

# Using Visual Analytics Tool for Improving Data Comprehension

Jan Géryk  
KD Lab Faculty of Informatics  
Masaryk University  
Brno, Czech Republic  
geryk@fi.muni.cz

## ABSTRACT

The efficacy of animated data visualizations in comparison with static data visualizations is still inconclusive. Some researches resulted that the failure to find out the benefits of animations may relate to the way how they are constructed and perceived. In this paper, we present visual analytics (VA) tool which makes use of enhanced animated data visualization methods. The time is an important variable that needs to be modeled in VA. VA methods like Motion Charts show changes over time by presenting animations in two-dimensional space and by changing element appearances. The tool is primarily designed for exploratory analysis of academic analytics and supports various interactive visualization methods which enhance the Motion Charts concept. We evaluate the usefulness and the general applicability of the designed tool with a controlled experiment to assess the efficacy of the described methods. To interpret the experiment results, we utilized one-way repeated measures ANOVA.

## Keywords

Animation; motion charts; visual analytics; academic analytics; experiment.

## 1. INTRODUCTION

Higher education institutions have a strong interest in improving the quality and the efficacy of the education. In [1], hundreds of higher education executives were surveyed on their analytic needs. Authors resulted that the advanced analytics should support better decision-making, studying enrollment trends, and measuring student retention. They also pointed out that management commitment and staff skills are more important in deploying academic analytics (AA) than the technology. In [2], authors concluded that the increasing accountability requirements of educational institutions represent a key for unlocking the potentials of AA in order to effectively enhance student retention and increase graduation levels. The authors also resulted that AA facilitate creation of actionable intelligence to enhance learning and student success, however, it is highly dependent on the quality of the accountability. The authors utilized AA for developing several predictive models of student enrollment and retention, and for identifying students being at the risk. They also highlighted three critical success factors—executives committed to decision-making based on the evidence, staff members with adequate data analysis skills and the flexible and effective technology platform. However, the authors also warned that more elaborated accountability can raise several privacy issues, faculty executive's involvement, and data administration.

The principal goals can be achieved by using educational data mining methods, as emphasized in [3]. The application of data

mining (DM) techniques in higher education systems have some specific requirements not present in other areas, as pointed out in [4]. Common DM methods were developed independently of visualization techniques. However, some key ideas influenced the research in the DM field. It resulted into the recent research topic called visual analytics (VA). Google Analytics, released in 2005, made a real progress in web-based interactive analytics. In 2007, Hans Rosling presented a TED talk demonstrating the power of animations to show the story in data. In 2009, Tim O'Reilly emphasized that data analysis, visualizations, and other techniques for searching patterns in data are going to be an increasingly valuable skill set [5]. While some researches resulted that animations appeared better than static visualizations in enhancing learning, an elaborate examination of the studies revealed a lack of equivalence between animated and static visualizations in content [6]. Also, the failure to ascertain the benefits of animations in learning may also relate to the way how they are constructed, perceived, and conceptualized [7].

Visualizations are common methods used to gain a qualitative understanding of data prior to any computational analysis. By displaying animated presentations of the data and providing analysts with interactive tools for manipulating the data, visualizations allow human pattern recognition skills to contribute to the analytic process. The most commonly used statistical visualization methods (e.g. line plots, or scatter plots) generally focus on univariate or bivariate data. The methods are usually used for tasks ranging from the exploration to the confirmation of models, including the presentation of the results. However, fewer methods are available for visualizing data with more than two dimensions (e.g. motion charts or parallel coordinates), as the logical mapping of the data dimension to the screen dimension cannot be directly applied. Data exploration and interactive visualizations of multivariate data without significant dimensionality reduction remains a challenge. Animations represent a promising approach to facilitate better perception of changing values. In [6], authors pointed out that animations help to keep the viewer's attention. Visualizations and animations can also facilitate the learning process [8].

We develop visualization methods for multivariate data analyses that are adapted for academic settings. In this paper, we show the importance of data visualizations for successful understanding of complex and large data. In the next section, we examine characteristics of changes using Motion Charts (MC). Subsequently, we present several papers successfully utilizing MC for data visualization and analysis. This is followed by the elaborate description of our VA tool. Further, we conducted an empirical study with 22 participants on their data comprehension to compare the efficacy of static and animated data visualizations. We then

discuss the implications of our experiment results. Finally, we draw the conclusion from the experiment and outline future work.

## 2. EXAMINE CHARACTERISTICS OF CHANGE

Although a snapshot of the data can be beneficial, presenting changes over time provides a more sophisticated perspective. The efficacy of animated transitions for common statistic data visualizations such as bar charts and scatter plots was examined in [9]. The authors extended the theoretical model of data visualizations and introduced the taxonomy of transition types. Subsequently, they proposed design principles for creating effective transitions and illustrated the application of these principles in a dynamic visual system. Finally, they conducted two controlled experiments to assess the efficacy of various transition types, finding that animated transitions can significantly improve the visual perception. The visualization challenge posed by each of these experiments was to keep the viewer's attention during transitions. The survey resulted that viewers found animations more helpful and engaging. Unlike transition animations, which primarily help users to stay in the context, trend animations convey the meaning. While a transition animation moves from a still view to a new still view, a trend animation moves continuously between states. One early use of animations in visualization was for an algorithm animation. Kehoe et al. [10] describe a study that demonstrated that animations could help and noted that it improved the motivation of making a difficult topic more approachable. The study suggested that using animations for trend understanding could be valuable.

Animations allow knowledge discovery in complex data and make it easier to see meaningful characteristics of changes over time. To reduce the cognitive load and improve tracking accuracy, the target states of all transitioning elements should be predictable after viewing a fraction of the animation. The proper use of the acceleration should also improve the spatial and temporal predictability. A perceptual study in [11] provides evidence that animations and divergence motions are easier to understand than rotations. Animations with unpredictable motion paths or multiple simultaneously changing elements result in the increased cognitive load. Contrarily, simple transitions reduce confusion and improve clarity. In [12], authors concluded that animation stages should be long enough for accurate change tracking as well as to decrease the number of errors. However, too slow animations can disproportionately prolong the analytic phase and subsequently reduce the engagement.

Generally, effective analyses depend on the consistent and high-quality data. In [9], authors concluded that the correctly designed animations significantly improve the visual perception at both the syntactic and the semantic level. Visualizations are often engaging and attractive, but a naive approach can confuse analysts. Visualizations are just representations of the data which may or may not represent the reality. As Few pointed out in [13], computers cannot make sense of the data, only people can. The perception of animations can also be problematic because of severe issues with timing and the overall complexity that can occur during transitions as pointed out in [14]. Misleading results can be obtained if animations violate the underlying data semantics.

MC is a dynamic and interactive visualization method that enables analysts to display complex and quantitative data in an intelligible way. The dynamic refers to the animation of rich multidimensional

data changing over time. The interactive refers to dynamic interactive features which allow analysts to explore, interpret, and analyze information concealed in complex data, as presented in [15]. MC displays changes of element appearances over time by showing animations in a two-dimensional space. An element is basically a two-dimensional shape representing one object from the dataset. The variable mapping is one of the most important parts of the exploratory data analysis and no optimal method for mapping the data to variables is available. Naturally, the data mapping have a significant impact on the data comprehension and analysts should be free to choose variable mapping according to their intentions. Both the data characteristics and the investigative hypothesis influence the variable mapping.

## 3. APPLICATIONS OF MOTION CHARTS

Visualization tools represent an effective way how to make statistical data understandable to analysts, as showed in [16]. MC methods proved to be useful for data presentation and the approach was verified that can be successfully employed to show a story in data [17] or support decision making [18]. In [19], authors utilized MC for both the interpretation of results for better comprehension and the analysis when detecting topics of tweets. Several web-based data analysis tools allowing analysts to interactively explore associations, patterns, and trends in data with temporal characteristics are available. In [20], authors presented a visualization of energy statistics using an existing web-based data analysis tools, including IBM's Many Eyes, and Google Motion Charts. In [15], authors presented a Java-based infrastructure, named SOCR Motion Charts, designed for exploratory analysis of multivariate data. SOCR is developed as a Java applet using object-oriented programming language. The authors successfully validated this visualization paradigm using several publicly available datasets containing housing prices or consumer price index.

A pair of online assessments designed to measure students' computational thinking skills were presented in [21]. The assessments represent a part of a larger project that brings computational thinking into high school STEM classrooms. Each assessment included interactive tools that highlight the power of computation in the practice of the scientific and mathematical inquiry. The computational tools including Google Motion Charts used in the assessments enabled students to analyze data with dynamic visualizations and explore concepts with computational models.

Successful visualizations of language changes using the diachronic corpus data were presented in [22]. In two case studies, authors illustrated recent changes in American English. In the first study, they visualized changes in a diachronic analysis of nouns and verbs. In the second study, they showed structural changes in the behavior of complement-taking predicates. They emphasized that MC are useful for the analysis of multivariate data over time and concluded that viewing the resulting data points in separate time slices offers a proper representation of the complex linguistic changes.

In [23], authors incorporated examples using recent business and economic data series and illustrated how MC can tell dynamic stories. They utilized a database of Bureau of Labor Statistics which publishes data on inflation, prices, employment, and many other labor related subjects. For the first analysis, they utilized the data about Current Employment Statistics and presented differences between the perception of common static tables and graphs, and the

dynamic nature of MC. They concluded that the static presentation style serves well the purpose of relaying accurate and non-biased quantitative data to analysts. Subsequently, they utilized the same data, but imported them to Google Docs. By loading the Motion Charts Gadget within the spreadsheet, they generated MC and visualized several areas of Labor Statistics. They emphasized that the benefit of MC lays in displaying complex multidimensional data changing over time on a single plane with the dynamic and interactive features. Users are then allowed to easily explore, interpret, and analyze the information in the data. They concluded that MC is an excellent and interesting way how to present valuable information that may be otherwise lost in the data.

The report on the implementation of AA in a new medical school can be found in [24]. Authors pointed out that analytics address two challenges in the curriculum: providing the evidence of the appropriate curriculum coverage and assessing the student engagement during the clinical placement. The paper describes tools and approaches applied on the data gained from their web-based clinical log system. The authors utilized common data visualization methods and examined their potentials to generate important questions. They also examined the value of a flexible approach to select the tools, the need for relevant skills, and the importance of keeping the viewer's attention. Subsequently, they utilized more sophisticated visualization methods, namely MC and Tree map. Using MC, they mapped several important variables including entry date, frequency of entries, clinical problems, the level of involvement, and the level of confidence. The authors appreciated the benefits of comparison of the variation of the frequency of entries, the confidence, and the level of involvement between students. The authors concluded that AA analysis using visualizations have already been a critical enabler of educational excellence, but there is undoubtedly further potential.

A beneficial feature for better visual perception of changes in time-series analysis is presented in [25]. Initially, the author highlighted the need for effective ways to examine quantitative data that changes over time and also noted that according to several studies, more than 70 percent of all business charts display time-series information. Then, the author emphasized both the benefits and the drawbacks of common data visualization methods, namely line plots and bar charts. Subsequently, the author described issues with the time-series analysis and presented capabilities of MC. The author pointed out that patterns of changes over time can take many meaningful forms and introduced a new feature, called visual trails, specially designed for MC. The feature allows seeing the full path for each variable from one point in time to another. It can be used for overcoming visual perception limitations of MC and allows analysts to examine degree of change, shape, velocity, and direction of change. Finally, the author conducted the experiment as an evaluation of the proposed improvement.

#### 4. THE EDAIME TOOL

The preliminary version of the EDAIME tool was presented in [26]. We also described the results originated from the analysis of AA data. We utilized the data stored in the Information System of Masaryk University. The motivation to develop an enhanced version of MC was to improve its expression capabilities, as well as to facilitate analysts to depict each student or study as a central object of their interest. Moreover, the implementation enhances the number of animations that express the students' behavior during their studies more precisely. We validated usefulness of the

developed methods with a case study where we successfully utilized the capabilities of the tool for the purpose of confirming our hypothesis concerning student retention. Although, we concluded that the methods proved to be useful for analytic purposes, more adjustments are needed.

Two main challenges are addressed by the presented VA tool. EDAIME enables visualization of multivariate data and the qualitative exploration of data with temporal characteristics. The technical advantages over other implementations of MC are its flexibility and the ability to manage many animations simultaneously. The Force Layout component of D3<sup>1</sup> provides the most of the functionality behind the animations and collision detection utilized in the interactive visualization methods. Technical aspects of enhanced MC methods are elaborately described in [27]. Investigated data can be imported directly using the tool. In cases where datasets have missing values at the beginning or the end, the missing values are extrapolated from nearby data. In other cases, gaps are filled with interpolated values. For the purposes of the MC analysis, it is not important that the data are not entirely accurate.

In two figures below, two examples of our enhanced MC methods can be seen. We already utilized the methods to verify a hypothesis concerning student retention. Figure 1 depicts a snapshot of the method captured in the second semester. Each element represents a field of study and consists of a pie chart. It allows analysts to investigate another data dimension easily. Each pie chart animates a relationship between finished and unfinished studies where the green sector quantifies the complete ones, and the red sector quantifies the others. Figure 2 represents a snapshot of the second method utilized for the same dataset also captured in the second semester. The large clusters of elements represent the particular field of study consisting of small elements that represent individual students. Therefore, the size of the cluster of elements corresponds to the number of students enrolled in the particular field of study. The size of the small elements determines the number of credits gained by students in the particular semester of the study. Besides the study progress, the animations are also utilized to express the study termination, the change of the mode of study and the change of the field of study. During the animation process, dropout students turn red and fall down the chart in the semester when they left the study. The stroke-width of the elements represents states of the study and the element color represents attributes of the study.

When animations are used for exploratory analysis of unfamiliar data, analysts do not know what elements are important and play the animation hoping that something emerges. Analysts may determine areas that look promising and replay the animation several times focusing on each of the potentially interesting areas in depth. This can become an issue, perhaps making trend animations slower and more error prone for analyses. If there is a lot of variability in the data, there will be a lot of random motions, making hard to perceive trends. If there are too many elements, a clutter and counter-trends can easily intricate an observation. In the next section, we describe several user interface features that may solve some of these issues. Naturally, all methods using animations have several limitations, but appropriately designed user interface features can considerably aid visual inspection of data.

---

<sup>1</sup> <http://d3js.org/>

## 4.1 User Interface Features

The EDAIME tool offers several beneficial configurable interactive features for a more convenient analytic process. User interface features are highly customizable and allow analysts to arrange the display and variable mapping according to his or her needs. Available features include a mouse-over data display, color and plot size representation, traces, animated time plot, variable animation speed, changing of axis series, changing of axis scaling, distortion, and the support of statistical methods.

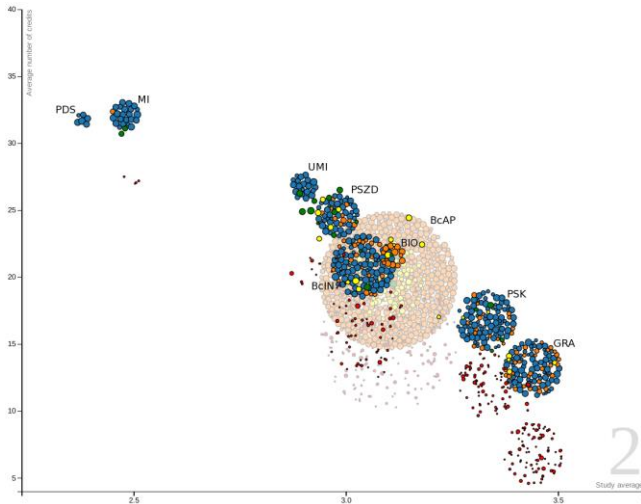


Figure 1. EDAIME snapshot: clusters of students.

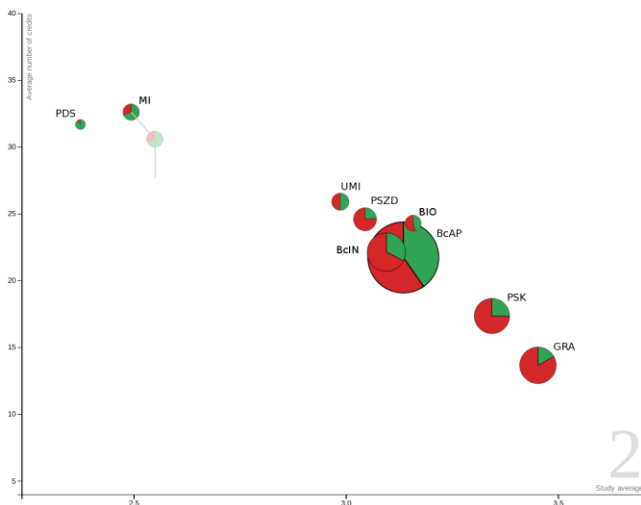


Figure 2. EDAIME snapshot: additional dimension using pie charts.

The focus-plus-context technique allows to interactively exploring objects of interest in detail while preserving the surrounding context. More precisely, if an analyst zooms in for detail, the chart area is too big to full overview. Contrarily, if an analyst zooms out the screen to see the overall chart area, the tiny but potentially important characteristics can disappear. Generally, distortions are particularly beneficial to overcome the aforementioned issues. The circular distortion magnifies the area around the mouse pointer, while leaving the chart area unaffected for the context. This

distortion is useful especially to distinguish individual elements in a cluster. However, the area near the circumference of the elements is then compressed. Therefore, it is not suitable for representing quantitative values. However, a function which magnifies the details continuously in order to avoid such local errors exists. It applies the distortion to each dimension separately which results in Cartesian distortion. If elements overlap each other during the animation, it will be more difficult to track their paths. Using the jitter feature, a better visual perception of data can be obtained by adding small random quantities to all elements' values before displaying them. As mentioned earlier, it is not important that the data are not entirely accurate for the purposes of a trend analysis.

Regardless of the power of a human brain, a memory is limited. It is difficult to reconstruct the past events from a memory, to recapture the sequence of events and details of each moment. The tool provides analysts with the ability to select particular elements and show a trace for each of the selected elements as it progresses. This is particularly useful in verifying apparent anomalies noticed during an animation. The traces show elements at each location and sizes for each time point. The traces are then connected with edges to help clarify their sequences. Analysts can observe any interesting element while the previous states are still fresh in their memory. Anomalies emerge and can be examined even without animations, so analyses may be faster and less error prone. Points that move continuously through a range of values appear as clear trends. One key challenge must be addressed in the design of this view. The trend line direction must be made visually expressive, because there is no animation to indicate the direction. We solved this problem by using element transparency, fading from mostly transparent in the earliest elements to mostly opaque in the latest elements in the sequence. In order to perceive the flow direction even for smaller elements we employed the same approach with lines connecting the elements. In addition, it was necessary to render larger bubbles first to avoid occluding smaller bubbles. As described in [25], traces are particularly useful to reveal the nature of change and can help to examine the magnitude, shape, velocity, and direction of changes.

The support of statistical methods is also useful for examining the nature of change. The statistics provide simple summaries that form the basis of the initial description of the data and also serve as a part of a more extensive analysis. We implemented several measures that are commonly used to describe a dataset, i.e. measures of central tendency or measures of variability. The measures may be beneficial when identifying meaningful data characteristics of changes over time. We utilized both the univariate and the bivariate statistical methods. Input parameters for statistical methods consist of investigated MC variables. When an animation is running, each statistical measure is computed for every element on the background. Any combination of measure and variable can be selected using the user interface. The list of univariate measures includes coefficient of variation, skewness, mean, variance, standard deviation, median absolute deviation, median, geometric mean, and interquartile range. The mouse-click event on any element will extract an interactive HTML table on the right side of the chart area. The table consists of the measure computed for every element sorted in the descending order of the specified variable. If analysts select a row, the corresponding element will be highlighted. More precisely, the other elements are either transparent or hidden. Bivariate measures can be applied to any pair of variables. The list of bivariate measures includes sample covariance, sample correlation, and paired t-test.

The layout of the EDAIME user interface is presented in Figure 3. Using control, analysts can pause and advance the animation or change the speed. The Play, Pause, and Restart buttons are situated in the upper right corner next to the chart area. Above the buttons, the time slider is situated. Analysts can grab the time slider control to adjust the playback speed. Traces control is situated beneath the control buttons and it allows selecting elements of the interest to show their traces. This makes the selected elements more distinguishable and solves clutter issues.

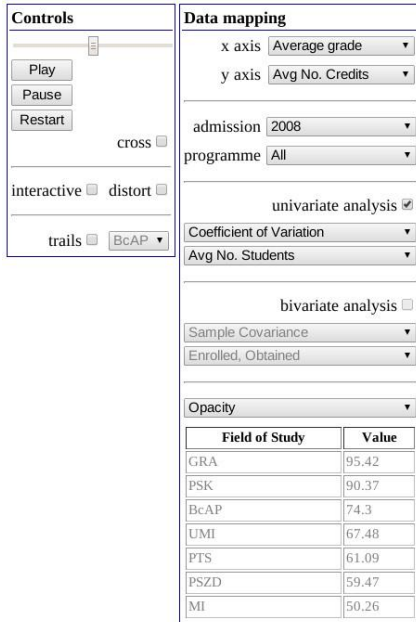


Figure 3. The EDAIME user interface layout.

## 5. EXPERIMENTATION

Any quantitative research of AA also requires a preliminary exploratory data analysis. Though useful, MC involves several drawbacks in comparison with common data visualization methods. Thus, empirical data is needed to evaluate its actual usability and efficacy.

In this section, we describe the experiment for the purpose of evaluating the efficacy of the enhanced MC methods implemented in EDAIME. We present the results including a detailed discussion. Twenty-two subjects (9 females, 13 males) with the average age of 31.6 (SD = 6.8) participated in our experiment. The participants ranged from 24 to 46 years of age. All participants came from professions requiring the use of data visualizations, including college students, analysts, and administrators. The experiment was conducted using standard desktop PCs. All subjects performed the experiment on an Intel Core i3 PC with 4 GB of RAM running Windows 7 or Fedora Core 20. Each PC had a 24" LCD screen running at the resolution of 1920 x 1080. We prefer Chrome as a web browser as it excellently supports HTML5 and CSS3 standards.

We performed a study to validate the usefulness and the general applicability of the enhanced version of MC in comparison with common data visualization methods when employed to analyze study related data. The experiment used a 4 (visualization) x 2 (size) within-subjects design. The visualizations varied between the static

and the animated methods. The static methods were represented by line plots (LP) and scatter plots (SP) which were generated for each semester. The animated methods were represented by the standard MC with the basic user interface (BMC) and the enhanced MC with advanced user interface features (EMC) described in the previous section. The size of datasets varied between small and large ones with the threshold of 500 elements. For the experiment, we utilized study related data about students admitted to bachelor studies of the Faculty of Informatics Masaryk University between the years of 2006 and 2012.

### 5.1 Hypotheses and Tasks

We designed the experiment to address the following three hypotheses:

- H1. BMC methods will be less effective than static methods when used for small datasets, and more effective when used for large datasets. In other words, the participants will be (a) faster and (b) make fewer errors when analyzing large datasets using BMC methods.
- H2. EMC will be more effective than the other methods for all datasets. In other words, the participants will be (a) faster and (b) make fewer errors when using EMC methods for all dataset sizes.
- H3. The participants will be more effective with small datasets than with large datasets. In other words, the participants will be (a) faster and (b) make fewer errors when analyzing small datasets.

In each trial, the participants completed 16 tasks, each with 1 to 5 required answers. Each task had students' IDs as the answer. Several questions have more correct answers than requested. The participants were asked to proceed as quickly and accurately as possible. In order to reduce learning effects, the participants were told to make use of as many practice trials as they needed. We also instructed them to practice until they had reached the desired performance level. Moreover, the participants had access to the tool several days before the experiment.

Sample of tasks:

- Select 4 students whose rate of enrolled credits was faster than their rate of obtained credits.
- Which student had the most significant decrease of the average grade?
- Select 5 students with the significant increase of the number of credits.
- Select 3 students whose average grade increased first and decreased later.
- Which student had the most significant increase in the number enrolled credits?

The participants selected answers by selecting student IDs in legend box located in the upper right from the chart area. In order to complete the task, two buttons can be used—either "OK" button to confirm the participant's choice or "Skip Question" button to proceed to the next task without saving the answer. There was no time limit during the experiment. For each task, the order of the datasets was fixed with the smaller ones first. This also allowed the participants to build their skills as they proceeded.

## 5.2 Study Method

The experiment used a 4 (visualization) x 2 (size) within-subjects design. Each experiment block was preceded with a training session in which we showed the subjects the correct answers after they confirmed it to allow participants to get familiarized with the settings and UI. It was followed by 16 tasks (8 small dataset tasks and 8 large dataset tasks in this order). After that, the subjects completed survey with questions specific for the visualization. Each block lasted about 2 hours. The subjects were screened to ensure that they were not color-blind and understood common data visualization methods. We also attempted to balance gender. The study results are divided into three sections: accuracy, completion time, and subjective preferences. To test for significant effects, we conducted repeated measures analysis of variance (ANOVA). Only significant results are reported. Post-hoc analyses were performed by using the Bonferroni technique.

## 5.3 Accuracy

Since some of the tasks required multiple answers, accuracy was calculated as a percentage of the correct answers. Thus, when a subject selected only three correct answers from five, we calculated the answer as 60 % accurate rather than an incorrect answer. The analysis revealed several significant accuracy results at the .05 level. The type of visualization had a statistically significant effect on the accuracy for large datasets ( $F(1.930, 40.535) = 25.655, p < 0.001$ ). Figure 4 illustrates graph of the mean accuracy of visualizations for large datasets including error bars that show the 95% confidence interval. Pair-wise comparison of the visualizations found significant differences showing that both animated methods were significantly more accurate than the static methods. EMC was more accurate than LP ( $p = 0.001$ ). EMC was also more accurate than the BMC ( $p < 0.001$ ). LP were more accurate than SP ( $p = 0.016$ ). For small datasets, visualizations were not statistically distinguishable, except for SP which had lower accuracy than other methods. Also, the subjects were more accurate with small datasets ( $F(1, 21) = 38.679, p < 0.001$ ) as can be seen in Figure 5.

## 5.4 Task Completion Time

An answer was considered to be incorrect if none of the correct answers was provided. In terms of time to task completion, we also observed a statistically significant effect ( $F(1.764, 37.044) = 43.875, p < 0.001$ ). Post-hoc tests revealed that BMC was the slowest for both dataset sizes. For large datasets, the LP was faster than the EMC ( $p < 0.001$ ). EMC and SP were not statistically distinguishable. The mean time for LP was 76.36 seconds compared to 85.95 seconds for the EMC—about 13% slower, 88.59 seconds for the SP—about 16% slower, and 91.64 seconds for the BMC—about 20% slower. For small datasets, static methods were significantly faster than animated. Pair-wise comparison of the visualizations found significant differences between all of them except for EMC and SP. LP were the fastest for all datasets. EMC was slower than the LP ( $p < 0.001$ ) and faster than the BMC ( $p < 0.017$ ). The mean time for BMC was 70.18 seconds compared to 67.6 seconds for the SP—about 3% faster, 66.55 seconds for the EMC—about 6% faster, and 61.36 seconds for the LP—about 14% faster.

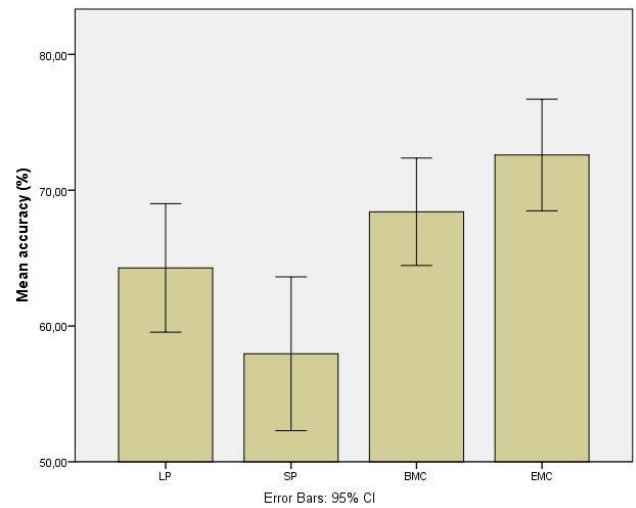


Figure 4. Mean accuracy of answers per visualization method.

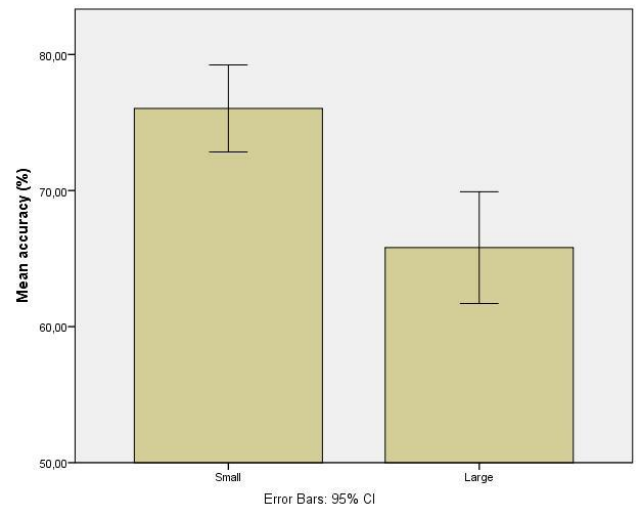


Figure 5. Mean accuracy of answers per dataset size.

## 5.5 Subjective Preferences

For each experiment block, the subjects completed a survey where the subjects assessed their preferences regarding analyses. The subjects rated the static and animated methods on a ten-point Likert scale (1 = strongly disagree, 10 = strongly agree). Using RM-ANOVA, we revealed statistically significant effects ( $F(1.696, 35.611) = 80.1332, p < 0.001$ ). Post-hoc analysis found that EMC was significantly more helpful than other methods, more precisely BMC ( $p < 0.001$ ) and LP ( $p < 0.001$ ). The obtained results are presented in Table 1, indicating the resulted mean values of the preferences for each question.

The significant differences indicate that animated methods were judged to be more helpful than the static methods. The subjects significantly preferred the LP to use for small datasets. However, animated methods were judged to be more beneficial than static methods for large datasets ( $p < 0.001$ ). The results also showed that

animated methods were more entertaining and interesting than the static methods ( $p < 0.001$ ).

**Table 1. The resulted mean values of the preferences.**

|   | LP   | SP   | BMC  | EMC  |
|---|------|------|------|------|
| The visualization was helpful in answering the questions. | 5.41 | 4.27 | 6.86 | 7.55 |
| I found this visualization entertaining and interesting.  | 5.36 | 5.14 | 7.14 | 8.05 |
| I prefer visualization for small datasets.                | 6.70 | 4.41 | 5.59 | 5.82 |
| I prefer visualization for large datasets.                | 5.90 | 5.18 | 7.41 | 8.32 |

## 6. DISCUSSION

Our first hypothesis (H1) was that BMC would outperform both the static methods for large datasets and will be less effective when used for small dataset. This hypothesis was confirmed only partially. BMC methods were more accurate than the static methods, but contrary to the hypothesis, the static methods proved to achieve better speed than the BMC for the both dataset sizes. Moreover, the methods were not statistically distinguishable in terms of accuracy for small datasets. The second hypothesis (H2) expected that EMC will be more effective than the other methods for all dataset sizes. The hypothesis was only partially confirmed as well. EMC was the most accurate method for all datasets. Contrary to the hypothesis, LP was the fastest method for all datasets. We also hypothesized that the accuracy will be higher for smaller datasets (H3). The hypothesis H3.a was supported, because the subjects were faster with small datasets. The mean time for large datasets was 85.64 seconds and for small datasets was 66.42 seconds. The hypothesis H3.b was also supported, because the subjects committed fewer errors with small datasets when compared with large datasets. Generally, the accuracy is the issue for static visualizations when large datasets were employed.

The EDAIME tool facilitates users to utilize the enhanced MC methods with advanced interactive features. After the experiment, multiple subjects reported that they make use of advanced user interface features and spent a lot of time exploring the data during the practice trials. In the final discussion, the several subjects reported that the animations were entertaining and interesting. Contrarily, several subjects reported that for large datasets as the number of elements rose they experienced increasing difficulty to identify and remember the element of their interest that they were following and without user interface features it would be hard to handle it. The overall accuracy was quite low in the study with average about 70%. However, only three questions were skipped.

The study supports the intuition that using animations in analysis requires convenient interactive tools to support effective use. The study suggests that EMC leads to fewer errors. Also, the subjects found MC methods to be more entertaining and exciting. They slightly preferred it to the static method. The evidence from the study indicates that the animations were more effective at building the subjects' comprehension of large datasets. However, the simplicity of static methods was more effective for small datasets. These observations are consistent with the verbal reports in which

the subjects refused to abandon the static visual methods generally. This finding illustrates that interest in animations does not preclude the subjects' appreciation of common methods. Overall, the participants would prefer to utilize both types of visual methods. Results supported the thoughts that MC does not represent a replacement of common statistic data visualizations but a powerful addition.

## 7. CONCLUSION AND FUTURE WORK

Commonly used static methods have principal limitations in terms of the volume and the complexity of the processed data. Animations are substantially transparent techniques that can present a good overview of the complex and large data. MC presents multiple elements and dimensions of the data on a single two-dimensional plane. The main contribution lies in enabling critical questions about data relationships and characteristics.

In the EDAIME tool, we enhanced the MC concept and expanded it to be more suitable for AA analyses. We also developed an intuitive, yet powerful, user interface that provides analysts with instantaneous control of MC properties and data configuration, along with several customization options to increase the efficacy of the exploration process. The tool provides a smart, convenient, and visually appealing way to identify potential correlations between different variables. We validate the usefulness and the general applicability of the designed tool with the experiment to assess the efficacy of the described methods in comparison with visual static methods.

The study suggests that animated methods lead to fewer errors for the large datasets. Also, the subjects find MC to be more entertaining and interesting. The entertainment value probably contributes to the efficacy of the animation, because it serves to hold the subjects' attention. This fact can be useful for the purpose of designing methods in learning settings. The more entertaining a method is, the easier it is to concentrate on the process and the more information can be acquired. The study also indicates that we need to appropriately adjust analytic tools when we begin to process time-varying, high-dimensional data. Especially, we need to focus on user interface features.

The current limitations of the tool are predominantly originated in the use of HTML5 standard, because there are still serious performance problems in several web browsers. Thus, only a certain number (generally less than 1000) of data points may be effectively visualized using animations. Features enabling effective data manipulation are essential. The additional representation of the data using enhanced MC methods gives analysts more possibilities in exploring the data.

We plan to create the synergy of EDAIME animated methods with common DM methods to follow the VA principle more precisely. We already implemented a standalone EDAIME method utilizing decision tree algorithm providing visual representation. We prefer decision trees because of their clarity and simplicity to comprehend.

## 8. ACKNOWLEDGMENTS

We thank all colleagues of IS MU development team and Knowledge Discovery Lab for their assistance. This work has been partially supported by Faculty of Informatics, Masaryk University.

## 9. REFERENCES

- [1] Goldstein, P. J. 2005. Academic analytics: The uses of management information and technology in higher education. EDUCAUSE. Retrieved from <https://net.educause.edu/ir/library/pdf/ers0508/rs/ers0508w.pdf>.
- [2] Campbell, J. P., DeBlois, P. B., and Oblinger, D. G. 2007. Academic analytics: A new tool for a new era. *EDUCAUSE Review*, 40-57.
- [3] Romero, C. and Ventura, S. 2013. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12-27. DOI: <http://dx.doi.org/10.1002/widm.1075>.
- [4] Delavari, N., Phon-Amnuaisuk, S., and Beikzadeh, M. R. 2008. Data Mining Application in Higher Learning Institutions. *Informatics in Education*, 31-54.
- [5] O'Reilly, T. and Battelle, J. 2009. Web squared: Web 2.0 five years on. DOI: <http://dx.doi.org/10.4304/jait.2.4.204-216>.
- [6] Tversky, B., Morrison, J. B., and Betrancourt, M. 2002. Animation: Can It Facilitate?. *International Journal Human-Computer Studies*, 247-262. DOI: <http://dx.doi.org/10.1006/ijhc.2002.1017>.
- [7] Margaret, S., Chan, J., and Black, B. 2005. When can animation improve learning?. Some implications on human computer interaction and learning. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, 933-938.
- [8] Le, D.-T. 2013. Bringing Data to Life into an Introductory Statistics Course with Gapminder. *Teaching Statistics*, 114-122. DOI: <http://dx.doi.org/10.1111/test.12015>.
- [9] Heer, J. and Robertson, G. 2007. Animated Transitions in Statistical Data Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 1240-1247. DOI: <http://dx.doi.org/10.1109/TVCG.2007.70539>.
- [10] Kehoe, C., Stasko, J., and Taylor, A. 2001. Rethinking the Evaluation of Algorithm Animations as Learning Aids: An Observational Study. *International Journal of Human-Computer Studies*, 265-284. DOI: <http://dx.doi.org/10.1006/ijhc.2000.0409>.
- [11] Bertamini, M. and Proffitt, D. R. 2000. Hierarchical motion organization in random dot configurations. *Journal of Experimental Psychology: Human Perception and Performance*, 1371-86.
- [12] Robertson, G., Cameron, K., Czerwinski, M., and Robbins, D. 2002. Animated Visualization of Multiple Intersecting Hierarchies. *Information Visualization*, 50-65.
- [13] Few, S. 2009. *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press. DOI: <http://dx.doi.org/10.1080/10543401003641225>.
- [14] Baudisch, P., Tan, D., Collomb, M., Robbins, D., Hinckley, K., Agrawala, M., Zhao, S., and Ramos, G. 2006. Phosphor: explaining transitions in the user interface using afterglow effects. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*. DOI: <http://dx.doi.org/10.1145/1166253.1166280>.
- [15] Al-Aziz, J., Christou, N., and Dinov, I. D. 2010. SOCR Motion Charts: an efficient, open-source, interactive and dynamic applet for visualizing longitudinal multivariate data. *Journal of Statistics Education*, 18(3), 1-29.
- [16] Grossenbacher, A. 2008. The globalisation of statistical content. *Statistical Journal of the IAOS. Journal of the International Association for Official Statistics*, 133-144.
- [17] Baldwin, J. and Damian, D. 2013. Tool usage within a globally distributed software development course and implications for teaching. *Collaborative Teaching of Globally Distributed Software Development*, 15-19. DOI: [10.1109/CTGSD.2013.6635240](http://dx.doi.org/10.1109/CTGSD.2013.6635240).
- [18] Sultan, T., Khedr, A., Nasr, M., and Abdou, R. 2013. A Proposed Integrated Approach for BI and GIS in Health Sector to Support Decision Makers. *Editorial Preface*.
- [19] Yoon, S., Elhadad, N., and Bakken, S. 2013. A Practical Approach for Content Mining of Tweets. *American journal of preventive medicine*, 122-129. DOI: <http://dx.doi.org/10.1016/j.amepre.2013.02.025>.
- [20] Vermeylen, J. 2008. *Visualizing Energy Data Using Web-Based Applications*. American Geophysical Union.
- [21] Weintrop, D., Beheshti, E., Horn, M. S., Orton, K., Trouille, L., Jona, K., and Wilensky, U. 2014. Interactive Assessment Tools for Computational Thinking in High School STEM Classrooms. *Intelligent Technologies for Interactive Entertainment*, 22-25. DOI: [http://dx.doi.org/10.1007/978-3-319-08189-2\\_3](http://dx.doi.org/10.1007/978-3-319-08189-2_3).
- [22] Hilpert, M. 2011. Dynamic visualizations of language change: Motion Charts on the basis of bivariate and multivariate data from diachronic corpora. *International Journal of Corpus Linguistics*, 435-461. DOI: <http://dx.doi.org/10.1075/ijcl.16.4.01hil>.
- [23] Battista, V. and Cheng, E. 2011. Motion Charts: Telling Stories with Statistics. *JSM Proceedings, Statistical Computing Section*, 4473-4483.
- [24] Olmos, M. and Corrin, L. 2012. Academic analytics in a medical curriculum: enabling educational excellence. *Australasian Journal of Educational Technology*, 1-15.
- [25] Few, S. 2007. *Visualizing Change: An Innovation in Time-Series Analysis*. In *Visual Business Intelligence Newsletter*, White paper SAS.
- [26] Geryk, J. and Popelinsky, L. 2014. Analysis of Student Retention and Drop-out using Visual Analytics. In *Proceedings of the 7th International Conference on Educational Data Mining*. International Educational Data Mining Society, 331-332.
- [27] Geryk, J. and Popelinsky, L. 2014. Visual Analytics for Increasing Efficiency of Higher Education Institutions. In *Proceedings of the 6th Workshop on Applications of Knowledge-Based Technologies in Business*, 117-127. DOI: <http://dx.doi.org/10.1007/978-3-319-11460-6>.