KIT

Karlsruhe Institute of Technology

Interactive Visualization of Correlations in High-Dimensional Streams

Bachelor's Thesis of

Yimin Zhang

at the Department of Informatics
Institute for Program Structures and Data Organization (IPD)

Reviewer: Prof. Dr.-Ing. Klemens Böhm
Second reviewer:
Advisor: M.Sc. Edouard Fouché

01. Mar 2019 – 30. Jun 2019

Karlsruher Institut für Technologie
Fakultät für Informatik
Postfach 6980
76128 Karlsruhe

Abstract

A fundamental task of Data Mining is to estimate the correlation between the attributes of a data set. Knowing the relationship between a set of variables, one can infer useful knowledge about external, a priori unknown outcomes.

In contrast to static data, the data is often available as a stream, i.e., it is an infinite, ever evolving sequence of observations. Concepts learned at a certain time cannot be expected to hold in the future. Therefore, correlation estimation should be a continuous process.

Also, the data is often high-dimensional, i.e., it contains hundreds or thousands of dimensions. Besides the computational burden to estimate the correlation between many subsets, it becomes difficult for a human observer to extract knowledge from the results. The task becomes even more difficult if one considers correlations between more than two variables, because the size of the result increases exponentially.

The aim of this bachelor thesis is to develop a graphical interface for data scientists, dedicated to the visualization of correlation in user-given data streams. With this interface, available for example as a web-service, users may provide their own data sets. Then, the system's back-end estimates the correlations and visualizations of the results. Users interact in several ways with the interface, by setting parameters to tune the visualization. We evaluate the benefits of our interface and determine which visualization is the most appropriate, depending on specific types of user query, via controlled user studies.

Zusammenfassung

Die Abschätzung der Korrelation von Attribute in einer Datenmenge ist einer der grundlegenden Aufgaben von Data Mining. Wenn man die Beziehung von Variablen kennt, dann kann man einige nützliche Ausgaben über zusätzliche und unbekannte Informationen folgern.

Normalerweise sind die Daten als Datenfluss verfügbar, d.h. Es ist unendlich und sogar evolutionär. Die zur-zeitigen schon erkennende Begriffe und Resultaten kann man in der Zukunft nicht mehr benutzen. Deshalb muss die Abschätzung der Korrelation ständig werden.

Ein anderes Problem ist die hohen Dimensionen. Die Daten enthalten oft mehr als 100 oder sogar 1000 Dimensionen, sodass es ist schwierig für das Rechnen der Daten. Es ist auch schwer für ein Mensch um Daten zu analysieren. Wenn es um die Korrelation über mehr als zwei Variablen geht, wächst das Rechnen der Datenmenge exponentiell an. Unsere Arbeit konzentriert sich zur Zeit nur auf Korrelation über zwei Variablen und das Problem auf die Korrelationen über mehr als zwei Variablen steht noch in der zukünftigen Arbeiten aus.

Das Ziel dieser Arbeit ist die Entwicklung von einer graphischen Schnittstelle für die Daten Wissenschaftler, um die Korrelation der Daten zu visualisieren. Mit dieser Schnittstelle, als zum Beispiel Web-Service, laden die Benutzer selbst Datenmenge hoch. Danach wird das Backend von System die Korrelationen von Attribute berechnen und einige Visualisierungen von Daten ausgeben. Es ist auch möglich für die Benutzer mit der Schnittstelle Parameter aufzustellen, um die Visualisierung zu verbessern. Zum Schluss bewerten wir die Vorteile und Nachteile dieser Schnittstelle und bestimmen die beste Visualisierung anhand verschiedene Situationen und Forderungen durch einige Anwendungsfälle.

Contents

List of Figures

1 Introduction

1.1 Motivation

Correlation analysis aims at discovering and summarizing the relationship between the attributes of a data set. Knowing the relationship between a set of variables, one can infer useful knowledge about external, a priori unknown outcomes.

For example, we can measure the stock relationship via correlation coefficients in stock markets that may undergo large fluctuations of stock prices. The representing of the correlation coefficients among the five funds' returns from 1980 to 1998 by Katrina Simons is shown in Figure 1.1. Correlation coefficients describe the extent to which asset returns "move together." Correlation coefficients range in value between negative one (completely negatively correlated) and positive one (completely positively correlated), while a correlation of zero means that there is no correlation between two attributes. Performing good correlation analysis could be of great help to a financial analysis.

Fund	U.S. Stock	European Stock	Pacific Stock	U.S. Bond	U.S. Money Market
U.S. Stock	1.00	.59	.33	.29	−.05
European Stock		1.00	.53	.22	−.13
Pacific Stock			1.00	.14	−.10
U.S. Bond				1.00	.14
U.S. Money Market					1.00

Figure 1.1: Correlations Among the Five Funds' Returns, Monthly Returns, from 1980 to 1998[8]

We can see that the stock markets are positive correlated between each other, which indicates a similar behavior during this period. If we are fully aware of our current stock market, we are likely to predict the behavior of other stock markets due to the positive correlation. As a result, if we know that one stock market is performing well, we can maximize our wealth by investing other stock markets, which are positive correlated to our current stock market. If we want to minimize the risk of investing, it's better to invest in no correlated stock markets.

Analyzing the correlation between different attributes helps us to understand their relationship. In general, the correlations between attributes remain same or change gradually. If the correlation structure changes brutally, which often indicates a sudden peak or valley, we can infer that one of the attributes may have an enormous change or the relationship between them may differ thoroughly. George Filis et al. analyzed in [3] the dynamic correlation between stock market index and the crude oil price, which is shown in Figure 1.2. During the period 1987 - 2009, 6 important events occurred:

- Iraq invasion in Kuwait/first war in Iraq

- Asian economic crisis

- Housing market boom

- Second war in Iraq

- Chinese economic growth

- Global financial crisis

These events signed the brutal changes in the correlation between markets and oil price, which are printed in blue circles in the Figure 1.2.

1.2 Challenges

The challenges of analyzing high-dimensional data streams is twofold: the evolving nature of streams and the high-dimensionality.

1.2.1 The evolving nature of streams

In contrast to static data, the data is often available as a stream, i.e., it is an infinite, ever evolving sequence of observations. As the concepts learned at a certain time cannot be expected to hold in the future, correlation analysis should be a continuous process. We can see from Figure 1.2 that the correlation between market and oil price is always changing throughout the time. The correlation values even have brutal changes, when meet up with important events. Therefore, it's quite hard for the data scientists to predict the correlation value at next timestamp.

1.2.2 The high-dimensionality

Also, the data is often high-dimensional, i.e., it contains hundreds or thousands of dimensions. In the case of streams with many dimensions, it is difficult to extract actionable insights from the correlation matrix, as the number of pairs of attributes increases quadratically and the coefficients evolve over time in unforeseen ways. For the pairwise correlation analysis of any data steam with n components, one need to compute the correlation between $\dfrac{n*(n-1)}{2}$ pairs, i.e., with $n = 5$, one needs to compute 10 pairs, with $n = 10$, one needs to compute 45 pairs. Therefore, it becomes difficult to visually keep track of correlation and impossible to understand the result of the correlation analysis as the number of attributes increases. The Figure 1.3 shows the visualization of different numbers of attributes. As an instance, with the developing number of attributes, it becomes even harder for users to compare the correlation value between two pairs using

Heatmap as the standard tool to visualize the data.

1.3 Goal of the thesis

The goal of this thesis is to propose and evaluate new tools for the interactive visualization of correlation in high-dimensional streams. We are going to compare different visualization methods. Our interactive interface aims at providing a visualization of correlations in streams, which may change arbitrarily over time, for people. Users are able to choose a certain period of time to perform the correlation analysis and visualization. In our thesis, we only focus on pairwise relationships and the correlations between more than two variables may remain to be discovered in the future work. We are going to answer the following three questions in the thesis:

- What visualization method is the most appropriate to visualize correlation for various specific user information needs?

- What visualization method is the most suitable to visualize characteristics of a data set?

- What are the desirable features of a correlation monitoring interface?

To answer these questions, we first search for three interactive visualization methods and then develop an interface to visualize correlations using these methods. This interface can be available in a browser and should be user friendly, which means that users provide their own data sets and the system's back-end calculates the correlations and then provides the visualization of the results. Users are also able to interact in several ways with this interface, such as setting parameters to tune the visualization. At last, we evaluate the benefits of our interface systematically via controlled user studies and discuss the advantages and disadvantages of each visualization method when meet up with different scenarios. Based on the results, we discover the appropriate visualization methods of correlations in high-dimensional streams and the desirable features of the interactive interface.

1.4 Thesis outline

This thesis is divided into three parts: the literature review of visualization methods, the design and implementation of interface, and the evaluation of this interface.
In Chapter 2, the first part of the thesis, we introduce three state-of-art methods for correlation visualization: Heatmap, Bar Graph and Force-Directed-Graph.
Chapter 3 is about the related work. In Chapter 3, we give an overview[1] of parallel coordinates and scatterplot matrices as examples of multidimensional visualization techniques. Also, we introduce an interactive Framework for Exploring and Understanding Multivariate Correlations "FEXUM"[2] created by Louis Kirsch et al., which uses Force-Directed-Graph as the visualization method for the data set.

Chapter 4 is the main part of this thesis. Section 4.1 is the design of the interface and Section 4.2 describes the implementation of this interface.

We evaluate the developed interface in Chapter 5. It contains not only the experimental settings of controlled user studies for evaluation in Section 5.1, but also the results of controlled user studies and feedbacks in Section 5.4.

The summary of the whole thesis comes in the end.
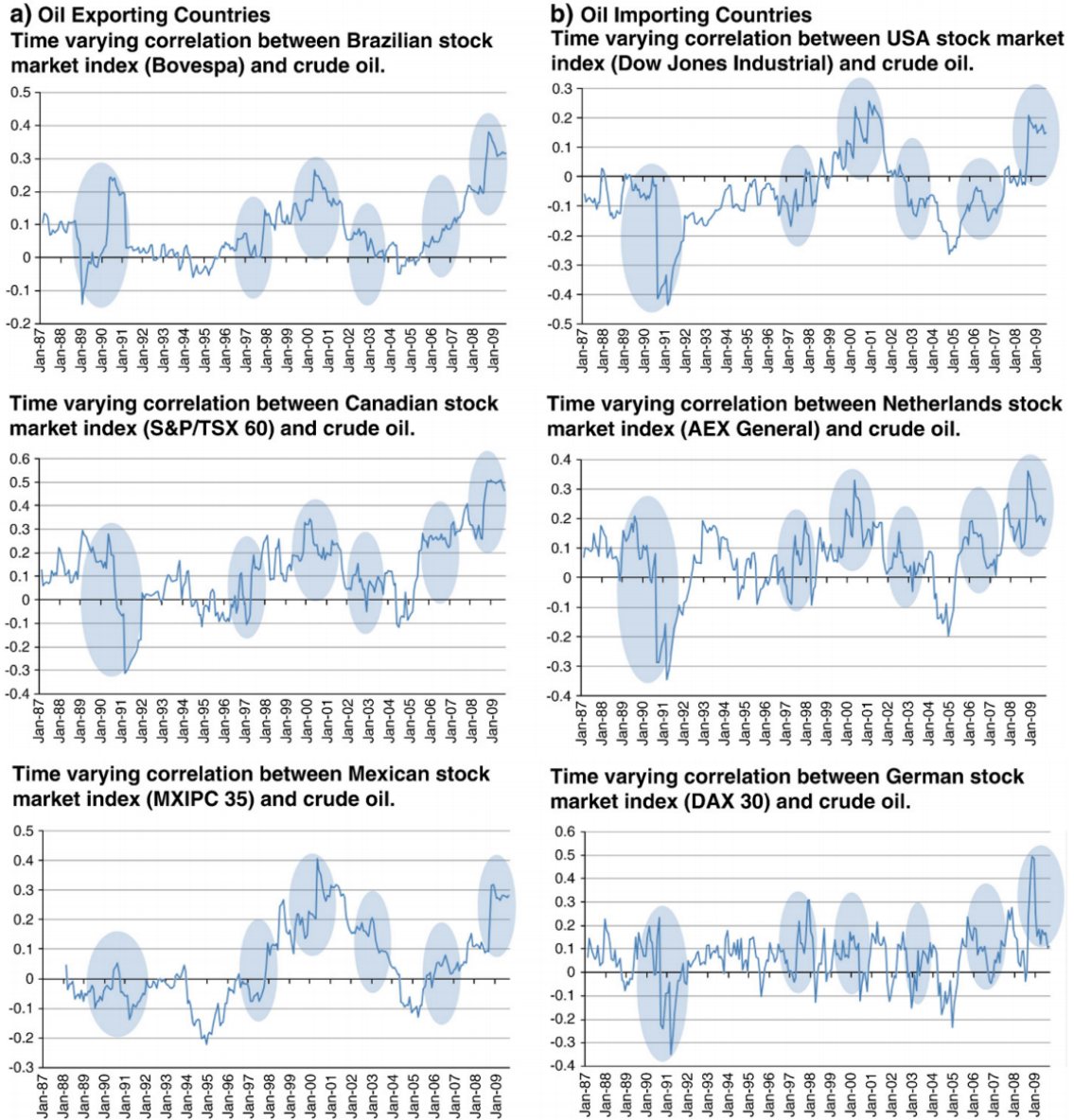
Figure 1.2: Dynamic correlation between stock market index and the crude oil price[3]

(a) Correlation matrix of 5 attributes

(b) Correlation matrix of 10 attributes

(c) Correlation matrix of 40 attributes
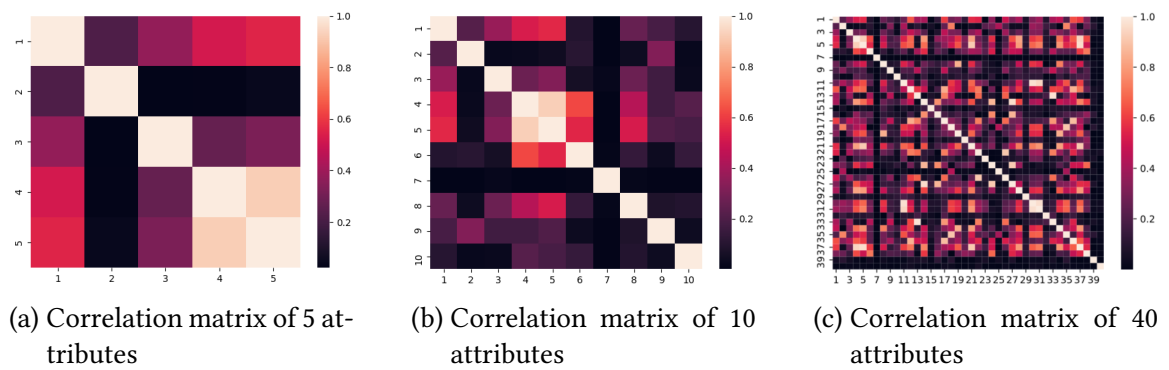
Figure 1.3: Visualizations of different numbers of attributes

2 Visualization Methods

In this chapter, we first introduce the correlation matrix in Section 2.1 and then introduce three useful interactive visualization methods Section 2.3 Heatmap, Section 2.4 Bar Graph and Section 2.5 Force-Directed-Graph, and give examples of each visualization method using the same data set in Section 2.2.

2.1 Correlation Matrix

The most familiar measure of dependence between two quantities is the Pearson product-moment correlation coefficient[7], known as "Pearson's correlation coefficient". It is obtained by dividing the covariance of the two variables by the product of their standard deviations. The population correlation coefficient $\rho_{X,Y}$ between two random variables X and Y and standard deviations σ_X and σ_Y is defined as

$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{2.1}$$

where cov means covariance, and corr is a widely used alternative notation for the correlation coefficient. The Pearson correlation is defined only if both of the standard deviations are finite and non-zero.
The standard tool of correlation analysis is the computation of a correlation matrix

$$\rho = \begin{bmatrix} \rho_{1,1} & \cdots & \rho_{1,n} \\ \cdots & \cdots & \cdots \\ \rho_{n,1} & \cdots & \rho_{n,n} \end{bmatrix} n \in \mathcal{N}^* \tag{2.2}$$

for n variables. The correlation matrix is used to investigate the dependence between multiple variables at the same time. In fact, we are only interested in the lower half of the matrix because of its symmetry and invariant diagonal line.

2.2 Data Set Mtcars

Data Set Mtcars represents Auto MPG Data Set, which can be found in UCI Machine Learning Repository by url: https://archive.ics.uci.edu/ml/datasets/auto+mpg. It is a data frame with 32 observations on 11 variables. This data set can be seen as a standard data set widely-used in the field of data analysis. Section 2.3, Section 2.4 and Section 2.5 all use this data set to visualize.

2.3 Heatmap

A common visualization is the heatmap[9], which is originated in 2D displays of the values in a data matrix. The Figure 2.1a is a heatmap of a correlation matrix, in which the variables with strong correlation (high values) are printed in light colour and those with low correlation are in dark colour.

2.4 Bar Graph

A bar graph[4] presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. We use vertical bar graph in our developed interface. In a vertical bar graph, the x-Axis shows the specific categories being compared, and the y-Axis represents a measured value, which is the correlation value in our situation. Bar graphs provide a visual presentation of categorical data, which are usually qualitative. Figure 2.1b is a bar graph, using the same data set in Figure 2.1a.

2.5 Force-Directed-Graph

Force-Directed-Graph[5] assigns forces among the set of edges and the set of nodes of a graph drawing. The purpose of it is to position the nodes of a graph in two-dimensional or three-dimensional space so that all the edges are of more or less equal length and there are as few crossing edges as possible. In such a simulation, the forces are applied to the nodes, pulling them closer together or pushing them further apart. This can be used to simulate the relationship of different attributes throughout the time. Figure 2.1c is the example of Force-Directed-Graph, which also uses Data Set Mtcars.

(a) Heatmap as the visualization method

(b) Bar Graph as the visualization method

(c) Force-Directed-Graph as the visualization method
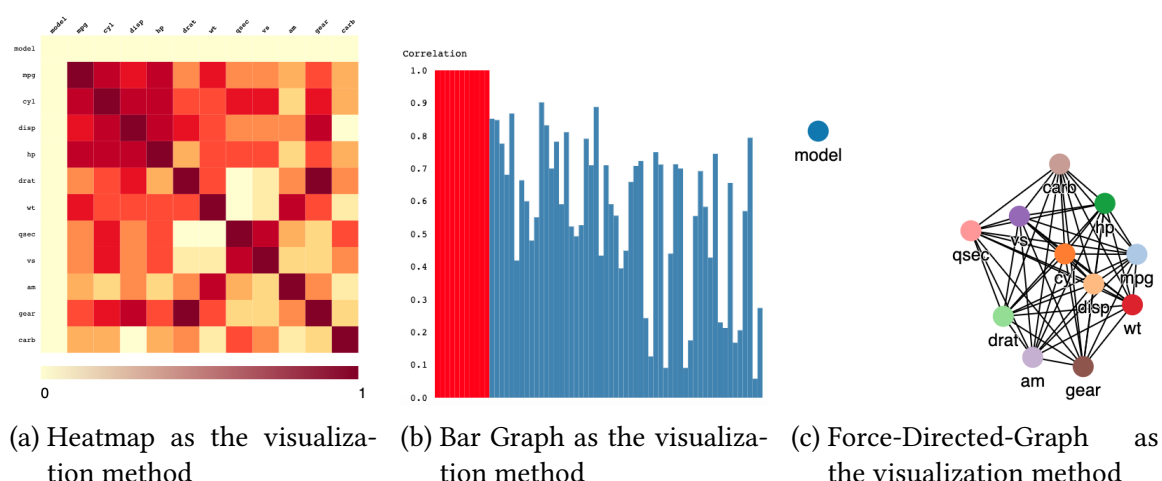
Figure 2.1: Three visualization methods of the Data Set Mtcars

3 Related Work

With the exponentially increasing amount of acquired multivariate data, several multi-dimensional visualization techniques have been proposed during the last decades. Parallel coordinates and scatterplot matrices are widely used to visualize multi-dimensional data sets. But these visualization techniques are insufficient when the number of dimensions grows. To solve this problem, different approaches to select the best views or dimensions in advance have been proposed in the last years.
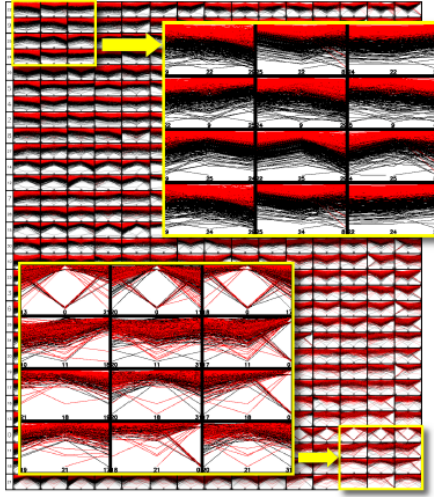
Figure 3.1: Parallel coordinates matrices[1] for the data set

Georgia Albuquerque et al. presented three new methods to explore multivariate data sets: a parallel coordinates matrix (Figure 3.1), in analogy to the well-known scatterplot matrix, a class-based scatterplot matrix that aims at finding good projections for each class pair (Figure 3.2), and an importance aware algorithm[1] to sort the dimensions of scatterplot and parallel coordinates matrices. They aims at providing a visualization of the whole data set, not the correlations of attributes in this data set. As we focus on the correlations only between each two attributes in our thesis, it is no need for us to have parallel coordinates in our framework.

In Section 1.1, we have pointed out the use and importance of correlation analysis between different attributes, which helps people to understand the relationship of attributes in a data set. Unlike the work of Georgia Albuquerque et al.[1], our interface provides the visualization of correlation values in a data set.

As a high-dimensional data set may contain many redundant features, feature selection becomes an essential step for correlation analysis. Louis Kirsch et al. developed
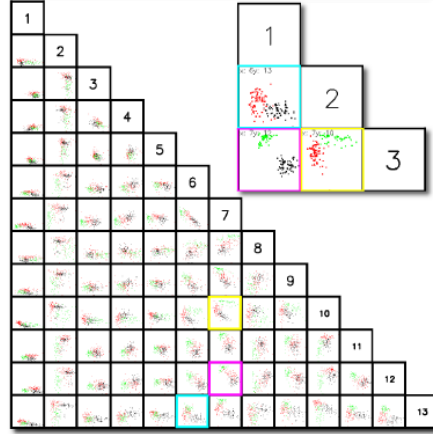
Figure 3.2: Class-based scatterplot matrices[1] for the data set

an interactive Framework for Exploring and Understanding Multivariate Correlations (FEXUM)[2] to simultaneously visualize all feature correlations to the target and pairwise correlations using Force-Directed-Graph. This visualization provides a layout in which a smaller distance between two features denotes a greater redundancy. In Figure 3.3, nodes represent features and weighted edges represent distances. FEXUM provides users with an understanding of how features interact with each other in terms of redundancy so that they can easily find influential value ranges in the analysis view and make feature selection.

In our developed interface, we give the users an overview of all correlations in the data set. Instead of choosing a target attribute to focus on, our system represents the whole correlations of the data set. Unlike FEXUM[2], Force-Directed-Graph is not the only visualization method in our system. Heat map and Bar graphs are alternative visualization method for the users so that the users can choose the most suitable visualization method in their opinion.

In Subsection 1.2.1, we have discussed that the correlation analysis should be a continuous process. FEXUM enables the users to upload their own data sets and visualize them. However, in our system, the uploaded data set by users can be a data set of data stream so that the users can choose a period of time to perform the correlation analysis and visualization. Our goal is to visualize a concise but useful summary of correlations in the stream over time.

Figure 3.3: Features drawn using a force-directed graph (right), with the target high-lighted in green. An analysis view of two features (left) for inspecting the correlations.[2]

4 Interface

To answer the questions we mentioned in the Section 1.3, we develop an interface in a browser available as a web-service. The entire framework is open source and available online via url https://github.com/yimin95/InteractiveVisualization. In Section 4.1, we describe the mock-up of this interface and its available functions. And we introduce the details of implementation in Section 4.2.

4.1 Design

Figure 4.1 is the mock-up of this interface. This interface can be used with a wide range of data sets as csv files, supplied through upload by the user. The first line of such file is the list of attributes' names. The corresponding values of each attributes are displayed in the following lines. The following table is an example of a csv file, while the blanking blocks are values of corresponding attributes.

Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6

After the calculation in the back-end, the visualization of data correlation is shown in the website, for example, via a Heatmap. In the mock-up, a sliding window with start point and step size is supposed to be used to represent the continuous process of data throughout the time. Also, we have some simple user settings, such as changing the window size, setting the minimum and maximum of correlation to be visualized.

In our interface, we focus on the absolute value of the correlation. **Minimum** is the minimal value of the correlation value and it is set to 0 as default. **Maximum** is the maximal value of the correlation value and it is set to 1 as default. **The window size** represents the size of the selected timestamps of the current data set. Timestamp represents the instance of data set. As we aim at data stream, each instance of data set is the representation for values of attributes at a certain timestamp. The width of panel on the slider indicates the size of the current window. With the sliding of panel, visualizations of correlations during different time periods are shown on the website, representing the visualization of correlation in data streams. **the step size** gives out the difference between two adjacent movements of slider. Combined with the step size, the user is able to get the visualization of a certain time period by sliding the sliding window. **The current point** represents the starting point of current selected group of timestamps. It is also possible for the user to input the starting point of timestamps.
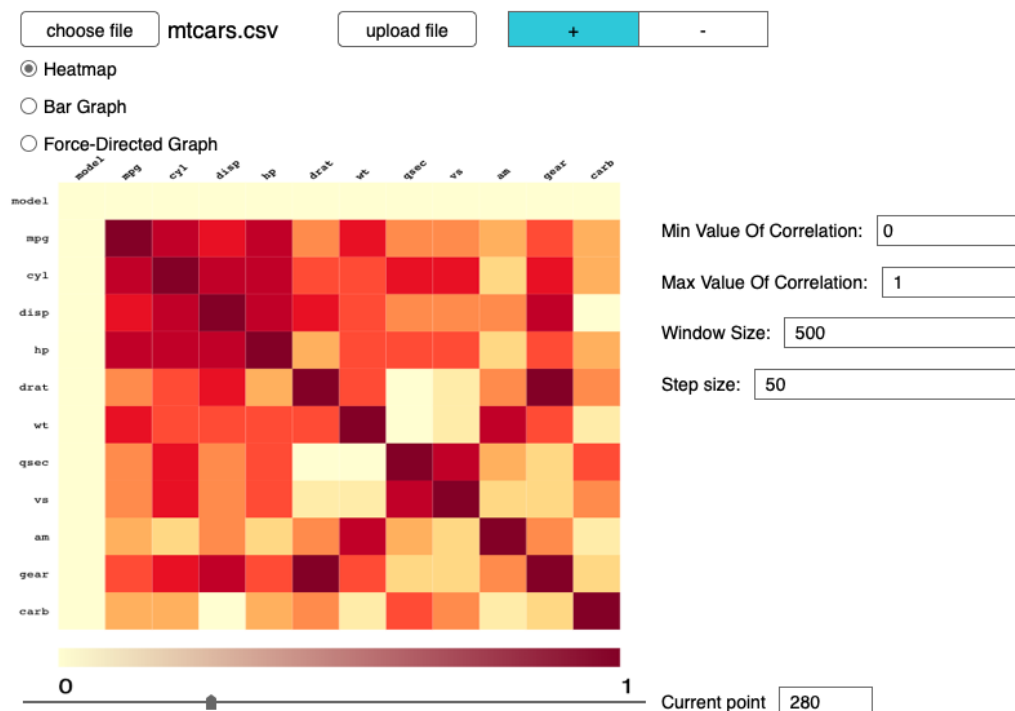
Figure 4.1: Mock-up of the interface

4.2 Implementation

As the interface is available by the web service, Javascript is the programming language for developing. In our project, we mainly use D3.js to implement the visualizations of correlations in high-dimensional data streams.

4.2.1 D3.js

D3.js (Data-Driven Documents)[6] is a data-driven JavaScript library for producing dynamic, interactive data visualizations in web browsers. It makes use of the widely implemented SVG, HTML5, and CSS standards and allows great control over the final visual result. SVG stands for Scalable Vector Graphics which is technically an XML based markup language. It is a great tool to display icons, logos, illustrations or charts, which is supported in all major browsers and requires no third-party lib because of owning the DOM interface.

4.2.2 Details

The Figure 4.2 is an overview of our developed interface, which is available in a browser as a web-service. Users can press the "Choose file" button to upload their own data sets.

Figure 4.2: Overview of the interface

After pressing the **Upload** button, the data set will be stored at the back-end. The system calculates the correlations of each pair of attributes and the corresponding elements for performing the visualizations. If the calculation is finished, a confirm window will pop up to inform the maximal window size of the current data set to the user. The maximal window size is the whole size of the uploaded data. Also, the step size will be set to 1 and the window size will be set to the maximal window size after uploading the data set, see Figure 4.3.

Figure 4.3: After uploading a csv file

Users are able to change the **minimum**, **maximum**, **window size**, **step size** and **current point**. After setting these values and pressing the **update** button, the visualization will be re-performed and the width of panel will be set to corresponding window size, seeing Figure 4.4.

Our interface provides three visualization methods: Heatmap, Bar Graph and Force-Directed-Graph. After selecting the corresponding radio button, a visualization will be drawn on the web site. The Figure 4.5 shows three visualizations based on the same data set we

Interactive Visualization of Correlations in High-Dimensional Streams

Figure 4.4: After pressing the "update" button

introduced in Section 2.2. Figure 4.5a is the Heatmap. Each box represents a pair of attributes and the color of the box represents the correlation value of this pair. Figure 4.5b is the Bar Graph, in which the height of each bar represents the correlation value of each pair of attributes. As 0 is not likely to be seen in case of many pairs of attributes, we paint the hole bar red and set the height to the maximal height. And Figure 4.5c is the Force-Directed-Graph, in which the length of link between two nodes represents the correlation value of this pair of attributes. The shorter the linked distance is, the bigger is the correlation value. Users can also change the minimal and maximal correlation value they want to visualize, and slide the slider to see different time period of the current data set.

(a) Select Heatmap as the visualization method

(b) Select Bar Graph as the visualization method

(c) Select Force-Directed-Graph as the visualization method
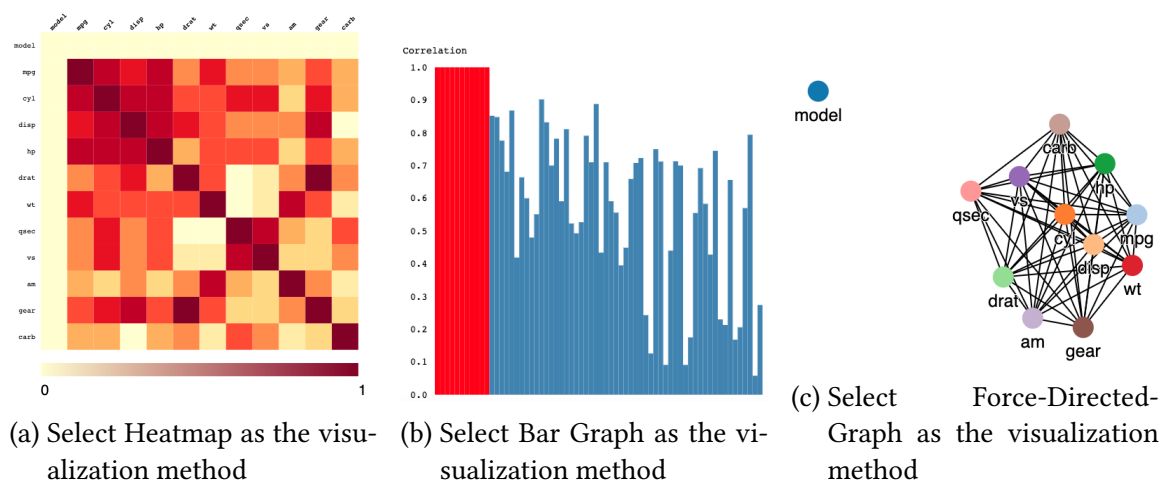
Figure 4.5: Select different visualization methods

5 Evaluation Via User Studies

5.1 Experimental Settings

In this thesis, we compare three different visualization methods, namely the so-called "Heatmap", "Bar Graph" and "Force-Direct-Graph". The participants are assessed with respect to different conditions, to which they will be assigned randomly. Participants are not aware of the condition they are assigned to:

- **Condition A**: Participants can only use the Heatmap and Figure 5.1 is its mock-up
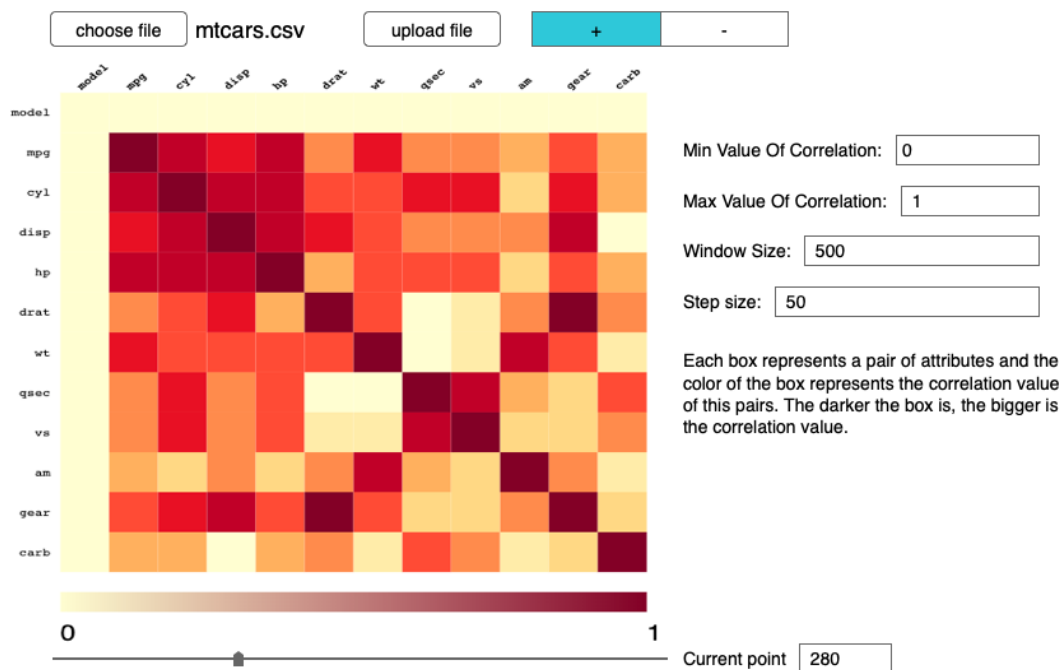
Figure 5.1: Mock-up for Condition A

- **Condition B**: Participants can only use the Bar Graph and Figure 5.2 is its mock-up

- **Condition C**: Participants can only use the Force-Directed-Graph and Figure 5.3 is its mock-up
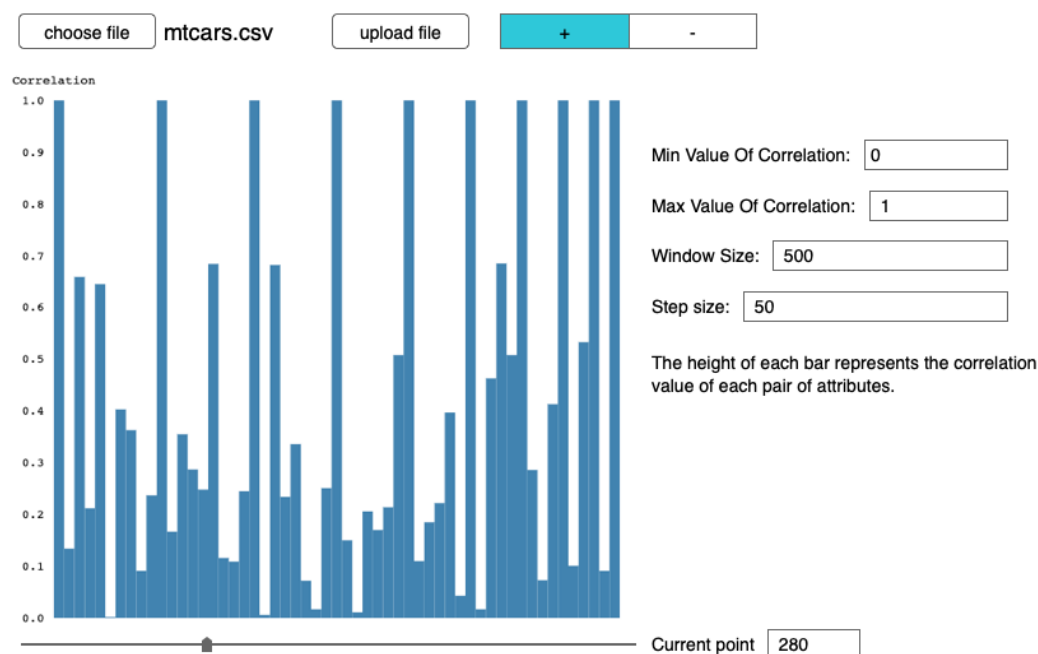
Figure 5.2: Mock-up for Condition B

- **Condition D**: Participants can use any visualization method they want,which are mentioned in Condition A,B and C, and Figure 4.1 is its mock-up

5.1.1 Participants Profiles

The participants we are looking for are adults who at least have the basic knowledge about browsing web page, typically 14 to 40 years old. We will sample a large share of our participants from the pool of students at KIT, so we may expect basic knowledge of computer science and data analysis.

Knowledge in data analysis and correlation analysis should not be required to use the prototype for the visualization of a data set. Still, we hypothesis that participants with prior exposure to data visualization will require less time to fulfill the tasks.

5.1.2 Data Set

The participant of user study are asked to use a data set taken at random from a pool of 3 data sets in total for completing the tasks. **Data Set 1 (DS1)** has 10 attributes within 2000 timestamps. **Data Set 2 (DS2)** has 20 attributes within 2000 timestamps. **Data Set 3 (DS3)** has 40 attributes within 2000 timestamps. These 3 data sets are actually the subsets of a real-world data set. For the evaluation, we make some modifications of this data set: reducing the number of attributes and selecting only 2000 instances of it. It is obvious
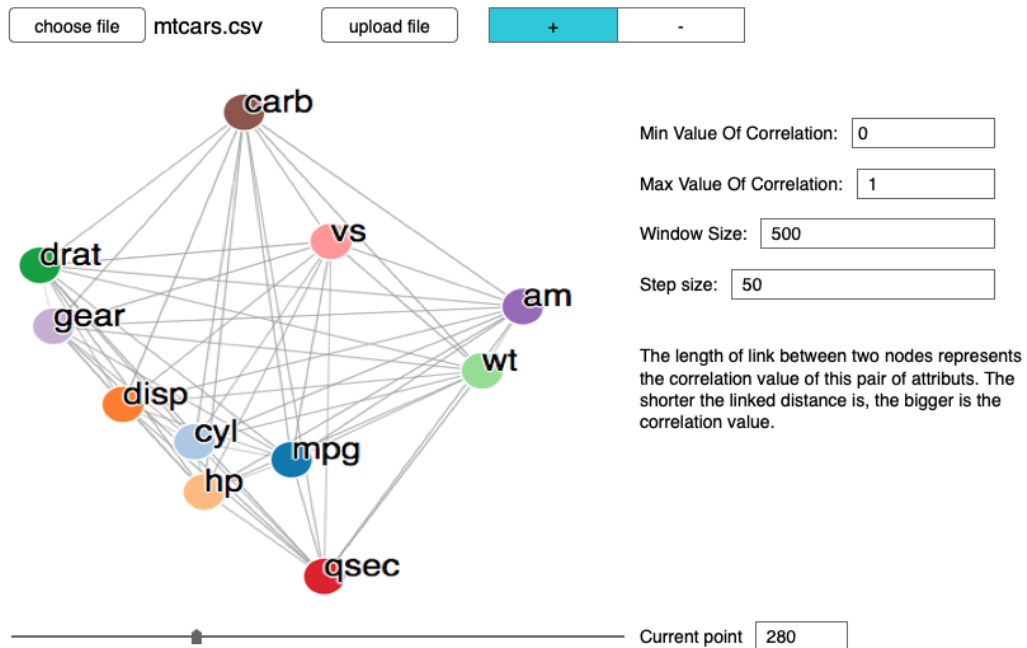
Figure 5.3: Mock-up for Condition C

that the difference between DS1, DS2 and DS3 are their level of difficulty according to their number of dimensions. In practice, this means that the same question will be harder to answer with DS3 than DS2 and DS1.

5.1.3 Settings of interface

The window size is set to 200 and the step size is set 50 for all three data sets, which cannot be changed by the participants.
During the user study, the participants are only able to change the minimum and maximum to filter the correlation values. Also, they can slide the sliding window or input "current point" to get the visualization graph of current timestamp.

5.2 Questionnaire

Before answering the questionnaire, all participants are asked to sign a Consent Form, shown in Section 7.1. The questionnaire is divided into 3 parts. In the first part of questionnaire Section 5.2, the participants are asked to give some basic information. In the second part of questionnaire, the participants have to answer some questions using different visualization methods. Subsection 5.2.2 shows the question types and each question type will be asked three time in the user study. Subsection 5.2.3 is the last part of question-
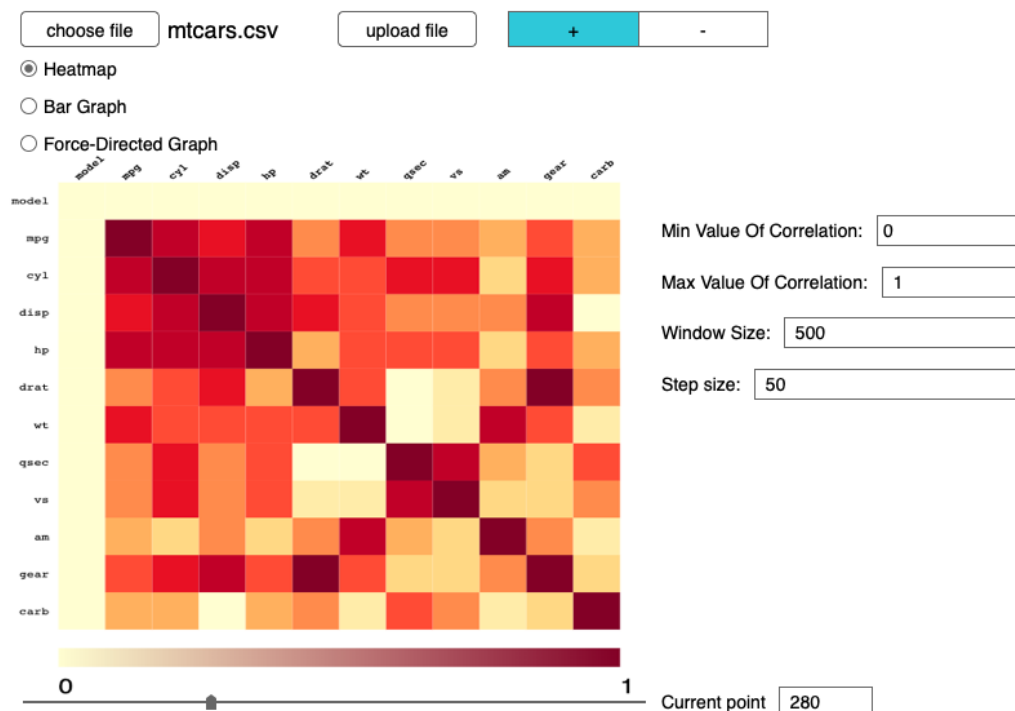
Figure 5.4: Mock-up for Condition D

naire, which introduces the feedback of using this interface. Participants under Condition A/B/C are asked to give feedback of the corresponding visualization methods they use, while participants under Condition D have to give feedback of each visualization method and the whole system.

5.2.1 Basic Information

- Field of study

- Number of semester

- Age:

| 18-24 | 25-30 | More than 30 |

- Gender:

| M | F | Do not wish to answer |

- How familiar are you with the following concepts?

	1 (Unfamiliar)	2	3	4	5	6	7 (Very Familiar)
Correlation analysis							
Data analysis							
Data visualization							

5.2.2 Visualization

- How many attributes are available in this data set?

- How many pairs of attributes are available in this data set?

- What is the correlation value between ***Attribute A*** and ***Attribute B*** at ***Timestamp T***, or the probable range?

- Which pair of attributes has the biggest correlation at ***Timestamp T***?

- Which pair of attributes has the smallest correlation at ***Timestamp T***?

- The following statement is true or false: "The correlation value between ***Attribute A*** and ***Attribute B*** remains the same at ***Timestamp T1*** and at ***Timestamp T2***"?

  | True | False |

- Which pair(s) of attributes has/have a correlation value that is not smaller than ***X*** at ***Timestamp T***?

- Which pair(s) of attributes has/have a correlation value that is not bigger than ***X*** at ***Timestamp T***?

5.2.3 Feedback

5.2.3.1 Participants Under Condition A/B/C

Participants under Condition A, B and C are asked to give feedback based on following questions:

- Please write down the visualization method you have used and rate for it.

	1 (Strongly Disagree)	2	3	4	5	6	7 (Strongly Agree)
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

- In your opinion, what are the strengths of this visualization system?

- In your opinion, what are the weaknesses of this visualization system?

- Do you have any suggestions for improving this visualization system?

5.2.3.2 Participants Under Condition D

Participants under Condition D, who are able to use either of 3 visualization methods, are asked to give feedback to the whole system, which is also quite similar to the one for participants under Condition A, B or C. In addition, they need to give ratings for each visualization method.

- As you are able to use all visualization methods during the research, which one of the method, in your opinion, is the most helpful method to fulfill the task?

- And which of the method did you use the most?

- Please rate for *Heat map*:

	1 (Strongly Disagree)	2	3	4	5	6	7 (Strongly Agree)
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

- Please rate for *Bar Graph*:

	1 (Strongly Disagree)	2	3	4	5	6	7 (Strongly Agree)
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

- Please rate for *Force-Directed Graph*:

	1 (Strongly Disagree)	2	3	4	5	6	7 (Strongly Agree)
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

- Please rate for *whole system*:

	1 (Strongly Disagree)	2	3	4	5	6	7 (Strongly Agree)
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

- In your opinion, what are the strengths of this visualization system?

- In your opinion, what are the weaknesses of this visualization system?

- Do you have any suggestions for improving this visualization system?

5.3 Script For Conducting The Experiment

The following information will be given to the participants, orally:
Thank you for participating to this experiment. My name is Yimin Zhang. I am doing my Bachelor thesis in the Institute for Program Structures and Data Organization (IPD Böhm).
The goal of the experiment is to evaluate an interface that I have developed for my thesis. First of all, please read and sign the consent form. If you have any problems during the study, please be free to talk to me.
(After signing)
Please fill in the blanks of the website about the basic information, which is the first part of the questionnaire: Section 1.
(After Part 1)
Your task from now on is to upload the data set and to answer the questions in Part 2. You are free to use all the elements on the web page to ease the task. Please cross a question off, if you cannot answer the question. The time you need to answer for each question will also be recorded.
(After Part 2)
The last task for you is to provide your personal feedback for this interface. This is the third part of the questionnaire.
(After Part 3)
Thank you for participating in my study. If you are interested in my thesis or may have further questions, please be free to contact me using the contact information in the consent form. I wish you a nice day.

5.4 Result

We invited 22 students studying various fields at KIT to participate the user study. They are all studying 6 or higher semester and below 30 years old, 12 female students and 10

male students. Figure 5.5 shows the statics of basic information of these participants. The most unfamiliar concept for them is the data visualization, which reaches the average of 2.1. They are more familiar with correlation and data analysis, both with the average of 3.1.

5.4.1 Statics about the visualization part

Questions about the number of attributes and pairs of correlations are answered correctly by all the participants. The time they used to answer the other questions are recorded and analyzed in Figure 5.6. Figure 5.7 shows the statics about the average time of participants using different data sets to finish each question type. It is obvious that both the average time of using Heatmap and Bar Graph is close whatever data set the participants use. When using Force-Directed-Graph, the participants need more time. The time for participants who can use all three visualization methods is also similar to the time using Heatmap/Bar Graph.

Figure 5.8 shows the average time of participants using different visualization methods to finish each question type. We can conclude from the figure that the time for the participants to finish the questions is mostly in direct proportion to the size of the data set, which is also related to the number of pairs for correlations.

Figure 5.9a shows the accuracy rate of each question type using different visulization methods. All the accuracy rates are over 50%. When being able to use Heatmap or 3 visualization methods at random, the accuracy rate even climbs to 75%. However, for questions asking about the precision, like **Q2.3**: What is the correlation value between ***Attribute A*** and ***Attribute B*** at ***Timestamp T***, or the probable range?, the participants are not able to give out their answers by using Force-Directed-Graph. We can infer that Force-Directed-Graph is more suitable to analyze the relationship between attributes than to have a precise value. The accuracy rate of each question type using different data sets is shown in Figure 5.9b. We can see that using the smallest data set: Data Set 1, always results in the highest accuracy rate.

5.4.2 Feedback

From the survey, we have found out that the Heatmap and Bar Graph are the most often-used visualization methods. We can infer that although they can use three visualization methods in random, they are not quite likely to use Force-Directed-Graph to answer the questions. The ratings for each visualization method and the interface is displayed on the following table:

	Heatmap	Bar Graph	Force-Directed-Graph	Interface
Intuitive	5.83	6.17	3.5	6
Convenient	5.33	5.67	3.17	5.83
Interactive	5.83	5.67	4.17	6.16
Useful	6	5.67	3.33	5.83
Complicated	2.67	3	4	3.17
Efficient	5.33	5	3.33	5.83
Effective	5.5	5.67	3.67	5.67

From the table, we can see that the Force-Directed-Graph is the most complicated visualization methods for participants. Most participants found Heatmap and Bar Graph very intuitive in data visualization. In the feedback, intuition is one of the strength of this visualization system. In addition, this system meets many demands for data processing. As we have three visualization methods, each method can be used for different use to ease the correlation analysis. It is also interesting to see the change of graphs by sliding the sliding window.

The maximal value is very easy to find out by the interface, but when it comes to a relative small correlation value, it is quite hard for participants. Although different color of blocks represents different value in Heatmap, the blocks looks the same when the difference of two values is small, e.g. 0.005. Also, when the number of correlations is big, it's labored to see through the bars to find the change of one pair of attributes. Thanks to the filtering of minimal and maximal values of correlation values, the participants can save some time and strength to analyze the data.

The participants are glad to see more visualization methods and to use more functions of this interface. It is suggested that the interface is not only used for correlation analysis, but also for data analysis and have functions like mean, variance and median. Also, it could be helpful to output the exact correlation value of one pair by inputting the names of attributes. For Bar Graph, sort functions to re-arrange the order is quite useful to find certain pairs. As the Force-Directed-Graph shows the relationships of attributes, it is interesting when one node is pointed, only the related links will be shown. Also, only showing small sub-groups of the attributes can be helpful.

(a) Field of study

(b) Number of semester

(c) Age

(d) Gender

(e) How familiar are you with the following concepts?
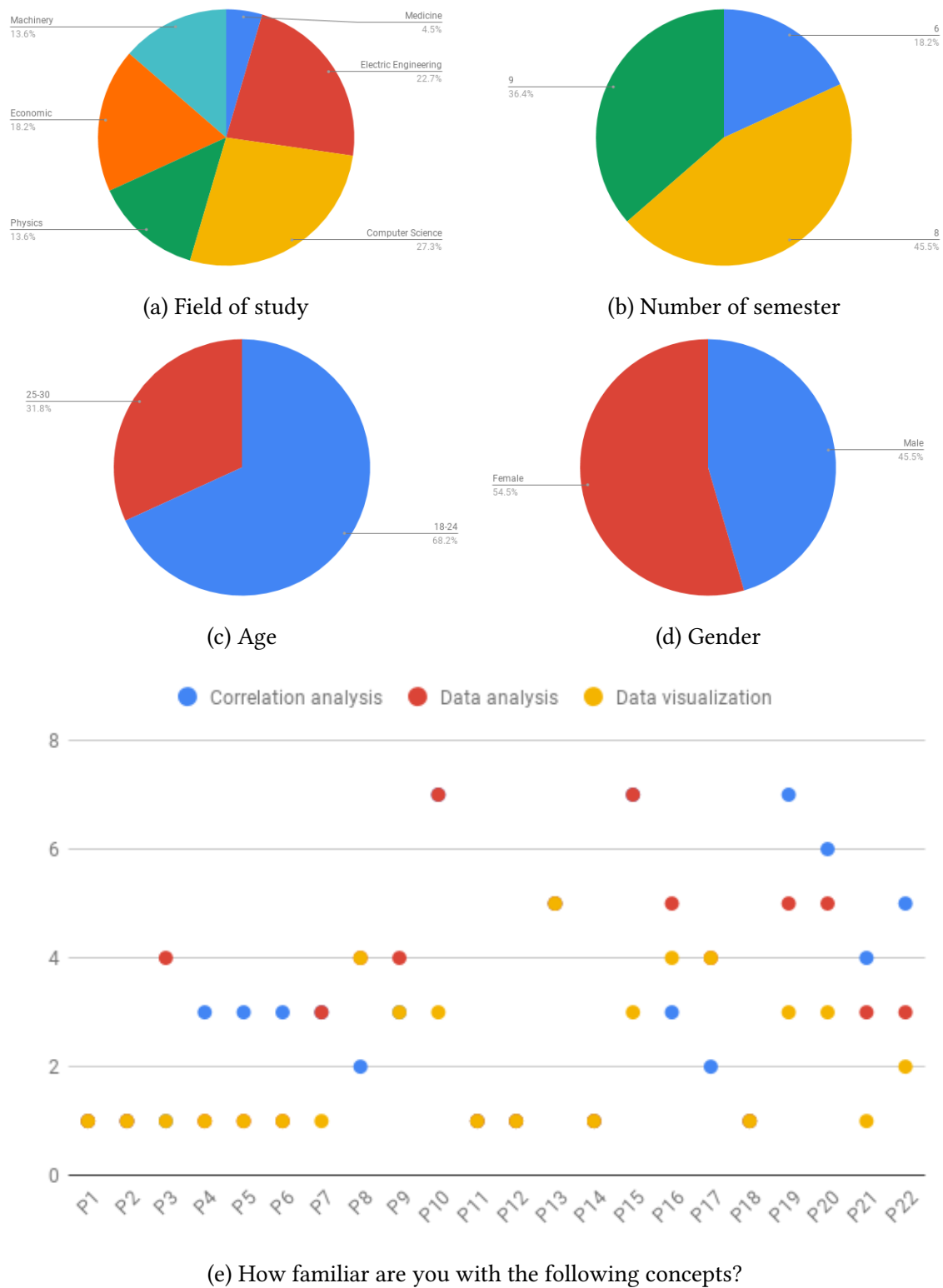
Figure 5.5: Statics of basic information
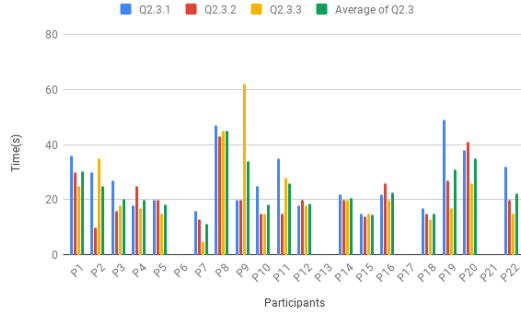
(a) Time for Question Q2.3
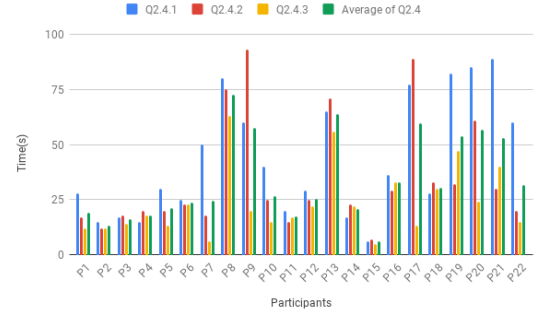
(b) Time for Question Q2.4
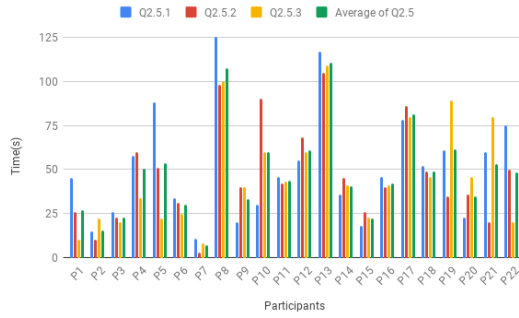
(c) Time for Question Q2.5
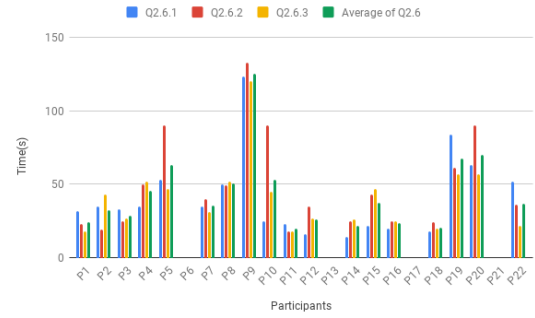
(d) Time for Question Q2.6
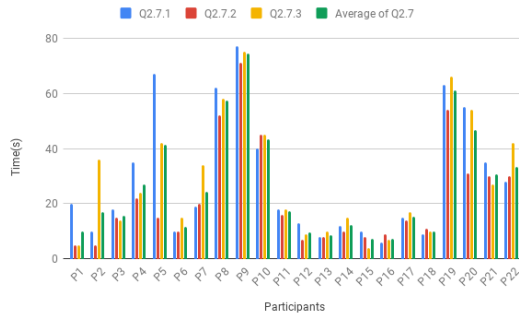
(e) Time for Question Q2.7
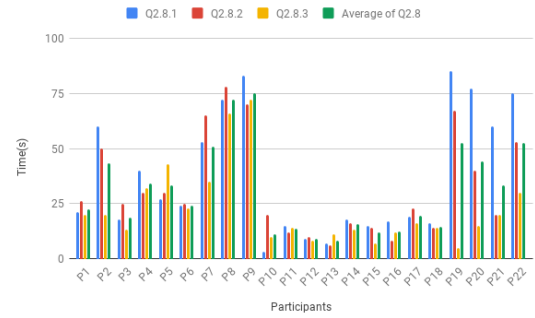
(f) Time for Question Q2.8

Figure 5.6: Time of each participant finishing questions
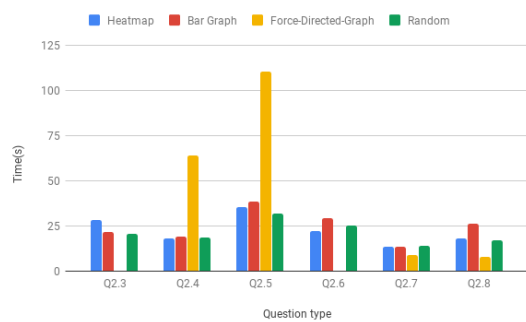
(a) Use Data Set 1
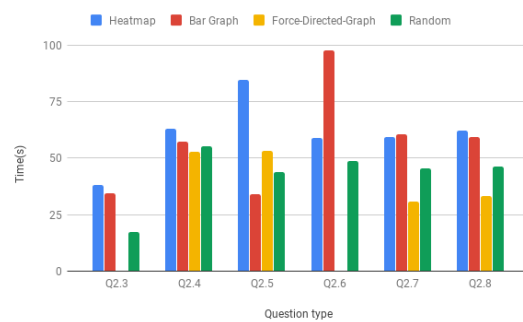
(b) Use Data Set 2
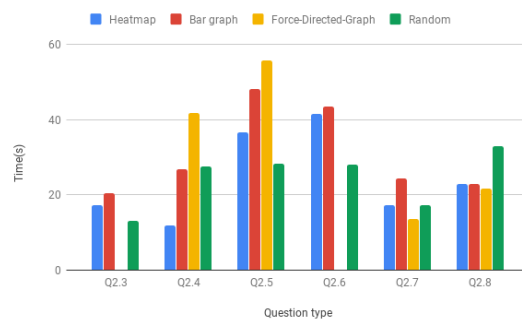
(c) Use Data Set 3

Figure 5.7: Average time of participants finishing each question type using different data sets
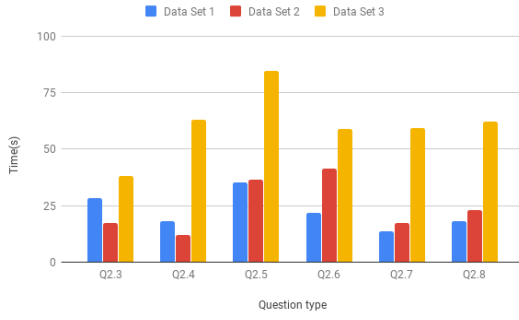
(a) Use Heatmap as visualization method
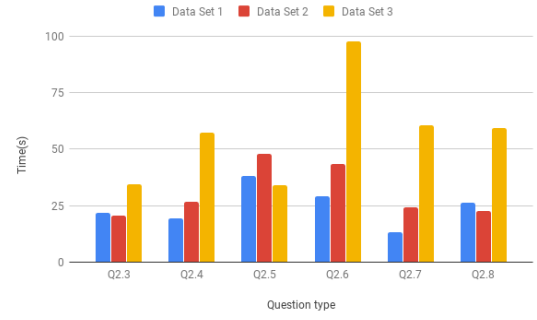
(b) Use Bar Graph as visualization method
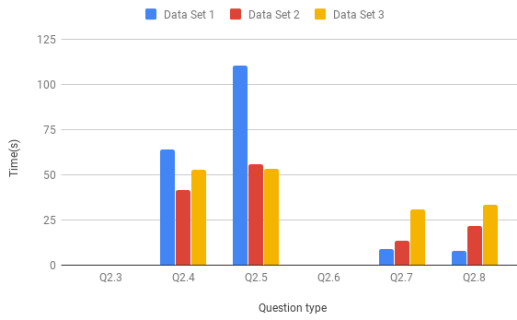
(c) Use Force-Directed-Graph as visualization method
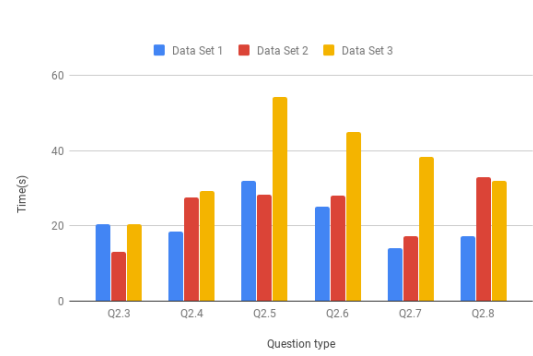
(d) Use random visualization method

Figure 5.8: Average time of participants finishing each question type using different visualization methods
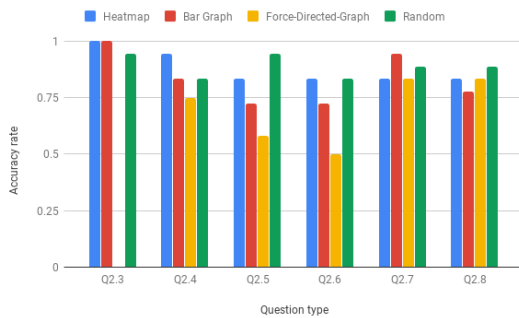
(a) Accuracy rate of each question type using different visualization methods

(b) Accuracy rate of each question type using different data sets

Figure 5.9: Accuracy rate of each question type

6 Conclusion

Correlation analysis is one of the fundamental task of Data Mining. It aims at discovering and summarizing the relationship between the attributes of a data set. Knowing the relationship between a set of variables, one can infer useful knowledge about external, a priori unknown outcomes. The evolving nature of streams and the high-dimensionality are two main challenges of analyzing high-dimensional data streams.

The goal of this thesis is the development of an interface for visualization of correlation in high-dimensional streams. We compare three different visualization methods through this interface: Heatmap, Bar Graph and Force-Directed-Graph. Our interactive interface reflects a tendency of the data correlation by visualization throughout the time. Users are able to choose a certain period of time to perform the correlation analysis and visualization. It is easy and quick for users to find pairs of attributes with strong correlations and also to see the evolution of data set during the time. This interface makes it possible to have a first glance at the data and provides some basic information before starting detailed analysis.

With the help of feedback from controlled user study, more useful functions can be added to our interface in order to ease the correlation analysis. Also, this interface can have more visualization methods than Heatmap, Bar Graph and Force-Directed-Graph and the current existing visualization methods can still be improved. In our thesis, we only focus on pairwise relationships and the correlations between more than two variables may remain to be discovered in the future work.

Bibliography

[1] ALBUQUERQUE, G., EISEMANN, M., LEHMANN, D. J., THEISEL, H., AND MAGNOR, M. Quality-Based Visualization Matrices. *In Proceedings of the Vision, Modeling and Visualization* (2009), 341–350.

[2] B, L. K., RIEKENBRAUCK, N., THEVESSEN, D., PAPPIK, M., STEBNER, A., KUNZE, J., MEISSNER, A., SHEKAR, A. K., AND EMMANUEL, M. Machine Learning and Knowledge Discovery in Databases. 404–408.

[3] FILIS, G., DEGIANNAKIS, S., AND FLOROS, C. Dynamic correlation between stock market and oil prices: The case of oil-importing and oil-exporting countries. *International Review of Financial Analysis 20*, 3 (2011), 152 – 164.

[4] KELLEY, W. M.; DONNELLY, R. A. The humongous book of statistics problems. *New York* (2009).

[5] KOBOUROV, S. G. Spring embedders and force-directed graph drawing algorithms. *eprint arXiv:1201.3011* (2012).

[6] MURRAY, S. Interactive data visualization for the web, an introduction to designing with d3. 272.

[7] RODGERS, J. L.; NICEWANDER, W. A. Thirteen ways to look at the correlation coefficient. *The American Statistician* (1988).

[8] SIMONS, K., ET AL. Should us investors invest overseas? *New England Economic Review* (1999), 29–40.

[9] WILKINSON, LELAND; FRIENDLY, M. The history of the cluster heat map. *The American Statistician* (2009).

7 Appendix

7.1 Consent Form

<div align="center">

Consent Form

</div>

STUDY: Interactive Visualization of Correlations in High-Dimensional Streams

INSTITUTION: Institute for Program Structures and Data Organization (IPD Böhm), Karlsruhe Institute of Technology, Am Fasanengarten 5,76133 Karlsruhe

PURPOSE OF STUDY: The evaluation of visualization methods for correlations in high-dimensional streams

PROCEDURE: Throughout the study, you will be asked to use a prototype interface for data visualization and to answer a set of questions about your experience with the prototype. Your answers will be collected anonymously and analyzed for the purpose of the evaluation of the prototype. The time you take to solve each task will also be recorded.

RISKS: To the best of our knowledge, the tasks you are going to perform do not yield to a higher risk of harm than you would experience in everyday life.

BENEFITS: You are not likely to experience any direct benefit from the results of this study. However, the results of the study may yield to new insights in the field of data visualization.

CONFIDENTIALITY

Please do not write any identifying information.

Every effort will be made by the institution to preserve your confidentiality including the following:

- Assigning code names/numbers to participants that will be used on all research notes and documents, to preserve their anonymity
- Keeping notes, interview transcriptions, and any other identifying participant information in a locked file.

Participant data will be kept confidential except in cases where the researcher is legally obligated to report specific incidents. These incidents include, but may not be limited to, incidents of abuse and suicide risk.

CONTACT INFORMATION

If you have questions at any time about this study, or you experience adverse effects as the result of participating in this study, you are free to ask them now. If you have questions regarding your rights as a research participant, or you experience adverse effects as the result of participating in this study, you can contact with the researcher directly by telephone at 017683581082 or via the following email address uhelt@student.kit.edu.

VOLUNTATION PARTICIPATION

Your participation in this study is voluntary. It is up to you to decide whether or not to take part in this study. If you decide to take part in this study, you will be asked to sign this consent form. After you sign the consent form, you are still free to withdraw at any time and without giving a reason. Withdrawing from this study will not affect the relationship you have, if any, with the researcher and the institution. If you withdraw from the study, your data will be returned to you or destroyed.

CONSENT
I have read and I understand the provided information and have had the opportunity to ask questions. I understand that my participation is voluntary and that I am free to withdraw at any time, without giving a reason and without cost. I understand that I will be given a copy of this consent form. I voluntarily agree to take part in this study.

Participant's Name (printed) _____

Participant's Signature _____

Date _____

7.2 Questionnaire

7.2.1 Questionnaire For Participants under Condition A/B/C

7.2.1.1 Participants using Data Set 1

Questionnaire:

1. Basic Information
   1.1. Field of study:
   1.2. Number of semester:
   1.3. Age:

   | *18-24* | *25-30* | *More than 30* |

   1.4. Gender:

   | *M* | *F* | *Do not wish to answer* |

   1.5. How familiar are you with the following concepts?

	1	2	3	4	5	6	7
Unfamiliar							*Very Familiar*
Correlation analysis							
Data analysis							
Data visualization							

2. Visualization
   2.1. How many attributes are available in this data set?

   2.2. How many pairs of attributes are available in this data set?

   2.3. a) What is the correlation value between **Attribute 1** and **Attribute 3** at **Timestamp T1**, or the probable range?

   b) What is the correlation value between **Attribute 1** and **Attribute 5** at **Timestamp T401**, or the probable range?

   c) What is the correlation value between **Attribute 6** and **Attribute 8** at **Timestamp T1501**, or the probable range?

   2.4. a) Which pair of attributes has the biggest correlation at **Timestamp T451**?

   b) Which pair of attributes has the biggest correlation at **Timestamp T1351**?

   c) Which pair of attributes has the biggest correlation at **Timestamp T1151**?

2.5. a) Which pair of attributes has the smallest correlation at **Timestamp T1051**?

b) Which pair of attributes has the smallest correlation at **Timestamp T1151**?

c) Which pair of attributes has the smallest correlation at **Timestamp T1701**?

2.6. a) The following statement is true or false: "The correlation value between **Attribute 1** and **Attribute 5** remains the same at **Timestamp T1751** and at **Timestamp T1801**"?

 True   *False*

b) The following statement is true or false: "The difference of correlation value between **Attribute 1** and **Attribute 3** at **Timestamp T51** and at **Timestamp T101** is smaller than 0.1"?

 True   *False*

c) The following statement is true or false: "The difference of correlation value between **Attribute 3** and **Attribute 9** at **Timestamp T1351** and at **Timestamp T1401** is bigger than 0.2"?

 True   *False*

2.7. a) Which pair(s) of attributes has/have a correlation value that is not smaller than **0.1** at **Timestamp T1251**?

b) Which pair(s) of attributes has/have a correlation value that is not smaller than **0.1** at **Timestamp T651**?

c) Which pair(s) of attributes has/have a correlation value that is not smaller than **0.1** at **Timestamp T501**?

2.8. a) Which pair(s) of attributes has/have a correlation value that is not bigger than **0.001** at **Timestamp T51**?

b) Which pair(s) of attributes has/have a correlation value that is not bigger than **0.001** at **Timestamp T951**?

c) Which pair(s) of attributes has/have a correlation value that is not bigger than **0.001** at **Timestamp T1301**?

3. Feedback
    3.1. Please write down the visualization method you have used and rate for it.

	1 Strongly Disagree	2	3	4	5	6	7 Strongly Agree
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

    3.2. In your opinion, what are the strengths of this visualization system?

    3.3. In your opinion, what are the weaknesses of this visualization system?

    3.4. Do you have any suggestions for improving this visualization system?

7.2.1.2 Participants using Data Set 2

Questionnaire:

1. Basic Information
    1.1. Field of study:
    1.2. Number of semester:
    1.3. Age:

    | *18-24* | *25-30* | *More than 30* |

    1.4. Gender:

    | *M* | *F* | *Do not wish to answer* |

    1.5. How familiar are you with the following concepts?

    | | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
    | | *Unfamiliar* | | | | | | *Very Familiar* |

Correlation analysis							
Data analysis							
Data visualization							

2. Visualization
    2.1. How many attributes are available in this data set?

    2.2. How many pairs of attributes are available in this data set?

    2.3. a) What is the correlation value between **Attribute 1** and **Attribute 13** at **Timestamp T801**, or the probable range?

    b) What is the correlation value between **Attribute 10** and **Attribute 18** at **Timestamp T151**, or the probable range?

    c) What is the correlation value between **Attribute 1** and **Attribute 3** at **Timestamp T901**, or the probable range?

    2.4. a) Which pair of attributes has the biggest correlation at **Timestamp T701**?

    b) Which pair of attributes has the biggest correlation at **Timestamp T1701**?

    c) Which pair of attributes has the biggest correlation at **Timestamp T1601**?

2.5. a) Which pair of attributes has the smallest correlation at **Timestamp T601**?

b) Which pair of attributes has the smallest correlation at **Timestamp T551**?

c) Which pair of attributes has the smallest correlation at **Timestamp T1451**?

2.6. a) The following statement is true or false: "The correlation value between **Attribute 5** and **Attribute 7** remains the same at **Timestamp T801** and at **Timestamp T851**"?
            True   *False*

b) The following statement is true or false: "The difference of correlation value between **Attribute 1** and **Attribute 15** at **Timestamp T1001** and at **Timestamp T1501** is smaller than 0.1"?
            True   *False*

c) The following statement is true or false: "The difference of correlation value between **Attribute 5** and **Attribute 6** at **Timestamp T1** and at **Timestamp T51** is bigger than 0.2"?
            True   *False*

2.7. a) Which pair(s) of attributes has/have a correlation value that is not smaller than **0.3** at **Timestamp T851**?

b) Which pair(s) of attributes has/have a correlation value that is not smaller than **0.3** at **Timestamp T1951**?

c) Which pair(s) of attributes has/have a correlation value that is not smaller than **0.2** at **Timestamp T601**?

2.8. a) Which pair(s) of attributes has/have a correlation value that is not bigger than **0.001** at **Timestamp T1051**?

b) Which pair(s) of attributes has/have a correlation value that is not bigger than **0.001** at **Timestamp T1901**?

c) Which pair(s) of attributes has/have a correlation value that is not bigger than **0.001** at **Timestamp T801**?

3. Feedback
   3.1. Please write down the visualization method you have used and rate for it.

	1 *Strongly* *Disagree*	2	3	4	5	6	7 *Strongly* *Agree*
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

   3.2. In your opinion, what are the strengths of this visualization system?

   3.3. In your opinion, what are the weaknesses of this visualization system?

   3.4. Do you have any suggestions for improving this visualization system?

7.2.1.3 Participants using Data Set 3

Questionnaire:

1. Basic Information
   1.1. Field of study:
   1.2. Number of semester:
   1.3. Age:

   | | *18-24* | *25-30* | *More than 30* |

   1.4. Gender:

   | | *M* | *F* | *Do not wish to answer* |

   1.5. How familiar are you with the following concepts?

	1	2	3	4	5	6	7
	Unfamiliar						*Very Familiar*

Correlation analysis							
Data analysis							
Data visualization							

2. Visualization
   2.1. How many attributes are available in this data set?

   2.2. How many pairs of attributes are available in this data set?

   2.3. a) What is the correlation value between **Attribute 10** and **Attribute 20** at **Timestamp T401**, or the probable range?

   b) What is the correlation value between **Attribute 8** and **Attribute 35** at **Timestamp T751**, or the probable range?

   c) What is the correlation value between **Attribute 7** and **Attribute 25** at **Timestamp T951**, or the probable range?

   2.4. a) Which pair of attributes has the biggest correlation at **Timestamp T301**?

   b) Which pair of attributes has the biggest correlation at **Timestamp T51**?

   c) Which pair of attributes has the biggest correlation at **Timestamp T451**?

2.5. a) Which pair of attributes has the smallest correlation at **Timestamp T351**?

b) Which pair of attributes has the smallest correlation at **Timestamp T551**?

c) Which pair of attributes has the smallest correlation at **Timestamp T751**?

2.6. a) The following statement is true or false: "The correlation value between **Attribute 1** and **Attribute 37** remains the same at **Timestamp T701** and at **Timestamp T751**"?

     True   *False*

b) The following statement is true or false: "The difference of correlation value between **Attribute 10** and **Attribute 20** at **Timestamp T551** and at **Timestamp T601** is smaller than 0.1"?

     True   *False*

c) The following statement is true or false: "The difference of correlation value between **Attribute 27** and **Attribute 30** at **Timestamp T801** and at **Timestamp T851** is bigger than 0.2"?

     True   *False*

2.7. a) Which pair(s) of attributes has/have a correlation value that is not smaller than **0.95** at **Timestamp T101**?

b) Which pair(s) of attributes has/have a correlation value that is not smaller than **0.995** at **Timestamp T801**?

c) Which pair(s) of attributes has/have a correlation value that is not smaller than **0.95** at **Timestamp T1**?

2.8. a) Which pair(s) of attributes has/have a correlation value that is not bigger than **0.001** at **Timestamp T151**?

b) Which pair(s) of attributes has/have a correlation value that is not bigger than **0** at **Timestamp T951**?

c) Which pair(s) of attributes has/have a correlation value that is not bigger than **0** at **Timestamp T551**?

3. Feedback

    3.1. Please write down the visualization method you have used and rate for it.

	1 *Strongly* *Disagree*	2	3	4	5	6	7 *Strongly* *Agree*
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

    3.2. In your opinion, what are the strengths of this visualization system?

    3.3. In your opinion, what are the weaknesses of this visualization system?

    3.4. Do you have any suggestions for improving this visualization system?

7.2.2 Questionnaire For Participants under Condition D

7.2.2.1 Participants using Data Set 1

Questionnaire:

1. Basic Information
   1.1. Field of study:_____
   1.2. Number of semester: _____
   1.3. Age:

   | *18-24* | *25-30* | *More than 30* |

   1.4. Gender:

   | *M* | *F* | *Do not wish to answer* |

   1.5. How familiar are you with the following concepts?

	1	2	3	4	5	6	7
	Unfamiliar						*Very Familiar*
Correlation analysis							
Data analysis							
Data visualization							

2. Visualization
   2.1. How many attributes are available in this data set?

   _____

   2.2. How many pairs of attributes are available in this data set?

   _____

   2.3. a) What is the correlation value between **Attribute 1** and **Attribute 3** at **Timestamp T1**, or the probable range?

   _____

   b) What is the correlation value between **Attribute 1** and **Attribute 5** at **Timestamp T401**, or the probable range?

   _____

   c) What is the correlation value between **Attribute 6** and **Attribute 8** at **Timestamp T1501**, or the probable range?

   _____

   2.4. a) Which pair of attributes has the biggest correlation at **Timestamp T451**?

   _____

   b) Which pair of attributes has the biggest correlation at **Timestamp T1351**?

   _____

   c) Which pair of attributes has the biggest correlation at **Timestamp T1151**?

   _____

51

2.5. a) Which pair of attributes has the smallest correlation at **Timestamp T1051**?

_____

b) Which pair of attributes has the smallest correlation at **Timestamp T1151**?

_____

c) Which pair of attributes has the smallest correlation at **Timestamp T1701**?

_____

2.6. a) The following statement is true or false: "The correlation value between **Attribute 1** and **Attribute 5** remains the same at **Timestamp T1751** and at **Timestamp T1801**"?

     True   *False*

b) The following statement is true or false: "The difference of correlation value between **Attribute 1** and **Attribute 3** at **Timestamp T51** and at **Timestamp T101** is smaller than 0.1"?

     True   *False*

c) The following statement is true or false: "The difference of correlation value between **Attribute 3** and **Attribute 9** at **Timestamp T1351** and at **Timestamp T1401** is bigger than 0.2"?

     True   *False*

2.7. a) Which pair(s) of attributes has/have a correlation value that is not smaller than **0.1** at **Timestamp T1251**?

_____

b) Which pair(s) of attributes has/have a correlation value that is not smaller than **0.1** at **Timestamp T651**?

_____

c) Which pair(s) of attributes has/have a correlation value that is not smaller than **0.1** at **Timestamp T501**?

_____

2.8. a) Which pair(s) of attributes has/have a correlation value that is not bigger than **0.001** at **Timestamp T51**?

_____

b) Which pair(s) of attributes has/have a correlation value that is not bigger than **0.001** at **Timestamp T951**?

_____

c) Which pair(s) of attributes has/have a correlation value that is not bigger than **0.001** at **Timestamp T1301**?

_____

3. Feedback

3.1. As you are able to use all visualization methods during the research, which one of the method, in your opinion, is the most helpful method to fulfill the task?

And which of the method did you use the most?

Please rate for *Heat map:*

	1 Strongly Disagree	2	3	4	5	6	7 Strongly Agree
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

Please rate for *Bar Graph:*

	1 Strongly Disagree	2	3	4	5	6	7 Strongly Agree
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

54

Please rate for **Force-Directed Graph:**

	1 **Strongly Disagree**	2	3	4	5	6	7 **Strongly Agree**
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

Please rate for **whole system:**

	1 **Strongly Disagree**	2	3	4	5	6	7 **Strongly Agree**
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

3.2. In your opinion, what are the strengths of this visualization system?

3.3. In your opinion, what are the weaknesses of this visualization system?

_____

_____

_____

3.4. Do you have any suggestions for improving this visualization system?

_____

_____

_____

7.2.2.2 Participants using Data Set 2

Questionnaire:

1. Basic Information
   1.1. Field of study:_____
   1.2. Number of semester: _____
   1.3. Age:

   | *18-24* | *25-30* | *More than 30* |

   1.4. Gender:

   | *M* | *F* | *Do not wish to answer* |

   1.5. How familiar are you with the following concepts?

	1	2	3	4	5	6	7
	Unfamiliar						*Very Familiar*
Correlation analysis							
Data analysis							
Data visualization							

2. Visualization
   2.1. How many attributes are available in this data set?

   _____

   2.2. How many pairs of attributes are available in this data set?

   _____

   2.3. a) What is the correlation value between **Attribute 1** and **Attribute 13** at **Timestamp T801**, or the probable range?

   _____

   b) What is the correlation value between **Attribute 10** and **Attribute 18** at **Timestamp T151**, or the probable range?

   _____

   c) What is the correlation value between **Attribute 1** and **Attribute 3** at **Timestamp T901**, or the probable range?

   _____

   2.4. a) Which pair of attributes has the biggest correlation at **Timestamp T701**?

   _____

   b) Which pair of attributes has the biggest correlation at **Timestamp T1701**?

   _____

   c) Which pair of attributes has the biggest correlation at **Timestamp T1601**?

   _____

2.5. a) Which pair of attributes has the smallest correlation at **Timestamp T601**?

b) Which pair of attributes has the smallest correlation at **Timestamp T551**?

c) Which pair of attributes has the smallest correlation at **Timestamp T1451**?

2.6. a) The following statement is true or false: "The correlation value between **Attribute 5** and **Attribute 7** remains the same at **Timestamp T801** and at **Timestamp T851**"?

    True   ***False***

b) The following statement is true or false: "The difference of correlation value between **Attribute 1** and **Attribute 15** at **Timestamp T1001** and at **Timestamp T1501** is smaller than 0.1"?

    True   ***False***

c) The following statement is true or false: "The difference of correlation value between **Attribute 5** and **Attribute 6** at **Timestamp T1** and at **Timestamp T51** is bigger than 0.2"?

    True   ***False***

2.7. a) Which pair(s) of attributes has/have a correlation value that is not smaller than **0.3** at **Timestamp T851**?

b) Which pair(s) of attributes has/have a correlation value that is not smaller than **0.3** at **Timestamp T1951**?

c) Which pair(s) of attributes has/have a correlation value that is not smaller than **0.2** at **Timestamp T601**?

2.8. a) Which pair(s) of attributes has/have a correlation value that is not bigger than **0.001** at **Timestamp T1051**?

b) Which pair(s) of attributes has/have a correlation value that is not bigger than **0.001** at **Timestamp T1901**?

c) Which pair(s) of attributes has/have a correlation value that is not bigger than **0.001** at **Timestamp T801**?

3. Feedback

   3.1. As you are able to use all visualization methods during the research, which one of the method, in your opinion, is the most helpful method to fulfill the task?

   And which of the method did you use the most?

   Please rate for *Heat map:*

	1 *Strongly* *Disagree*	2	3	4	5	6	7 *Strongly* *Agree*
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

   Please rate for *Bar Graph:*

	1 *Strongly* *Disagree*	2	3	4	5	6	7 *Strongly* *Agree*
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

Please rate for **Force-Directed Graph:**

	1 **Strongly** **Disagree**	2	3	4	5	6	7 **Strongly** **Agree**
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

Please rate for **whole system:**

	1 **Strongly** **Disagree**	2	3	4	5	6	7 **Strongly** **Agree**
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

3.2. In your opinion, what are the strengths of this visualization system?

3.3. In your opinion, what are the weaknesses of this visualization system?

3.4. Do you have any suggestions for improving this visualization system?

7.2.2.3 Participants using Data Set 3

Questionnaire:

1. Basic Information
   1.1. Field of study: _____
   1.2. Number of semester: _____
   1.3. Age:

18-24	***25-30***	***More than 30***

   1.4. Gender:

M	***F***	***Do not wish to answer***

   1.5. How familiar are you with the following concepts?

	1	2	3	4	5	6	7
	Unfamiliar						***Very Familiar***
Correlation analysis							
Data analysis							
Data visualization							

2. Visualization
   2.1. How many attributes are available in this data set?

   _____

   2.2. How many pairs of attributes are available in this data set?

   _____

   2.3. a) What is the correlation value between **Attribute 10** and **Attribute 20** at **Timestamp T401**, or the probable range?

   _____

   b) What is the correlation value between **Attribute 8** and **Attribute 35** at **Timestamp T751**, or the probable range?

   _____

   c) What is the correlation value between **Attribute 7** and **Attribute 25** at **Timestamp T951**, or the probable range?

   _____

   2.4. a) Which pair of attributes has the biggest correlation at **Timestamp T301**?

   _____

   b) Which pair of attributes has the biggest correlation at **Timestamp T51**?

   _____

   c) Which pair of attributes has the biggest correlation at **Timestamp T451**?

   _____

2.5. a) Which pair of attributes has the smallest correlation at **Timestamp T351**?

b) Which pair of attributes has the smallest correlation at **Timestamp T551**?

c) Which pair of attributes has the smallest correlation at **Timestamp T751**?

2.6. a) The following statement is true or false: "The correlation value between **Attribute 1** and **Attribute 37** remains the same at **Timestamp T701** and at **Timestamp T751**"?

True   *False*

b) The following statement is true or false: "The difference of correlation value between **Attribute 10** and **Attribute 20** at **Timestamp T551** and at **Timestamp T601** is smaller than 0.1"?

True   *False*

c) The following statement is true or false: "The difference of correlation value between **Attribute 27** and **Attribute 30** at **Timestamp T801** and at **Timestamp T851** is bigger than 0.2"?

True   *False*

2.7. a) Which pair(s) of attributes has/have a correlation value that is not smaller than **0.95** at **Timestamp T101**?

b) Which pair(s) of attributes has/have a correlation value that is not smaller than **0.995** at **Timestamp T801**?

c) Which pair(s) of attributes has/have a correlation value that is not smaller than **0.95** at **Timestamp T1**?

2.8. a) Which pair(s) of attributes has/have a correlation value that is not bigger than **0.001** at **Timestamp T151**?

b) Which pair(s) of attributes has/have a correlation value that is not bigger than **0** at **Timestamp T951**?

c) Which pair(s) of attributes has/have a correlation value that is not bigger than **0** at **Timestamp T551**?

3. Feedback

    3.1. As you are able to use all visualization methods during the research, which one of the method, in your opinion, is the most helpful method to fulfill the task?

    And which of the method did you use the most?

    Please rate for *Heat map:*

	1 Strongly Disagree	2	3	4	5	6	7 Strongly Agree
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

    Please rate for *Bar Graph:*

	1 Strongly Disagree	2	3	4	5	6	7 Strongly Agree
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

Please rate for **Force-Directed Graph:**

	1 **Strongly Disagree**	2	3	4	5	6	7 **Strongly Agree**
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

Please rate for **whole system:**

	1 **Strongly Disagree**	2	3	4	5	6	7 **Strongly Agree**
Intuitive							
Convenient							
Interactive							
Useful							
Complicated							
Efficient							
Effective							

3.2. In your opinion, what are the strengths of this visualization system?

3.3. In your opinion, what are the weaknesses of this visualization system?

3.4. Do you have any suggestions for improving this visualization system?