated mechanism for selecting queries for which the algorithms produce distinct results. These cases are especially interesting because (i) they better capture the differences of the algorithms, and (ii) they represent the most difficult cases, for which careful visual inspection might be more critical. This approach also allows us to observe the effect of the underlying invariance assumptions more clearly within an experimental setting.

Another differentiation of our approach compared to previous studies is that we also measure how different participants agree on their assessments. Measuring agreement is important for assessing similarity perception as it enables us to evaluate the level of subjectivity and diversity in participants' answers in an objective way.

5 EXPERIMENTS

We conducted two experiments to study if using different time series visualizations, Line Charts (LC), Horizon Graphs (HG), and Colorfields (CF), changes whether time series are perceived as similar. And if invariances in the data effect this perception. Exp-1 investigated *time-warping* invariance by asking participants to compare the results of ED and DTW. Exp-2 investigated *amplitude* and *offset* invariance by asking participants to compare the results of ED with and without z-normalization. Aspects in the setup and procedure are common in both experiments, so we present them together unless explicitly stated.

5.1 Participants & Apparatus

A total of 36 volunteers, 23 to 42 years old (M = 29, SD = 5.6), participated in the two experiments without monetary compensation. We recruited from a local university mailing list 18 participants (seven women) for Exp-1 and 18 additional participants (three women) for Exp-2. Our participants came from different scientific backgrounds, including students and researchers in Computer Science, Electrical Engineering, Physics, and Finance. As our study is perceptual in nature, we opted for a general pool of participants rather than experts.

For both experiments, we used a 24" DELL monitor set to 1920 $\times$ 1080 resolution.The user interface was implemented with Javascript and D3.js and was set to full screen.

5.2 Visualization Techniques

Similarity search likely involves both point comparisons, such as finding maxima, and overview comparisons. It is thus unclear how position- or color-based visualizations would affect it (see Sec. 2.2). We thus focused on three visualization techniques that rely on position (LineCharts - LC), color (Colorfields - CF), or both (HorizonGraphs - HG). These visualizations can also scale when arranged in small multiples [37, 50, 53], e.g., in order to support context (see Sec. 3). We explain how we represented time series with these visualizations.

`Line Charts` (LC) map time to the horizontal axis, and value to the vertical. In our implementation, the y-axis was not visible to prevent participants from trying to read exact values. Nevertheless, all time series had a common scale to aid participants compare time series. The zero value was at the middle of the area allocated to each time series. We chose the line variation rather than filled area charts because it is commonly used by EEG visualization tools [35] and our own experts. It has also been used in previous studies on time series similarity [21, 41] and thus acts as a baseline.

`Horizon Graphs` (HG) Horizon graphs utilize space most efficiently with baselines that are specific to each time series, e.g., when the baseline is the average of the time series value range. Nevertheless, different baselines would make comparisons for similarity challenging, that is why we used a common baseline in our experiment for all time series, set to zero. The performance of these graphs seems to deteriorate when increasing the number of bands [29], thus we used a variation of two positive bands and two negative ones, similarly to previous studies [34]. We also followed the convention of using variations of red (#ff9999, #b30000) to indicate negative, and of blue (#bdd7e7, #08519c) to indicate positive values [29, 50], with darker hues assigned to the bands furthest form the baseline (most negative and positive).

`Colorfields` (CF) Previous work considers color scales of two [3, 45] or more colors [53]. We opted for a simple two color scale in our experiment. We again chose red tones (#ff0000) for the most negative and blue (#0000ff) for the most positive value. Pure tones were used to maximize the distance between the two extreme colors [2].

[2]We used a linear RGB interpolation in both experiments. In a follow-up experiment, we used the exact same tasks to compare linear to CIE L*a*b* interpolation. As Sec. 8 discusses, CIE L*a*b* interpolation might be a worse choice. For details, see our technical report [25] and our supplementary material.

The three visualizations utilize space differently. For our experiments, we allocated the same amount of vertical space per time series for all techniques, which is consistent with previous studies [34]. It is important to first understand how humans' similarity perception is affected by the actual visual encoding before considering additional factors, such as the vertical space.

We chose a fairly large vertical size (60 pixels) to ensure that time series were clearly visible in all visualizations. For LC, we fixed the position of the time axis at the middle of its available space since our data includes both positive and negative values. Due to their encoding, HG can utilize the vertical space more efficiently, as they superimpose negative and positive values in the same space. CF do not necessarily require as much vertical space [37], nevertheless this size ensures that colors are large enough to be seen clearly [58].

We fixed the horizontal size of the time series to 501 pixels, encoding one time point per pixel. In practice users, e.g., medical experts, explore their data at different granularities, by keeping the vertical space fixed and compressing or decompressing the time axis. Nevertheless, we decided to avoid factors, such as over-plotting and aggregation, that might also affect similarity perception.

5.3 Algorithms for Measuring Similarity

Participants had to assess the similarity of time series extracted from the dataset (Sec. 4.2). For each trial, we determine one time series that serves as the query and four additional ones as possible matches. These matches were extracted from the data using automatic similarity algorithms. Both experiments used the simple Euclidean Distance (`ED`) as control, but each investigated a different invariance:

Exp-1 (Time Warping): We examined time-warping invariance by contrasting `ED` to `DTW`. A main parameter of `DTW` is the warping size, i.e., the x-offset window size in which the algorithm searches for the best matching point. According to Ding et al. [18], constraining the warping size increases the speed of the algorithm by reducing the computation cost and enabling effective pruning. We set the warping window size to 10% of the time series length as this is the most common size used in the literature and larger sizes can hurt accuracy results [49].

Exp-2 (Z-Normalization): We examined amplitude and y-offset invariance by contrasting the results of Euclidean Distance without (`ED`) or in conjunction with z-normalization (`NormED`) [26]. For the second case, time series are z-normalized to acquire similar amplitude and y-offset, while maintaining the shape of their patterns. Then, ED computes the distance between the two normalized time series.

Both the query and its resulting matches were visualized without any deformations, such as the ones the algorithms perform to access similarity.

5.4 Task

In both experiments participants had to make subjective similarity judgments using one of the three visualizations. They were shown five time series, one of which was marked as "Query". Their task was to select which of the other four time series was the most similar to the query (Figure 4). Those four possible choices were results returned from the similarity algorithms presented above. In Exp-1, two choices came from `ED`, and two choices came from `DTW`. In Exp-2, two choices came from `ED`, and two choices came from `NormED`. Details on the trial generation are described in c Participants gave their answer by clicking on the time series of their choice, which became highlighted, and they rated their confidence on a 5-point scale ("very low" to "very high"). Although there was no time limit for the task, we instructed participants to be as fast and accurate as possible.

Participants performed the same tasks across all visualizations, but we randomized the vertical order of the five time sequences, so as to not favor one measure by presenting its results always closer to the query. We also ensured that time series were not directly one below the other to ensure that certain similarity algorithms, in particular `DTW`, were not penalized. This way participants could not perform a low-level point-by-point comparison of horizontally aligned data series. Instead, they made a more high-level subjective judgement of whether the time series

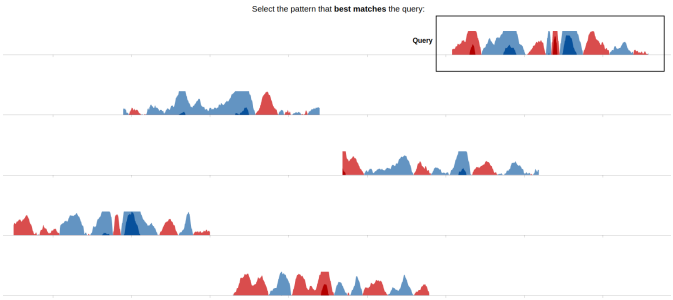


Fig. 4. Experimental screen for the Horizon Graph condition. The answer order and horizontal shift was randomized across visualizations. From the top the series are: Query, Out-ED, Top-ED, Top-DTW, Out-DTW.

were similar or not. The fact that the sequences were not vertically nor horizontally aligned is consistent with the practices of our domain experts, who often compare patterns across sensors or trials that appear in varying vertical positions, and patterns that appear in different times and at different frequencies for different patients (horizontal positions).

Notice that the task was a subjective assessment of similarity, so there was no correct or wrong answer. Our goal was to understand if some visualizations favor some automatic similarity measures and their invariances in terms of perceived similarity.

5.5 Trial Generation

All trials were generated from the annotated dataset described in Sec. 4.2. For each trial, we had to extract a time series that serves as the query and four additional sequences as possible matches. Two of these sequences were *top answers* of the two different algorithms that each experiment studied: `ED` vs. `DTW` (Exp-1), and `ED` vs. `NormED` (Exp-2). The other two sequences were *outsiders* that resulted from the same two algorithms but in a lower rank.

As discussed (Sec. 4.3) one challenge was how to differentiate between the similarity search algorithms, given that they may return similar results. We thus opted for a query-extraction process that ensures that the algorithms return top answers that are distinct.

`Step 1:` *Creating Candidate Queries.* We started from the manually annotated markers to extract possible queries. Epileptiform discharges last less than 250ms [56], but we extracted a larger window of 401ms around each marker (200ms left and right). This ensured that the full pattern of interest was included in the query, and that the sequence includes background activity (context), which can be important for assessing similarity. From 205 annotations, we extracted a pool of 202 candidate queries. We excluded three that were very close to the beginning or the end of a recording (and thus of smaller size).

`Step 2:` *Finding Similar Subsequences.* For each candidate query, we ran similarity searches by using the two search algorithms of interest: `ED` vs. `DTW` in Exp-1, and `ED` vs. `NormED` in Exp-2. We collected the first 100 Nearest Neighbor (NN) answers for each algorithm. We focused our searches on the same iEEG sensors as the query, but answers could be part of different recordings. We extended an optimization algorithm for early subsequence pruning [48] to support k-NN instead of 1-NN search. Its average time complexity for comparing two series of the same length (n points) is less than $O(n)$ for all distance measures and is the fastest algorithm known in the literature.

`Step 3:` *Selecting the Final Queries.* We then checked if the best results returned by each algorithm were unique. We considered the top five answers of the two measures we compared each time. Those were generally not common: an average of 62% of the top five answers of the two measures was different in Exp-1, and this percentage was 55% in Exp-2. We wanted to select answers that clearly highlight the differences of the two measures. In addition, we had to avoid biases that may arise when picking top answers for one measure that are also highly ranked for the other measure (and therefore more probable to be selected). Thus, we looked at queries where the top five answers of one measure did not appear within the top ten of the other. This resulted in a set of 30 queries for Exp-1 and a different set of 31 queries for Exp-2, from which we randomly picked 30 queries.

`Step 4:` *Choosing the Answers to each Query.* The experimental trials were formed from those 30 queries. Two of the four possible answers presented to participants were the highest ranked answers of each algorithm from Step-3 (referred to as `Top-ED`, `Top-DTW`, and `Top-NormED`, respectively for each algorithm). Another two answers were produced in a way similar to Step-3, but looking at answers between the lower 20-30 of each algorithm (we refer to them as `Out-ED`, `Out-DTW`, and `Out-NormED`). Outsiders were expected to be perceived as less similar than top answers, but were still valid answers to the query. They provided a control for assessing the accuracy of participants' answers with respect to the underlying algorithms. And acted as distractors to make the task more realistic, given that analysts may search through many subsequences to find a match.

5.6 Experimental Design

We followed a within-participants design – all participants were exposed to all three visualization techniques. The order of appearance of the three techniques was fully counterbalanced. For each technique, participants completed 5 practice and 20 main trials.

For each experiment, we generated a different set of 30 distinct trials (see Sec. 4.2). To make use of the full set of trials, we divided the trials in 3 bins of 10, and each participant saw one bin during training and the other two during the experiment (counterbalanced across participants). Overall, each trial was tested by exactly 12 participants. Each participant performed the same 20 trials for all three visualizations, but we randomized the vertical order of the five time sequences, including the query. This ensured that participants could not recognize the queries or their choices between conditions.

In summary, each experiment consisted of:

18 participants
× 3 visualizations (LC, HG, CF)
× 20 query-answer trials
= 1080 trials per experiment

5.7 Procedure

Before starting, participants completed a short color blindness test using the Isihara plates. They then signed a consent form and continued with a training session on how to read the respective visualization technique. Before the main experiment, participants had to pass three readability tests, where they compared values of different points in a time series.

As we were interested in participants' intuitive perception of similarity across visualizations, we gave no instructions about how to interpret similarity, did not mention invariances, and did not provide any guidelines about how to assess similarity with each technique. A similar approach was used by Correll and Gleicher [16]. Furthermore, we did not explain what the data represented or how the queries and their candidate answers were generated.

After the experiment, participants completed a questionnaire to provide background information and evaluate the three visualization techniques. The experiment lasted from 45 to 80 minutes.

5.8 Measures

We use a mix of measures that assess the types of answers given by participants, their accuracy with respect to the similarity algorithms that we tested, and their agreement among participants. In addition, we measure participants' confidence about their answers, time performance, as well as their subjective assessment of the three visualizations.

`Type of Answers:` We count the number of occurrences of each type of answer. For Exp-1, we count `Top-ED`, `Top-DTW`, `Out-ED`, and `Out-DTW`. For Exp-2, we count `Top-ED`, `Top-NormED`, `Out-ED`, and `Out-NormED`. Counts provide raw information about participants' choices and are used to construct our ratio measures (next).

`DTW vs.ED and NormED vs.ED:` We assess participants' tendency to select the top answers of one similarity algorithm over the other by calculating the ratio of their counts. For Exp-1, we take the ratio

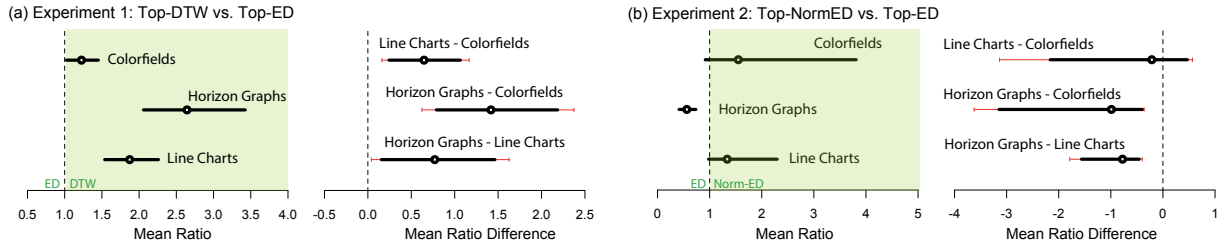


Fig. 5. Interval estimates comparing the mean ratios of (a) `Top-ED` to `Top-DTW` answers (Exp-1) and (b) `Top-NormED` vs. `Top-ED` answers (Exp-2) with the three visualization techniques. Error bars represent 95% CIs. For mean ratio differences, we also show (in red) CIs adjusted for three pairwise comparisons with Bonferroni correction. The dotted vertical lines show the values of reference.

of the counts of `Top-DTW` to those of `Top-ED`. A ratio greater than 1 indicates a preference for the top answers of `DTW`. For Exp-2, we take the ratio of the counts of `Top-NormED` to those of `Top-ED`. Here a ratio greater than 1 indicates a preference for the top answers of `NormED`. We compare the difference of these ratios between techniques, a difference greater or smaller that zero provides evidence that the techniques differ.

`Outsiders vs.Top Answers:` We assess the accuracy of participants' answers with respect to the answers of the similarity algorithms by calculating the ratio of the counts of their outsiders to the counts of their top answers. A large ratio indicates a relatively large number of outsiders in participants' choices.

`Agreement:` We assess the level of consensus in participants' choices with agreement coefficients, which are commonly used in the context of inter-rater reliability studies [28]. High agreement demonstrates low subjectivity in participants' choices. In contrast, low agreement indicates high uncertainty when making decisions. It may also imply that similarity perception is highly subjective.

We choose the κ_q coefficient of Brennan and Prediger [9]. The coefficient assumes that all q categories are selected by chance with the same probability $p_e = 1/q$. This assumption is valid in our case, since the $q = 4$ alternative answers were presented in a random order to participants, which avoided problems of bias [62]. In addition to overall agreement, we assess agreement *specific to categories* [55]. This allows us to assess how agreement is divided across different types of answers.

`Time Performance:` We measured the time it takes participants to complete a task, from the moment the time series are shown on the screen to the moment participants select their final answer. Although assessing time performance was not a primary goal of our experiments, this measure allows us to compare how easy or difficult it was to perform similarity tasks with each visualization technique.

`Subjective Measures:` We recorded participants' self-reported level of confidence on their answers to each query. We use this measure of confidence in conjunction with agreement measures.

5.9 Expected Outcomes

LC$\wedge\!\!/$ is extensively used in practice, so one could expect that it is the most appropriate technique for determining similarity of time series. HG and CF have not been studied in the context of perceived similarity tasks before, thus existing evidence about how they would perform compared to LC$\wedge\!\!/$ is limited. Previous studies have shown that HG were faster than line charts for discrimination tasks, but slower for peak and trend detection tasks [34]. Whereas CF has been shown to be a promising representation for overview tasks [15]. Similarity search likely requires both low-level (i.e., detecting picks) and overview tasks.

In terms of similarity algorithms, Dynamic Time Warping (`DTW`) is widely considered to give better results than Euclidean Distance (`ED`). For LC$\wedge\!\!/$, Eichmann and Zgraggen [21] found that `DTW` generally produce rankings that are closer to human-annotated ranking, so we expected to find similar results. Z-normalization is a recommended practice for all similarity measures [18], thus one could predict that it would produce more similar answers. However, we also expected that color encodings might be sensitive, i.e., non-invariant, to y-offset and amplitude transformations.

6 Results

We present the results of the two experiments. Our statistical analysis is largely based on interval estimation [19], as this approach better supports future replication efforts. All analyses reported were planned before data were collected.

6.1 Invariances: Time-Warping and Z-Normalization

We first examine how the three visual encoding techniques affected participants' choices in favor or against the two invariances of interest. Our analysis relies on ratios of counts, where counts are not independent. The sampling distribution of such measures can be complex and hard to approximate with analytical methods. We thus use bootstrapping methods to construct 95% confidence intervals (CI) of the mean. We apply Efron's [20] bias-corrected and accelerated (BCa) bootstrap method as implemented by *R*'s *boot* package [12]. For our analyses, we construct confidence intervals with 10000 bootstrap iterations.

Exp-1 (`DTW` vs. `ED`): Figure 5a presents interval estimates for individual means (left) and their differences (right). For all three techniques, we observe that participants considered as similar the `Top-DTW` answers more. This trend is however different across visualization techniques. It is especially pronounced for HG, where `Top-DTW` answers were on average 2.64 (SD = 1.49) times more frequent than `Top-ED` answers. The mean ratio of `Top-DTW` to `Top-ED` answers drops to 1.87 (SD = 0.80) for LC$\wedge\!\!/$, and 1.23 (SD = 0.48) for CF.

Exp-2 (`NormED` vs. `ED`): Figure 5b presents interval estimates for both means (left) and their differences (right). We observe a strong tendency in HG for participants to not find as similar `Top-NormED` answers, where their mean ratio to `Top-ED` answers is equal to 0.56 (SD = 0.36). In contrast, with the other visualizations they lean towards z-normalized answers, with mean ratios equal to 1.33 (SD = 1.18) for LC$\wedge\!\!/$and 1.55 (SD = 2.41) for CF. However, due to large variance, this trend is not clearly supported by statistical evidence. We see that HG favor `Top-ED` answers more than the other techniques, but we observe no clear difference between LC$\wedge\!\!/$ and CF.

6.2 Outsiders vs. Top Query Answers

We further analyze the ratio of outsiders to top query answers by using a similar analysis procedure.

Exp-1: Figure 6 shows interval estimates for Exp-1. Clearly, the top answers of the two algorithms dominated participants' choices. However, in many cases, participants perceived outsiders as more similar than top answers. Their ratio was 0.39 (SD = .20) for HG, 0.49 (SD = .22) for LC$\wedge\!\!/$, and 0.63 (SD = .36) for CF. The difference is clearer between HG and CF. The latter resulted in a relatively large number of outsiders.

Exp-2: Figure 7 presents interval estimates for Exp-2. We now observe the opposite trend, but differences between the techniques are less clear. The ratio of outsiders to top answers was 0.40 (SD = .27) for HG, 0.31 (SD = .21) for LC$\wedge\!\!/$, and 0.27 (SD = .16) for CF. CF now resulted in a lower ratio than HG.

Combined with the results of Section 6.1, these results seem to suggest that CF are less appropriate for DTW, while HG are less appropriate for z-normalized answers.

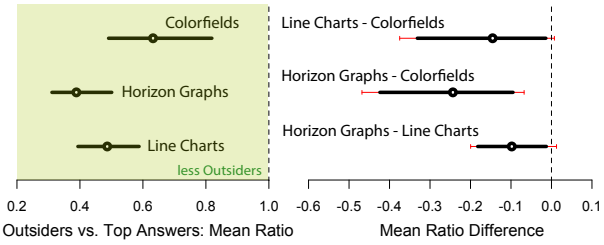


Fig. 6. Experiment 1: Interval estimates comparing the mean ratios of outsiders to top query answers. Error bars represent 95% CIs. Red extensions show the adjustment for three pairwise comparisons.

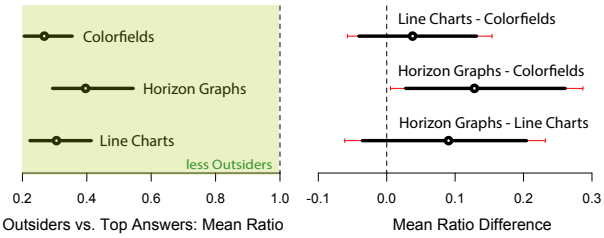


Fig. 7. Experiment 2: Interval estimates comparing the mean ratios of outsiders to top query answers. Error bars represent 95% CIs. Red extensions show the adjustment for three pairwise comparisons.

6.3 Agreement

To construct confidence intervals for our agreement estimates, we use the jackknife technique [28, 62] by assuming that raters, i.e., participants, are randomly sampled from a larger population, whereas the set of queries is fixed.

Exp-1: Table 1 summarizes the results of Exp-1. Overall, agreement is higher than zero for all three techniques. This verifies that similarity perception was not fully subjective and that participants' choices were not random. However, agreement values are generally low for HG and CF, which implies a higher subjectivity of participants' choices with these techniques. Overall, we observe a higher agreement for the choice of `Top-DTW` answers. This is especially the case for HG - this further shows the tendency of the technique towards `DTW`, as `Top-ED` answers were chosen with no consistency among participants. We observed a positive linear correlation between agreement values and the average confidence level reported by participants for each task (Pearson's moment correlation was $r = .45$, 95% CI $= [.27, .60]$). This result is not surprising – agreement or disagreement is largely due the confidence or uncertainty with which participants make choices.

Table 1. Experiment 1: Specific and overall agreement values (Brennan-Prediger κ_q). Brackets show 95% jackknife CIs.

	Line Charts	Horizon Graphs	Colorfields
`Top-ED`:	.42 [.22, .62]	.04 [−.07, .16]	.28 [.10, .47]
`Top-DTW`:	.54 [.41, .68]	.41 [.28, .55]	.35 [.22, .49]
`Outsider-ED`:	.14 [−.09, .36]	−.06 [−.20, .07]	.21 [.00, .42]
`Outsider-DTW`:	.39 [.26, .52]	−.01 [−.19, .17]	.07 [−.07, .22]
Overall:	.44 [.36, .52]	.21 [.13, .29]	.26 [.18, .33]

Exp-2: Table 2 summarizes the results of Exp-2. Again, overall agreement is higher than zero for all techniques. Agreement values are now more balanced across techniques. We note that HG resulted in low agreement values for z-normalized answers. This further shows that the technique may not be invariant to z-normalization. For this experiment, Pearson's moment correlation between participants' self-reported confidence level and agreement was $r = .59$, 95% CI $= [.43, .71]$.

6.4 Time Performance

Time measures are well-known to follow lognormal distributions [5,39], thus we log-transform time values and analyze them with standard parametric methods that assume normal distributions. According to this approach, comparisons between techniques are based on the ratio of their *median times* rather than their mean time differences [19].

Table 2. Experiment 2: Specific and overall agreement values (Brennan-Prediger κ_q). Brackets show 95% jackknife CIs.

	Line Charts	Horizon Graphs	Colorfields
`Top-ED`:	.38 [.19, .57]	.48 [.32, .63]	.41 [.27, .55]
`Top-NormED`:	.43 [.29, .57]	.14 [.01, .27]	.43 [.28, .59]
`Outsider-ED`:	.03 [−.23, .29]	−.03 [−.19, .12]	−.01 [−.27, .24]
`Outsider-NormED`:	.05 [−.09, .20]	.06 [−.14, .25]	.05 [−.12, .21]
Overall:	.33 [.21, .45]	.27 [.20, .35]	.34 [.23, .45]

Exp-1: Mean completion time was 20.5 sec (SD = 13.9 sec) for LC⩛, 23.7 sec (SD = 9.1 sec) for HG, and 15.6 sec (SD = 7.5 sec) for CF. Figure 8a shows interval estimates for median times (left) and their ratios of medians (right). We observe that CF was the fastest technique. And we have some evidence that HG were on average 33.6% slower than LC⩛.

Exp-2: Mean task-completion time was now 21.1 sec (SD = 12.6 sec) for LC⩛, 28.8 sec (SD = 15.8 sec) for HG, and 21.5 sec (SD = 13.2 sec) for CF. Figure 8b shows interval estimates for median times (left) and their ratios of medians (right). We found no evidence of a difference between LC⩛ and CF. HG was again the slowest, on average 40% slower than the two other techniques.

7 Discussion and Design Implications

Results from both experiments suggest that humans may perceive similarity differently, depending on the visualization, and that different visual encodings are invariant to specific signal parameters.

In *Exp-1* participants preferred results returned by Dynamic Time Warping (DTW), i.e., subsequences that can be shifted in the x-axis and locally stretched or compressed. This finding corroborates previous evidence [18, 21] that DTW is superior to Euclidean Distance (ED). Nevertheless, this effect differs across visualization techniques. It is stronger for horizon graphs, likely due to this technique's double encoding. Color variations often communicate high-level patterns (spikes/valleys, positive/negative ranges), while shape and position reveal details. Participants may have focused on the high-level patterns in color to determine similarity, considering shape and position (which encode warping and x-axis shifting) as secondary factors. Line charts favored DTW but to a lesser degree, and the trend was even weaker for colorfields. Colorfields aid the detection of ranges of similar color [2] so it is probable that participants considered both the color of the spikes and the width of the color ranges formed around them. Thus, they were likely to avoid candidates that were too stretched or compressed. The example in Figure 9:Left demonstrates this issue.

In *Exp-2* we observed a clear difference between horizon graphs and the two other visualizations. Horizon graphs strongly favored the answers of ED without z-normalization. The opposite trend was observed for line charts and colorfields. In horizon graphs, small amplitude and y-offset changes can fall on different sides of a band and have different colors. Thus, if participants tried to match colors rather than shape, they likely disregarded subsequences whose prominent characteristics fell on different bands (see Figure 9:Right). For line charts and colorfields, the exact amplitude and offset values can be less critical, as people seem to focus on relative values and overall shapes.

Overall, agreement scores were lower in horizon graphs and time performance was slower, which indicates this encoding can be difficult to visually identify patterns and make decisions when using it.

In both experiments, participants tended to select the top answers of the algorithms rather than their outsiders, irrespective of the visualization technique. This confirms that the rankings of these algorithms capture real differences in perceptual similarity.

Design Implications: Overall, our work indicates that the choice of visualization affects what temporal patterns people consider as similar, i.e., *the notion of similarity in time series is not visualization independent*. Visualization designers need to consider what invariances are important in the data domain [18] and suggest visualizations appropriately. Similarly, if designers use algorithmic distance measures, they should consider visualizations that match the invariances of those measures, or viewers could lose confidence in their results.

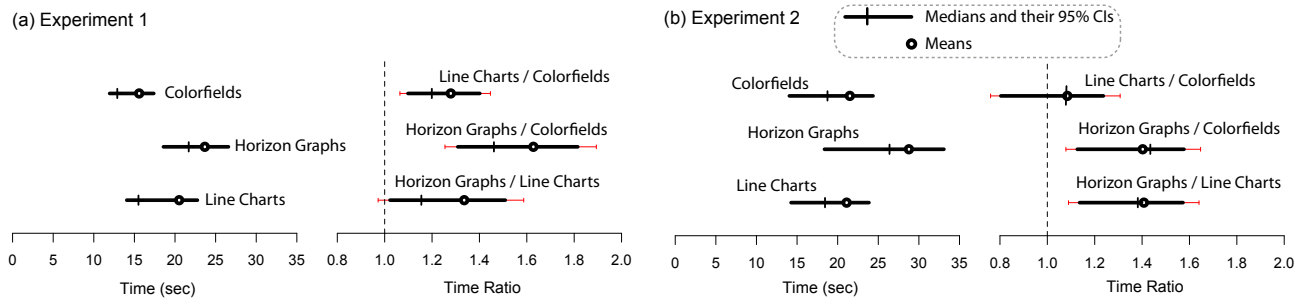


Fig. 8. Interval estimates comparing the median task completion time for each technique. Error bars represent 95% CIs. Red extensions (right) show adjustments for three pairwise comparisons. **Note:** The official version incorrectly states that CIs are for means (rather than medians).

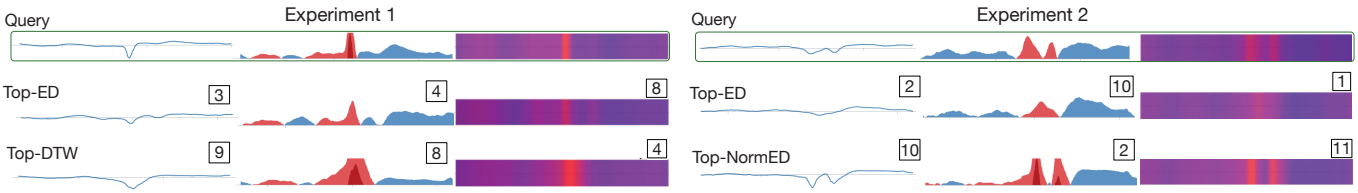


Fig. 9. Two queries for which different visualizations resulted in different choices. Boxes show the number of participants (out of 12) who chose the specific answer. Left: This example shows that Colorfields can be more sensitive than Line Charts and Horizon Graphs to stretching deformations along the time axis. Right: A strong preference for `Top-NormED` under Line Charts and Colorfields and a strong preference for `Top-ED` under Horizon Graphs. Overall, Horizon Graphs seem to exaggerate flat signals and are more sensitive to deformations along the y-axis.

Our results suggest that *colorfields are less appropriate for domains that require invariance to temporal warping*, as they are sensitive to temporal warping and shifting. Here, horizon graphs are a viable alternative to line charts, as they are less sensitive to warping. Nevertheless, designers should consider the visual complexity of time series visualizations. Agreement was lower for horizon graphs and time performance was slower, while participants reported they found it more difficult to visually identify patterns and make decisions when using it.

In turn, *horizon graphs are less appropriate when amplitude and y-offset invariance is important*, as they are sensitive to value transformations along the y-axis due to the explicit limits of their bands. Finally, as in previous work using line charts [18, 21], our results support that DTW, an algorithm that is invariant to temporal warping, is likely closer to what we perceive as similar in temporal patterns, and thus *DTW could be considered as a good default* unless otherwise indicated by the data domain [18].

8 Limitations and Future Work

There are several limitations to our work. First, we focused on a small number of similarity measures. The data-mining literature has studied measures for other types of invariance [6]. Future work needs to determine what visualizations best match such measures. Furthermore, our dataset consists of EEG data that have specific pattern characteristics, such as spikes followed by rapid discharges. Although we believe that our high-level results will hold for other types of signals, the sensitivity of visual perception to certain signal deformations may be less or more pronounced. Further studies are needed to validate our findings in a wider range of patterns and datasets from other domains.

Our implementation of colorfields used a naive, linear RGB interpolation. This approach leads to a color space that is not perceptually uniform, i.e., differentiating variations may be harder for one of the two color extremes. On the other hand, it may extend the differences near the central range of the color space, in magenta tones which humans are more sensitive to [38]. This central range is where low-amplitude variations and spikes (which might be important for EEG signals) are located. We conducted a follow-up experiment (N = 18 participants) that compared linear RGB interpolation to a perceptually uniform CIE L*a*b color space [25]. Accuracy and agreement scores were very similar for the two techniques, while most participants (10 vs. 6) found that is was easier to identify patterns with linear RGB interpolation. CIE L*a*b resulted in less pronounced differences between similarity measures, but we found no statistically significantly differences between the two interpolation techniques. We report the detailed results of this experiment as supplementary material. Nevertheless, it is possible that differences in these color mappings exist in other types of temporal patterns. Moreover, in domains where similarity comparison is the only task of interest, one could also consider dynamic mapping variations (e.g., difference color maps, or ones based on equi-depth or equi-width binning of time series values to provide wider color ranges for the most frequent values), that nonetheless distort the original signals. The effect of color in time series similarity is an exiting future research direction.

We focused on a small number of time series to compare, with a generous vertical drawing area. While we hypothesize that our results will hold for larger number of time series, their size might affect these results. For example, we expect that colorfields will scale well, but it is known that the choice of the aspect ratio affects readability in line charts [61]. Thus, for line charts and to a lesser degree for horizon graphs, a reduced vertical space could lead to a loss of small patterns and reinforce large structures (peaks, valleys) altering similarity perception.

Finally, we plan to compare additional visual encodings or variations of the ones studied in this paper, such as composite visualizations that go beyond horizon graphs [33], and area charts with alternative designs, e.g., designs based on single or dual fill color, and mirroring.

9 Conclusion

We presented two laboratory experiments that compare how three visualizations (line charts, colorfields, and horizon graphs) affect how we perceive similarity in time series. Specifically, we studied if some deformations in the data, detected by automatic similarity measures, are perceived in a different manner depending on the visualization. Our findings indicate that all three visualizations, favor similarity results from algorithmic measures that allow flexibility in local deformations in temporal position or speed (i.e., dynamic time warping). This is the case most notably for horizon graphs. On the other hand, this visualization does not promote results from algorithms that are invariant to y-offset shifts and amplitude rescaling (i.e., z-normalization).

Our work provides evidence that the notion of time series similarity is visualization dependent, and that when choosing visual representations, we should consider what deformations the underlying data domain considers as similar. This should be consistent with the similarity measures used in each domain. In the future, we plan to investigate how choosing appropriate visualizations to communicate similarity can affect agreement of what is similar among domain experts, and if this increases trust on the results of similarity search algorithms.

Acknowledgments

We thank Petra Isenberg for feedback on the paper, and Katia Lehongre and Denis Schwartz for the access to the MUSE tool and data.

REFERENCES

[1] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of Time-Oriented Data*. Springer Publishing Company, Incorporated, 1st ed., 2011.

[2] D. Albers, M. Correll, and M. Gleicher. Task-driven evaluation of aggregation in time series visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pp. 551–560. ACM, New York, NY, USA, 2014. doi: 10.1145/2556288.2557200

[3] D. Albers, C. Dewey, and M. Gleicher. Sequence surveyor: Leveraging overview for scalable genomic alignment visualization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2392–2401, Dec. 2011. doi: 10.1109/TVCG.2011.232

[4] J. Aßfalg, H.-P. Kriegel, P. Kröger, P. Kunath, A. Pryakhin, and M. Renz. Similarity search on time series based on threshold queries. In *Proceedings of the 10th International Conference on Advances in Database Technology*, EDBT'06, pp. 276–294. Springer-Verlag, Berlin, Heidelberg, 2006. doi: 10.1007/11687238_19

[5] T. Baguley. *Serious Stats: A guide to advanced statistics for the behavioral sciences*. Palgrave Macmillan, 2012.

[6] G. E. Batista, E. J. Keogh, O. M. Tataw, and V. M. Souza. Cid: An efficient complexity-invariant distance for time series. *Data Min. Knowl. Discov.*, 28(3):634–669, May 2014. doi: 10.1007/s10618-013-0312-3

[7] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, AAAIWS'94, pp. 359–370. AAAI Press, 1994.

[8] E. Bertini, P. Hertzog, and D. Lalanne. Spiralview: Towards security policies assessment through visual correlation of network resources with evolution of alarms. In *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology*, VAST '07, pp. 139–146. IEEE Computer Society, Washington, DC, USA, 2007. doi: 10.1109/VAST.2007.4389007

[9] R. L. Brennan and D. J. Prediger. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and psychological measurement*, 41(3):687–699, 1981.

[10] P. Buono and A. L. Simeone. Interactive shape specification for pattern search in time series. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '08, pp. 480–481. ACM, New York, NY, USA, 2008. doi: 10.1145/1385569.1385666

[11] L. Byron and M. Wattenberg. Stacked graphs – geometry & aesthetics. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1245–1252, Nov. 2008. doi: 10.1109/TVCG.2008.166

[12] A. Canty and B. D. Ripley. *boot: Bootstrap R (S-Plus) Functions*, 2017. R package version 1.3-20.

[13] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista. The ucr time series classification archive, July 2015. `www.cs.ucr.edu/~eamonn/time_series_data/`.

[14] Y. Chen, M. Nascimento, B. C. Ooi, and A. K. Tung. Spade: On shape-based pattern detection in streaming time series. In *Proceedings of the IEEE 23rd International Conference on Data Engineering*, ICDE'07, pp. 786–795. IEEE, 2007. doi: 10.1109/ICDE.2007.367924

[15] M. Correll, D. Albers, S. Franconeri, and M. Gleicher. Comparing averages in time series data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pp. 1095–1104. ACM, New York, NY, USA, 2012. doi: 10.1145/2207676.2208556

[16] M. Correll and M. Gleicher. The semantics of sketch: Flexibility in visual query systems for time series data. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 131–140, Oct 2016. doi: 10.1109/VAST.2016.7883519

[17] M. de Curtis, J. G. R. Jefferys, and M. Avoli. Interictal epileptiform discharges in partial epilepsy: Complex neurobiological mechanisms based on experimental and clinical evidence. *Jasper's Basic Mechanisms of the Epilepsies [Internet]. 4th edition*, pp. 303–325, 2012.

[18] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh. Querying and mining of time series data: Experimental comparison of representations and distance measures. *Proc. VLDB Endow.*, 1(2):1542–1552, Aug. 2008. doi: 10.14778/1454159.1454226

[19] P. Dragicevic. Fair statistical communication in hci. In *Modern Statistical Methods for HCI*, pp. 291–330. Springer, 2016. doi: 10.1007/978-3-319-26633-6_13

[20] B. Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987. doi: 10.1080/01621459.1987.10478410

[21] P. Eichmann and E. Zgraggen. Evaluating subjective accuracy in time series pattern-matching using human-annotated rankings. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, pp. 28–37. ACM, New York, NY, USA, 2015. doi: 10.1145/2678025.2701379

[22] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time-series databases. In *Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data*, SIGMOD '94, pp. 419–429. ACM, New York, NY, USA, 1994. doi: 10.1145/191839.191925

[23] T.-c. Fu. A review on time series data mining. *Eng. Appl. Artif. Intell.*, 24(1):164–181, Feb. 2011. doi: 10.1016/j.engappai.2010.09.007

[24] J. Fuchs, F. Fischer, F. Mansmann, E. Bertini, and P. Isenberg. Evaluation of alternative glyph designs for time series data in a small multiple setting. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pp. 3237–3246. ACM, New York, NY, USA, 2013. doi: 10.1145/2470654.2466443

[25] A. Gogolou, T. Tsandilas, T. Palpanas, and A. Bezerianos. Comparing time series similarity perception under different color interpolations. . Research Report RR-9189, Inria, 06/2018.

[26] D. Q. Goldin and P. C. Kanellakis. On similarity queries for time-series data: Constraint specification and implementation. In *Proceedings of the First International Conference on Principles and Practice of Constraint Programming*, CP '95, pp. 137–153. Springer-Verlag, London, UK, UK, 1995.

[27] M. Gregory and B. Shneiderman. Shape identification in temporal data sets. Master's thesis, Master's thesis, University of Maryland, 2009.

[28] K. L. Gwet. *Handbook of Inter-Rater Reliability, 4th Edition: The Definitive Guide to Measuring The Extent of Agreement Among Raters*. Advanced Analytics, LLC, 2014.

[29] J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pp. 1303–1312. ACM, New York, NY, USA, 2009. doi: 10.1145/1518701.1518897

[30] H. Hochheiser and B. Shneiderman. Dynamic query tools for time series data sets: Timebox widgets for interactive exploration. *Information Visualization*, 3(1):1–18, Mar. 2004. doi: 10.1145/993176.993177

[31] C. Holz and S. Feiner. Relaxed selection techniques for querying time-series graphs. In *Proceedings of the 22Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '09, pp. 213–222. ACM, New York, NY, USA, 2009. doi: 10.1145/1622176.1622217

[32] K. Indiradevi, E. Elias, P. Sathidevi, S. D. Nayak, and K. Radhakrishnan. A multi-level wavelet approach for automatic detection of epileptic spikes in the electroencephalogram. *Computers in Biology and Medicine*, 38(7):805 – 816, 2008. doi: 10.1016/j.compbiomed.2008.04.010

[33] A. Jabbari, R. Blanch, and S. Dupuy-Chessa. Composite visual mapping for time series visualization. In *2018 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 116–124, April 2018. doi: 10.1109/PacificVis.2018.00023

[34] W. Javed, B. McDonnel, and N. Elmqvist. Graphical perception of multiple time series. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):927–934, Nov. 2010. doi: 10.1109/TVCG.2010.162

[35] J. Jing, J. Dauwels, T. Rakthanmanon, E. Keogh, S. Cash, and M. Westover. Rapid annotation of interictal epileptiform discharges via template matching under dynamic time warping. *Journal of Neuroscience Methods*, 274:179 – 190, 2016. doi: 10.1016/j.jneumeth.2016.02.025

[36] S. juan Orta D, C. KH, Q. AZ, C. DJ, and C. AJ. Prognostic implications of periodic epileptiform discharges. *Archives of Neurology*, 66(8):985–991, 2009. doi: 10.1001/archneurol.2009.137

[37] R. Kincaid and H. Lam. Line graph explorer: Scalable display of line graphs using focus+context. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '06, pp. 404–411. ACM, New York, NY, USA, 2006. doi: 10.1145/1133265.1133348

[38] H. Levkowitz and G. Herman. The design and evaluation of color scales for image data. *Computer Graphics and Applications)*, 12(1):82–89, 1992.

[39] E. Limpert, W. A. Stahel, and M. Abbt. Log-normal distributions across the sciences: Keys and clues. 51:341–, 05 2001. doi: 10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2

[40] E. K. S. Louis and L. C. Frey. *Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants*. American Epilepsy Society, 2016.

[41] M. Mannino and A. Abouzied. Expressive time series querying with hand-drawn scale-free sketches. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 388:1–388:13. ACM, New York, NY, USA, 2018. doi: 10.1145/3173574.3173962

[42] P. McLachlan, T. Munzner, E. Koutsofios, and S. North. Liverac: Interactive visual exploration of system management time-series data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pp. 1483–1492. ACM, New York, NY, USA, 2008. doi: 10.1145/1357054.1357286

[43] W. Müller and H. Schumann. Visualization for modeling and simulation: Visualization methods for time-dependent data - an overview. In *Proceedings of the 35th Conference on Winter Simulation: Driving Innovation*, WSC '03, pp. 737–745. Winter Simulation Conference, 2003.

[44] P. K. Muthumanickam, K. Vrotsou, M. Cooper, and J. Johansson. Shape grammar extraction for efficient query-by-sketch pattern matching in long time series. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 121–130, Oct 2016. doi: 10.1109/VAST.2016.7883518

[45] D. Nadalutti and L. Chittaro. Visual analysis of users performance data in fitness activities. *Computers & Graphics*, 31(3):429 – 439, 2007. doi: 10.1016/j.cag.2007.01.032

[46] T. Palpanas. Data series management: The road to big sequence analytics. *SIGMOD Record*, 44(2):47–52, 2015. doi: 10.1145/2814710.2814719

[47] C. Perin, F. Vernier, and J.-D. Fekete. Interactive horizon graphs: Improving the compact visualization of multiple time series. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pp. 3217–3226. ACM, New York, NY, USA, 2013. doi: 10.1145/2470654.2466441

[48] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh. Searching and mining trillions of time series subsequences under dynamic time warping. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pp. 262–270. ACM, New York, NY, USA, 2012. doi: 10.1145/2339530.2339576

[49] C. A. Ratanamahatana and E. Keogh. Everything you know about dynamic time warping is wrong. In *Third Workshop on Mining Temporal and Sequential Data*. Citeseer, 2004.

[50] H. Reijner. The development of the horizon graph. available online at `http://www.stonesc.com/Vis08_Workshop/DVD/Reijner_submission.pdf`, 2008.

[51] H. G. Rey, C. Pedreira, and R. Quian Quiroga. Past, present and future of spike sorting techniques. *Brain Research Bulletin*, 119(Pt B):106–117, Oct 2015. doi: 10.1016/j.brainresbull.2015.04.007

[52] K. Ryall, N. Lesh, T. Lanning, D. Leigh, H. Miyashita, and S. Makino. Querylines: Approximate query for visual browsing. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '05, pp. 1765–1768. ACM, New York, NY, USA, 2005. doi: 10.1145/1056808.1057017

[53] T. Saito, H. N. Miyamura, M. Yamamoto, H. Saito, Y. Hoshiya, and T. Kaseda. Two-tone pseudo coloring: Compact visualization for one-dimensional data. In *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, INFOVIS '05, pp. 23–. IEEE Computer Society, Washington, DC, USA, 2005. doi: 10.1109/INFOVIS.2005.35

[54] T. Siddiqui, A. Kim, J. Lee, K. Karahalios, and A. Parameswaran. Effortless data exploration with zenvisage: An expressive and interactive visual analytics system. *Proc. VLDB Endow.*, 10(4):457–468, Nov. 2016. doi: 10.14778/3025111.3025126

[55] R. L. Spitzer and J. L. Fleiss. A re-analysis of the reliability of psychiatric diagnosis. *The British Journal of Psychiatry*, 125(587):341–347, 1974. doi: 10.1192/bjp.125.4.341

[56] K. J. Staley and F. E. Dudek. Interictal spikes and epileptogenesis. *Epilepsy Currents 6.6*, pp. 199–202, 2006. doi: 10.1111/j.1535-7511.2006.00145.x

[57] K. J. Staley, A. White, and F. E. Dudek. Interictal spikes: Harbingers or causes of epilepsy? *Neuroscience letters 497.3*, pp. 247–250, 2011. doi: 10.1016/j.neulet.2011.03.070

[58] M. Stone. In color perception, size matters. *IEEE Computer Graphics and Applications*, 32(2):8–13, March 2012. doi: 10.1109/MCG.2012.37

[59] B. Swihart, B. Caffo, B. James, M. Strand, B. Schwartz, and N. Punjabi. Lasagna plots: a saucy alternative to spaghetti plots. *Epidemiology (Cambridge, Mass.)*, 21(5):621–625, 2010. doi: 10.1097/EDE.0b013e3181e5b06a

[60] I. Snchez Fernndez, T. Loddenkemper, A. S. Galanopoulou, and S. L. Mosh. Should epileptiform discharges be treated? *Epilepsia*, 56(10):1492–1504, 2015. doi: 10.1111/epi.13108

[61] J. Talbot, J. Gerth, and P. Hanrahan. An empirical model of slope ratio comparisons. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2613–2620, Dec 2012. doi: 10.1109/TVCG.2012.196

[62] T. Tsandilas. Fallacies of agreement: A critical review of consensus assessment methods for gesture elicitation. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(3):18:1–18:49, June 2018. doi: 10.1145/3182168

[63] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA, 1986.

[64] J. J. Van Wijk and E. R. Van Selow. Cluster and calendar based visualization of time series data. In *Proceedings of the 1999 IEEE Symposium on Information Visualization*, INFOVIS '99, pp. 4–. IEEE Computer Society, Washington, DC, USA, 1999. doi: 10.1109/INFVIS.1999.801851

[65] M. Wattenberg. Sketching a graph to query a time-series database. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '01, pp. 381–382. ACM, New York, NY, USA, 2001. doi: 10.1145/634067.634292

[66] M. Weber, M. Alexa, and W. Müller. Visualizing time-series on spirals. In *Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01)*, INFOVIS '01, pp. 7–. IEEE Computer Society, Washington, DC, USA, 2001. doi: 10.1109/INFVIS.2001.963273

[67] J. Zhao, F. Chevalier, and R. Balakrishnan. Kronominer: Using multi-foci navigation for the visual exploration of time-series data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pp. 1737–1746. ACM, New York, NY, USA, 2011. doi: 10.1145/1978942.1979195

[68] J. Zhao, F. Chevalier, E. Pietriga, and R. Balakrishnan. Exploratory analysis of time-series with chronolenses. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2422–2431, Dec. 2011. doi: 10.1109/TVCG.2011.195

[69] K. Zoumpatianos, S. Idreos, and T. Palpanas. RINSE: interactive data series exploration with ADS+. *PVLDB*, 8(12):1912–1915, 2015. doi: 10.14778/2824032.2824099

[70] K. Zoumpatianos, S. Idreos, and T. Palpanas. ADS: the adaptive data series index. *VLDB J.*, 25(6):843–866, 2016. doi: 10.1007/s00778-016-0442-5