Visual Intelligence

RESEARCH Open Access

Colorslope: a balanced visualization of overview and details on ranks over time

Hao Wang[1], Xingyu Jiang[1], Apurva Nagarajan[2], Xiaolei Guo[1], Lu Ding[1], Dayu Wan[1], Junhan Zhao[3] and Yingjie Chen[1*]

Abstract

Users are often interested in exploring ranks over time data to compare the performance or ranking of multiple observations with respect to each other. However, predominant visualization techniques suffer from a high cognitive load due to visual clutter. We propose Colorslope, a hybrid of Tufte's slope graph and temporal heatmap, to depict ranks over time in one graph while maintaining an overview and details with scalability. Colorslope combines both canonical visualization methods' complementary benefits: depicting overall trends and enabling the estimation of detailed values. We evaluated the efficacy and effectiveness of Colorslope by comparing it with a standard bump chart and temporal heatmap on various data sizes. We conclude that Colorslope contributes by (1) allowing users to identify extremes of the data and rate of change effectively in a relatively large number of series; (2) allowing the visualization to have better scalability in a larger amount of data (e.g., 30 ∼ 50) than a bump chart; and (3) allowing users to gain a better estimate of data values than a heatmap. For a certain size of ranks over time data, Colorslope provides an alternative solution to visualize multiple time series simultaneously that provides both an overview and a certain level of detail.

Keywords: Visualization, Visualization design and evaluation methods, Human-centered computing, Information visualization

1 Introduction

Users are often interested in exploring ranks over time data to compare the performance or ranking of multiple observations with respect to each other. Users examine sports teams' rankings, university rankings, or countries' GDP rankings to make choices or decisions. Many digital collections record ranking information over time. Users often compare the rankings of many observations with respect to each other. For example, when applying to a university, high school students compare many universities' yearly rankings to evaluate the competitiveness of the universities, as well as the university's potential trend (growth) in the future. In sports, fans compare teams' rankings and enjoy their preferred team's gains or losses on the position. In general, people want to obtain their current positions, historical positions, and overall long-term trends. The patterns (up, down, or variability) presented by these overtime rankings provide rich information and can be of great help for the general public's or domain experts' decisions.

Ranks over time data are primarily time-serial data, with most values evenly distributed across the entire range. Occasionally, some items have the same rankings. For comparing multiple ranks over time data, traditional forms such as a slope chart, bump chart, or temporal heatmap have been produced to help better grasp the graph details. Each technique has its unique niche and goal but also carries certain pitfalls.

Bump chart, as one of the predominant visualization techniques, is capable of delivering detailed ranking information over time for all items. However, when there are many ranked items, it suffers from bottlenecks such

*Correspondence: victorchen@purdue.edu
[1] Department of Computer Graphics Technology, Purdue University, 401 N. Grant Street, KNOY Hall, West Lafayette, Indiana 47907, USA
Full list of author information is available at the end of the article

Springer

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

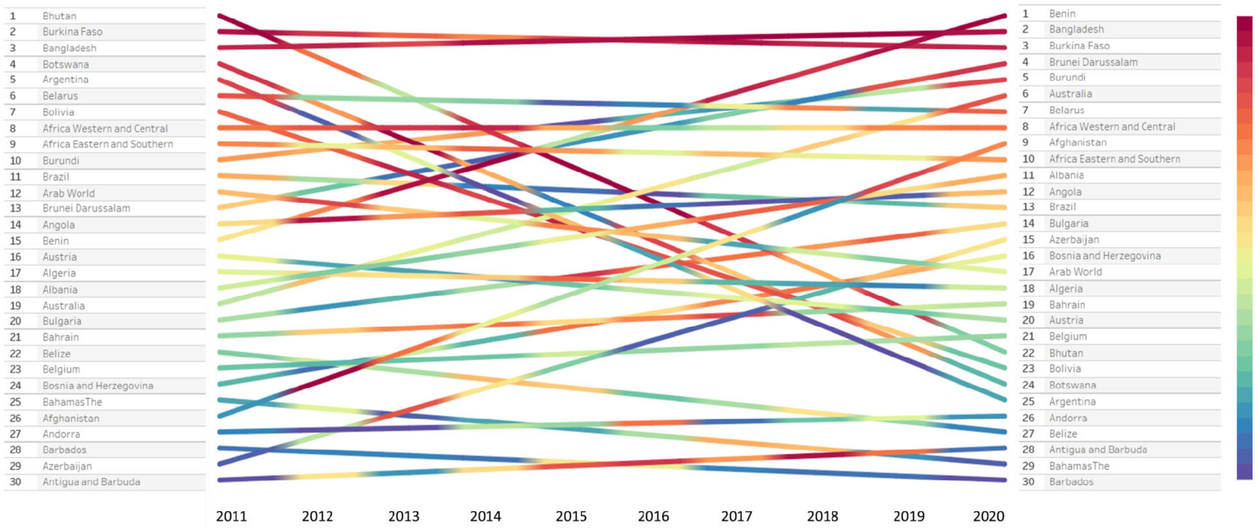


Figure 1 Colorslope to visualize 30 countries' GDP rankings generated by Tableau

as over-plotting, resulting in visual clutter and high cognitive load. According to the work of Saket et al., [1], the threshold of the number of visual marks after a visualization becomes illegible is 50. Apparently, a bump chart is not suitable for the visualization of too many data items (e.g., > 30), even for users to carry out basic tasks [1].

The slope graph [2] provides a simple visual form that is capable of showing a large number of ranking changes. A slope graph demonstrates the trends with precise values at the beginning and ending rankings. A slope graph can depict many lines at once due to its simple form of straight lines. However, the slope graph cannot offer details of temporal change in the middle of the period due to its over-simplified structure.

A temporal heatmap uses horizontal (or vertical) color-encoded lines to depict values over time. This method has linear scalability - a heatmap can add new lines for new data times. However, this method has two major problems. Due to the limitation of human perception, users lack the ability to precisely link color/shade with value and ratio [3], and it is difficult for the user to perceive the trends (or rate of change) from the color. The user has to mentally translate and compute the color change to trends and ratios.

When users encounter ranking data, they often want to find their current position, history positions, and trends. This raises a design challenge for us. How do we reduce visual clutter while preserving the details of the data when visualizing multiple ranks over time simultaneously? The challenge led us to our investigation and we resolved the issues by utilizing a simple and intuitive visual form to enable the extension of visualizing multiple ranks simultaneously and allowing users to quickly obtain enough details of rankings in history. We combined these two conventional methods, slope graph and temporal heatmap, while maintaining the advantages of both.

This paper proposes Colorslope, a visualization technique that aims to simultaneously visualize many ranks over time data. Colorslope projects multiple temporal heatmap lines onto their corresponding slope lines (Fig. 1). Simple straight lines allow Colorslope to display multiple lines simultaneously, with their slopes showing the accurate start, end, and trend. The color coding of these lines shows changes in the middle, which allows the user to estimate values at different times. We further designed and conducted two quantitative experiments to evaluate the effectiveness and scalability of Colorslope, compared to two benchmark-related visualizations: bump chart and heatmap. Our analysis indicates that Colorslope arguably performed better than the chosen visualizations when visualizing multiple observations for visual comparisons. Users of Colorslope remarkably outperformed others in identifying data general and detailed information with accuracy and scalability.

2 Design goals - key qualities

Few studies [4] have listed six key insights that users want to derive from time-series data: trend, variability, rate of change, co-variability, cycles, and exceptions. To identify what key information users want to extract from ranking data, we interviewed eight users and asked them what they care about the most when encountering ranks over time data, including sports, universities, or countries' developments. Here are the several items they all agree with that might be of user interest in the context of comparing ranks of many items:

- T1: Find the start or current (end) rank of one or several observations.

- T2: Find the highest/lowest ranking of observations during history.
- T3: Examine the variability and history of an observation - if the observation always ranks high/low in history or experienced some ups and downs?
- T4: Rate of change - If the rank changes, how fast does one observation improve/decline over time?

Users noted that rank data typically lack significant cycles. Additionally, co-variance is not a concern because people focus more on one item's rank over time. Generally, there do not exist strong associations among different observations in terms of ranking.

To help users see the above information from the data, we identified the following qualities that Colorslope should satisfy:

Macroscopic: One must show the overall trends and patterns of the data rankings. This depicts the overall progression of ranks over time, including the ranking at the start of time and at the end of time.

Scalability: Typically, there are many observations in the ranking data. The visualization should scale well with a relatively large (e.g., 32 soccer teams in Worldcup, 30 NBA teams, top 50 universities) number of items.

Detailability: Detailed historical ranks should be visible. An observation's history ranking can help the user understand the current ranking situation in context and can make better predictions and decisions.

While the above qualities may universally fit different forms of visualizations, neither the bump chart, temporal heatmap, nor slope graph can satisfy the above three qualities. The slope graph does not provide any details in the middle, the bump chart is very poor on scalability, and the temporal heatmap is poor at communicating the overall up/down trend. The design and evaluation of Colorslope were conducted with these key qualities in mind.

3 Related work

3.1 Visualizations for ranks over time

Primarily ranks over time are a kind of time-series data. Time-series data visualization has an extensive history. In 1786, William Playfair applied the line graph to represent time-series data [2]. The line graph and its variants are still the most popular technique for visualizing time-series data [5–7]. One particular line graph variation, bump chart, was specifically designed to depict the evolution of quantitative rankings for categories. As a kind of time-series data, many time-series data visualizations can be applied to ranks over time data.

Many time-series visualization methods have been developed based on the conventional line plots or bar charts. Tufte's sparklines [2] first integrated the line chart with small multiples to review multiple time series. This approach enables the visualization of multiple time series; nevertheless, tasks often involve concurrent series [8]. To address this issue, Heer and his colleagues compared horizon graphs with line charts and found the trade-off curve of speed accuracy based on subjects' estimations of different charts. They later proposed optimal graphical perceptional approaches [9]. Considering that Heer's work was limited to the use of only two time series, Javed et al. [10] conducted an investigation using their braided graphs to compare graphical perception to both line graphs and horizon graphs. In [11], the authors proposed a visualization named CloudLines, which allowed users to detect visual clusters in a compressed view within a limited space. The visualization encoded a sequence of event data on a linear timeline as circles and mapped the importance of the events by adjusting the sizes and opacity of the circles. These methods aim to untangle visual mess by better managing the space.

Other than using x/y locations of visual marks to depict temporal value, colors are another effective way to help users estimate values. A temporal heatmap uses color to present data values over time [12]. Pixel visualizations with color encoding are used to visualize large time-series data. Lolla et al. [13] introduced a simple parameter-light method that allows users to navigate large time-series datasets. Hao et al. [14] generated multi-resolution layouts encoded by color cells. This approach allows users to quickly perceive the importance of data features. Swihart et al. proposed Lasagna plots [15], which use small but constant color blocks instead of lines. These methods all represent time-series data details through colors. However, users will not be able to accurately estimate the value of color due to poor human perception capability of relating colors and shades to values and ratios [3].

Aigner et al. [16] conducted a systematic review of temporal data visualization techniques. They suggested that the visualization of time-series data should consider both representational and perceptional difficulties. In the case of a static graph, line graphs often have scalability issues due to over-plotting [17]. Various statistical aggregators have been used on temporal data [5] to overcome this issue of over-plotting.

3.2 Graphical perception

Mapping from data values to graphical elements [18] is the fundamental component of data visualization. Researchers have examined the impact of visual encoders on users' capacity to interpret and evaluate data portrayed in visualizations through human-subject studies [3]. Graphical perception is defined as users' abilities to comprehend visual encoding and thereby decode the information presented in the graph [19]. Early works mainly investigated the effectiveness and merit of different types of graphical representations in performing visual tasks. Peterson et al. [20] compared the accuracy of reading eight types of common graphs. Simkin and Hastie [21] compared the simple bar chart, and divided bar chart and pie chart, discussing the way in which graph type and judgment type interact to determine the speed and accuracy of quantitative information extraction.

Furthermore, many studies on graphical perception have focused on how visual coding affects human comprehension of the dataset. Bertin [22] revealed the first ranking of visual encoding effectiveness tested on different tasks. Cleveland and McGill [23] conducted a set of elementary perceptual experiments to examine and refine the ranking of different visual variables, making the ranking of visual coding more rigorous and scientific. In the current research, we now have a better understanding of the effects of visual coding variables (i.e., size, position, color) on the accuracy and/or response time of data estimates.

In ranks over time visualization perception, prior studies have investigated several visualizations using a variety of tasks. Javed et al. [19] compared four types of multiple ranks over time visualizations that split or share the space under maximum, slope and discrimination tasks. The results show that shared-space techniques work well for comparisons over smaller visual spans. For color encoding, Correll et al. [24] studied the efficiency of using position or color representations for ranks over time. Their experiment confirms that viewers are better at estimating averages when using color encoding. Albers et al. [25] conducted a set of tasks to compare eight different ranks over time visualizations. The results suggest that color encoding is effective for summary comparisons.

4 Visualization design

We designed the Colorslope technique by taking inspiration from Tufte's slope graph [2] and temporal heatmap. A slope graph contains two vertical axes with straight lines connecting points on the two axes. The first axis presents the starting ranks. The second axis presents ranks at the end (or current). The form is simple but leads to a loss of details in the middle. To solve this issue, we incorporate color encoding into the slope lines to represent intermediate points by using a color scale to reflect the absolute values of the data. Horizontally we put a time scale at the bottom to indicate different periods. Figure 2 illustrates the mapping relationship between the colored line plot (curve) and the Colorslope (straight line).

Colorslope can use any color palette that is meaningful and suitable for a temporal heatmap. Similar to the temporal heatmap, we use a fixed absolute color scale to represent the number of rankings. One color represents one number. As the primary purpose of this research study is to visualize many ranks over time, using a single-hue color palette may make it harder for readers to detect variations on minor scales. Moreover, Colorslope will have many overlapped slope lines. It is necessary to use a color scale that has a significant visual difference. Therefore, we used a color scale from Color Brewer 2.0 [26][1] that has a broad spectrum that ranges from blue to green, orange, and red. The color scale used in this visualization is not color-blind friendly due to its broad spectrum. It is possible to use other color scales, including color-blind safe options. In the experiments, we used a warmer tone to represent a higher value and a cooler tone to represent a lower value. The color does not mean that there are ups and downs. We described this clearly to participants in our experiments.

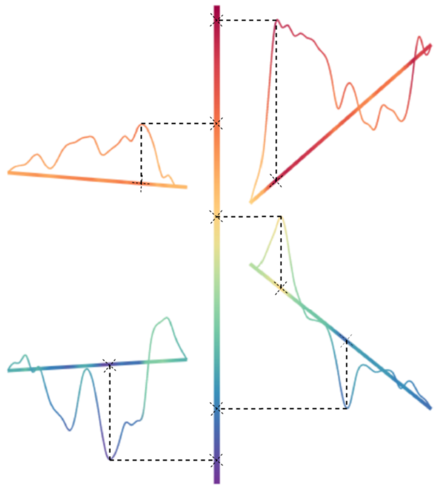


Figure 2 Visual interpretation of Colorslope

4.1 Algorithm

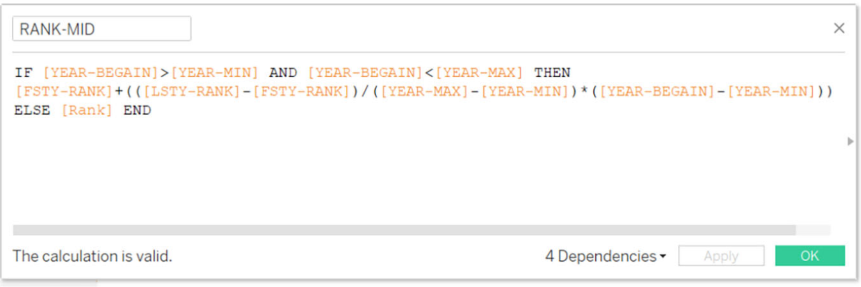
We construct a Colorslope graph using the following algorithm. For each observation, we construct a straight slope line with many segments. Each line's start and end point positions represent the real value of the observation's start and end ranking. Points along the line are assigned with colors representing the observation's rankings at different times. A segment's color is rendered using its two end points' color.

1) For each observation, we have $(0, 1, \ldots, n$ total $n + 1)$ observation points with i_{th} point as $p_i(x_i, y_i)$ (y as ranking, x as time).
2) To make a Colorslope, we adjust p_i's y_i value to y_i' to align all the points on the straight line between the start point $p_0(x_0, y_0)$ and end point $p_n(x_i, y_i)$ while keeping the point's x value. The adjusted y_i' is computed as:

$$y_i' = \frac{(y_n - y_0)}{(x_n - x_0)} \times (x_i - x_0) + y_0. \tag{1}$$

3) Plot these points $p_i(x_i, y_i')$ on the graph.

[1] https://colorbrewer2.org



```
IF [YEAR-BEGAIN]>[YEAR-MIN] AND [YEAR-BEGAIN]<[YEAR-MAX] THEN
[FSTY-RANK]+(([LSTY-RANK]-[FSTY-RANK])/([YEAR-MAX]-[YEAR-MIN])*([YEAR-BEGAIN]-[YEAR-MIN]))
ELSE [Rank] END
```

The calculation is valid.

Figure 3 Apply formula in Tableau (A screenshot from Tableau. The formula text is: "IF [YEAR-BEGIN] > [YEAR-MIN] AND [YEAR-BEGIN] < [YEAR-MAX] THEN [FSTY-RANK] + (([LSTY-RANK]-[FSTY-RANK])/([YEAR-MAX]-[YEAR-MIN])*([YEAR-BEGIN]-[YEAR-MIN])) ELSE [Rank] END")

4) Assign a color to each point p_i based on their corresponding point p_i's ranking values y_i.
5) For each pair of neighbor points $p_{i-1}(x_{i-1}, y'_{i-1})$ and $p_i(x_i, y'_i)$, draw a line segment and assign a smooth color transition using the two end points' colors.
6) Repeat the above for each of the observations.

4.2 Graph interpretation

A user can interpret a Colorslope graph from two perspectives, overall trend, detailed ranks at a time, and special items different from other items.

Trend Reading: Overall trend reading is just reading a regular slope graph. By looking at an item's placement on the left and right axes, users can determine its start and end ranking. Additionally, users can use the slope to estimate the speed of change throughout the time period.

Detail Reading: Users can read the color of the appropriate point on the selected slope for a particular year on the X-axis, use the color legend to confirm the ranking that the color denotes, and then determine the year's rank number. For example, in Fig. 1, if a user wants to check a country's GDP in a particular year, he or she can first locate the point on the slope using its year in the X direction based on the start (left axis) and end (right axis), and then compare the color to the color legend or the colors at left/right axes to obtain the rank data.

Special items: Looking at the sloped lines, users can easily find items that are quickly advancing or declining. Horizontal (or close to horizontal) lines are the ones that do not change much on ranks. Lines going up are advancing while lines going down mean declining in ranks.

4.3 Generating color slopes with Tableau

In addition to programming, it is also able to create a Colorslope graph with Tableau[2]. We first arrange ranks over time data as a table, with rows for observations and columns for different times. Each cell holds the ranking value of one observation at a time. By default, Tableau now has many points with x/y values representing observations' rankings (Y) at different times (X). Simply connecting these points will create a bump chart.

In Tableau, the user can adjust the points' positions by applying a formula. Figure 3 shows the formula in Tableau based on (1). Tableau can then mark each point to a color based on its actual Y value using its built-in color scale. Then we ask Tableau to plot line segments connecting neighbor points. By default, Tableau will assign a gradient color to the segment using its two end points' colors. For the points, the user can choose whether to display them in different sizes and shapes and adjust the line thickness.

In addition to internal color scales, Tableau also allows the user to import an external color scale via its script[3]. Figure 1 shows a Colorslope generated by Tableau with the dataset of 30 countries' GDP rankings from 2011 to 2020 using a color scale from Color Brewer 2.0.

We tested the design of the Colorslope on a 9.7" Ipad in horizontal orientation. The width of the line had to be thick enough to allow users to read color from the slope line. We tested several settings and determined an 8-pixel width (0.77 mm) line width. Thicker made the color reading easier but caused overlaps when there were multiple lines. On other monitors with different resolutions, the pixel width of Colorslope should be adjusted accordingly to obtain the best result.

5 Experiments

We designed two within-subject experiments to evaluate whether Colorslope meets our design goals. The study compared Colorslope to two predominant visualizations for rankings, bump charts and temporal heatmaps (Fig. 4).

The study consisted of a training process and two human subject quantitative experiments. From the experiments, we want to evaluate 1) the user's capability to observe specific ranks (T2), general trends (Macroscopic),

[2] https://www.tableau.com/

[3] https://help.tableau.com/current/pro/desktop/en-us/formatting_create_custom_colors.htm

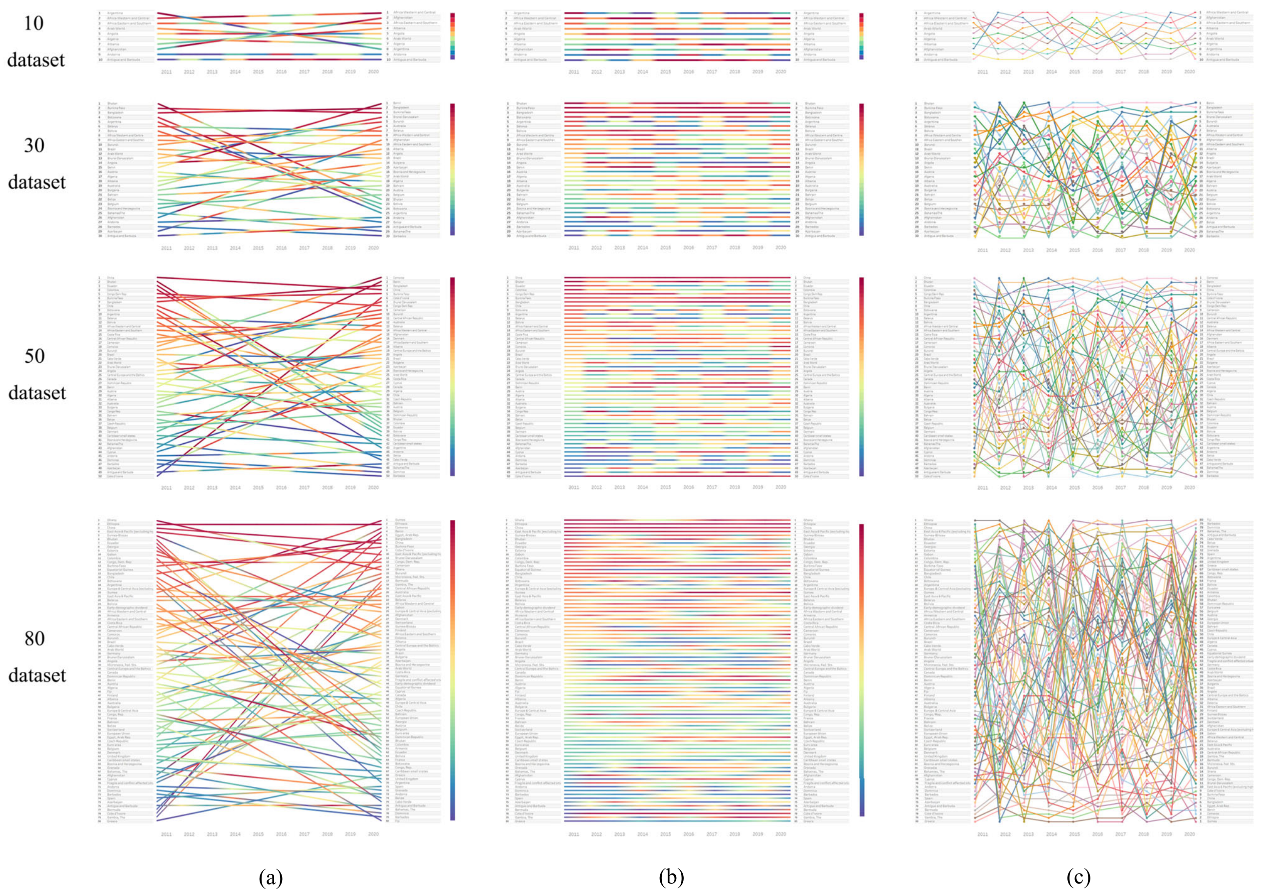


Figure 4 (**a**) Colorslope *vs.* (**b**) temporal heatmap *vs.* (**c**) bump chart to visualize 10, 30, 50 and 80 datasets

and rank variability (T3 and T4) with Colorslope. Additionally, we studied how many data items a Colorslope can handle within a regular display space without causing visual cluttering (Scalability); 2) How accurately a user can read the detailed history rankings from a Colorslope graph (Detailability).

A bump chart can show all the detailed information when there is a small number of ranks, but will suffer visual clutter once there are too many items. On the other hand, a heatmap can scale better by adding more lines (vertically or horizontally) to accommodate more data, but the sole color coding may make it hard for people to see the trend or accurately read details. We hypothesize that Colorslope could help users better read trends and details than heatmap when the number of rankings is more than a bump chart can handle. Hence we conducted two experiments:

Experiment One was designed to make a comparison within bump chart, temporal heatmap, and Colorslope to see their performances in general information with their corresponding scalabilities and certain detailed information. In this experiment, we tried to find whether there were a critical number of data items that could be handled by Colorslope.

Experiment Two was designed to make a comparison between the temporal heatmap and Colorslope and check for accuracy.

We recruited 35 college students (9 undergraduate and 26 graduate, 21 male, 14 female, age 28 ± 4.2 years) from a variety of majors (STEM, finance, and liberal arts). Participants used an iPad (9.7 inch display size) in the experiments.

5.1 Training

At the beginning of the study, we briefed participants regarding the purpose of the research and how to interpret ranks over time data using a temporal heatmap, Colorslope, and bump chart. Participants learned how to construct and interpret a Colorslope graph. Then, the participants were asked to sketch corresponding line graphs based on given Colorslope lines and temporal heatmap strips. They repeated the sketches 5 times to ensure that they understood both the Colorslope and the temporal heatmap. During the process, we superimposed the sketched curve over the original line curve (ground truth). We presented the combined lines to the participants to illustrate how well they could perceive the ranking data from the color-encoded visualization.

5.2 Datasets

In Experiment One, we retrieved GDP data from 195 countries from 2011 to 2020 from Kaggle[4]. The data allow us to arrange many combinations of ranks for the experiments. We randomly selected different numbers (10, 30, 50, and 80) of countries and ranked them based on their GDP.

For Experiment Two, we intended to evaluate how accurately a user can read a Colorslope. To obtain a desired result, the data's value should vary in a large range and with relatively strong variants (ups and downs). Flat or steady incline/decline data will be too simple which may result in a biased outcome. Hence, we chose 10 stock data points with large variations within 30 days of the closing price. We retrieved the data from Kaggle[5].

5.3 Experiment One: macroscopic, scalability, and certain tasks

Experiment One focused on studying the effectiveness of providing the overview and certain tasks. In this experiment, all the visualizations were placed on the same size display screen (9.7" ipad) that fit the display at maximum width and corresponding height. These visualizations were static with no interactions. We also ensured that each visualization had the same characteristics such as line thickness, line spacing, legend scale and size, text size, and font.

In Experiment One, we randomly produced Colorslope, bump chart, and temporal heatmap for 10, 30, 50, and 80 countries' 10-year GDP rankings. Within each size, we asked the participants to read Colorslope, bump chart, and temporal heatmap with the same questions. After answering all task questions in one dataset, we let the users quickly evaluate three types of graphs by telling their subjective feelings of frustration using a Likert scale from 1 to 7.

We asked the participants to maintain a good balance between speed and accuracy. We requested that they quickly estimate the answer using their eye instead of carefully examining the graph and mentally computing the result. Otherwise, a data table would be the most accurate way to answer all the questions.

Find Extremes: For a given country, we asked the participants to determine its best or worst ranks over the period (*Question 1 and Question 4*).

Find High Variability: Variability measures how much the data fluctuate, i.e., the value increases and decreases rapidly. We examine this since high variability means a high risk for prediction. We asked participants to identify the country with the highest variability (*Question 2*).

Find Rate of Change: Rate of change measures how quickly/slowly the data change over time. Users are always interested in the item with the fastest growth rate of rankings. For this task, we asked the participants to identify the country with the maximum increase or decrease during the period (*Question 3*).

Find General Trend: In this task, we wanted to identify the general trend of all countries. As per Few [4], time-series data have three possible trends: either increasing, decreasing, or remaining constant. For this task, we asked participants to identify the percentage of increasing and decreasing countries at a single glance (*Question 5*).

Questionnaire

1) Identify a given country's best ranking during the whole period.
2) Identify the country that has the highest variability.
3) Identify the country that increases/decreases the most from the 1st year to the 10th year.
4) Identify the country that has the lowest ranking in the 6th year.
5) Identify the ratio of countries increasing *vs.* decreasing.

Survey After participants finished the tasks under one data size, we asked them to take a quick survey about their preferences on the three methods of visualization in the test. The following questions were asked:

1) Give us your general feeling and first impression of visual clutter without reading much of the details in the graph.
2) Give us your feelings about understanding the overall variations and up/down ratio. After answering Question 2, Question 3, and Question 5, we asked the user to evaluate the difficulty/frustration level they experienced while answering them.
3) Tell us your feelings about finding out the country's ranking in the given year. After answering Question 1 and Question 4, we asked the user to evaluate the difficulty/frustration level they experienced while answering them using a Likert scale from 1 (comfortable/easy) to 7(frustrated/hard).

5.4 Experiment Two: reading details in graphs

Experiment Two examined how many details users can obtain from ColorSlope (Detailability in the design goals). Experiment Two was also a within-subject study. We quantitatively examined 35 participants' reading accuracy while reading ColorSlope. For detailed information other than a data table, a bump chart provided the best details due to its

[4] https://www.kaggle.com/datasets/zackerym/gdp-annual-growth-for-each-country-1960-2020

[5] https://www.kaggle.com/datasets/ehallmar/daily-historical-stock-prices-1970-2018

nature. Hence, we only compared Colorslope and temporal heatmap. A regular monochrome slope graph is simply not capable of showing any details in the middle.

Participants had already learned how to draw a line graph from a Colorslope line or temporal heatmap during the training (Sect. 5.1). In this experiment, we used data from a total of 10 stocks, with stocks each on Colorslope and temporal heatmap. We asked participants to draw 5 line graphs each from a Colorslope and temporal heatmap on the screen with Apple Pencil. To assist with drawing, a color legend of ranks was placed on the right side of each visualization, and a reference grid was placed in the background. Both visualizations used the same pixel height and width to ensure fairness.

6 Results

To validate the study, we filtered out the participants who spent less than 20% of the average time consumption and missed all the top five answers for visualizations in Experiment One. We believe that such participants do not pay attention to the experiments. Finally, 33 participants' results were kept for data analysis.

6.1 Experiment One

In Experiment One, participants answered a set of five task questions on three different visualizations for four different data sizes (Fig. 4). There is only one correct answer to each question. To analyze, we first defined scoring metrics for measuring the users' inputs. After scoring the participants' inputs and considering the data conditions (including test goals, normality, sample size, and variances), we conducted Welch's t-test on each pair of visualizations. We examined the data to ensure that Welch's model fits the data. We provided a matrix to check for significance between each pair of visualization methods.

Our analysis results indicated that Colorslope outperformed in most comparisons with the limitations of scalability. The specifics are further explained below.

6.1.1 Scoring metrics

Experiment One tested the participants' abilities to identify the countries' GDP with extremes, variability, rate of change, and trends. One ground truth (best answer) existed in each question, with several runners-up close to the best answer. We asked the users to pick the correct country from the group. For each of the questions, we counted the number of correct answers. The maximum possible correct number is the number of participants. Additionally, the user may pick up a different answer that is close to the best answer. We could not treat all not-best answers wrong since answers close to the ground truth are still helpful. The answer to each question can be quantified. If he or she picked the correct answer, the user received a 0 (zero) score. For answers other than the best, the score is the absolute numerical distance of the answer to the best answer. Thus, we obtain a non-uniform scoring system to quantify the answers.

Find Extremes (Q1 and Q4): Ground truth is the highest/lowest ranking. The score for the participant for a treatment group with the corresponding dataset was calculated as the absolute distance of the participant's chosen value from the ground truth. The best score is zero if the person chooses the correct answer.

Find High Variability (Q2): Ground truth is the country with the maximum variance value. A participant's score is computed using the variance value of their chosen country minus the ground truth.

Find Rate of Change (Q3): The country with the most significant change is the ground truth. Each country has a rate of change. A participant's score was calculated using the distance from the selected country's change rate to the ground truth.

Find General Trend (Q5): Each dataset was partitioned into three groups for this task: increased, decreased, and no changes. The ratio of increased *vs.* decreased was recorded as the ground truth. The users were given multiple options, and their scores were counted by subtracting the selections from the ground truth.

6.1.2 Analysis results

We conducted an approximate t-test to compare each pair of visualizations (sample size ≥ 30, moderately skewed distribution is allowed, and there are no outliers or similar shape). For the three visualizations, we had three pairs in total (Table 3). Our hypothesis was that users using Colorslope would produce better results than those utilizing any of the other two visualization techniques.

Using the aforementioned scoring system, we obtained the two matrices of t-test p-values ($\alpha = 0.05$) and statistical analysis results (means and standard deviations).

With 10 datasets, the bump chart generally performed better in the mean of the question scores, but in terms of p-value, there was no significant difference between the three visualization methods, indicating that all examined techniques are clearly legible with this data size.

With 30 $\sim$ 50 datasets, as shown in Table 1, Table 2, and Table 3, Colorslope outperformed on the tasks of identifying rates of change (Q3) and finding the general trend (Q5). This finding confirmed our expectations because of the advantage of the slope graph: precise readings of the start and end points.

Additionally, Colorslope also led the performance of determining extreme values (Q4) at datasets of 50. This evidence may result from the fact that when the data size grows, it becomes difficult for users to read the details from the bump chart due to the clutter. For Q1, we can see that Colorslope does not perform better than heatmap in this same type of task. This may be caused by the fact that when comparing the multiple colors of a specific year, the

Table 1 The number of correct answers/mean and standard deviation for each visualization method, 30 datasets

	Q1	Q2	Q3	Q4	Q5
Colorslope - Mean	26/2.232	7/9.380	27/2.336	25/2.272	28/0.597
Colorslope - STD	1.214	5.852	8.212	4.060	4.161
Heatmap - Mean	27/2.028	1/11.260	13/7.080	26/2.239	10/6.749
Heatmap - STD	1.292	9.616	8.936	5.568	9.723
bump chart - Mean	23/2.742	4/12.270	5/10.763	27/2.014	7/9.365
bump chart - STD	3.544	12.940	8.442	5.197	8.782

Table 2 The number of correct answers/mean and standard deviation for each visualization method, 50 datasets

	Q1	Q2	Q3	Q4	Q5
Colorslope - Mean	24/3.739	6/11.669	26/3.165	22/2.427	27/1.366
Colorslope - STD	1.1	6.732	9.17	4.196	4.566
Heatmap - Mean	24/3.748	1/13.574	9/10.684	18/5.575	8/8.132
Heatmap - STD	1.295	9.814	9.449	5.347	10.322
bump chart - Mean	17/8.133	2/13.950	4/13.384	9/4.739	5/11.852
bump chart - STD	3.412	8.343	9.363	6.628	4.148

Table 3 *P*-values of Welch's *t*-tests on pair-wise comparison, 30 and 50 datasets

		Q1	Q2	Q3	Q4	Q5
30 datasets	Colorslope - Heatmap	0.745	0.171	**<0.05**	0.511	**<0.001**
	Colorslope - bump chart	**<0.001**	0.05	**<0.001**	0.119	**<0.001**
	Heatmap - bump chart	**<0.001**	0.358	**<0.05**	0.568	0.125
50 datasets	Colorslope - Heatmap	0.488	0.181	**<0.001**	**<0.01**	**<0.001**
	Colorslope - bump chart	**<0.001**	0.109	**<0.001**	**<0.05**	**<0.001**
	Heatmap - bump chart	**<0.001**	0.433	0.121	0.716	0.03

temporal heatmap gives a clear version, while some colors will inevitably be covered by other lines in Colorslope. The advantage of Colorslope, however, is revealed because it combines color and slope to provide ranking perception while heavily reducing the clutter caused by bumps.

In the p-value analysis, we can see that with the data size increased to 50, other than Q2 and part of Q1, Colorslope has a significant difference ($p < 0.05$) from the other two methods and Q2 is not significant.

We can see that none of the tested methods is better than the others in helping users determine the highest variability (Q2). Overlooking all similar performances, Colorslope still returned a convincing performance, with smaller means and standard deviations (Table 1 and Table 2). These results correspond to the initial motivation that multiple time-series visualizations overload observers' perception and cognition abilities.

With 80 datasets, all three visualization methods scored low on the task questions except Q3 and Q5 for Colorslope, and the pair-wise p-values are $p = 0.00$. There was no significant difference between other tasks in terms of p-value, implying that this data size is too large for all of them to be legible.

The frustration level results also support our findings. We can see that in Fig. 5, from a scalability perspective, at current settings, bump chart is legible for 10. However, when at or more than 30 datasets, it becomes difficult for users to read details and starts to cause frustration due to clutter. At 50 or 80, it is simply not acceptable. The temporal heatmap has similar patterns to bump chart, but causes less frustration when users read the overall variations and up/down ratio information.

In comparison, Colorslope performs acceptably for 30 $\sim$ 50 datasets. For 50 $\sim$ 80 datasets, Colorslope began to feel cluttered and people needed more effort to read the details. At 80 datasets, people started to feel the graph was on the difficult-to-read side, but still way less than the bump chart and temporal heatmap.

We also used Fisher's exact test to compare the number of correct answers for Colorslope *vs.* other methods on each question. Fisher's exact test is a statistical significance test to examine the significance of association (contingency) between two kinds of classification. Our analysis tells us if the numbers of correct answers are significantly different from each other (the lower the value, the more significant the difference) between the two visualization methods. From Table 4, we can see that at 30 and 50 datasets, for Q1, Q3, Q4, and Q5, Colorslope's numbers of correct answers are significantly higher than the others, and Q2's difference is not significant.

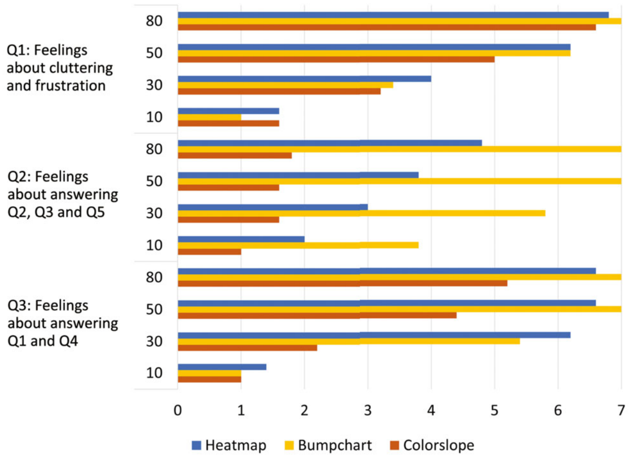


Figure 5 Frustration levels of Colorslope *vs.* others. The lower number is better

Table 4 Fisher's exact tests on pair-wise comparison of numbers of correct answers

		Q1	Q2	Q3	Q4	Q5
30 datasets	Colorslope - Heatmap	1	0.054	0.001	1	0.000
	Colorslope - bump chart	0.072	0.335	0.000	0.556	0.000
	Heatmap - bump chart	0.037	0.357	0.028	0.763	0.405
50 datasets	Colorslope - Heatmap	1	0.105	0.000	0.450	0.000
	Colorslope - bump chart	0.051	0.145	0.000	0.001	0.000
	Heatmap - bump chart	0.051	1	0.128	0.025	0.363

6.2 Experiment Two

In Experiment Two, we asked participants to sketch five line graphs based on Colorslope and temporal heatmap. In essence, this experiment evaluates how accurately users read values from the Colorslope/temporal heatmap each time. We compared the participants' sketches with the ground truth line graph. Each ground truth has 30 data points. The smaller the difference, the more accurately the participants read from the Colorslope/temporal heatmap. In total, we obtained 4950 (5 Colorslope/heatmap × 30 data points × 33 participants) data points. Figure 6 shows sketch samples, mean quartiles, and max/min values of the sketches.

6.2.1 Analysis results

After data processing, we summarized and used box plots to demonstrate the participants' overall performance in 10 visualizations for two tasks. Figure 6 shows the two results from the Colorslope method and the temporal heatmap method, respectively. The red lines represent the ground truth value. For both the Colorslope and temporal heatmap methods, most of the ground truth in red sits within the interquartile range of the observed values. This evidence gives us the confidence to conclude that the average 50% of all observed values of the points are very close (within ± 5%) to the ground truth, and 95% of the values are within 8%.

Furthermore, we conducted a two-sample t-test to compare participants' performance on Colorslope and heatmap (Table 5). We found that the user's estimation from reading Colorslope is consistently more accurate than reading using heatmap (Table 5, p-value < 0.01). Since the start point and end point of a Colorslope line are accurately positioned on the Y-axis, the error of these points on sketches should be close to zero. To make a fair comparison without these endpoints, we extracted the center 20 points to compare the sketches' accuracy on intermediate points in both visualizations. Again, users performed significantly better on Colorslope than on heatmap.

7 Discussion

Our experiments revealed that Colorslope has more legibility than the bump chart and temporal heatmap in general and detail readings, and it has superior scalability when facing a more considerable number of items within an ideal range between 30 $\sim$ 50 datasets. Colorslope performs nicely with less visual cluttering and in some ways much better, e.g., understanding the overall information when the dataset increases. The results met our expectations. The Colorslope graph not only combines the advan-

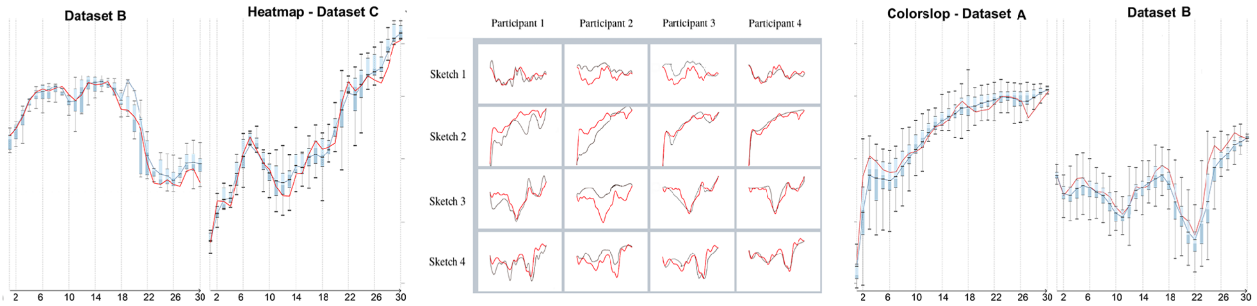


Figure 6 Visual summary of users' inputs (center: sketched curves by participants) with box plots (left: heatmap, right: Colorslope) shows the drawing range at each temporal point. The red line is the ground truth. On the left and right graphs, the thick blue segments indicate that 50% of people marked the error range, while the thin blue lines mark the 95% error range

Table 5 Reading accuracy (error) of Colorslope *vs.* Heatmap

	Observation of all 30 points		Observation of middle 20 points	
	Top 50%	95%	Top 50%	95%
Colorslope	±3.77%	±7.11%	±3.981%	±7.9%
Heatmap	±4.09%	±10.21%	±4.466%	±12.18%
	$P < 0.01$			

tages of both slope graph and temporal heatmap but also enhances both usage and effectiveness.

In Experiment One of reading general information with scalability, we found that the effectiveness of reading ranking data is related to the size of the dataset. When the data quantity was too small (e.g., 10 datasets) or too large (e.g., 80 datasets), all examined approaches performed well or poorly enough to make a no-noticeable difference. When the data volume is in the range of 30 to 50, the difference becomes significant.

Users can use Colorslope to quickly determine the rate of change (Q3) and the general trend (Q5) because they only need to identify the line that tiltes the most. This is actually pre-attentive since human perception is sensitive to tilts and parallels. On this task, other visualizations did not perform well. In a temporal heatmap, users must first determine the start and end points of each country, and then either imagine lines between two points or compute the value differences. With many items in a graph, checking and comparing colors back and forth between two ends makes it cognitively challenging to process the change. A bump chart with many fluctuating lines is too cluttered to be legible.

Equally important, as the data size increased, we found that Colorslope significantly outperformed others in finding extremes (Q1 and Q4). Finding the maximum/minimum value at a given time may be obvious with a bump chart, but when asked to search a specific data series in the set, the user has to track the line. When many lines are clustered together, the user can easily get lost or frustrated. The advantage of Colorslope against heatmap may indicate that the user did not solely rely on color to find the extreme value. With Colorslope, the user can use the combination of the vertical position and color, first identifying a small range of lines based on vertical position and then using color-encoding to aid in the identification of the extreme values.

None of the methods perform well in identifying variability (Q2). Identifying variability is a cognitively intensive task that requires complete attention and a less cluttered display. Although users may gain a general idea of variability, it is difficult to derive the statistical value, especially when considering many items.

In terms of scalability, at a larger number of datasets, Colorslope outperforms the bump chart on all three survey questions. Due to the straight lines, it is not surprising that users feel Colorslope is less cluttered. Additionally, because of its pre-attentive capability for both slope and color, Colorslope performs well on survey Q2 on all four numbers even when there are many (50 or 80) data points. For a heatmap, the user has to visually query and compute the locations of the start, end, and middle points to obtain the trends of particular items. Since the positions of the bump chart nodes make it simpler to identify the relevant rankings than the color of the spectrum, the bump chart should perform better than Colorslope in recognizing a ranking. However, the node's advantage is mitigated by bump chart's visual cluttering with the increasing data size. The temporal heatmap seems clutter-free as the data amount grows. Nevertheless, without slope, it requires extra effort from the user to compare the color changes between the ends of the horizontal line when answering questions about change rate, making its scores on survey Q2 better than bump chart but less than Colorslope.

In Experiment Two of reading details, Colorslope provides better accuracy than the heatmap, although they are both using the same color scale and line width. Although both Colorslope and heatmap incorporate color encoding, Colorslope takes advantage of color, position, and line tilt. Evidently, users did not need to guess where the lines began and ended from the Colorslope. Colorslope's start/end values are double encoded by position and color, which makes it easier for the user to distinguish colors in other midway regions. Users can compare the colors within the graph directly instead of consulting the legend on the side. Position and tilt help to reduce the range of values, which in turn reduces the number of color spectrum comparisons. The end points, which correctly correspond to a color on the color scale, can also be identified by their position. The outcomes of this experiment confirm our findings from Experiment One in relation to the recognition of extremes and rate of change.

7.1 Pre-attentive visual features

Colorslope takes advantage of two pre-attentive visual features to enable rapid information reading by humans. The slope of lines is the most evident that a user can easily identify any data items that are going up, down, or unchanged in the overall trend. In addition, users can quickly see items that go with a similar trend (parallel). The 2nd pre-attentive feature comes from the color. When one or a small number of series drop in a group of series data performing similar trends, these outliers will show different colors, which makes them evident and pre-attentive. For example, in the Colorslope figures from Fig. 4, we can see multiple places where one part of the line's color is different from its surroundings. This hints to the user that these parts have abnormal increases or declines in a given time period compared to their neighbor lines, even though such lines share similar trends with the other slopes.

By adding color encoding to the slope graph, we depicted details of temporal changes. Furthermore, when adding colors into the context of vertical location, users are better able to comprehend the values represented by different colors. Therefore, combining two simple visualization methods can greatly empower both and expand the field of use.

7.2 Generalization and limitation

In addition to representing ranks over time, Colorslope could be extended to visualize other types of time serial data that carry the same characteristics, e.g., stock values, ticket box sales, and growth of regions' population. However, there are certain limitations to this type of data, such as having the data evenly distributed along values to avoid overlapping.

While inheriting the pros of slope graphs, Colorslope also inherits limitations from traditional Slope visualization. For example, when the start value or end value is very close, then the slope segments will tend to overlap or be cluttered together. Colorslope is not suitable for visualizing periodical time-series data, e.g., daily temperature data, since users will rely on color change to depict cycles. If data are cyclic with an up/downtrend, the color changes will be even harder to track.

In Colorslope, a line needs to carry certain width (e.g., 0.77 mm/8 pixel for the 9.7" Ipad screen used in this experiment) to make sure users can read the color correctly. The wider the line, the more accurate the color can help a user read, but it will also increase the chance of visual clutter or lines overlapping. The designer needs to balance the number of items, the display space, and the line width to achieve the best effect.

Colorslope can use different color palettes. A broad color spectrum (e.g., from blue to red) is recommended for the static figure in order to provide a more significant visual difference so that readers can differentiate values easily. However, this color scheme is not colorblind-friendly. We surely can employ a color vision impairment safe palette in Coloslope, but typically such a color palette has limited color ranges. For a small dataset, a color-blind-safe pallet should work well since the color difference between ranks may be significant enough. For a dataset with many items, it may make the users difficult to differentiate between two items that rank next to each other. Add interaction to the graph, e.g., mouse over to see detailed ranks, could potentially solve the problem.

Our experiments did not allow interactions, and the users only looked at static images. The result will be different in situations where a graph can be interactive, e.g., with highlight/filter/zoom. Especially the highlight/filter interaction could make reading any of the visualizations much easier and more accurate.

8 Conclusion

Incorporating both slope graph and temporal heatmap, we proposed a design of temporal visualization, Colorslope, for simultaneously visualizing ranks over time data that could provide both an overview and a certain level of detail. We carried out two quantitative human-subject experiments to evaluate Colorslope against a bump chart and temporal heatmap. The current predominant traditional bump chart and the heatmap graph both experience various problems, from visual clutter to heavy cognitive load once there are many ranking data in the display. We found that at a relatively large number of ranks (e.g., $30 \sim 50$), Colorslope provides satisfactory performance on overview general trends for comparing multiple ranks over time and shows the detailed ranks at an acceptable degree of accuracy.

Colorslope takes advantage of the benefits from the pre-attentive visual features from both slope graphs and colored heatmap. Additionally, the double encoding (color and position) of end points reduces the user's cognitive load while reading the graph. We believe for temporal data which share similar characteristics and needs as ranks over time, Colorslope can contribute as an alternate visualization choice that could provide both overall trends with a certain level of detail.

Acknowledgements
We express our gratitude to all the individuals who participated in our human-subject study for this project, and to the members of the Purdue Intelligent Visualization and Interaction Lab (PIVIL) for their support.

Funding
This work was supported by the Department of Computer Graphics Technology at Purdue University and the Purdue Intelligent Visualization and Interaction Lab (PIVIL).

Availability of data and materials
The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request. The GDP data was obtained from Kaggle[6].

Declarations

Competing interests
The authors declare that they have no competing interests.

Author contributions
All authors contributed to the study's conception and design. The first draft of the manuscript was written by AN and JZ. HW and XJ collected data and conducted the experiment. AN, XG, LD, and JZ contributed to the analysis and manuscript preparation. HW carried out the data analyses and finalized the manuscript. All authors revised previous versions of the manuscript and have read and approved the final manuscript. YC conceived the original concept and supervised the study. JZ assisted in supervising this study.

Author details
[1]Department of Computer Graphics Technology, Purdue University, 401 N. Grant Street, KNOY Hall, West Lafayette, Indiana 47907, USA. [2]Google Inc., Mountain View, CA, USA. [3]Medical School, Harvard University, Boston, MA 02115, USA.

Received: 13 December 2022 Revised: 7 February 2023
Accepted: 3 March 2023 Published online: 01 June 2023

References

1. Saket, B., Endert, A., & Demiralp, Ç. (2019). Task-based effectiveness of basic visualizations. *IEEE Transactions on Visualization and Computer Graphics*, *25*(7), 2505–2512.
2. Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire: Graphics Press.
3. Ware, C. (2004). *Information visualization: perception for design* (2nd ed.). San Francisco: Elsevier.
4. Few, S. (2009). *Now you see it: simple visualization techniques for quantitative analysis* (1st ed.). Oakland: Analytics Press.
5. Sorenson, E., & Brath, R. (2013). Financial visualization case study: correlating financial timeseries and discrete events to support investment decisions. In *Proceedings of the 17th international conference on information visualization, IV '13* (pp. 232–238). Los Alamitos: IEEE.
6. Statman, M. (1987). How many stocks make a diversified portfolio? *Journal of Financial and Quantitative Analysis*, *22*(3), 353–363.
7. Zacks, J., & Tversky, B. (1999). Bars and lines: a study of graphic communication. *Memory & Cognition*, *27*(6), 1073–1079.
8. Hochheiser, H., & Shneiderman, B. (2004). Dynamic query tools for time series data sets: timebox widgets for interactive exploration. *Information Visualization*, *3*(1), 1–18.
9. Heer, J., Kong, N., & Agrawala, M. (2009). 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). New York: ACM.
10. Javed, W., McDonnel, B., & Elmqvist, N. (2010). Graphical perception of multiple time series. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 927–934.
11. Krstajic, M., Bertini, E., & Keim, D. (2011). Cloudlines: compact display of event episodes in multiple time-series. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2432–2439.
12. Zhou, L., & Hansen, C. D. (2015). A survey of colormaps in visualization. *IEEE Transactions on Visualization and Computer Graphics*, *22*(8), 2051–2069.
13. Kumar, N., Lolla, V. N., Keogh, E., Lonardi, S., Ratanamahatana, C. A., & Li, W. (2005). Time-series bitmaps: a practical visualization tool for working with large time series databases. In H. Kargupta, J. Srivastava, C. Kamath et al. (Eds.), *Proceedings of the 2005 SIAM international conference on data mining (SDM)* (pp. 531–535). Philadelphia: SIAM.
14. Hao, M., Dayal, U., Keim, D., & Schreck, T. (2007). Multi-resolution techniques for visual exploration of large time-series data. In K. Museth, T. Moeller, & A. Ynnerman (Eds.), *Eurographics/IEEE-VGTC symposium on visualization*. Eindhoven: The Eurographics Association.
15. Swihart, B. J., Caffo, B. S., James, B. D., Strand, M., Schwartz, B. S., & Punjabi, N. M. (2010). Lasagna plots: a saucy alternative to spaghetti plots. *Epidemiology*, *21*(5), 621–625.
16. Aigner, W., Miksch, S., Müller, W., Schumann, H., & Tominski, C. (2007). Visualizing time-oriented data-a systematic view. *Computer Graphics*, *31*(3), 401–409.
17. Keim, D., Kohlhammer, J., Ellis, G., & Mansmann, F. (2010). *Mastering the information age – solving problems with visual analytics*. Eindhoven: The Eurographics Association.
18. Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, *229*(4716), 828–833.
19. Javed, W., McDonnel, B., & Elmqvist, N. (2010). Graphical perception of multiple time series. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 927–934.
20. Peterson, L. V., & Schramm, W. (1954). How accurately are different kinds of graphs read? *Audio Visual Communication Review*, *2*(3), 178–189.
21. Simkin, D., & Hastie, R. (1987). An information-processing analysis of graph perception. *Journal of the American Statistical Association*, *82*(398), 454–465.
22. Bertin, J. (1967). *Sémiologie graphique*. Paris: Gauthier-Villars.
23. Cleveland, W. S., & McGill, R. (1984). Graphical perception: theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, *79*(387), 531–554.
24. Correll, M., Albers, D., Franconeri, S., & Gleicher, M. (2012). Comparing averages in time series data. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1095–1104). New York: ACM.
25. Hernandez, J., Paredes, P., Roseway, A., & Czerwinski, M. (2014). Under pressure: sensing stress of computer users. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 51–60). New York: ACM.
26. Harrower, M., & Brewer, C. A. (2014). ColorBrewer.org: an online tool for selecting colour schemes for maps. In A. J. Kent & K. Field (Eds.), *Landmarks in Mapping: 50 Years of the Cartographic Journal* (pp. 184–200). Leeds: Maney Publishing.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[6] https://www.kaggle.com/datasets/zackerym/gdp-annual-growth-for-each-country-1960-2020