Social Impact of Time Series Visualization

An Interactive Qualifying Project submitted to the faculty of

Worcester Polytechnic Institute

February 23, 2017

By:
Cuong Nguyen
Charles J. Lovering

Submitted To:
Professor Gabor Sarkozy
Worcester Polytechnic Institute

Advised by:
Professor Gabor Sarkozy
Professor Elke Rundensteiner
Rodica Neamtu
Ramoza Ahsan

Contents

1 Abstract

In this Interactive Qualifying Project we explore the social impact of providing visual technology for supporting time series mining. Based on literature research, we develop a visual analytics system for time series mining. Our system enables users to explore and interact with time series datasets, while also offering guidance for parameter tuning and for selecting similarity measures. Together the powerful interactions and the rich visual displays empower users to find insights in time series datasets. Built as a web service, the system increases accessibility to public datasets. Evaluation based on user studies with over 400 subjects as well as interviews with domain experts led to improvements in user experience and insight into the social impact of time series analysis.

2 Executive Summary

In § 4, we explore the motivation of building this system, give detailed description of data mining concepts that are critical for understanding the context of the system, and explore potential visualization techniques as well as related time series analytic systems. We then explore more in-depth the ONEX technology in § 5. We explain the design of our system as well as the methodology leading us to such design in § 6. Next, we show a case study of how the system is used in a real-word scenario in § 7. Finally, we describe our procedure for user studies and interviews and their results in § 8.

3 Introduction

We are entering an age where business, research, and medical fields are fueled by staggering amount of data. It is increasingly difficult to understand and interpret this exploding amount of information. This is true not only for massive corporations trying to understand the nature of their consumers, but also for helping students find the best subject to study, and for new start-ups trying to discover new markets and better locations to work at. There is daily need for people to easily access and understand this data. While the titans of industry have experts working to untangle the deep hidden meaning of data continually being spooled together, not everyone has the resources to do so.

This dire need for people to be able to easily understand patterns and information present in data is the critical reason we undertook this project. One type of data becoming incredibly pervasive is time series data. Time series data is simply data ordered over time. Nearly every domain, field, and industry generates immense amounts of time series that hold a wealth of information. In addition to being helpful for large companies, this kind of data is also useful for medical researchers and practitioners. For example, understanding ECG classification, and reactivity to drugs over time could help in diagnosing and reading heart condition.

Here is a real-time example illustrating the crucial role that time series data plays. In 2013 in Massachusetts the MHTC organization set out to repeal the Sales and User Tax on computer and software services because it was perceived that it had a negative impact on the economic health of the state. Vast amounts of data in form of time series was studied and analyzed. It was found that there were similarities between tax rates and fluctuations social and economic factors obtained from a wide range of public governmental websites including the Tax Policy Center [1], the Census Bureau [2], and the Bureau of Economic Analysis [3]. The capacity to find and interpret the similarities between these diverse economic indicators represented as time series were central to this process. This process was incredibly arduous. The analysts were faced with numerous difficulties and needed to overcome a variety of challenges. We will detail these in § 4.1, but in short the analysts were inhibited by not being able to use simple, traditional similarity methods, having to handle data from different domains, and not being able to analyze this huge amount of data. Our work provides an answer to some of these challenges.

As alluded to before, this also applies to individuals. In fact, the need for more intuitive and powerful visualizations and explanations of time series data may be even more vital for individuals. Anyone who ignores statistical patterns, historical trends, or even recent tax rates is going to have a hard time making informed decisions. Not analyzing time series data when making important decisions and incorrectly understanding the time series data can be costly. Our work aims to help diminish this opportunity cost for the average person by increasing the accessibility of an easy to use data analytics system. It is difficult to enumerate the ways in which being able to analyze time series can help an aspiring student or modern-day worker. But it is also important to also consider the communal cost for a person not being able to analyze time series data. With further review, this cost becomes ore apparent. Education and awareness operate like "herd immunity" in our modern society. If portions of the population do not connect time series patterns and thus the real effects policy on key *results* like unemployment rates and health care costs, it becomes increasingly difficult to constructively improve upon these policies. Without an intuitive understanding of the effects of policies in the social and economic space, peoples' *gut opinions* - while perhaps noble and righteous, are completely blind. This applies to every political spectrum and ideology; we must examine data and patterns over time to have hope for positive, constructive policy. It is for this reason that easily accessible and understandable visualizations of data are invaluable. Not everyone has background in time series analysis, and this is completely understandable and expected. However, it is possible to make understanding of time series comparisons and analysis easier for everyone.

We created a powerful platform with an intuitive, interactive and easily-accessible web-based interface that will enable people with limited training to better understand time series analysis. Specifically, it helps people compare time series with different lengths and misalignment. It allows people to visually compare sub-sequences of time series. This way people can find a sequence of time series data that is similar to another one even if it is stretched or shifted. For example, if one were to start a small business in the software industry, one could study tax and employment trends of the states that successful startups incubated in - and then search for the most similar trends among states and countries today. This shows how it is useful to compare across different sections of time. However, it is important to note that our system can also search for the shape of the pattern. Specifically, if one were to examine the tax and employment rate environment that propelled Slack or AirBnb to success and searched for a time period of five years, the system could find a similar pattern even if it spans over three or seven years. Again, this applies across domains, and is critical to interpreting time series similarity in a sensible manner. We examine this in greater detail below in § 4.2.

We have already demonstrated this system at *2016 IEEE MIT Undergraduate Research Technology Conference* [4]. Here we were able establish the unique and interactive nature of our work, showing how a refinement of a well-designed architecture (6), real data [5], and well-researched and experimental graphing techniques (4.4) improve understanding of time series for experts and novices alike. A demonstration paper of our work has also been accepted at the leading academic conference Sigmod2017 [6]. The system has helped direct work on the underlying comparison algorithm's research (5) and access to the system has been requested by several other research groups to help understand new datasets.

Thus, the overarching goal of this interactive qualifying project is to build a system for time series comparisons and exploration. This system bridges the growing disparity between the volume of time series data produced and the current capacity of domain experts to understand this data by augmenting the power of ONEX technology [7] with an intuitive, interactive and feature-rich interface. We then gauge the impact of such system on societal domains by evaluating it with the MATTER dataset and users with background in data mining and human computer interaction. This system has potential to have a large societal impact; if our local state government is able to better make decisions based on time series using statistical and proven methods, demonstrated via clear visualizations and powerful analytics we can all directly benefit. We have already seen how this can be applied to policy tax decisions, but we can also see how this could be applied within the context of the environment. We can show demonstrably how the temperatures and climates are changing by analyzing time series that describe these behaviors in a negative, and unprecedented manner. Analysis and visualization of this nature may make the difference in our very own lives, and without doubt the lives of our children. Yes, this has already been done, but clearly we need a certain amount of repetition and perhaps presenting it with further intuitive visualizations can help persuade more fringe minorities. In the age where the importance of data is increasing exponentially, so too must its sanctity, and a step in this direction is creating a space to easily analyze and visualize it.

This work impacts vital and diverse of domains. Our project is focused on building a system to enable time series understanding, and determine how people in general understand such data. We researched time series and how people can understand them via various visualization techniques, implementing a system to help people understand similarity between time series - a critical task for time series analysis; and finally, evaluating this platform on its utility and intuitiveness through feedback from interviews and user surveys. From this strong base, we look to future works to extend and apply our findings.

4 Background

4.1 Motivation

The amount of time series data generated and the demand for fast time responses are both rapidly increasing. This demand for immediate summary and analysis is present in almost all domains including: finance, astronomy [8, 9], computational biology [10], medicine [11], etc. In fact, many enterprises generate trillions of data points: hospitals generate trillions of data points in both EEGs and ECGs [12]. In many domains, fast analysis is an important factor for success or even viability. In stock market trading, small fractions of seconds can result in millions lost or gained. In a medical setting, for diagnosis and treatment immediate similarity processing of a patient's ECG can have a tremendous impact. These examples suggest a need for applications that integrate actionable insights into time series datasets. Furthermore, these applications should incorporate intuitive interfaces to aid a wide range of audiences from diverse domains.

4.1.1 Massachusetts' Policy

As we mentioned above in § 3, in 2013 the Massachusetts High Technology Council (MHTC) led a movement to repeal the Sales and User Tax. Analysts worked to ascertain the negative effects that this tax could cause. However, they faced numerous difficult challenges. Analysts had to answer complex questions that were not always based on traditional similarity searches. For example, they had to look for recurring similarity patterns in the growth or unemployment rate of a state over a few years. This type of question is not easily answered because the sheer amount of data makes it difficult to process in a timely manner. The presence of data from different domains reported over specific intervals required comparisons of time series of different lengths and alignments, as the impact of a tax change might play out with different time durations. Again, there was a dire need to be able to compare and analyze specific parts of time series of different lengths.

During this process analysts used specific indicators, like the growth rate, to evaluate the potential impact of introducing a new tax. For example, they looked to find similarities between the growth rates of the states having already implemented such a tax. Furthermore, they handcrafted a sample growth rate time line indicative of a positive impact of the tax and searched for matches among all states. As it is unlikely that analysts will *luckily* guess the length of the pattern, there is a high demand for the capacity to compare sequences of different lengths. Furthermore, these comparisons can be complicated so an intuitive interface to explore the similarity would greatly assist this type of analysis.

MATTERS datasets: We showcase our system on a complex, real dataset, namely MATTERS, created by the MHTC. The Massachusetts Technology, Talent, and Economic Reporting System website maintains data from widespread, trusted, websites of various domains such as technology, talent and economic metrics from all 50 U.S states. The system aims to provide policy makers with crucial information for comparing states to develop better policy to attract and retain business; also, the policy should improve the economical health of the state in general. Our project uses datasets extracted from MATTERS to evaluate the impact of our application in further leveraging the data exploration process, especially for time series data.

4.2 Time Series Data

A **Time Series** X is an ordered list $X = [x_1, x_2, x_3 ... x_{n-1}, x_n]$ where each x_i is a value; i.e. a time series is a list of data points in a specified order over time. We are often interested in specific parts of time series, called **subsequences**. A **subsequence** $\mathbf{X}_{i,k}$ is the time series starting at the i^{ith} data point in time series X with length k. For example, a time series could be a the stock values of Microsoft over a year, and an analyst may be interested in a specific subsequence of that time series, perhaps the stock value from June to July.

4.3 Dynamic Time Warping

We first describe Euclidean distance, a ubiquitous distance measure for time series. Euclidean distance is fast and simple, but this simplicity prevents it from capturing the subtleties in comparing time series. It is a pairwise operation; it directly compares corresponding points and does not consider the shape of the time series directly.

To find the **Euclidean distance** (ED) of two time series of equal lengths Q and C, we calculate the square root of summation of the squared difference between each data point, defined as:

$$\sqrt{\sum_{i=1}^{n}(Q_i - C_i)^2} \tag{1}$$

Fig. 1 offers a visualization of this function. The strength of the ED is in its simplicity; it is has a linear complexity.

However, this method is brittle and fails to handle time series that are misaligned on their temporal axis. As seen in Fig. 2, although the two time series in Fig. 2a hold a resemblance in their shape, the misalignment in their peak leads ED to consider them

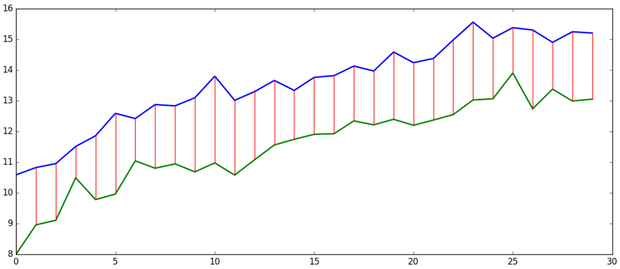
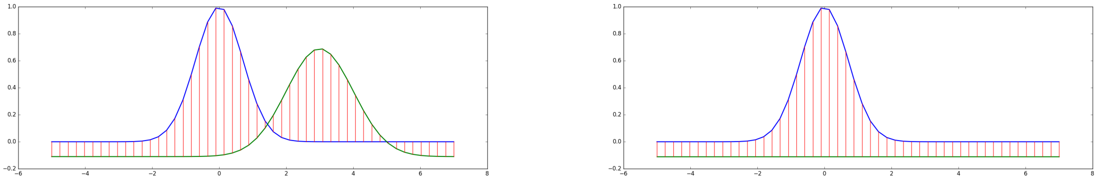


Figure 1: Visualization of Matched Sequences using Euclidean Distance

as completely different. There is a lower similarity than that of the two time series in Fig. 2b, which consists of one time series from Fig. 2a and a horizontal line. Another inherent weakness of ED is that it cannot compare time series of different lengths.



(a) Euclidean distance: 3.11

(b) Euclidean distance: 2.70

Figure 2: Example of Euclidean's inflexibility

To address these problems, **Dynamic Time Warping** (DTW) algorithm aligns the two time series so that their similarity, despite being distorted along the time axis, is better captured. Intuitively, instead of performing a rigid one-to-one mapping from a data point in one time series to its corresponding data point in another time series, the DTW algorithm maps multiple consecutive points in one time series to a single point in the other time series in such a way that it minimizes the difference between the two time series. This is captured in the visualization of the matching in Fig. 3.

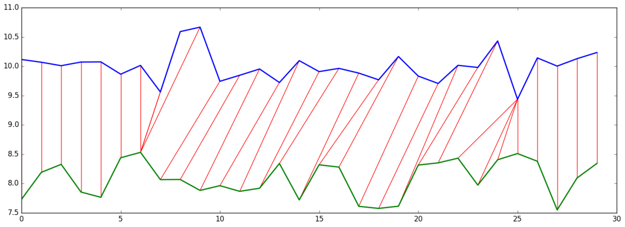
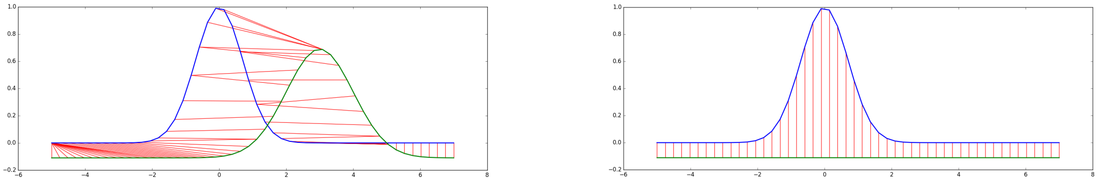


Figure 3: The mapping produced by Dynamic Time Warping algorithm

In Fig. 4, the red lines indicate a more flexible mapping between the time series in that it maps the two peaks with each other, resulting to a more meaningful distance. DTW compares the similar parts of time series sequences.



(a) DTW distance: 0.79

(b) DTW distance: 2.70

Figure 4: Example of DTW capturing the similarities of two misaligned time series.

To produce such elastic measurement, DTW stretches and compresses the sequences to best match each other. This allows time series of different lengths to be compared. Suppose we have a time series Q of length m and another time series C of length n. We construct an m-by-n matrix D where $D_{i,j}$ corresponds to the alignment of Q_i and C_j and equals to the square difference of Q_i and C_j. In other words,

$$D_{i,j} = (Q_i - C_j)^2 \tag{2}$$

To find the optimal mapping between two time series, we search for a path through the matrix, from $D_{1,1}$ to $D_{m,n}$, that minimizes the cumulative sum of all values along the path. This path can be found by using the following recurrence relation

$$F_{i,j} = D_{i,j} + min(F_{i-1,j}, F_{i,j-1}, F_{i-1,j-1}) \tag{3}$$

Where $F_{i,j}$ denotes the minimum cumulative distance up to cell (i, j) and $F_{m,n}$ holds the resultant DTW distance. Elements outside the bound of the matrix D are set to 0 to complete the base cases of the recurrence. At each step of this calculation, we memorize which of the three preceding cells is used so we can trace the path after the calculation. Fig. 5, illustrating an optimal path found in a DTW matrix, shows that the relation (3) also implicitly poses the following constraints: 1) An optimal path always starts at $F_{1,1}$ and ends at $F_{m,n}$. 2) For every data point in one time series, there always exists a mapping to at least one data point in the other time series. 3) In each step of forward calculation, the path can only go upward, to the right or upward and to the right simultaneously.

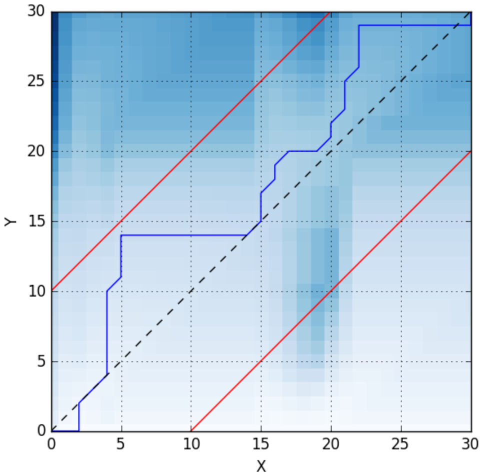


Figure 5: The Warping Path Matrix

Dynamic Time Warping is one of the most efficient distances for time series classification [13]. Its complexity is $O(mn)$, where m and n are the lengths of the compared time series. This is greater than ED's linear complexity. This additional complexity challenges efficient mining, especially when searching for a similar time series in a large dataset. Many pruning strategies and optimization techniques have been proposed throughout the years. As shown in the Fig. 5, for example, one can limit the search to a window - in this case that restriction is 10 and demonstrated by the lines parallel to the diagonal. This helps speed up the algorithm because less values need to be computed and avoid pathological paths. Along with this strategy, a large number of candidates in a dataset can be pruned away if the DTW distance between them exceeds an appropriate chosen lower bound [14]; other techniques include early abandoning and cascading lower bound effectively speeds up searching time series in very big datasets [13]; the state-of-the-art ONEX approach, transfers most of the computational time to its one-time clustering phase using the cheap Euclidean distance, can perform DTW-based searching at a phenomenal speed [7]. ONEX is the data mining technique that powers the application proposed in this project.

4.4 Visualization Techniques for Time Series Analysis

Our purpose in researching and creating visualizations for time series analysis is to express the results of the similarity comparisons. Different techniques highlight similarity in different ways. Furthermore, clear visualizations help illustrate the underlying ideas more clearly than raw numbers. This will help experts and novices alike.

4.4.1 Multiple Line Charts

The primary graphing technique we use is very basic. We will plot a query, a user defined subsequence of a time series Q, and its closest match in a user selected dataset C. With restricted space, a simple linear plot provides very high understanding [15]. As our platform focuses specifically on comparing two time series, and currently uses dynamic time warping as our primary measure, we enhance the graph with lines that connect data points in Q and C. These lines, dashed in Fig. 6, are determined by the warping path of DTW. Being able to discern information and patterns from data will increase the social impact of this application.

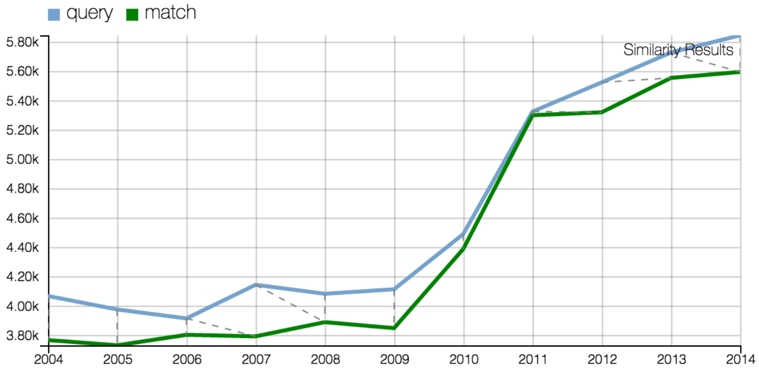


Figure 6: Two time series plotted on a line chart. Correspondence of points produced by DTW is visualized by dashed lines.

4.4.2 Radial Charts

Radial visualization is the technique of displaying data in a circular pattern [16]. One example of such visualization is the radar chart, which plots each variable of multivariate data onto different axes starting from the same point. Fig. 7 shows an example of such chart.

For time series data, we can view each data point as a variable and plot them likewise on a radar chart. We call a radar chart plotted with time series a “radial chart”, since having a big number of axes makes it more circular and the time series looks as if it emerges radially outward from the center of the circle. Visualizing time series with radial visualization is common practice, especially in periodic patterns detection [16].

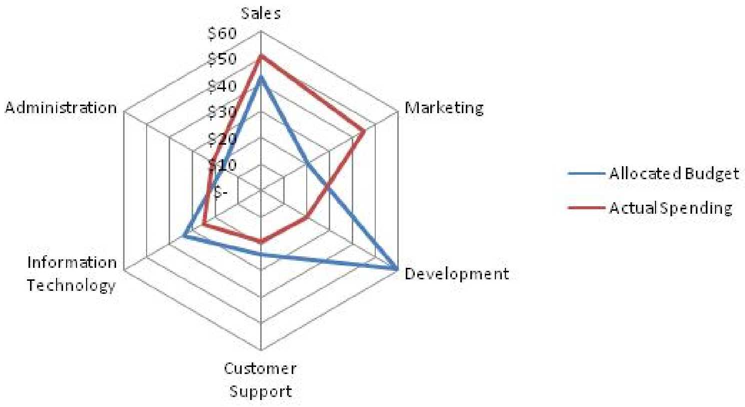


Figure 7: An example of radar chart (Illustration from Wikipedia)

In many datasets, radial charts reveal new insights of time series comparisons. For example, the Leaf dataset [12] contains a collection of shapes extracted from digital images of leaf specimens. These shapes are stored as sequences of data points specifying the distance from the center of the leaf to points on its edge, thus revealing a meaningful picture only when plotted on radial chart. Fig. 8 shows an example from the Leaf dataset.

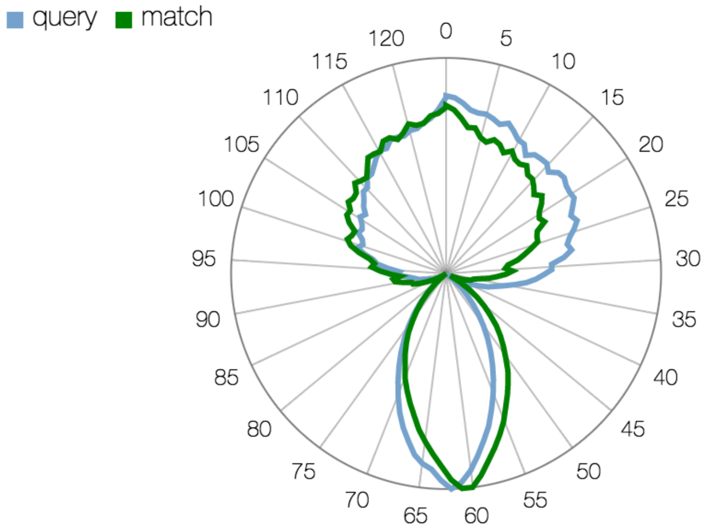


Figure 8: The shape of two leaves are plotted on a radial chart for comparison.

4.4.3 Connected Scatter Plots

A connected scatter plot (CSP) shows how the data changes over time, and highlights the how similar respective values of the warping path are. Essentially, the CSP is a scatter plot. A scatter plot displays data points by, usually, two different attributes. For example, each data point is displayed by one attribute value on the x axis, and the the other attribute on the y axis. For a CSP, the attributes are the values in the matching subsequences. Each data point consists of the values of corresponding points from the warping path. To maintain the ordering of a dataset one connects the points with directed arrows. Although this concept is often unfamiliar to analysts, there are studies showing that people quickly grasp how understand this graph quickly [17]. Intersections have additional meaning: both time series have similar repeating patterns. This occurs because the only thing that dictates where a point is in the graph are the values of the corresponding points. CSP also measures similarity as the distance from a $45°$ diagonal. This is due to the fact that X and Y axis are the values of the two time series Q and C so if a point (i, j) is on the diagonal, then the values of the data points Q_k and C_k where k is the index of the data point in the match, and thus they must be equal.

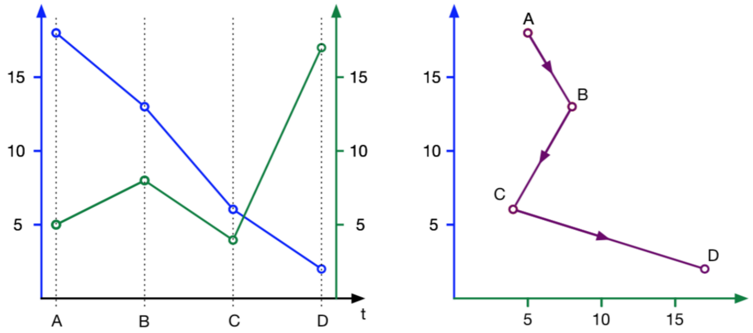
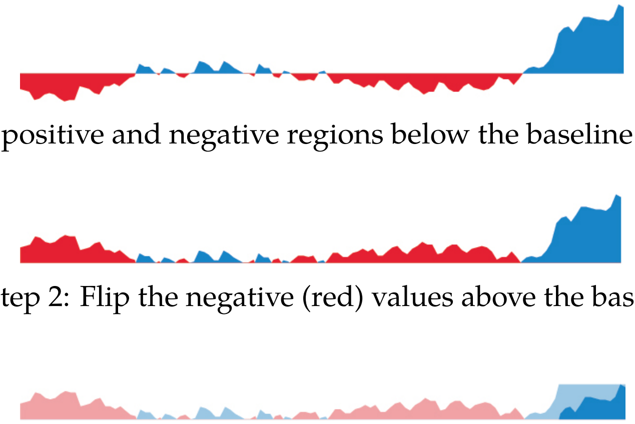


Figure 9: Corresponding graphs of multiple time series (left) and a connected scatter plot (right). (Illustration taken from [17])

4.4.4 Horizon Graphs

Horizon Graphs are an extension of the regular line graphs. These graphs work to maximize the **data density** which is the amount of information encoded per pixel of display space. They do so through a clever folding and coloring technique. After determining a baseline, generally 0 - all positive curves are colored one color, generally blue, and all the negative curves below the baseline are colored red. All negative curves are flipped over the baseline. This reduces the size of the graph by half, maintaining 100% of the data. The idea is to use different color to represent what would have used

100% increased space. The process of creating folds in the data can be repeated as shown in Fig. 10. The question then is how well do users understand these graphs? There have been extensive user studies investigating how well users interpret horizon graphs. Its generally been found they perform as well or nearly as well as line graphs for larger charts, and greatly outperform line graphs when allotted smaller regions of space [15]. There is also strong evidence that more than two additional folds quickly results in deteriorating comprehension, but for quick references the 1-fold and 2-fold horizon graphs are excellent candidates for tight spaces. They can be utilized effectively when comparing multiple time series [18].



(a) Step 1: Color both positive and negative regions below the baseline (respectively blue and red in this example)

(b) Step 2: Flip the negative (red) values above the baseline.

(c) Step 3: Repeat steps 1 and 2, creating more folds and further compressing the data. Note that while this step is not necessary, but can be repeated more than once.

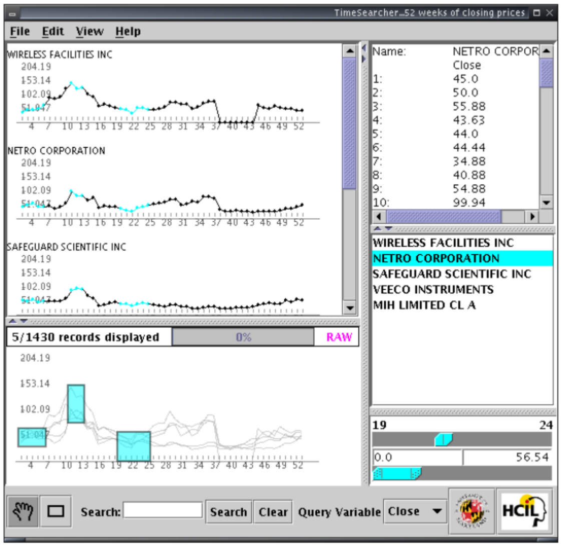
Figure 10: How to construct horizon graphs (Illustration taken from [15])

4.5 Related Work on Time Series Analytics

We investigated many time series analytic tools and describe them below.

4.5.1 TimeSearcher 1

TimeSearcher 1 [19] uses the concept of **timeboxes** as the primary tool for querying a subset of time series. Specifically, timeboxes are rectangular regions created by clicking and holding at a point then dragging and releasing at another point on a time-series-plotted canvas. Let (x_1, y_1) and (x_2, y_2) respectively be the coordinate of the upper-left corner and the lower-right corner of the rectangular region; only time series with all data points that lie in the temporal range $[x_1, x_2]$ and simultaneously have values in the range $[y_1, y_2]$ are displayed. One can also construct multiple timeboxes to specify a conjunctive query (AND relation) as shown in Fig. 11.



Name:	NETRO CORPOR
	Close
1:	45.0
2:	50.0
3:	55.88
4:	43.63
5:	44.0
6:	44.44
7:	34.88
8:	40.88
9:	54.88
10:	99.94

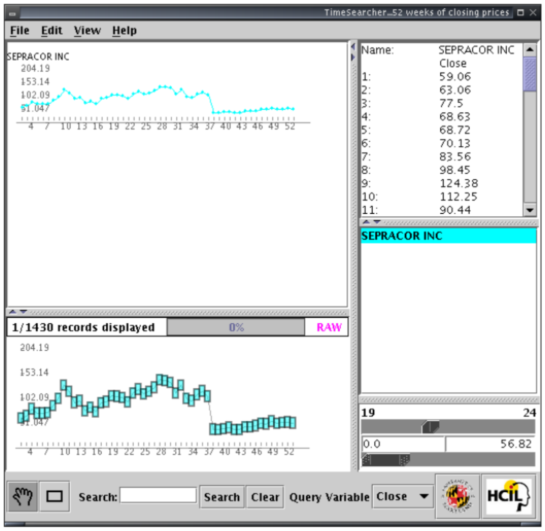
WIRELESS FACILITIES INC
NETRO CORPORATION
SAFEGUARD SCIENTIFIC INC
VEECO INSTRUMENTS
MIH LIMITED CL A

5/1430 records displayed

Figure 11: Interface of TimeSearcher 1 with 3 constructed timeboxes (Illustration taken from [19]).

Moreover, the conjunctive query can be exploited further to perform operations such as query-by-example. As shown in Fig. 12, dragging a time series from a dataset and dropping to the query window automatically generates an array of timeboxes that approximates the selected time series. Consequently, this action is akin to querying for time series in the dataset that are similar to the selected time series.

TimeSearcher 1 is innovative for its timeboxes concept, which is versatile and simple, both to implement and to use. The tool, however, lacks the ability to perform similarity comparisons between time series of different lengths and temporally not-aligned. Timeboxes are also sensitive to outliers: having only one data point outside of a timebox can have a high impact on how the similarity is evaluated.



Name:	SEPRACOR INC
	Close
1:	59.06
2:	63.06
3:	77.5
4:	68.63
5:	68.72
6:	70.13
7:	83.56
8:	98.45
9:	124.38
10:	112.25
11:	90.44

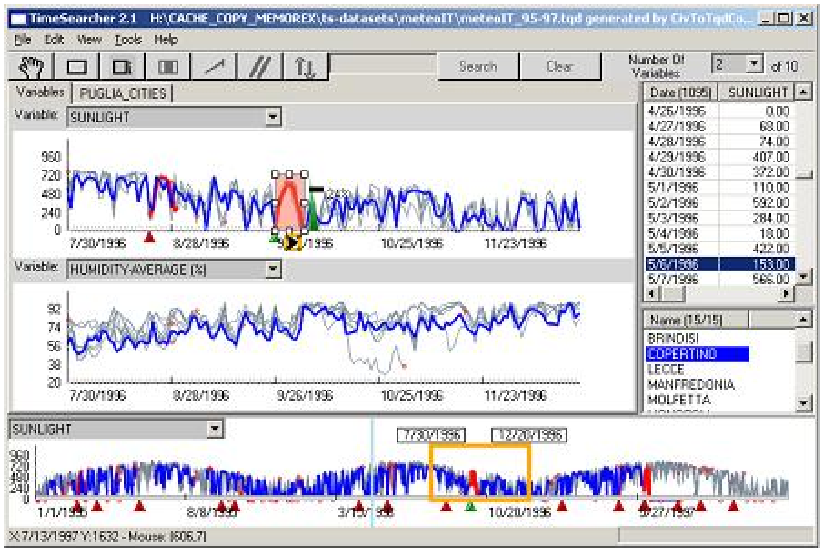
SEPRACOR INC

Query Variable: Close

Figure 12: Finding similar time series in the dataset using an array of timeboxes (Illustration taken from [19]).

4.5.2 TimeSearcher 2

TimeSearcher 2 [20], as the name suggests, is a successor of TimeSearcher 1. It inherits the timebox tool from TimeSearcher 1 and extends it for pattern finding in a long time series. Fig. 13 shows the interface of TimeSearcher2. This application uses a procedure called three-step interactive search. In the first step, users use timeboxes to extract a subset of interested time series. Multiple timeboxes can be specified to perform a conjunctive query. In the second step, users select a time series in the dataset then draw a single box to highlight a specific pattern. After that, the initial search is completed with the chosen pattern being used as a query and the scope is limited to the subset in step 1. In the last step, users adjust similarity tolerance by dragging a slider appearing next to the pattern box. TimeSearcher 2 uses ED measure similarity. The strength of this tool searching procedure, its use of ED is its responsiveness, as discussed in § 4.3.



Number Of Variables: 2 of 10

Variable: SUNLIGHT

Variable: HUMIDITY-AVERAGE (%)

SUNLIGHT

Date (1095)	SUNLIGHT
4/26/1996	0.00
4/27/1996	68.00
4/28/1996	74.00
4/29/1996	407.00
4/30/1996	372.00
5/1/1996	110.00
5/2/1996	592.00
5/3/1996	284.00
5/4/1996	18.00
5/5/1996	422.00
5/6/1996	153.00
5/7/1996	566.00

Name (15/15)
BRINDISI
COPERTINO
LECCE
MANFREDONIA
MOLFETTA

Figure 13: User interface of TimeSearcher 2. Illustration taken from www.cs.umd.edu/hcil/timesearcher/

4.5.3 KronoMiner

KronoMiner [21] is designed to facilitate interactive exploration and modification of multiple subsequences for fine analysis, rather than multiple long time series. Its main feature is a multi-foci hierarchy tree represented by a circular sequential layout, shown in Fig. 14. The entire dataset is plotted in the central ring. A region of interest (ROI) can be selected by directly brushing on any existing series. This action creates a child segment in relation to that series. Repeating this action builds a hierarchy tree of segments where segments further down the tree have more refined view of the time series. The ROI of each segments can easily be moved, expanded or removed.

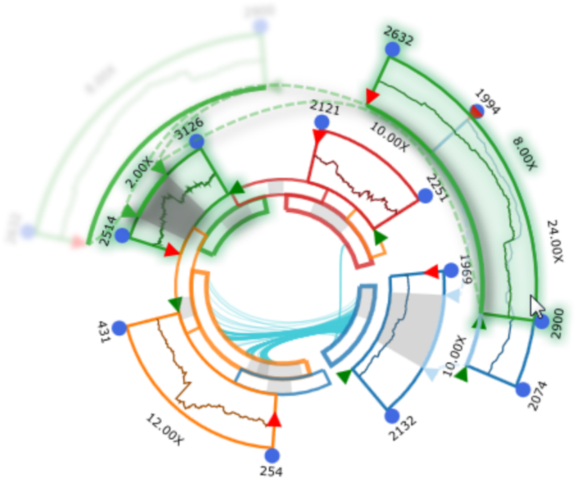


Figure 14: Circular layout of the multi-foci hierarchy tree of KronoMiner. (Illustration taken from [21]).

Another feature of KronoMiner is finding the Best Match. In Best Match mode, the current selected segment is used as a query pattern. When users hover on another segment, it is used as a target and an arch is drawn to link the query to its best match on the selected target as shown in Fig. 15.

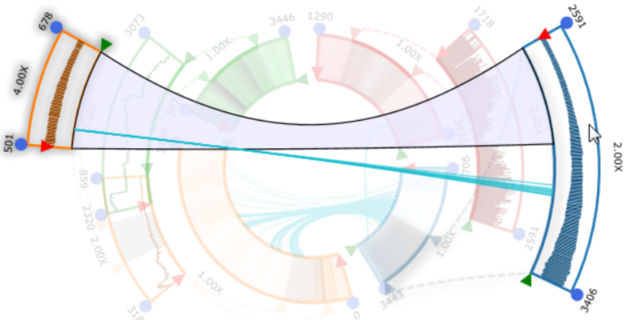
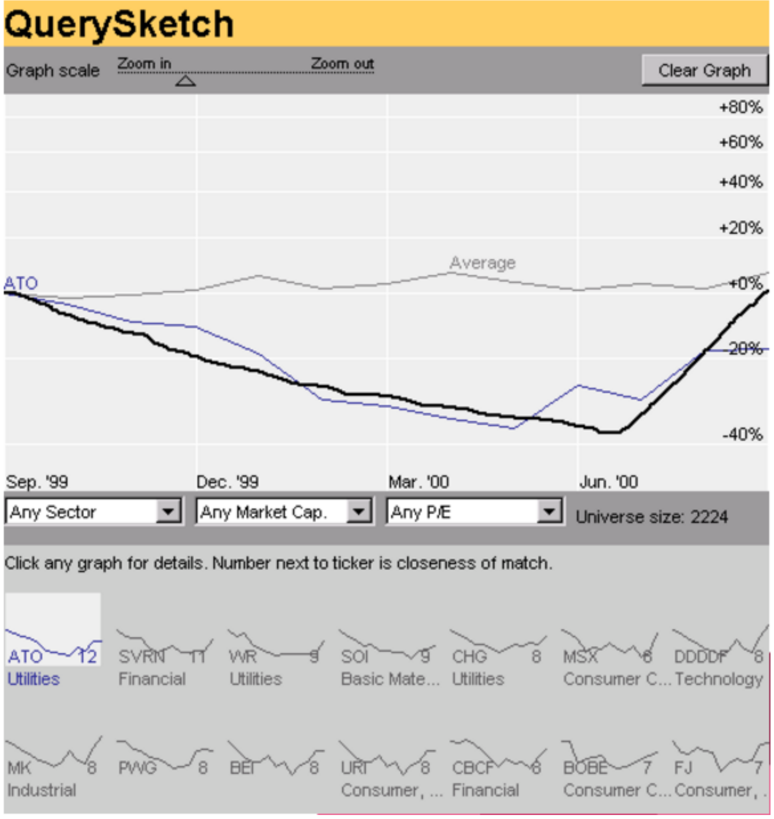


Figure 15: KronoMiner in Best Match mode. An arch is drawn to link the query to its best match in the targeted series. (Illustration taken from [21]).

KronoMiner excels in its clever use of circular hierarchical view to solve the multi-resolution view problem of a large dataset. Its combination of colors, glowing effects and adaptive fading in and out provides an engaging and visually pleasing experience. On the other hand, although its best match mode is no doubt useful in multiple scenarios, it still lacks robustness when it comes to misaligned time series.

4.5.4 QuerySketch

QuerySketch [22] provides an easy-to-use interactive time series searching tool by allowing users to freely draw a line graph as a query. In [22], the author tested the tool on a historical stock price dataset. Everytime a user finishes his drawing action by releasing the mouse, a search is performed and the results are displayed at the bottom part of the interface. For example, Fig. 16 shows the companies whose stock price declined from September 1999 and made a return in 2000. This capability to produce queries graphically and interactively as opposed to entering digits in a text file unlocks a huge amount of freedom for users. Nonetheless, the matching algorithm used in this tool is based on the Euclidean distance, which limits the tool to only searching for time series with a fixed length.



Any Sector

Any Market Cap.

Any P/E

Universe size: 2224

Click any graph for details. Number next to ticker is closeness of match.

Figure 16: Results from QuerySketch after drawing a query.

5 ONEX

Our system utilizes ONEX, Online Exploration of Time Series, is developed by a research group at WPI. It is an answer to the question posed by the difficulty of comparing time series. Its a high-performance system with high accuracy and speed for computing time series comparisons [7].

Time Series Comparison algorithms have now largely fallen into the camp of Dynamic Time Warping [23]. As illustrated in § 4.3, Dynamic Time Warping (DTW) is critical for being able to correctly compare time series of different lengths, as alignments and stretching of patterns inhibits the faster Euclidean Distance (ED). Even still, there are applications that solely rely on ED, and have to pay the cost of accuracy and thus, utility [24, 25]. For various reasons explained in § 4.3, DTW has difficulty as time series get very long and datasets large. Essentially, it is a quadratic algorithm and as such, it becomes increasingly more expensive as the size of the data increases. There have been very impressive improvements and optimizations to this algorithm as described by [13]. ONEX improves upon these impressive speeds by using the novel idea of performing a one-time preprocessing step that clusters the data based on critical similarity relationships and then saving these relationships compactly. ONEX increases the speed of the subsequent similarity comparisons to a near real-time performance. This approach has been proven to have an increased matching accuracy of up to 20% while being also several times faster than the state-of-the-art similarity search tools [7, 13]. As alluded in § 4.3, INSIGHT uses ONEX as its algorithmic suite for computing similarity matches. The end-result matchings are based on DTW, whereas the internal clustering is based on ED.

5.1 Related Works & State of the Art

Most other systems face the trade-off between accuracy and time response especially when dealing with high volumes of data. In the face of this difficulty, some systems provide an exact or a highly accurate solution [26, 27, 28] at the expense of responsiveness. Others [29] use preprocessing steps to improve the timely responsiveness, but their requirement for setting many different parameters limits their efficiency [13]. Clearly, a dilemma exists between choosing complex similarity distances and time responsiveness, namely, while shorter time responses are guaranteed by the use of fast-to-compute distances like the Euclidean Distance [30], such distances cannot handle sequences with different time alignments. Meaningful comparisons of such sequences require the use of time-warping distances like DTW, whose computational complexity [31] leads to slow responsiveness and poor scaling as data grows.

5.2 Algorithmic Outline

5.2.1 Pre-Processing

The dataset is first preprocessed. This begins by grouping every subsequence of every length into groups. The key idea is that every subsequence in a group will be similar to each other. Specifically, all the sequences in a group will be within half the threshold from the representative of that group. The algorithm iteratively builds up groups by going through all the subsequences, creating a new group when a time series doesn't belong to one of the current groups. These comparisons are performed using normalized Euclidean Distance, which works very quickly - and well, as only subsequences of equal length are compared.

5.2.2 Similarity Search

To find the best match, ONEX explores the similarity groups by comparing the representative of each group to the query. It selects the group with the most similar representative, and then searches within that group for the most similar time series. Both of these comparisons use the elastic DTW so it can compare the query to the representatives of different lengths.

5.3 Results

Tables 1 and 2 show the results comparing ONEX and Trillion highlighting speed and accuracy [7, 13]. The datasets are public, and used commonly benchmarking time series comparison algorithms.

	Italy Power	ECG	Face	Wafer	Symbols	Two Pattern
ONEX	0.01	0.024	0.028	0.042	0.176	0.109
Trillion	0.04	0.063	0.11	0.189	0.439	0.585

Table 1: Search Times

	Italy Power	ECG	Face	Wafer	Symbols	Two Pattern
ONEX	97.77	99.48	97.82	97.87	97.2	99.2
Trillion	82.97	74.58	71.87	87.67	96.99	88.04

Table 2: ONEX Accuracy

6 Methodology

6.1 Goals

Before designing and implementing the system, we set concrete goals for our system. They include:

- Create a system for time series exploration. Specifically, the system, given a time query or a sub-sequence of a time series, finds the most similar sequence in a dataset within a threshold. The sample query and the result can be different in length.
- Provide intuitive interface to enable any type of user to utilize this system. Analysts should be able to use the interface with little guidance.
- Visualize the similarity between two time series, specifically to give insight into how DTW determined the similarity between the sequences.
- Leverage intrinsic structure of ONEX to provide additional context to the result.

6.2 Overview

Fig. 17 shows the components that comprise the system, and the data flows through these components. The flow starts with an expert analyst uploading her normalized time series dataset to the system. The data then goes through an offline preprocessing step of ONEX, where it is clustered into multiple groups using the cheap Euclidean distance. Once the groups are ready in the memory, the expert analyst can start querying the dataset. There are two types of queries: **Similarity Search** and **Seasonal Similarity**. Similarity search looks for a similar time series in the dataset to a provided query time series. Seasonal similarity helps finding various patterns in a time series. Both of these queries use dynamic time warping as their similarity measure. The whole system appears to the analyst as a friendly and intuitive interface. Here, the analyst can effortlessly perform dataset grouping, perform the two mentioned queries and receive the result in diverse kind of plots.

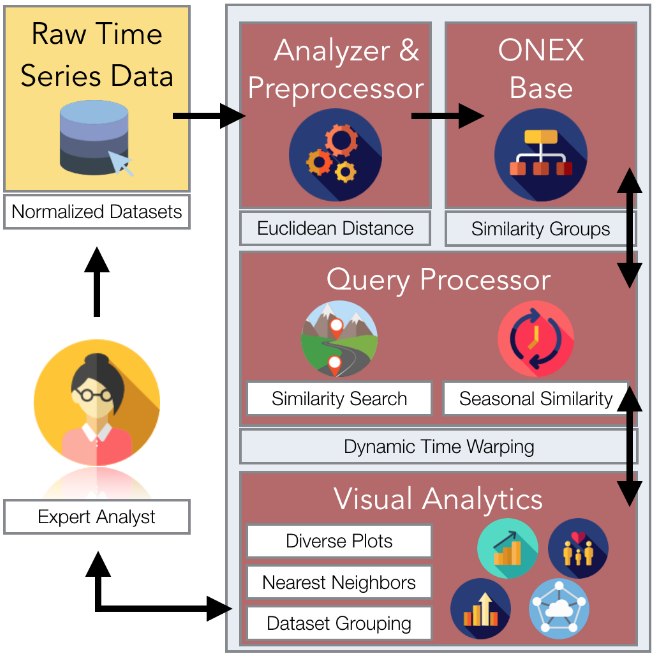


Figure 17: Components and data flow inside the system.

6.3 Design and Implementation

Now we describe the system and its funcitonality. Fig. 18 shows a high-level view of the architecture of the system. We use a client-server architecture. The system is built as a web application and consists of two parts: a server and a client. The server is run on a more powerful machine so that it is able to perform computation at a better speed; whereas the client can be run on average personal computers since it does not have to do any heavy work. The server and the client communicate with each other over the network via a RESTful API [32].

6.3.1 The Server

As shown in Fig. 18, the server contains two components: ONEX and the router. ONEX the algorithmic engine we utilize for the similarity search, is written in C++. However, the router, built on top of the light-weight Flask [33] web development framework, is written in Python. To resolve this difference, we use the Boost.Python library that enables interoperability between Python and C++. The router provides a set of API to facilitate communication between the client and the server. For example, a typical communication is shown in Fig. 19. In this example, the client sends to the server an HTTP request GET /dataset/list, saying that it needs a list of names of the available

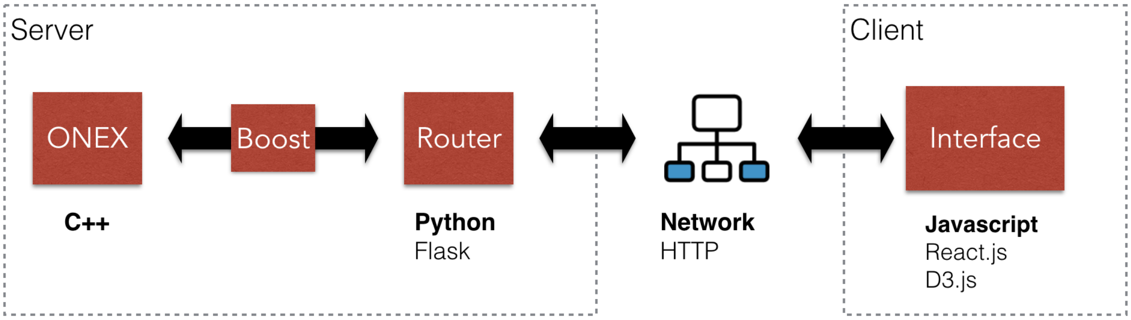
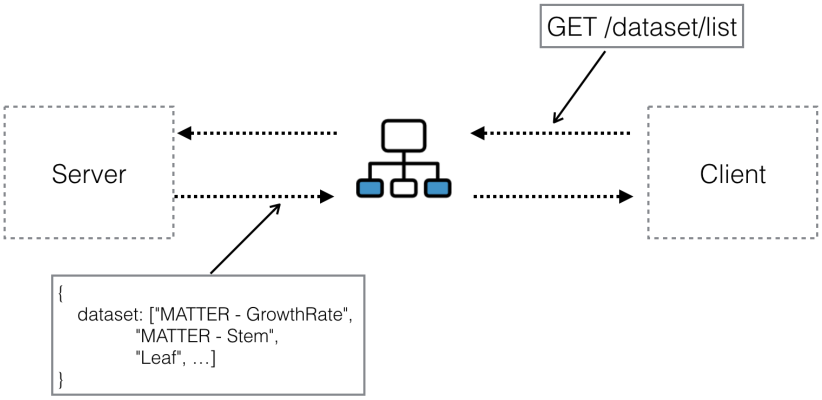


Figure 18: Architecture of the system.

datasets. The server responses with a JSON object containing an array of names of available datasets.



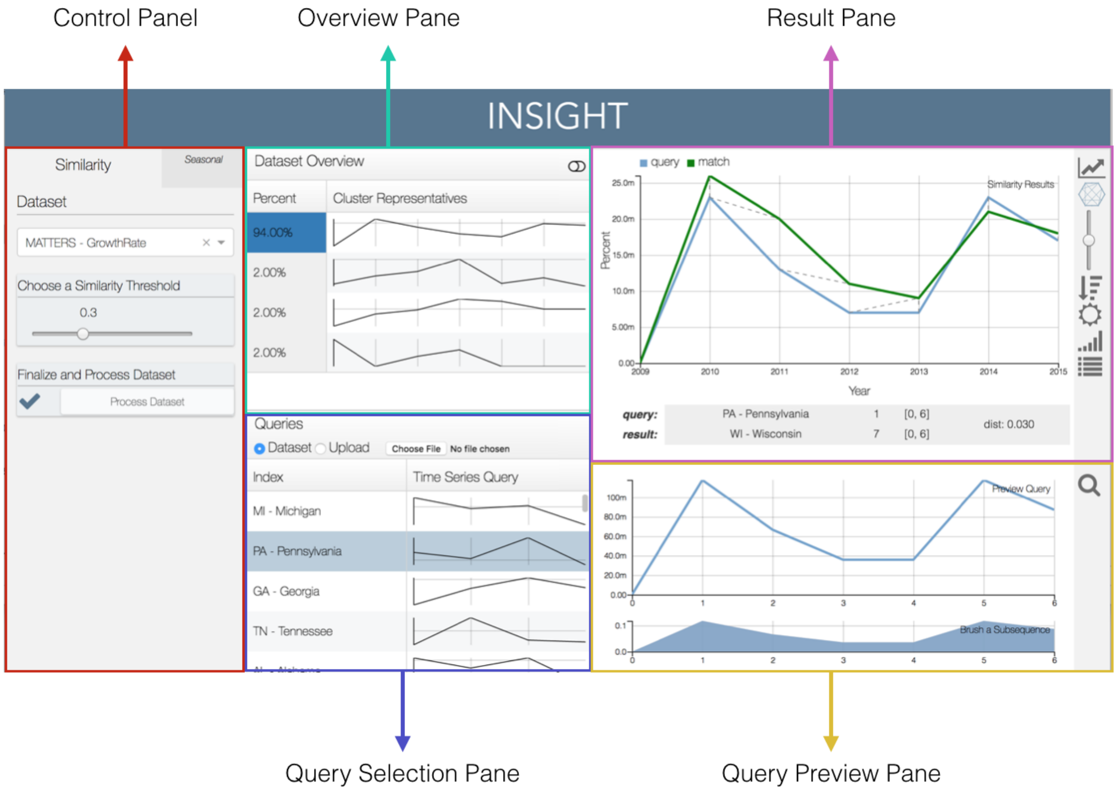
```
{
   dataset: ["MATTER - GrowthRate",
             "MATTER - Stem",
             "Leaf", ...]
}
```

Figure 19: An example of a client sending a request to the server and the server responding.

6.3.2 The Client

The server architecture and implementation fulfils the system's first goal. The client side of the system, the user interface, fulfills the other goals. We begin by designing the interface with Shneiderman's widely accepted mantra [34] in mind: ”overview first, zoom and filter, details on demand”. This principle acts as a framework for designing interface of an application dealing with large datasets. The interface using the React [35] library from Facebook to render our different components and D3.js [36] for constructing the graphs.

As mentioned earlier, the system enables two types of exploratory operations: Similarity Search and Seasonal Similarity; which are accessible via the interface's two view modes: Similarity view and Seasonal view. Fig. 20 shows the interface in Similarity view with annotations. On the far left is the Control Panel, where users can select a dataset, choose a similarity threshold and start processing the dataset. The right side of the interface is partitioned into 4 panes with well-defined functionality. Now, we go over each pane in counter-clockwise order.



Dataset: MATTERS - GrowthRate

Choose a Similarity Threshold: 0.3

Finalize and Process Dataset

Percent	Cluster Representatives
94.00%	
2.00%	
2.00%	
2.00%	

Index	Time Series Query
MI - Michigan	
PA - Pennsylvania	
GA - Georgia	
TN - Tennessee	

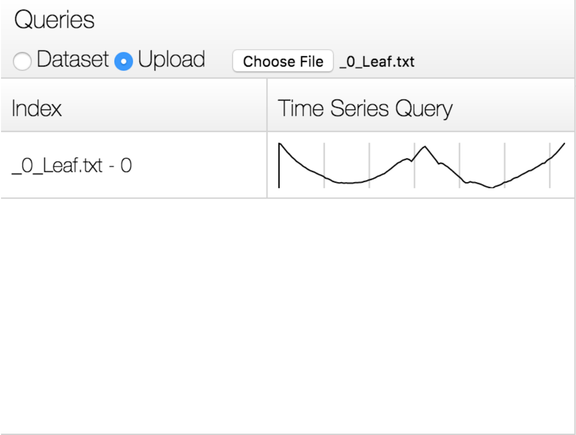
query:	PA - Pennsylvania	1	[0, 6]	dist: 0.030
result:	WI - Wisconsin	7	[0, 6]	

Figure 20: Interface in Similarity view mode.

First is the *overview pane*. This pane provides a general view of the processed dataset by showing representatives of different kinds of shape and the percentage that each of these shapes takes up of the dataset. The background of each percentage is color-coded over a gray-blue gradient where the bluer the group, the more of the dataset it represents.

Next is the *query selection pane*; it contains a scrollable list of all the raw data. Each time series is represented by its name and a thumbnail that gives an idea of its shape.

Moreover, users have the choice to upload their own queries by clicking on the file browsing button, find and open their queries. After this action, the pane shows the uploaded data instead as shown in Fig. 21.



Queries

○ Dataset ● Upload

_0_Leaf.txt

Index	Time Series Query
_0_Leaf.txt - 0	

Figure 21: User can upload a custom query in selection browse pane.

The third pane is the *query preview pane*. Clicking on a time series in the query selection pane transfers it to this pane for a full view and better interpretation. In addition, users are able to zoom into a sub-sequence of the time series by brushing on the smaller graph at the bottom. This is illustrated in Fig. 22. The time series or sub-sequence shown in the larger graph of the *query preview pane* is used as a query for similarity search. The search is triggered by clicking on the magnifying glass icon on the right of the query preview pane.

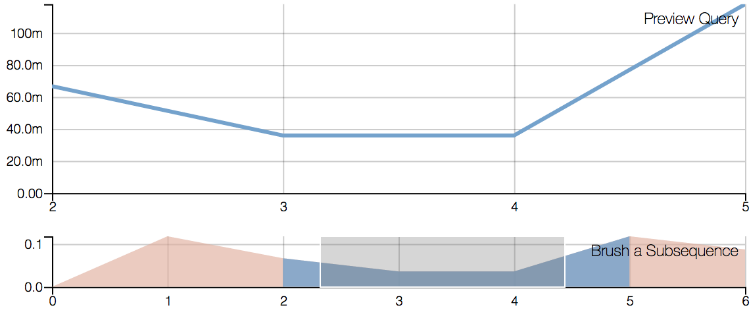
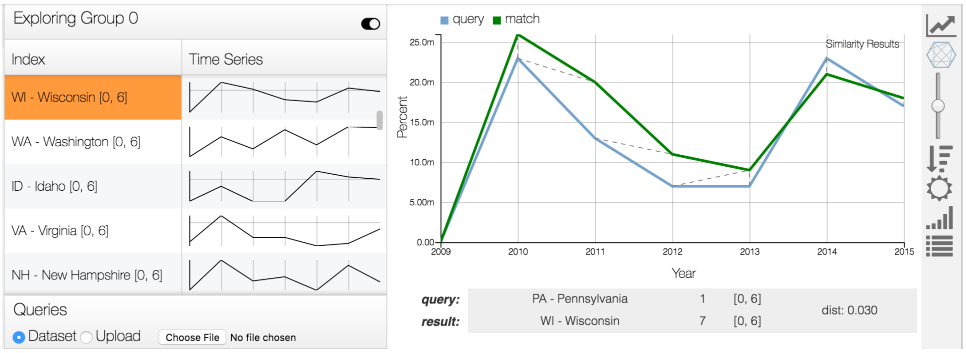


Figure 22: Zoom into a time series by brushing on the smaller graph.

The results of the search are nearly immediately displayed in the *result pane*. Information of the selected query, its match and distance between the query and the match is displayed at the bottom of this pane. On the right side, there is a strip of buttons; each buttons is a different options of visualization that are explained more in

§ 6.3.3. The overview pane is also affected by the search action; it is switched to a *group explore* pane as shown in Fig. 23. We are exploiting the ONEX's integral data structures to provide more context than just the result of the similarity query. Specifically, this pane now shows every time series in the same group as the resultant time series. This means that these time series, although not the best match among the group, have a proximity within the selected threshold to the query time series. Additionally, users can click on each of these time series to compare it with the query. We switch this pane back to overview pane by clicking on the switch icon at the top right corner.


Exploring Group 0

Index	Time Series
WI - Wisconsin [0, 6]	
WA - Washington [0, 6]	
ID - Idaho [0, 6]	
VA - Virginia [0, 6]	
NH - New Hampshire [0, 6]	

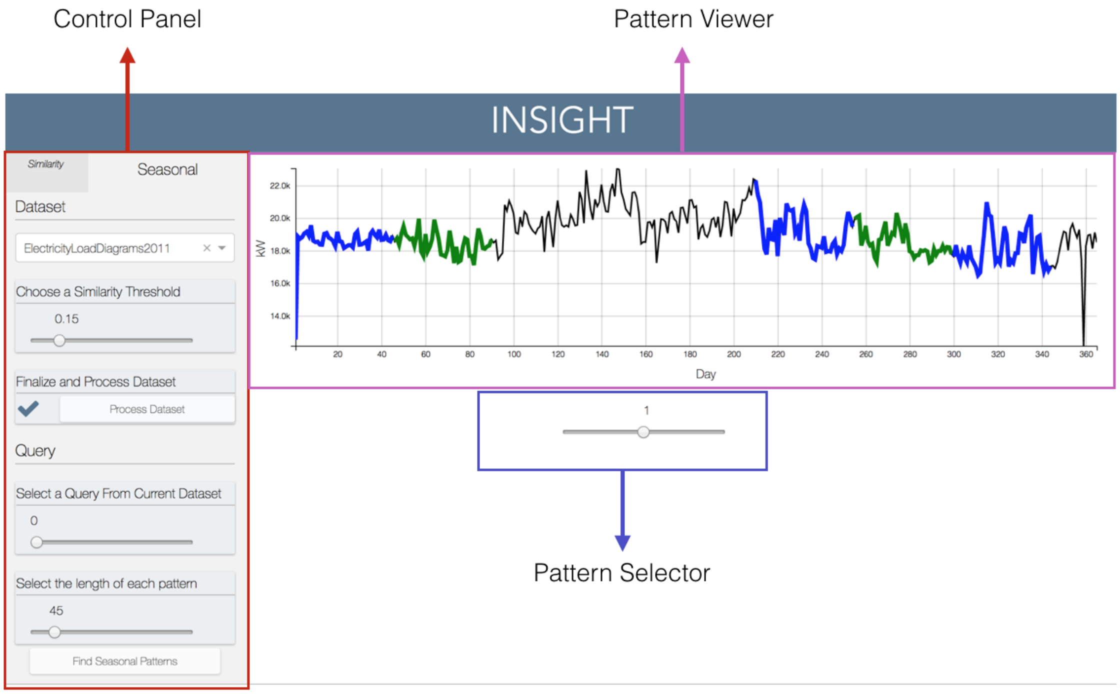
Queries

● Dataset ○ Upload Choose File No file chosen

query:	PA - Pennsylvania	1	[0, 6]	dist: 0.030
result:	WI - Wisconsin	7	[0, 6]	

Figure 23: Result of a similarity search.

The Seasonal view is shown in Fig. 24. This view allows users to find seasonal pattern in a time series: a group of non-overlapping sub-sequence with the same length that are in the same similarity group. The left side of this view is also a Control Panel, but now has further options inside. After processing a dataset, users continue with selecting a query from the dataset using a slider; the pattern viewer is updated accordingly as users select a new time series. Next, length of the pattern is specified; and finally, the seasonal patterns are found by clicking on the button at the bottom. The results are returned from the server in the form of colored sub-sequences of the current time series. Two colors are alternately used in denoting the sub-sequences so that two sub-sequences standing right next to each other can be easily differentiated. Apart from this reason, the difference of colors serves no other purpose. Below the graph is a pattern selector slider where users can drag to switch between different patterns having the specified length and occurring in the the same time series. This slider is hidden if only one pattern can be found.



Dataset
ElectricityLoadDiagrams2011

Choose a Similarity Threshold
0.15

Select a Query From Current Dataset
0

Select the length of each pattern
45

Figure 24: Interface in Seasonal view mode.

6.3.3 Visualization

The Similarity view of the interface comes with a variety of choice for time series presentation. In this section, we look closely to the *result pane* and different types of chart it provides; these charts are summarized in Fig. 25. There are 6 different types of chart that a user can choose from: line chart, warped line chart, radial chart, connected scatter plot, stacked line chart, and difference chart. Generally, in each chart, two time series, a query and a match, are displayed. The y-axis on each chart is shared by both time series, whereas the x-axis only describes the query since the match comes from another location in the dataset; this limitation on x-axis does not apply for stacked line chart since it plots two time series on separate charts.

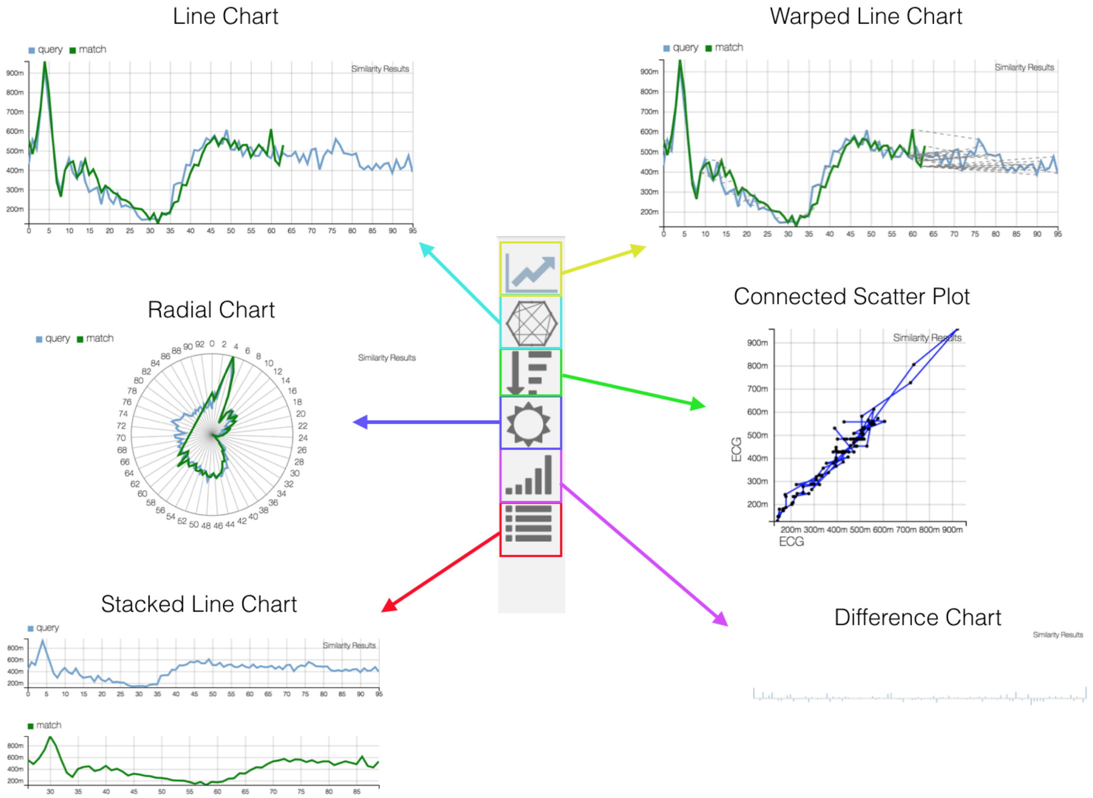


Figure 25: Different types of charts provided by the interface.

Line chart and **warped line chart** are closely related. The only difference is that warped line chart has lines that connect between data points to signify their relationship in DTW distance. These lines might be occluded if the two time series are too close to each other, so we provide a slider below the warped line chart button;

dragging this slider up and down moves the found match toward the corresponding direction and stretches the distance between the two time series. This action reveals all of the obstructed connecting lines; in addition, the line color turns to red and a note is added to the line's legend to remind that this is not the original position of the line; this idea is illustrated in Fig. 26.

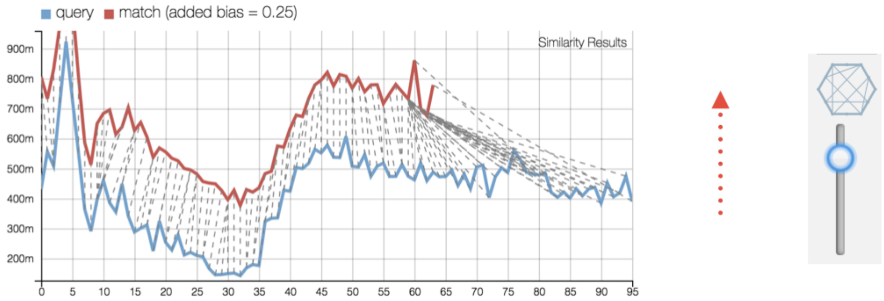
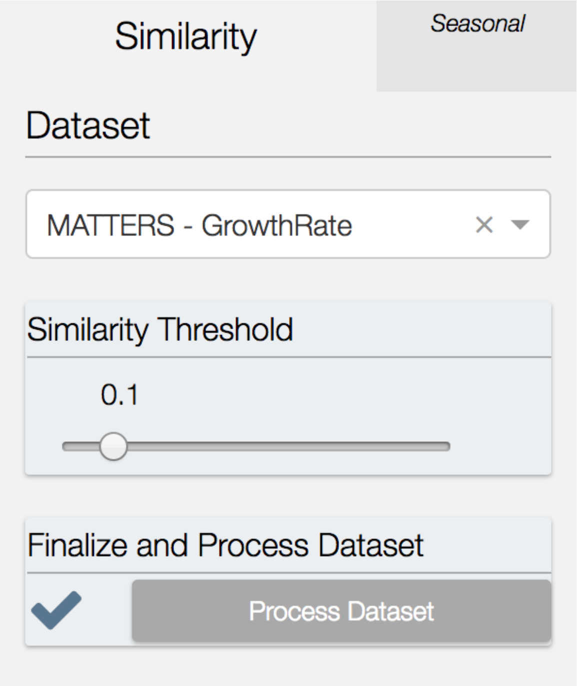


Figure 26: Dragging the slider upward subsequently shifts one of the two time series upward, revealing the occluded connecting lines.

Beside **radial chart** and **connected scatter plot**, which are implemented following the characterizations in § 4.4, the only two charts that we have not mentioned up to this point are difference chart and stacked line chart. Despite their simplicity, they profusely provide additional convenience. Firstly, **difference chart** is a bar chart where each bar denotes the squared difference of a pair of data point. We can think of a pair as two data point connected by a line in the warped line chart. This chart supply an effortless way to evaluate the closeness of two time series at a glance: if every bar reduce to a thin line, we can assure that two time series are very close to each other; otherwise, the two time series differ at the part where the bars are high. Lastly, **stacked line chart** is merely a stack of two line charts. However, since it plots two time series on two different charts using their own axes, it removes the limitation regarding to axes when two lines are plotted on the same chart as in line chart and warped line chart.

7 Case Study

A Demonstration of the Interface:
In order to demonstrate the utility of our system we will show how our system could have been helpful in solving the predicament that the 2013 in Massachusetts Sales and User Tax posed. We will assume an audience of analysts that first struggled with the issues without our system. The audience is able to directly interact with ONEX via an intuitive visual web interface to understand how it assists analysts in addressing complex societal and economical questions, such as the ones described in our motivating examples. We use real datasets from diverse domains such as economic, census and tax datasets from MATTERS [5] and a power usage dataset ElectricityLoad [12]. Our interface empowers analysts to draw insights with ease from these collections, as described below.



Dataset

MATTERS - GrowthRate

Similarity Threshold

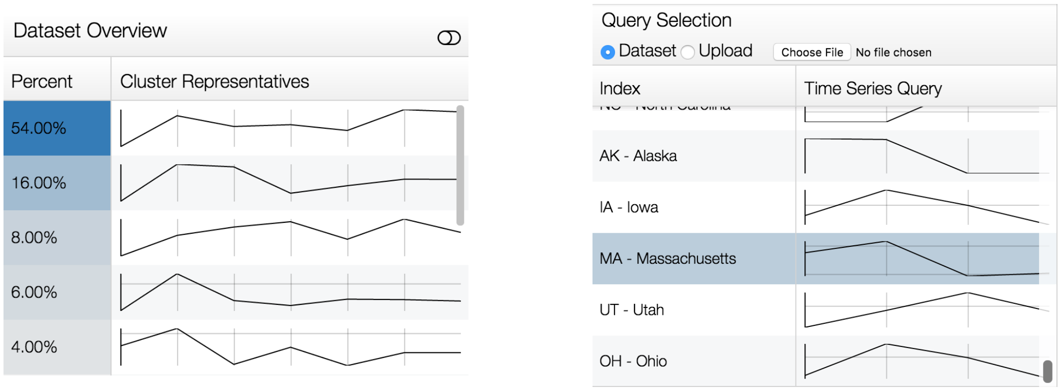
0.1

Finalize and Process Dataset

Figure 27: Choosing a dataset

Data Loading into ONEX. With a click of a button, analysts can load new datasets into ONEX, Fig 27. Loading a new dataset, such as the MATTERS GrowthRate, triggers the preprocessing of this data at the server side and its loading into the respective ONEX Base. Thereafter, the ONEX server provides near real-time responsiveness to the analyst exploring the data via a client-server architecture described above in § 6.3 and Fig. 18.

Making Sense of Overall Time Series Trends. To offer an overview of the data, the



Dataset Overview

Percent	Cluster Representatives
54.00%	
16.00%	
8.00%	
6.00%	
4.00%	

(a) *Overview Pane*

Query Selection

● Dataset ○ Upload Choose File No file chosen

Index	Time Series Query
NC - North Carolina	
AK - Alaska	
IA - Iowa	
MA - Massachusetts	
UT - Utah	
OH - Ohio	

(b) *Query Selection Pane*

Figure 28: Choosing a Time Series to investigate with context - Selecting Massachusetts

Overview Pane (Fig. 28a) displays the representatives of the similarity groups, color-coded such that the color intensity increases proportional with the cardinality of sequences in the group. This gives an immediate sense of the typical patterns within the dataset as well as the overall data distribution to the audience. Each representative is shown as a small graph that captures the general shape of the group. As described in § 5, this is the shape of the centroid - or the mean of the group. This supports analysts in finding the states with similar growth rates. Drilling down into specifics, the audience scrolls through the states in the *Query Selection Pane* (Fig. 28a), each visualized by its name and a small line graph displaying the growth rate over the last 6 years. Our audience, being interested in policy decisions in Massachusetts (MA) will select MA from the list as shown in Fig. 28b.

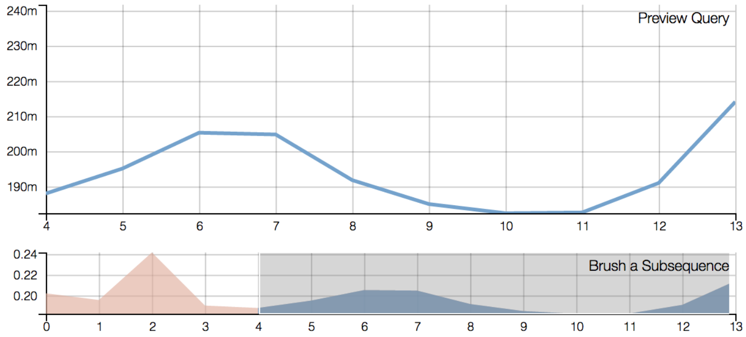
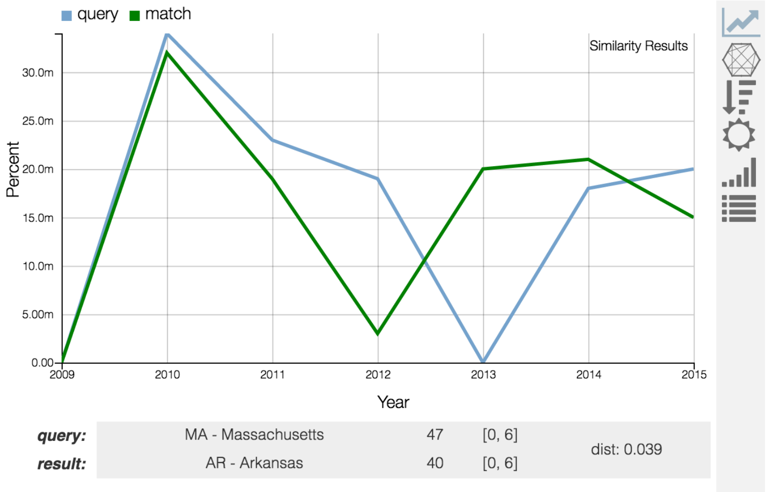


Figure 29: *Query Preview Pane*: Selecting a Subsequence

Precise queries, previewing selections. The *Query Preview Pane* displays the chosen sample query in more detail. Brushing the second half of the graph will focus the attention on the recent trends in MA. As the first preview graph is brushed, the upper chart is updated to show the selected subsequence in more detail.



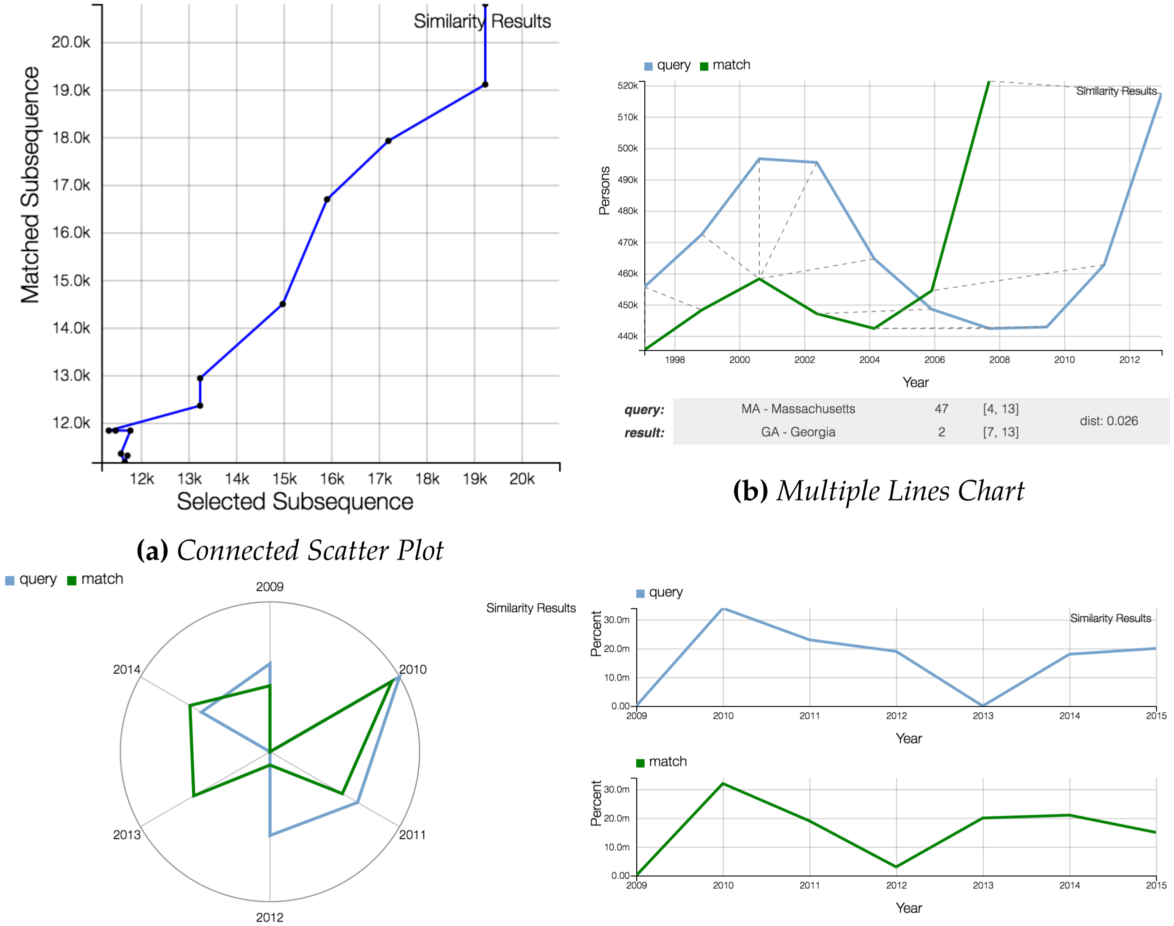
query:	MA - Massachusetts	47	[0, 6]	dist: 0.039
result:	AR - Arkansas	40	[0, 6]	

Figure 30: *Result Pane*: Exploring Similarity

Honing in On Specific Temporal Trends. The *Results Pane* shown in Fig. 30 on the right assists the analyst in finding states with a similar economic growth rate to that of MA, while the *Seasonal View* in Fig. 32 enhances the understanding of a specific time series by highlighting repeating patterns.

Highlighting Time-Warped Shape Matching. When the analyst performs a similarity search, the best match sequence in the dataset is displayed along with the sample query subsequence in the *Results Pane* (Fig. 30 top right). The default *Multiple Lines Chart* displays both time series on a single graph as in Fig. 31b. The "matched points" are connected with dotted lines helping the analyst get a better intuition of how similar the time series shapes are and their relative warping. These connections are shown explicitly in the *Multiple Lines Chart* chart, but these connections are also used internally to generate the other graphs.

Contrasting Trends Across Multiple Linked Perspectives. Different visualizations illustrate different aspects of similarity. For example, to get a richer understanding of the similarity between MA and ARK, the analyst can switch to different visuals by selecting the mode via the right menu bar of the *Results Pane* as above in Fig. 30. The same pair of time series can now be viewed in a compacted *Radial Chart* (Fig. 31c). This view allows a consistent compression of the data, providing the analyst with alternative views to compare sequences. Further, in the *Connected Scatter Plot* (Fig. 31a), the shape is close to a 45 degree angle. This indicates that the match is extremely close – when a point in such plot lies on the diagonal, it has the exact same value in both series. This observation coupled with the fact that all values are very close in range indicates that the subsequences are a close match. The Stacked Line Chart in Fig. 31d allows for the



query:	MA - Massachusetts	47	[4, 13]	dist: 0.026
result:	GA - Georgia	2	[7, 13]	

(a) *Connected Scatter Plot*

(b) *Multiple Lines Chart*

(c) *Radial Chart*

(d) *Stacked Line Chart*

Figure 31: Exploring Similarity of Social Trends with Multiple Views

analyst to separate the sequences to examine them individually, as needed.

Exploring Re-occurrence of Motives Within Time Series. Our ONEX exploratory tool can work with data from diverse domains. We showcase now its use in exploring electrical usage data using the ElectricityLoad collection. The *similarity view* provides a wealth of information about repeated patterns in electricity usage.

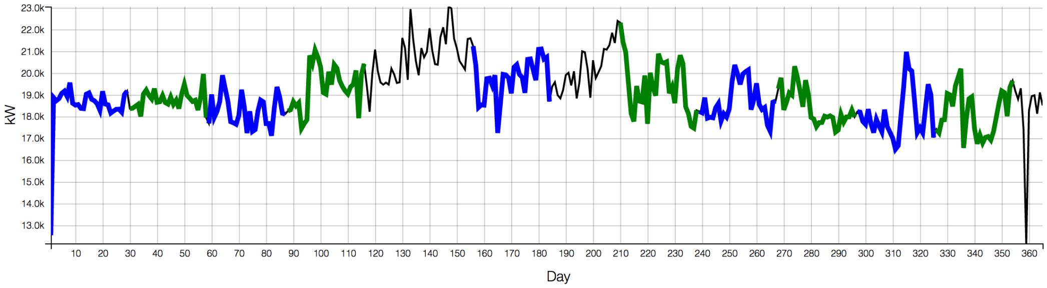


Figure 32: *Seasonal View*: Discovering Repeated Patterns in a Time Series

Focusing on the electrical consumption of a single household, Fig. 32 shows a single time series across one year in Portugal and finds repeated patterns within it. The alternating blue and green coloration are used to clarify instances of consecutive segments. The top graph displays a monthly pattern indicating that this household tends to use electricity in a consistent manner throughout the summer months. The bottom pattern shows that winter months too have similar trends, empowering the analyst to determine that a few small habit changes could have a large savings impact.

8 Experiments & Procedure

We designed and administrated two experiments to generate both quantitative and qualitative data in order to determine the efficacy of the platform, the techniques we employed, and get a baseline for comparison we ran a survey and a series of interviews. We designed the survey using practiced techniques to ask unbiased questions and while also make it easy and quick enough to entice people to spend their time on it [37, 38, 39]. Further, we aim on exploiting the format of the interview to leverage experts' knowledge - those who would end up actually using the system. Broadly, the survey aims to find general trends and patterns among a wide range of people; we did not focus on a particular major, education level, or denomination. We used these results together to determine the potential social impact an increased understanding and awareness of time series could have; from our results we found a connection between perceived importance and comprehension. As we discuss in § 11, those who do not value or utilize time series analysis in their decision making progress sacrifice their capacity to make fully informed decisions. This necessarily leads to worse decisions, and this can have monetary and quality-of-life repercussions [40]. Again, this applies to both organizations, companies, and individuals.

8.1 Survey

Our survey was designed to generate quantitative information that evaluates the utility of the platform, intuitive perceptions of time series similarity, and to gauge the relative utilities of different experimental graphing techniques [41]. Our questions for the survey are derived from a small pool of questions, each pooled into three main categories. These are *Similarity and Visualization Techniques*, *Human Computer Interaction*, and their *Social Implications and Usage*. We review the sections of the survey, and additionally provide the reader with context that was **not** accessible to the survey's population.

8.1.1 Goals

- Survey general population opinions on the ubiquity of time series data and the usefulness of exploring time series similarity in the given context of the MATTERS dataset.
- Inquire about the effectiveness of different graphs and features of the interface.
- Determine the expected social impact of time series understanding

- Keep the average survey time be under 10 minutes [42]
- Ensure all questions are concise and clear [43]

8.1.2 Procedure

We utilized Mechanical Turk to get a large number of responses in order to discriminate between patterns and noise [44]. This platform allows researchers to connect with a large population of respondents; its crowdsourcing for survey and similar *human intelligence tasks*. We were able to validate user completion and uniqueness by adding a random number generator embedded into the survey, available at the end of survey and ask the users to input that number into Amazon Mechanical Turk website. It was relatively straightforward to organize and set up, and easy to validate the completion of tasks by your users.

The survey was run via WPI's qualtrics subscription [45]. Qualtrics is an established and professional platform that allows for easy formulation of surveys. It also allows for easy distribution of the survey and collection of the data. It enabled us to directly download the results after we closed the survey. It allowed for a secure, and easy manner with which we were able to distribute the survey to students. We further distributed the survey via Amazon Mechanical Turk website.

8.1.3 Questions

In order to avoid clutter, we relegate a detailed review of questions to A, but shall still describe the types of questions here.

1. *Social Implications and Usage*

   A key part of this study and project is to highlight the social implications of understanding of time series comparison analysis. In order to better determine respondent awareness of time series, it is important to see if awareness includes the cognizant realization that time series are pervasive in modern life; the news, weather, and economics, etc.

2. *Measuring Similarity and Visualization Techniques*

   One of the primary goals of our IQP and specifically our system is both to discover the best way to measure similarity between time series, and also to find and define what it means to people. Intuitively, is whatever is most mathematically similar, and what is similar to the human eye and mind, closely aligned? We have several of questions to give us insight into these issues.

3. *Human Computer Interaction* A proper and intuitive environment to work in is critical to understanding time series and their context, is. We evaluate the intuitiveness of the interface, and furthermore determine whether the underlying theory we used is backed by real human opinion.

8.2 Interview

The interviews target skilled, knowledgeable study population. We interview a group of graduate students who are familiar with time series, several skilled, respected researchers who use time series data, a couple industry computer science-based leaders, and domain experts in computational chemistry - who have to analyze time series (albeit in different forms.) For privacy reasons we do not name them as per the restrictions defined by the Instituinal Review Board.

8.2.1 Goals

The goal of the interview is to answer the following questions:

- Is the interface intuitive and easy to use and to what extent? (No prior knowledge, need some prior knowledge, need extensive knowledge).
- What type of graph best visualizes the data in different scenarios?
- What do experts think of the possible utility and scope of our system?

8.2.2 Procedure

Each interviewee was given access the interface on their personal machines. We did this so that they were comfortable with the machine they used to interact with the survey. The experiment starts with a brief introduction about the time series data, the MATTERS datasets, similarity finding problem, seasonal finding problem and how the interface helps in solving these problems. We also used a qualtrics question form to guide the interview, setting tasks and problems for the interview to solve using the interface. In order to use this form to help structure the limited time we allotted with each interviewee, we run it on an IPad next to the interviewee as they use their laptop to interact with the interface.

8.2.3 Questions

There were three distinct parts to the interview: we would first test the intuitiveness of the interface, then evaluate different graph types for their capacity to convey similarity between time series in a clear manner, and then open up the interview to suggestions, critiques and ideas. Again, we will leave the exact questions for the interested reader in Appendix § A.

9 Results

9.1 Survey

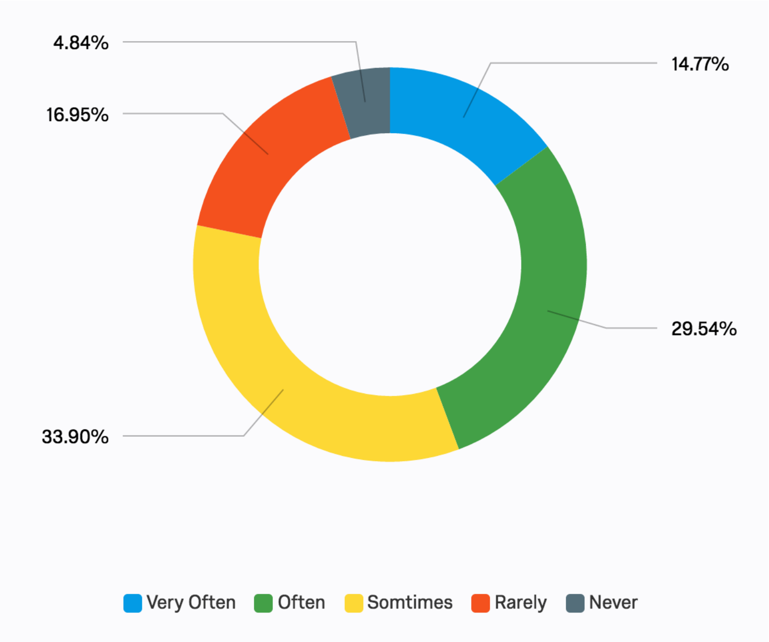
For some questions of the survey, there really was no **correct** answer. In these cases, we were looking for what people naturally were most inclined to favor and use this information to appraise our approach and critically evaluate how time series are socially inspected. However, for the questions were there was a definitive or algorithmically correct answer - for example which two time series are in fact the most similar, it is interesting to see that on the whole, the survey applicants did quite well. We have **413** responses from the participants in the survey.

For each question we will briefly go over the question, the purpose of the question - and if relevant, the correct answer. For a full details of any question please refer to § 8.1.

9.1.1 Data and Analysis

Question 1.
The first question looked to determine how familiar the subject was with time series. As expected there was a range of responses, but a majority were to some degree aware of time series as 75% of respondents either dealt with time series sometimes, often or very often via their work or education.

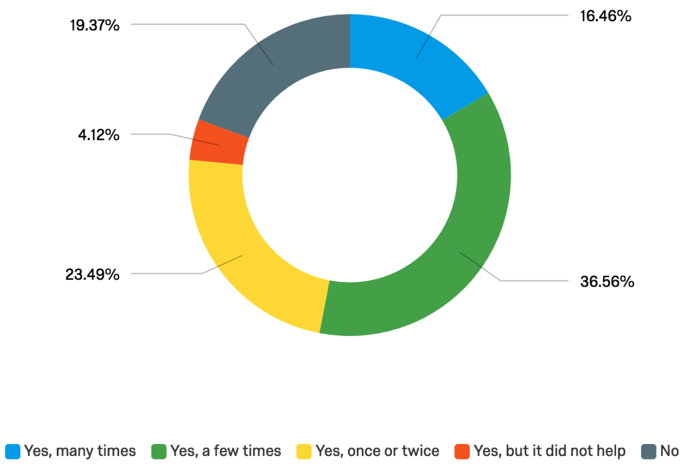
Data.



Analysis.
This relatively high amount of familiarity bodes well for the relative validity and potential feedback we can get from this survey. We have analyzed the responses of this survey with respect to their self-professed familiarity with time series but we do not here include those graphs as they are cluttered and largely show the same information.

Question 2.
This question looks to determine if they use time series to help determine important decisions in their life. We admit that this question has its limitations as some respondents reached out and said that while they only looked time series data once or twice to make the decision of which major or job to take, but once the decision has been made they do not feel the need to continue to review their decision. So, the difference between *yes, a few times* and *yes, once or twice* is not very significant.

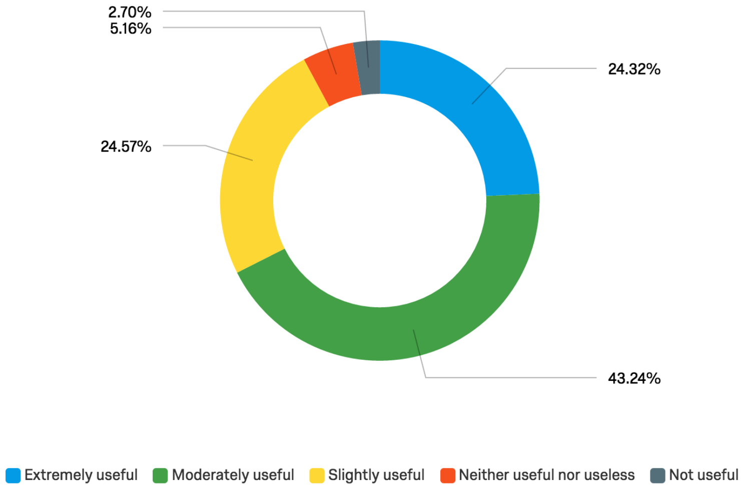
Data.



Analysis.
We can here see that while a large majority 76% used time series to inform them to tackle important life decisions, and found them useful. While the percentage of people who did not find them useful was very small 4.84%, nearly a fifth of the survey population did not use time series at all to help them make important decisions. This shows us that while an increased capacity to compare time series similarity would further help people, an increase in awareness of time series analysis would have a large social impact.

Question 3.
Here we evaluate the populations opinion on the utility of time series comparison.

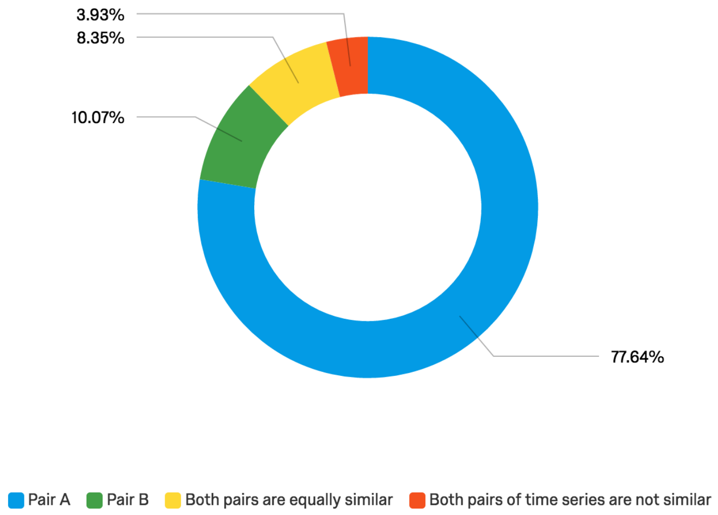
Data.



Analysis.
The majority of the respondents found that it would be useful, although a quarter of the responses thought it would only be slightly useful. This indicates that people believe there is some limitation to this type of analysis or perhaps they think it is domain specific. This is an equal percentage to the percentage of people who were unfamiliar with time series. Perhaps those who are unfamiliar fail to see the utility others do, and thus lose whatever information the other people think are useful. This cost could be very significant in the real world; if people make incorrect or unadvised decisions they will likely fail to make the optimal decision.

Question 4.
Here we assess several things: we test the capacity of people to determine the similarity between time series just by intuition. By familiarizing the respondents with analysis using these graphing techniques we can see which method is more effective when comparing the average correctness of respondents of this question and the following question. Specifically, we are testing the utility of the **Stacked Line Chart**. The correct answer is pair **Pair A**.

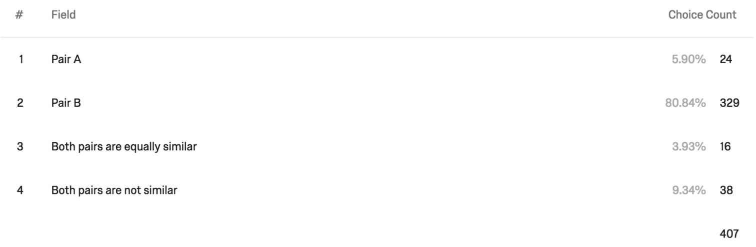
Data.



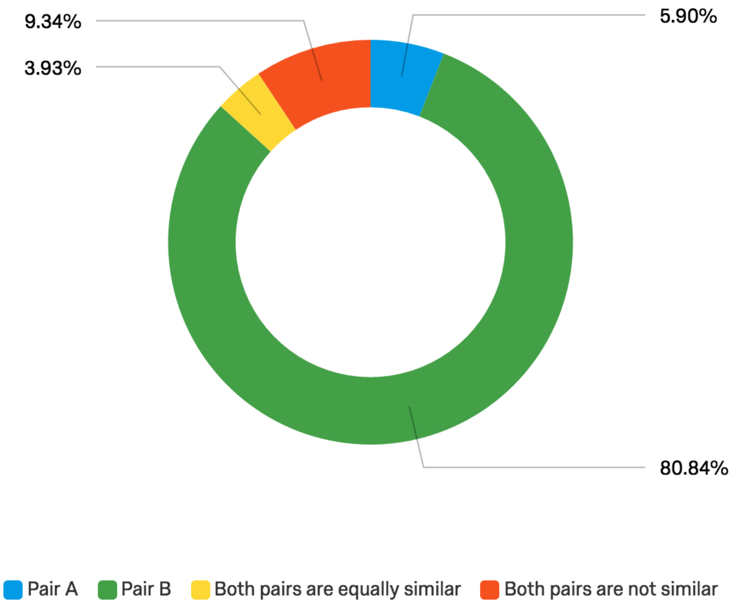
Analysis.
The survey population does quite well, 78% of the respondents correctly determine the more similar pair. In full fairness, excluding the second and third option, we would expect 50% accuracy if they were guessing randomly. Even with this in mind, we can see that this is significant number of the respondents answer this question correctly.

Question 5.
We continue to assess the respondent's capacity to interpret the similarity of time series. We also are specifically testing the **Multiple Lines Chart**. The correct answer is **Pair B**.

Data.



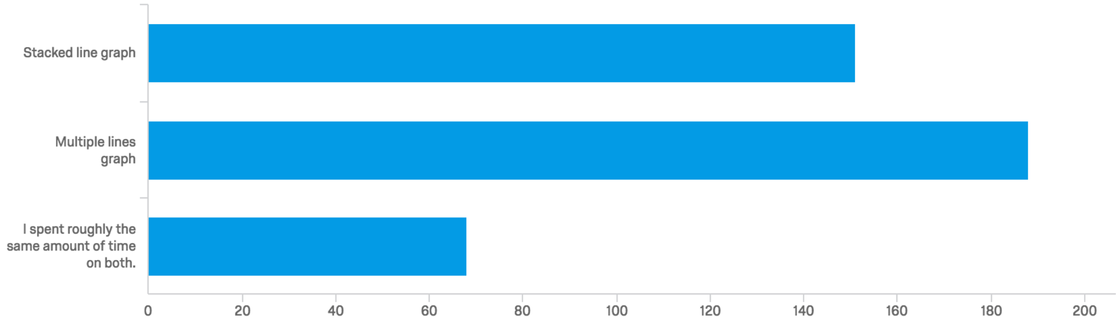
#	Field	Choice Count	
1	Pair A	5.90%	24
2	Pair B	80.84%	329
3	Both pairs are equally similar	3.93%	16
4	Both pairs are not similar	9.34%	38
			407

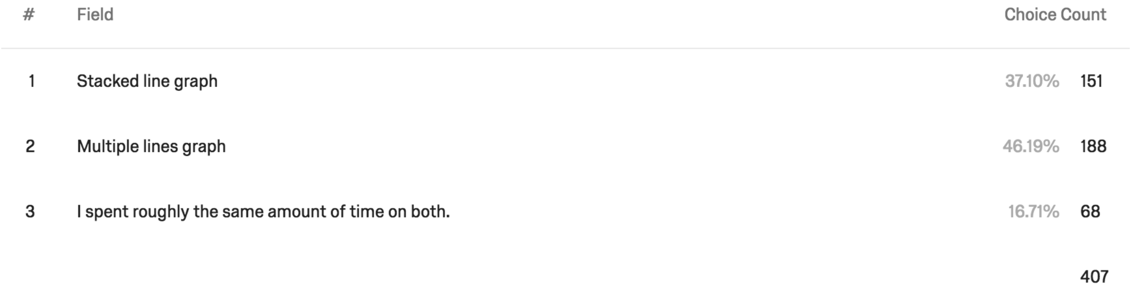


Analysis.
A similarly high percentage of the respondents (80%) correctly determine the more similar pair. The data in **Pair B** is in fact the same exact data as in **Pair A** from Question 4. There is not a large difference in correctness between the percent correctness for **Stacked Line Chart** and **Multiple Lines Chart**. There is some variation in the other options. For this Question 8 we see that a larger proportion of the population consider both pairs as not being similar.

Question 6.
This question asks the users which graph they think made it easier to compare. Keeping in mind that the users performed equally well with both graphs, this question will largely measure peoples preference.

Data.



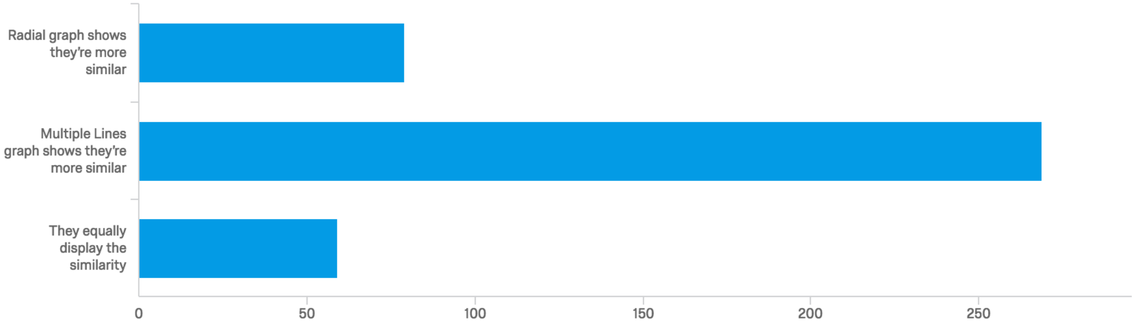


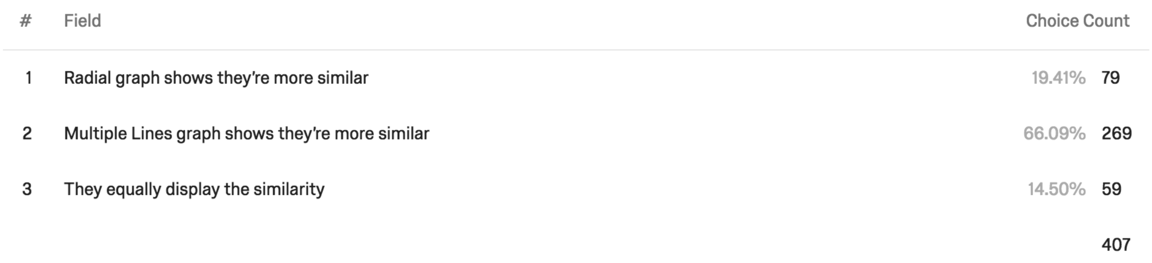
#	Field	Choice Count	
1	Stacked line graph	37.10%	151
2	Multiple lines graph	46.19%	188
3	I spent roughly the same amount of time on both.	16.71%	68
			407

Analysis.
The classic **Multiple Lines Chart** was favored by a larger percentage of respondents, 48%.

Question 7.
This question determines whether people prefer **Radial Charts** over **Multiple Lines Chart** when the subsequences are **not** of equal length.

Data.



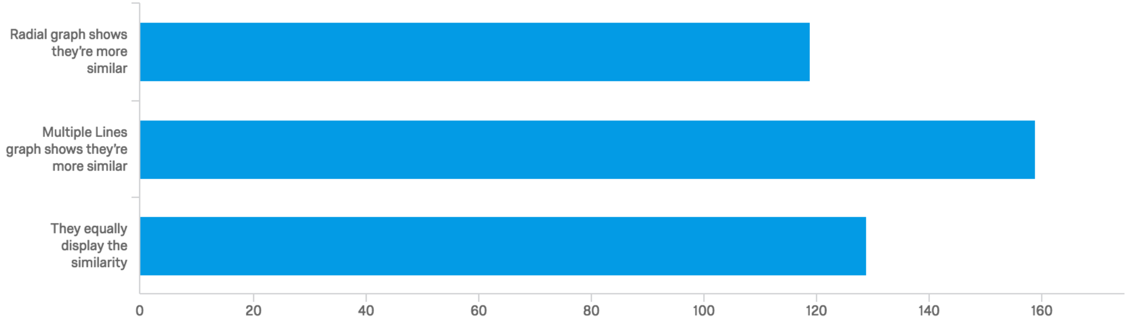


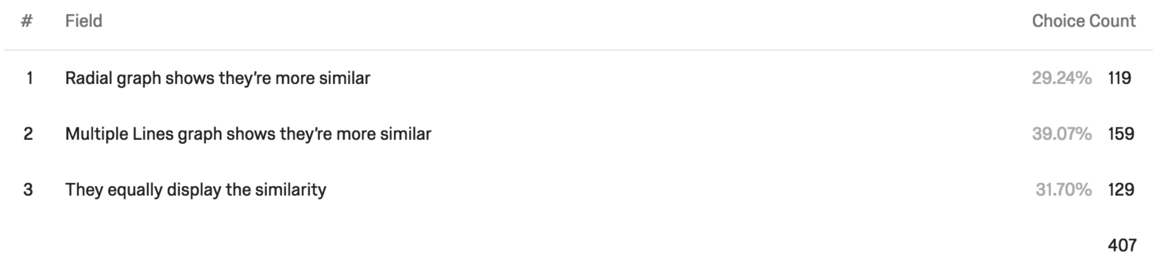
#	Field	Choice Count	
1	Radial graph shows they're more similar	19.41%	79
2	Multiple Lines graph shows they're more similar	66.09%	269
3	They equally display the similarity	14.50%	59
			407

Analysis.
67% of the respondents favored the multiple lines chart as the better graphing technique when the subsequences were of different lengths. This indicates that it is difficult to to interpret the radial charts when the sequence lengths are different.

Question 8.
This question determines whether people prefer **Radial Charts** over **Multiple Lines Chart** when the subsequences are of equal length.

Data.





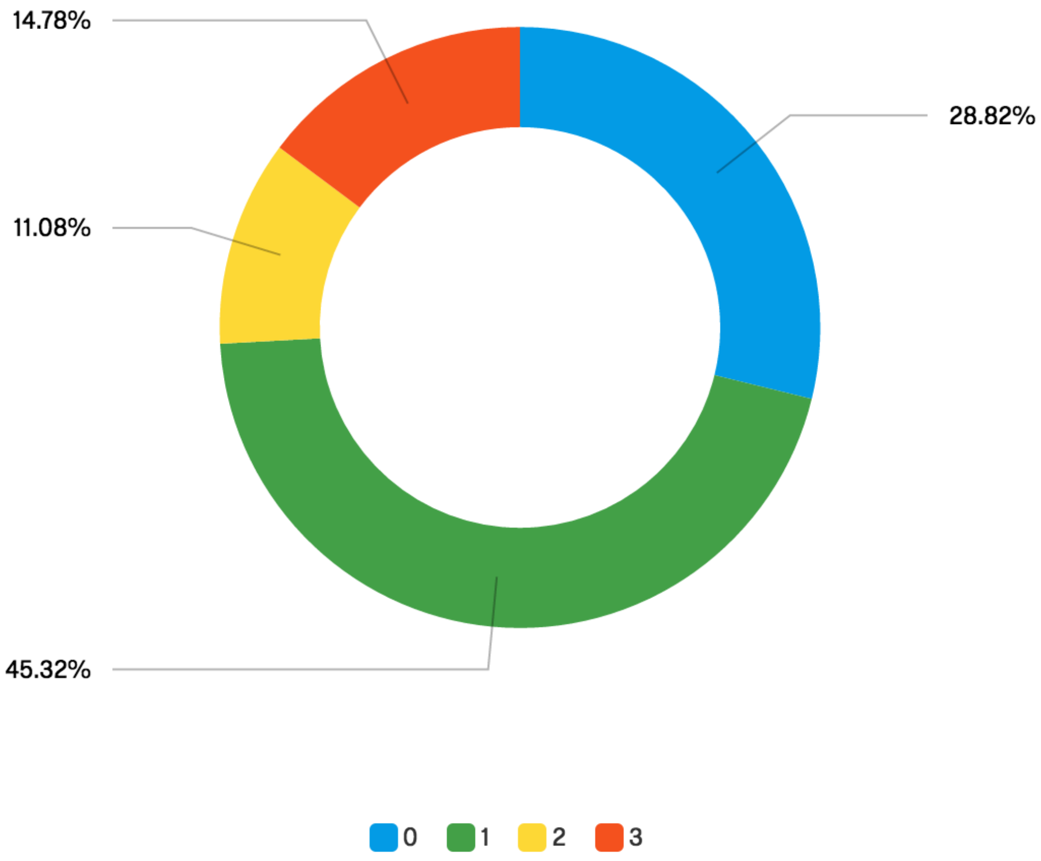
#	Field	Choice Count	
1	Radial graph shows they're more similar	29.24%	119
2	Multiple Lines graph shows they're more similar	39.07%	159
3	They equally display the similarity	31.70%	129
			407

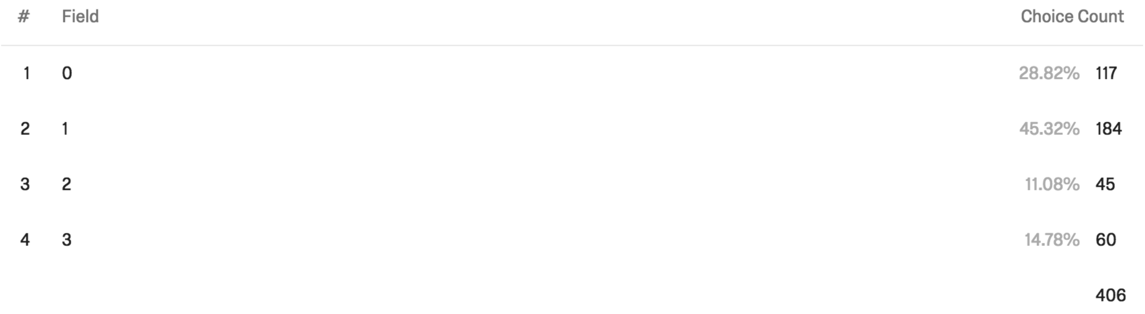
Analysis.
Approximately equal proportions of respondents preferred **Radial Charts** over **Multiple Lines Chart** or have no preference. There is a still a slight preference for the **Multiple Lines Chart**, but when this question is considered with the fact that

the **Radial Charts** had a much lower preference for when the subsequences had different lengths shows both that the **Radial Charts** have some utility, but may be limited or need to be improved upon, and that the **Multiple Lines Chart** are powerful as they are effective in a wider range of scenarios.

Question 9.
Here we abstract our problem of determining a user friendly way of both showing proportionality by color, while also making it aesthetically pleasant. We do so by directly informing the respondents of the question in a theoretical sense, we do not give them a concrete example.

Data.



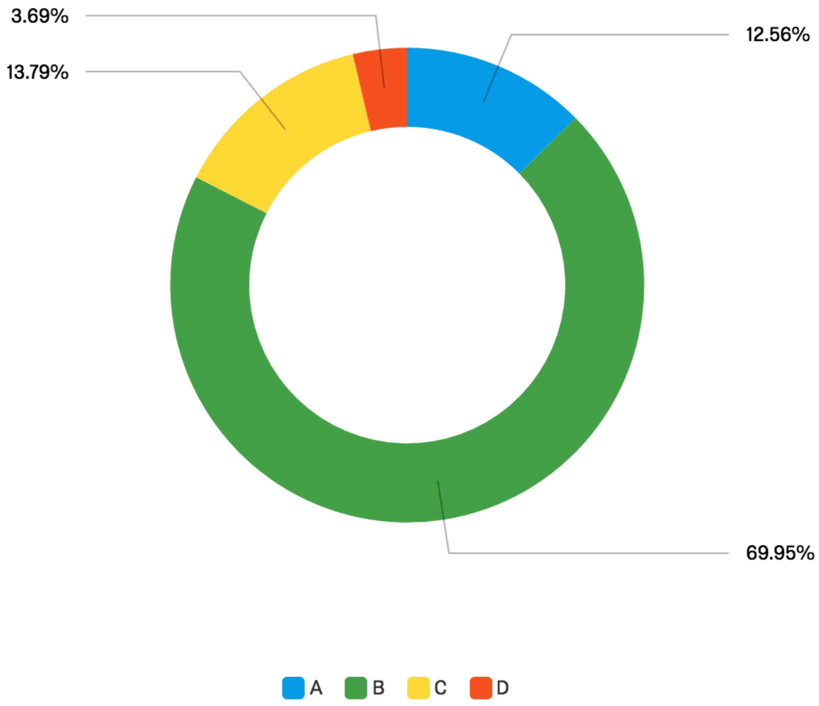


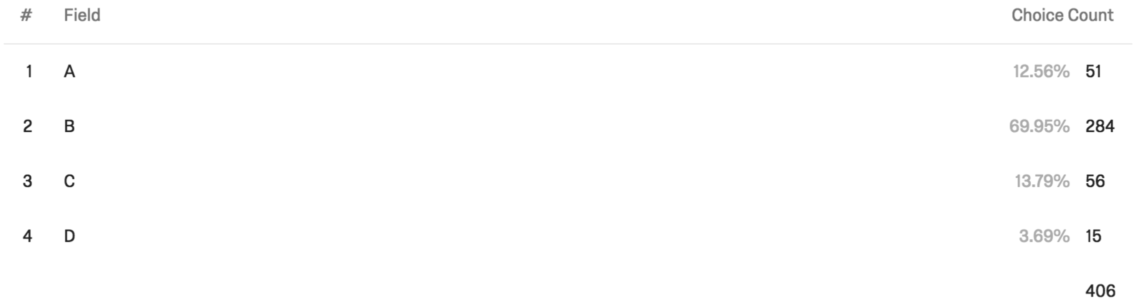
#	Field	Choice Count	
1	0	28.82%	117
2	1	45.32%	184
3	2	11.08%	45
4	3	14.78%	60
			406

Analysis.
The most votes were for the color function displayed in set 1 at nearly 50%.

Question 10.
We continue the line of questioning we began in Question 9, but give the users context - we specifically ask them which color scheme they prefer for differentiating clusters. These are live screen-shots from our interface, using the Matters dataset.

Data.





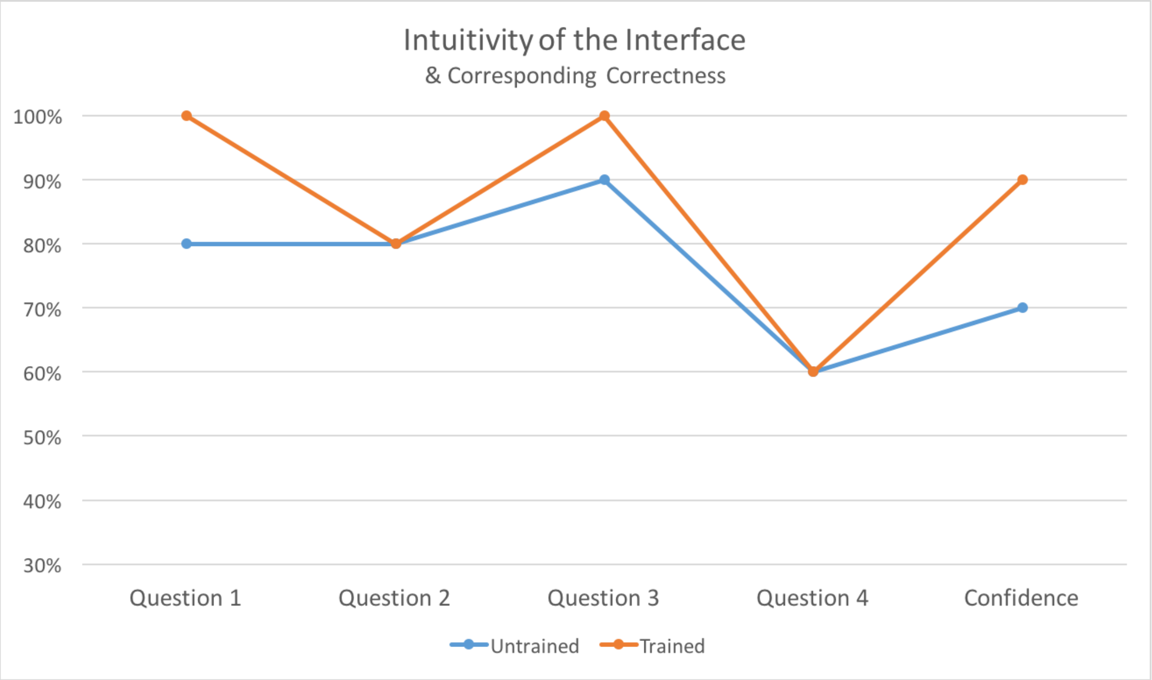
#	Field	Choice Count	
1	A	12.56%	51
2	B	69.95%	284
3	C	13.79%	56
4	D	3.69%	15
			406

Analysis.
70% of the respondents chose the second group, which is the same function that had the highest percentage in the previous question. The increase in preference confirms that asking the question in a concrete sense helps users determine which option they prefer as it applies to something tangible, and thus more immediately important. This function is in fact the option we had deemed best in our preliminary tests (personal preference and experimentation), but confirming its widespread preference among a large group of over 400 respondents further increases our confidence.

9.2 Interview

9.2.1 Intuitiveness of the Interface

The first part of the interview was designed to determine how intuitive and user friendly the interface is - we want to discover if people can operate the interface and fully use its functionality. As described in § 8.2, the first round the users were **untrained** and then they were **trained** and answered similar questions. Although we can see there was some confusion with a relatively minor portion of the interface, the percentage correct was high at 80% and improved after training. At times, when people did not understand the purpose of the interface - or misunderstood it - they would attempt to complete tasks manually. After instruction, we found that the users were guided by their understanding of the purpose of the interface and the purpose their task in addition to specific instructions. We find that the interface was usable with no instruction whatsoever, but it was effective with some training. We shall discuss the ramifications of this further in the feedback below (§ 9.2.3).



9.2.2 Comparing Graphing Techniques

When evaluating the graph types with expert interviewers we ran each through three different examples. These examples used real, varying datasets. The diversity of these datasets helped reveal that the most effective graph depends on the data itself. This validates the utility of our system with its capacity to represent data with a wide range of graphing types.

The graph, Fig. 33, shows the interviewers' mean ratings for the different graphs for each example, rating the graphs on their capacity to convey the similarity of the data. In every case the **Multiple Lines Chart** either has the highest or tied for highest rating. The **Connected Scatter Plot** does poorly, although it does have some variance showing that some find it useful in some scenarios.

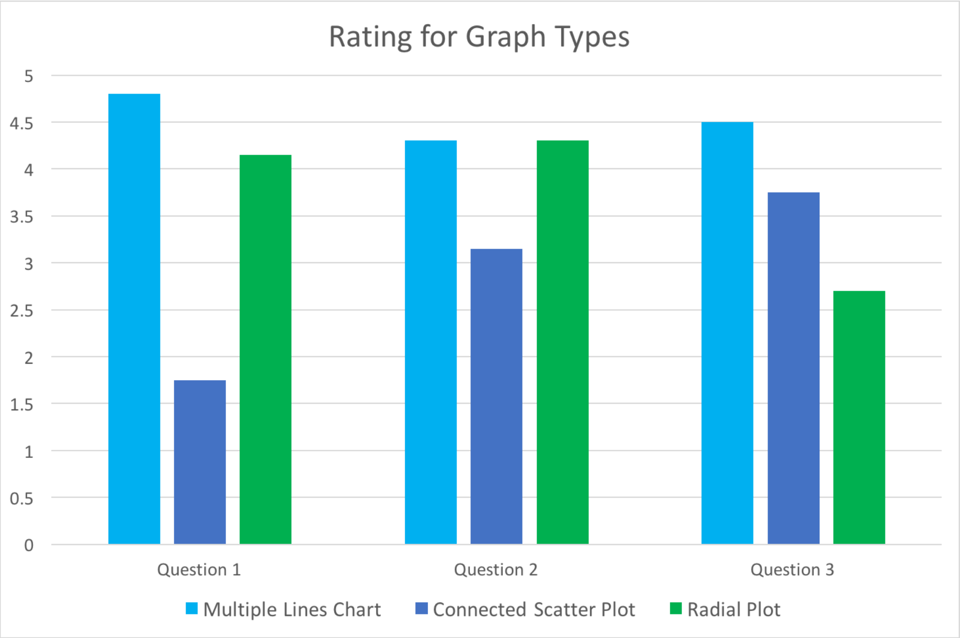


Figure 33

We also asked the interviewers which graph type they found to convey the information most quickly; we asked them to determine which graph illustrated the degree of similarity between the sequences fastest. In some domains speed of understanding is critical.

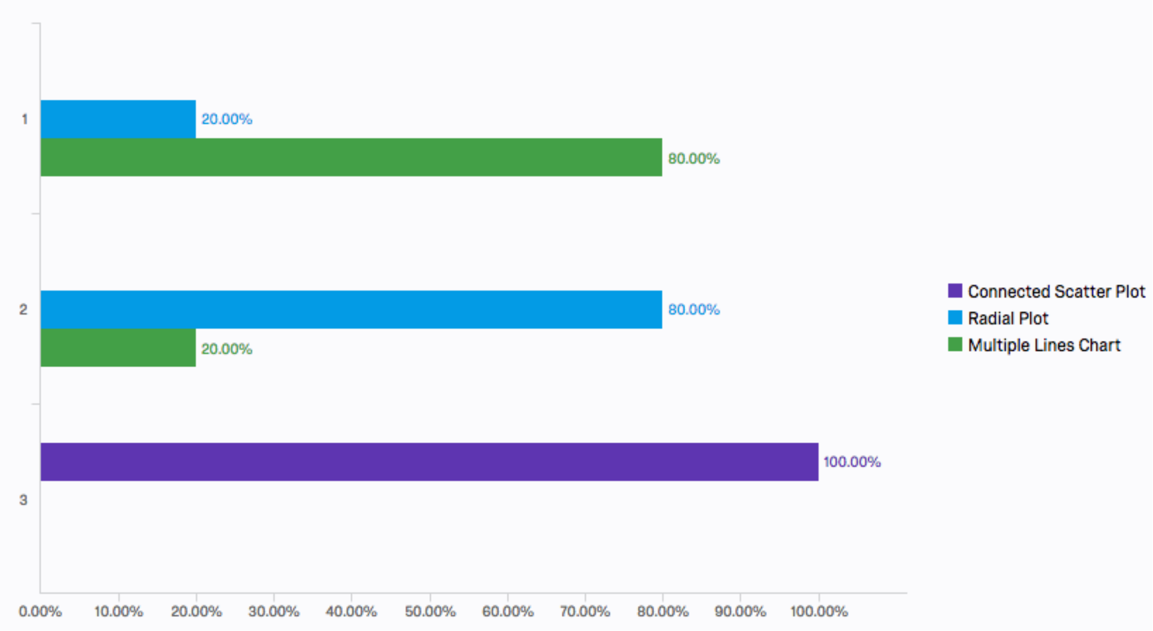


Figure 34: Question 1

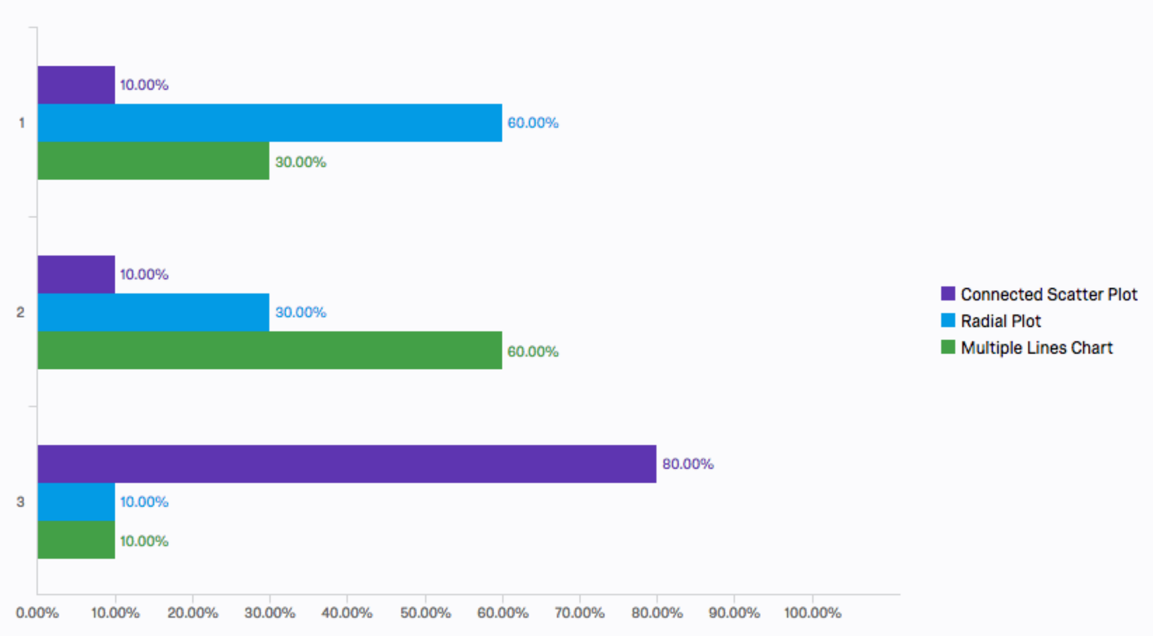


Figure 35: Question 2

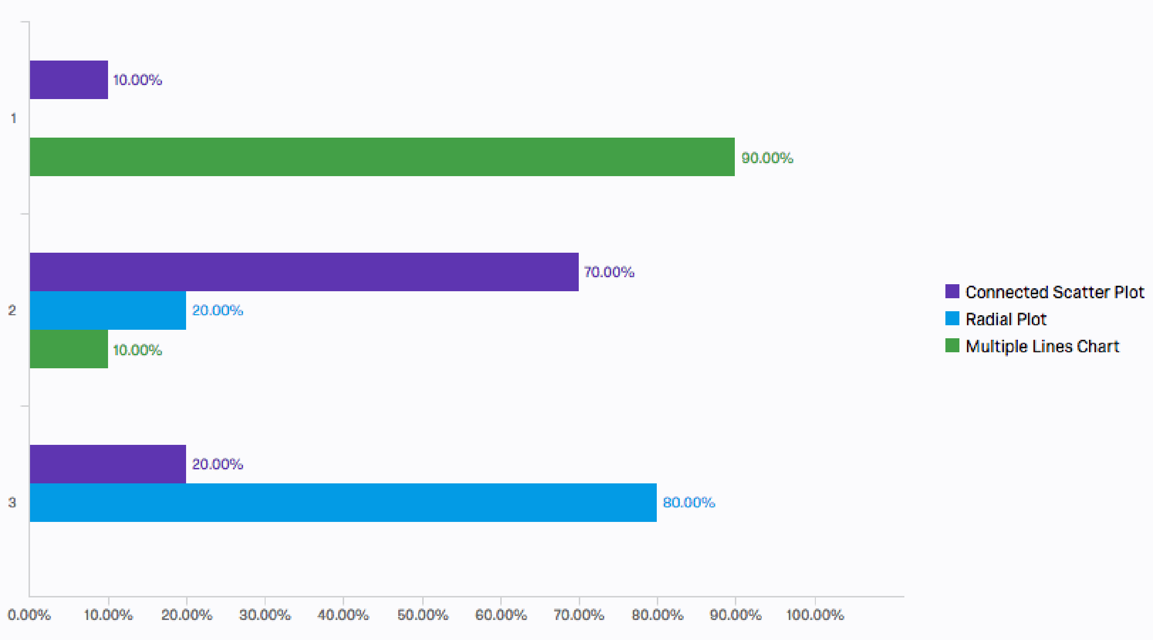


Figure 36: Question 3

We can see that the responses capture in Figs. 34, 35, and 9.2.2 mirror the ratings shown in Fig. 33. There is some variation between the different graphs, but the **Multiple Lines Chart** always performs well. The **Radial Plot** and the **Connected Scatter Plot** are not useful in all cases. In fact, the third example in Fig. 9.2.2 10% of the interviewers found the **Connected Scatter Plot** the method to convey information the fastest, with 70% finding it to be the fastest after the **Multiple Lines Chart**.

9.2.3 Feedback

We break up the feedback that the interviewees gave us into three sections.

Similarity View.
This feedback was pointed mostly at the **Similarity View** and consists of advice and comments for improving our interface.

1. Improve the visibility of the search button
2. Try reordering the panes of the interface
3. A built in help or walkthrough would help users figure out how to accomplish some of the more complicated tasks
4. Search queries by typing in their labels

5. Optional type-entering mechanism for the threshold
6. Alphabetize labeled queries
7. Attach metadata to the datasets and have the capacity to search for a dataset based on what type of data it contains

Although a lot of this falls into **Future Work** (§ 10), we briefly address some of these comments. We are working on increasing the visibility of the search button, considering highlighting it right after a query is chosen. As we mention specifically below in **Ideas and Comments**, an expert in data visualization told us that some minor issues in intuitiveness of your interface is not a major issue. People would spend a long time learning how to use a tool that would make their future work far more efficient.

Seasonal View.
This feedback was pointed mostly at the **Seasonal View** and consists of advice and comments for improving our interface.

1. Use more than two colors for the repeating patterns.
2. Highlight background for patterns instead of the time series itself.
3. Add further explanatory text for the colors on the plot.

Ideas and Comments.

1. Provide supporting statistical analysis for the time series and datasets.
2. Provide the ability to select a subsequence with exclusion, for example if a time series consisted of four points, choosing the first and the last as the subsequence.
3. Provide the ability to handle extremely large datasets.
4. Consider building the backend with a distributed framework.
5. Several experts in the field demonstrated strong interest in using the tool to help assist in their research.
6. People will spend whatever time necessary to fully learn how to use an effective tool that offers a service not offered anywhere else.

10 Future Works

10.1 Promotion

The promotion of time series analysis is critical to making systems such as ours most effective. We must work to elevate larger numbers of people to be able to understand intuitively how time series are compared meaningfully. One way to do this would be to create a short, clear, and introductory course on time series analysis and distribute it for free. This type of online course is called a MOOCs [46]. In order to get people to take it, one could promote it via public figures on social media. Although it would be challenging to convince people on the importance of this course, one could highlight the costs of not doing so.

10.2 Exploration of Social Impact

We found that there is a high percentage of people who are generally aware of time series and can correctly determine which time series are similar to each other. We also saw that 20% did not have this familiarity, and further they did not look at time series data when making important decisions. Thus, it would be valuable to evaluate exactly this cost and how to lower it. In future works, exploring promotion and advertisement methods that would help people realize the value in consulting data and time series would be a constructive venue.

10.3 Studies

There is a wealth of potential studies that could be constructed using the capabilities of the interface. For example, one could specifically test to what extent humans are able to detect similarity for the varying graph types, not just preference in using them. In our work we used real datasets to convey the real-life importance of this work, and potential social impact, but now that that has been determined - one could specifically generate data to help run these studies. In this iteration of the system we relied and focused on the opinions of experts, but it may be beneficial to test these assumptions by constructing specific experiments.

10.4 Features

Largely, the experts we interviewed found our system complete. There were recommended minor touches, and further support that we could add. For example, increasing the usability of the system by organizing the interface in a different manner. There were some suggestions to implement **K-nearest neighbors**. This is difficult to achieve in an efficient manner given the underlying structure of our supporting system ONEX and the properties of the grouping structure we use. However, we have implemented an alternative way to see multiple similar results. The suggestion of a distributed system is very appealing, but really out of the scope of our work. However, if done it may enable the interface to handle very large datasets.

11 Conclusions

There is a growing need to be able to analyze data, including time series data. This includes easy to understand visualizations, both of time series, and the complex queries that operate on time series. Companies, organizations, and individuals interested in anything from medicine, policy, to the environment will directly benefit from being able analyze time series similarity in an accurate and fast manner. Again, this accuracy and speed have no meaning if the results are not presented in an understandable and intuitive manner.

We have built a sleek and powerful system that is an answer to many of these challenges. It is set up as a web-interface that connects to the algorithmic suite ONEX via a server. Its current implementation allows users to explore time series similarity in a number of dynamic and interactive ways. Further, it analyzes a dataset and informs the analyst of the common trends and shapes of the data. The clean visuals, responsive feel, and intuitive layout help experts and novices in time series comparison analysis quickly gain deep insight into their data. It has already began to receive some recognition as researchers request access to it and we are giving a demonstration of it at Sigmod2017.

We had strong results indicating the high utility of our system, and indicators that people in general could understand time series comparisons well. People have good intuitions. Unfortunately, there was a sizable portion of our survey population that were not familiar with time series and this subset of respondents did not perform as well on the analysis questions. This confusion can have a steep personal cost to that particular person across his or her life because they may make uninformed decisions, and these choices can also negatively affect others.

In the future the focus of this work would be to continue improving the system, and perhaps enable it work in a more distributed design. This would allow for it to more easily handle mass amounts of users. More importantly however would be a focus on the promotion of understanding time series. One of the steps to helping the public and socially impact our country is to build a system enabling improvement and exploration. But this alone will do as little as high accuracy and speed will do without understandable visualizations. Avenues to accomplish widespread understanding of time series are diverse, but one route may include the design, setup and distribution of free online courses that enable anyone to quickly understand time series. From here, it would be a matter of promoting the powerful need and benefit of understanding time series. This would be difficult, but courses online are becoming increasingly popular and via social media we could easily promote it with an extensive reach.

References

[1] T. P. Center, "Tax Policy Center," Feb 2017. http://www.taxpolicycenter.org/.

[2] U. C. Bureau, "Bureau, US Census," Feb 2017. www.census.gov.

[3] U. D. of Commerce, "Bureau of Economic Analysis," Feb 2017. www.bea.gov/.

[4] "2017 ieee mit undergraduate research technology conference." http://ieee.scripts.mit.edu/conference/.

[5] "Matters - massachusetts technology, talent, and economic reporting system." http://matters.mhtc.org/.

[6] "Sigmod/pods 2017." http://sigmod2017.org/.

[7] R. Neamtu, R. Ahsan, E. Rundensteiner, and G. Sarkozy, "Interactive time series exploration powered by the marriage of similarity distances," *Proceedings of the VLDB Endowment*, vol. 10, no. 3, pp. 169–180, 2016.

[8] E. Keogh, L. Wei, X. Xi, M. Vlachos, S.-H. Lee, and P. Protopapas, "Supporting exact indexing of arbitrarily rotated shapes and periodic time series under euclidean and warping distance measures," *The VLDB Journal—The International Journal on Very Large Data Bases*, vol. 18, no. 3, pp. 611–630, 2009.

[9] U. Rebbapragada, P. Protopapas, C. E. Brodley, and C. Alcock, "Finding anomalous periodic time series," *Machine learning*, vol. 74, no. 3, pp. 281–313, 2009.

[10] J. Aach and G. M. Church, "Aligning gene expression time series with time warping algorithms," *Bioinformatics*, vol. 17, no. 6, pp. 495–508, 2001.

[11] N. A. Chadwick, D. A. McMeekin, and T. Tan, "Classifying eye and head movement artifacts in eeg signals," in *Digital Ecosystems and Technologies Conference (DEST), 2011 Proceedings of the 5th IEEE International Conference on*, pp. 285–291, IEEE, 2011.

[12] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista, "The ucr time series classification archive," *note www. cs. ucr. edu/˜ eamonn/time_series_data*, 2015.

[13] T. Rakthanmanon *et al.*, "Searching and mining trillions of time series subsequences under dynamic time warping," in *ACM SIGKDD*, pp. 262–270, ACM, 2012.

[14] E. Keogh, "Exact indexing of dynamic time warping," in *Proceedings of the 28th international conference on Very Large Data Bases*, pp. 406–417, VLDB Endowment, 2002.

[15] 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*, pp. 1303–1312, ACM, 2009.

[16] G. M. Draper, Y. Livnat, and R. F. Riesenfeld, "A survey of radial methods for information visualization," *IEEE transactions on visualization and computer graphics*, vol. 15, no. 5, pp. 759–776, 2009.

[17] S. Haroz, R. Kosara, and S. L. Franconeri, "The connected scatterplot for presenting paired time series," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 9, pp. 2174–2186, 2016.

[18] W. Javed, B. McDonnel, and N. Elmqvist, "Graphical perception of multiple time series," *IEEE transactions on visualization and computer graphics*, vol. 16, no. 6, pp. 927–934, 2010.

[19] H. Hochheiser and B. Shneiderman, "Visual queries for finding patterns in time series data," *University of Maryland, Computer Science Dept. Tech Report, CS-TR-4365*, 2002.

[20] P. Buono, A. Aris, C. Plaisant, A. Khella, and B. Shneiderman, "Interactive pattern search in time series," in *Electronic Imaging 2005*, pp. 175–186, International Society for Optics and Photonics, 2005.

[21] 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*, pp. 1737–1746, ACM, 2011.

[22] M. Wattenberg, "Sketching a graph to query a time-series database," in *CHI'01 Extended Abstracts on Human factors in Computing Systems*, pp. 381–382, ACM, 2001.

[23] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," *KDD*, vol. 10, no. 6, pp. 359–370, 1994.

[24] M. P. E. Keogh, K. Chakrabarti and S. Mehrotra., "Locally adaptive dimensionality reduction for indexing large," *ACM SIGMOD Record*, vol. 30, no. 2, p. 151–162, 2001.

[25] B.-K. Yi and C. Faloutsos, "Fast time sequence indexing for arbitrary lp norms," *VLDB*, 2000.

[26] R. Agrawal *et al.*, *Efficient similarity search in sequence databases*. Springer, 1993.

[27] E. J. Keogh and M. J. Pazzani, "Scaling up dynamic time warping for datamining applications," in *Sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 285–289, ACM, 2000.

[28] Y. Sakurai *et al.*, "Stream monitoring under the time warping distance," in *ICDE*, pp. 1046–1055, IEEE, 2007.

[29] V. Athitsos *et al.*, "Approximate embedding-based subsequence matching of time series," in *ACM SIGMOD*, pp. 365–378, ACM, 2008.

[30] C. Faloutsos, R. M, *et al.*, *Fast subsequence matching in time-series databases*. ACM, 1994.

[31] S. Chu, E. Keogh, *et al.*, "Iterative deepening dynamic time warping for time series.," in *SDM*, pp. 195–212, SIAM, 2002.

[32] Wikipedia, "Representational state transfer — wikipedia, the free encyclopedia," 2017.

[33] A. Ronacher, "Flask." http://flask.pocoo.org/.

[34] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pp. 336–343, IEEE, 1996.

[35] Facebook, "React." https://facebook.github.io/react/.

[36] M. Bostock, "D3.js data-driven document." https://d3js.org/.

[37] C. B. Cohen D, *Qualitative Research Guidelines Project*. 2006. http://www.qualres.org/HomeInte-3595.html.

[38] G. E. Bauman, LJ., *The use of ethnographic interviewing to inform questionnaire construction*. Health Education Quarterly., 1992.

[39] G. E. Bauman, LJ., *Research methods in cultural anthropology*. Sage Publications, 1988.

[40] L. B. O'Connell, "A Brief History of Decision Making," Feb 2017. https://hbr.org/2006/01/a-brief-history-of-decision-making/.

[41] C. Briggs, *Learning how to ask: A sociolinguistic appraisal of the role of the interview in social science research*. Cambridge University Press, 1986.

[42] J. Chirban, *Interviewing in depth: The interactive relational approach.* Thousand Oaks, CA. Sage Publications, 1996.

[43] N. Britten, "Qualitative research: Qualitative interviews in medical research," *BMJ*, pp. 251–253, 1995.

[44] Amazon, “Amazon mechanical turk.” https://www.mturk.com/mturk/welcome.

[45] “The world’s leading research & insights platform.” https://www.qualtrics.com/.

[46] C. Central, “Massachusetts Institute of Technology • Free Online Courses and MOOCs — Class Central..”

Appendices

A Survey Questions

1. *Social Implications and Usage*

   As aforementioned in 8.1, a key part of this study and project is to highlight the social implications of increased capacity and understanding of time series comparison analysis. In order to better determine student awareness levels, it is important to see if awareness includes the cognizant realization that time series are pervasive in modern life; the news, weather, and heavily so in economics.

   I. This question is meant to determine how familiar you are with Time Series. Time series is a series of data points ordered by time. For example, the following graph represents the number of people employed in Technology-related fields in Massachusetts over 14 years.

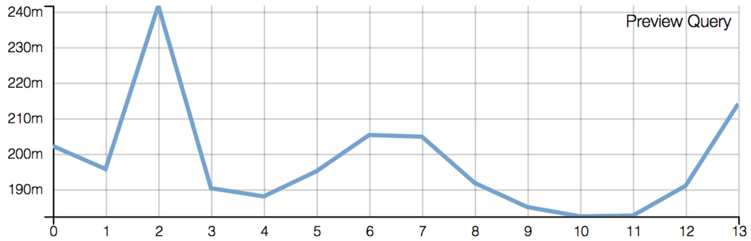


Figure 37: *Question I Image*

   How often do you see time series plots for classes in your major (or in internships)?

   i. Very Often

   ii. Often

   iii. Sometimes

   iv. Rarely

   v. Never

   II. Have you looked at time series data, such as employment trends by job title and state income taxes, to help choose your career path, or help you plan where you may to live?

   i. Yes, many times

ii. Yes, a few times

iii. Yes, once or twice

iv. Yes, but it did not help

v. No

III. The following figure shows the line graphs of Annual Gross State Product Growth Rate of Massachusetts (top) and Arkansas (bottom). These two lines share very similar shape. How useful do you think it is to be able to, given a time series, quickly search for another time series with similar shape?

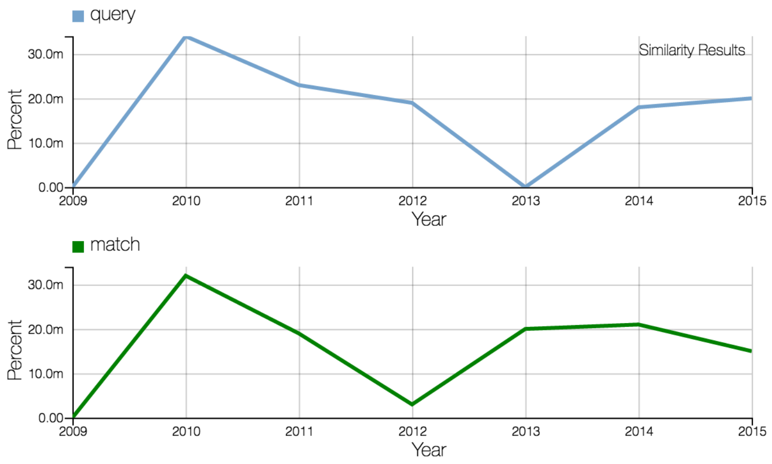


Figure 38: *Question III Image*

i. Extremely useful

ii. Moderately useful

iii. Slightly useful

iv. Neither useful nor useless

v. Not useful

2. *Measuring Similarity and Visualization Techniques*

As aforementioned in 8.1, one of the primary goals of this platform is both to discover what the best way to measure similarity between time series, but also to find and define what it means to people. Intuitively, is whatever is most mathematically similar, and what is similar to the human eye and mind, closely aligned? We have a several of quick questions to help us get insight into these questions.

IV. Below are two pairs of “Stacked line graph” (A and B). Each pair consists of plots of 2 series stacked vertically. Which pair of series looks more similar to each other?

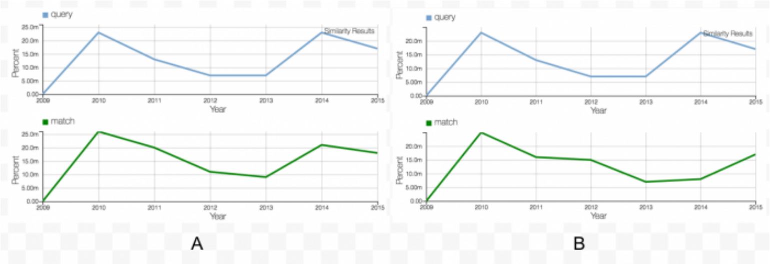


Figure 39: *Question IV Image*

i. Pair A

ii. Pair B

iii. Both pairs are equally similar

iv. Both pairs of time series are not similar

V. Below are two “Multiple lines graph” (A and B), each graph consists of 2 series plotted on the same axes. Which pair of series looks the more similar to each other?

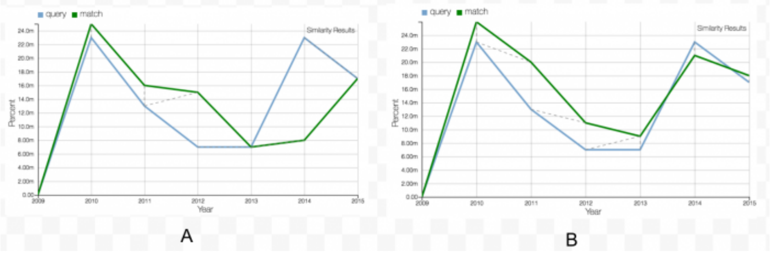


Figure 40: *Question V Image*

i. Pair A

ii. Pair B

iii. Both pairs are equally similar

iv. Both pairs of time series are not similar

VI. Which kind of graph in the previous 2 questions made it easier to compare (which one did you spend less time to answer the question)?

i. Stacked line graph

ii. Multiple lines graph

iii. spent roughly the same amount of time on both

VII. Which graph better shows how similar (or dissimilar) these two subsequences are? The left is a multiple lines graph consisting of 2 series plotted on the same axes, and the right graph is a radial chart which compacts both series, graphing them into a circle.

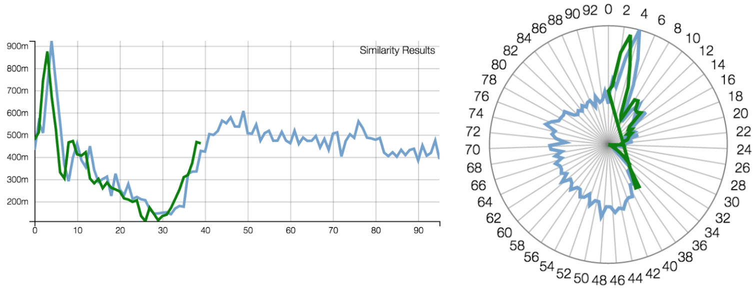


Figure 41: *Question VII Image*

i. Radial graph shows they're more similar

ii. Multiple Lines graph shows they're more similar

iii. They equally display the similarity

VIII. Which graph better shows how similar (or dissimilar) these two subsequences are? The left is a multiple lines graph consisting of 2 series plotted on the same axes, and the right graph is a radial chart which compacts both series, graphing them into a circle.

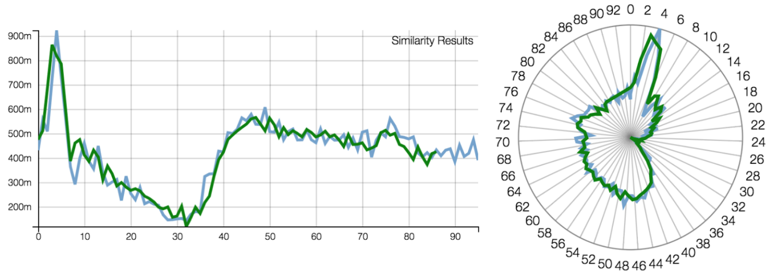


Figure 42: *Question VIII Image*

(a) Radial graph shows they're more similar

(b) Multiple Lines graph shows they're more similar

(c) They equally display the similarity

3. *Human Computer Interaction* As aforementioned in 8.1, critical to understanding time series and their context, is a proper and intuitive environment to work in. We here work to ascertain the intuitivity of the platform, and furthermore whether the underlying techniques we employed were sound.

IX. We use color to highlight information. Which set of these color values (sets span from left to right) makes each block the most distinguishable?

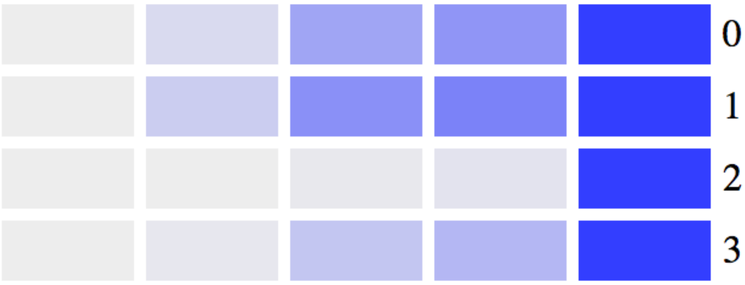


Figure 43: *Question IX Image*

i. 0

ii. 1

iii. 2

iv. 3

X. Below are lists of groups of time series. Each group has a representative displayed as a line graph. The size of each group is encoded over a blue-gray gradient such that more blue means bigger size. Which set of colors allows you to see the difference in size most easily?

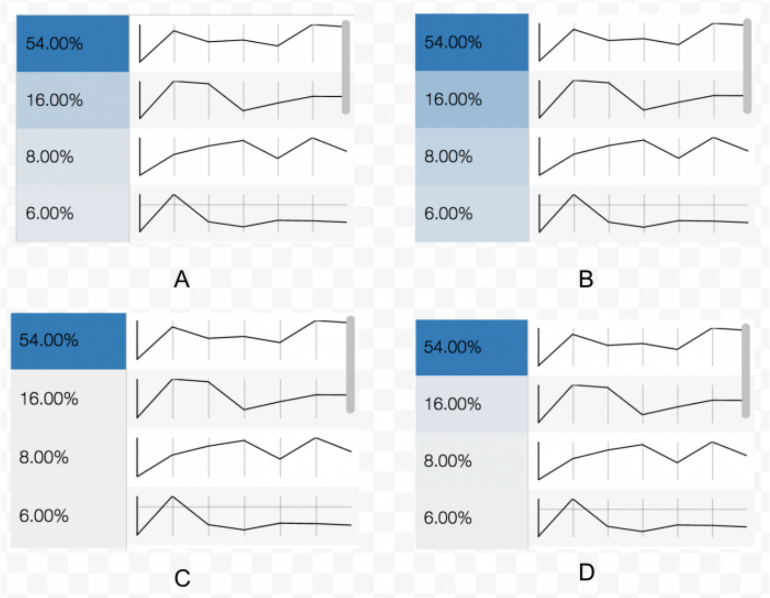


Figure 44: *Question X Image*

i. A

ii. B

iii. C

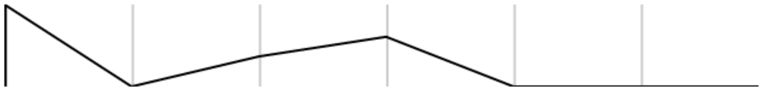
iv. D

B Interview Questions

Part 1: Intuitiveness Assessment

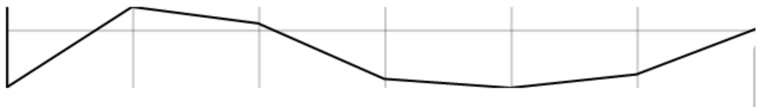
I. Which of the following shape is the most prevalent in the MATTERS - GrowthRate dataset at similarity threshold of 0.15?

i.

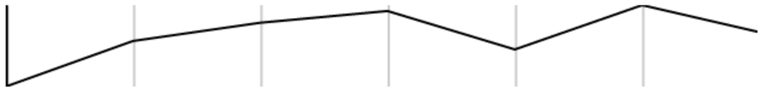


png

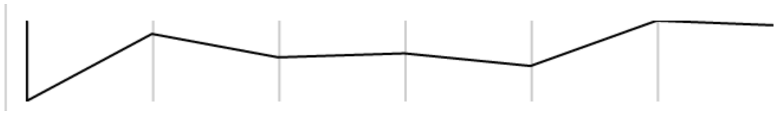
ii.



iii.

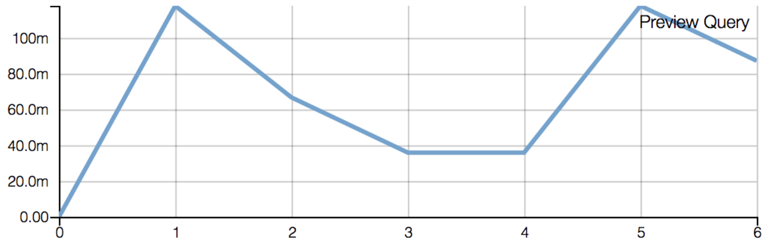


iv.

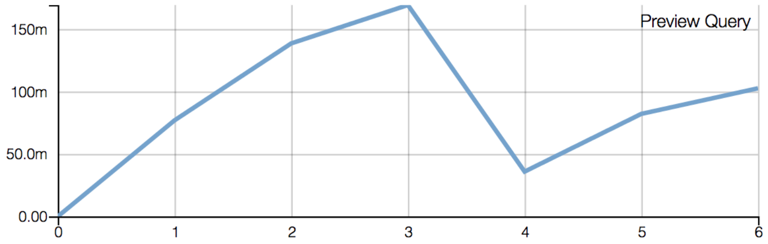


II. Which one of the following graphs represents the growth rate of Massachusetts?

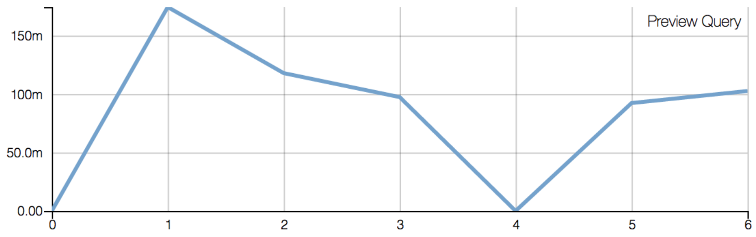
i.



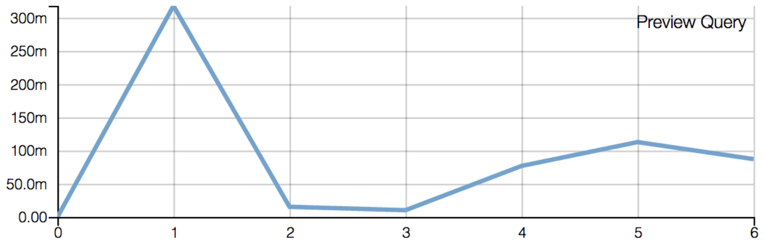
ii.



iii.



iv.



III. Which state has the GrowthRate graph that is the most similar in shape with that of Massachusetts?

  i. Akansas

  ii. Montana

  iii. Iowa

  iv. Utah

IV. How many seasonal patterns of length 60 are there in the 0th time series of the ElectricityLoadDiagram2011 dataset under Similarity Threshold of 0.3? (Not how many times the pattern repeats.)

  i. 1

  ii. 2

  iii. 3

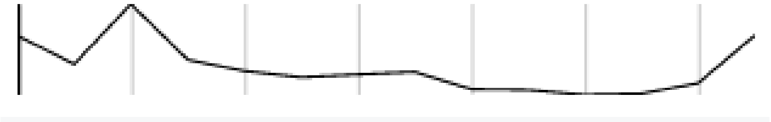
  iv. 4

V. To what extent do you think that you have understood the interface's features and functionalities? (0, it was very confusing. 5, it was very clear.)

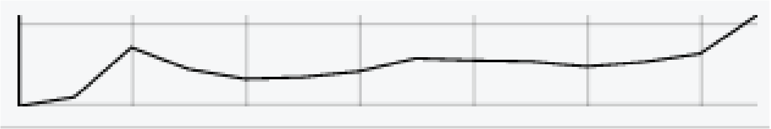
We here give the interviewees a message: ”Great work! We will now give some explanation on the structure of the interface, and see if slight guidance clears up any issues by running over similar questions again.”

We then proceed to given them a tutorial on how to use the interface. We do this to see if brief training would have an impact on performance, should the interface prove difficult to use without instruction.

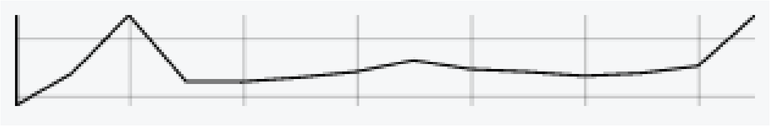
VI. Which of the following shape is the most prevalent in the MATTERS - TechEmployment dataset at similarity threshold of 0.10? i.



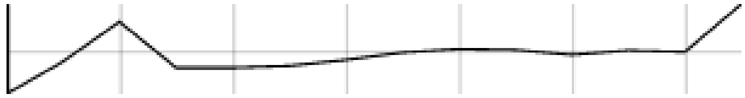
ii.



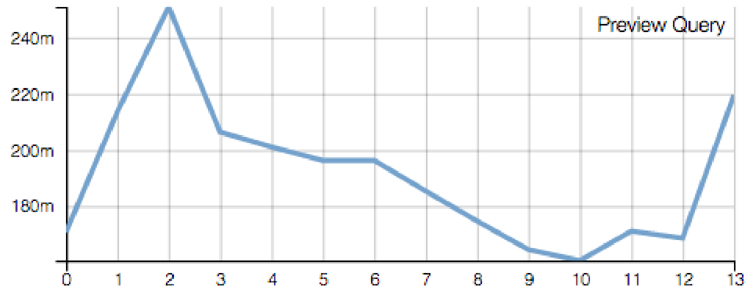
iii.



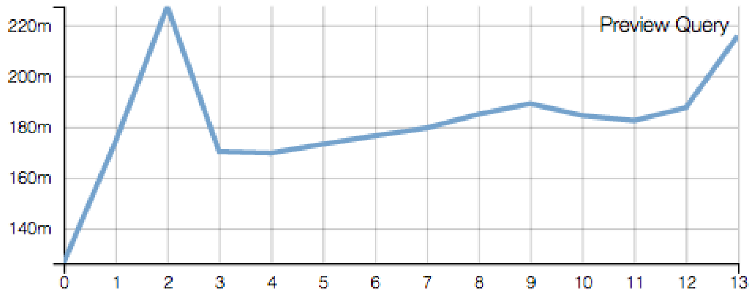
iv.



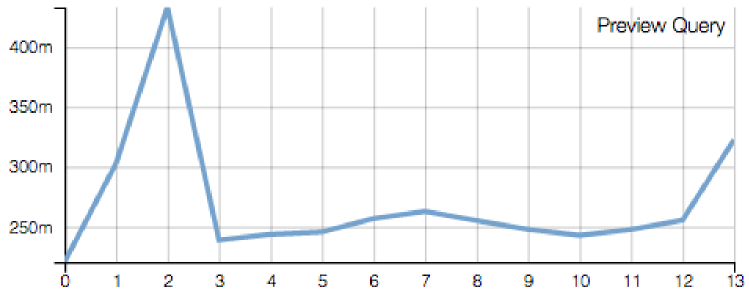
VII. Which one of the following graphs represents the TechEmployment rate of California? i.


Preview Query
240m
220m
200m
180m
0
1
2
3
4
5
6
7
8
9
10
11
12
13

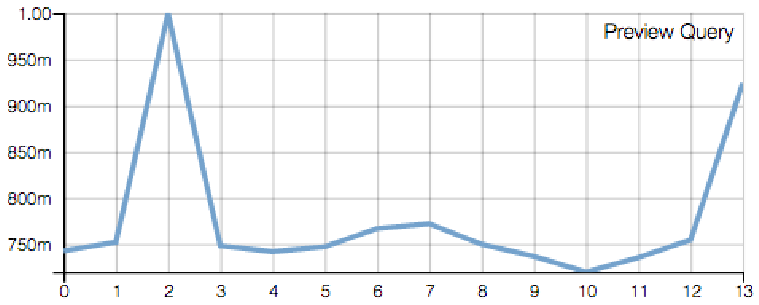
ii.


Preview Query
220m
200m
180m
160m
140m
0
1
2
3
4
5
6
7
8
9
10
11
12
13

iii.


Preview Query
400m
350m
300m
250m
0
1
2
3
4
5
6
7
8
9
10
11
12
13

iv.



VIII. Which state has the TechEmployment graph that is the most similar in shape with that of Michigan?

i. Georgia

ii. Florida

iii. California

iv. Massachusetts

IX. How many times does the first seasonal patterns of length 60 repeat for the 0th time series of the ElectricityLoadDiagram2011 dataset under Similarity Threshold of 0.3? (Not how many patterns there are of length 60.)

i. 1

ii. 2

iii. 3

iv. 4

X. To what extent do you think that you have understood the interface's features and functionalities after reviewing it a second time? (0, it was very confusing. 5, it was very clear.)

XI. How easy was it to process a dataset? (0, it was very confusing. 5, it was very clear.)

XII. How easy was it to locate and select a time series?

XIII. How easy was it to find the most similar time series to a selected time series?

Part 2: Graph type assessment
We here gave the interviewees additional background information on three of the different graphing techniques we implemented. This included the *multiple lines chart*, the *radial plot* and the *connected scatter plot*. The information is omitted here,

but very similar to the background given above in § 4.4.1-§ 4.4.4. We give the interviewee three examples to run on the interface and then to rank the different graphs on both efficacy and speed of understanding. The three examples (below) are plotted with their most similar sequences, and the interviewee compares the different graph types.

Series Label	Dataset	Threshold	Range
MI - Michigan	MATTERS - TechEmployment	0.15	0-13
Leaf - 127	Leaf	0.2	0-127
CA - California	MATTERS - GrowthRate	0.1	0-4

XVII. How well did the graph types show how similar the two sequences are for the results? (Score Each 1-5)

XVIII. Which graph do you think you understood the quickest? (Order Methods 1-3)

Part 3: Feedback
We did not limit the interviewees to the following questions below, but rather used them as an outline or direction with which to inspire critiques and ideas. Most interviewees thus ended up not answering these directly, but gave us numerous comments - detailed below in the results.

I. Any significant limitations of the interface?

II. Any particular part you liked or do you think it could help you in your work?

III. Is there a specific feature you have in mind that would help you?