Improving Comprehension Efficiency of High Content Screening Data Through Interactive Visualizations

Wienand A. Omta,[1–3] Jacob de Nobel,[3] Judith Klumperman,[1] David A. Egan,[3] Marco R. Spruit,[2] and Matthieu J.S. Brinkhuis[2]

[1]Department of Cell Biology, Center for Molecular Medicine, UMC Utrecht, Utrecht, Netherlands.
[2]Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands.
[3]Core Life Analytics B.V., Utrecht, Netherlands.

ABSTRACT

In this study, an experiment is conducted to measure the performance in speed and accuracy of interactive visualizations. A platform for interactive data visualizations was implemented using Django, D3, and Angular. Using this platform, a questionnaire was designed to measure a difference in performance between interactive and noninteractive data visualizations. In this questionnaire consisting of 12 questions, participants were given tasks in which they had to identify trends or patterns. Other tasks were directed at comparing and selecting algorithms with a certain outcome based on visualizations. All tasks were performed on high content screening data sets with the help of visualizations. The difference in time to carry out tasks and accuracy of performance was measured between a group viewing interactive visualizations and a group viewing noninteractive visualizations. The study shows a significant advantage in time and accuracy in the group that used interactive visualizations over the group that used noninteractive visualizations. In tasks comparing results of different algorithms, a significant decrease in time was observed in using interactive visualizations over noninteractive visualizations.

Keywords: interactive visualizations, VDM, data mining

INTRODUCTION

The enormous increase in data generation that has occurred in the last few decades has greatly challenged researchers. "Never before in history have data been generated at such high volumes as it is today. Exploring and analyzing the vast volumes of data becomes increasingly difficult."[1] This statement is as true today as it was in 2002; more data have been created in the last 2 years than in the entire previous history of human race. The rate at which data creation is happening is ever increasing and it is estimated that by the year 2020, about 51 TB of data will be created per year for every person.[2] This may be truer for biology than for any other branch of science, as with the advent of Next Generation Sequencing, high content screening (HCS), and other high-throughput technologies, life scientists are producing more and more data.

HCS is a technology that combines automated fluorescence microscopy and image analysis to measure phenotypic response of cells to bioactive molecules. Using image analysis, changes in cell morphology are detected. Because multiple features are measured at the same time, this technique can be used for complex tasks such as drug candidate target prediction. It has recently been documented, however, that most HCS experiments do not exploit their full potential, as 60%–80% of all HCS screens only use one or two measured variables, even though hundreds of variables can be measured per individual cell.[3] The fact that the larger part of the data generated with high-throughput technologies (high-throughput screening) is being produced in an automated manner makes the analysis of the data even more important.

Usually many features are recorded, leaving the researchers with highly dimensional data. The data produced in these experiments are often highly heterogeneous, which adds to its complexity. Data mining of biological high-throughput data is therefore much more difficult because prior knowledge is required to understand the patterns in the data.[4]

There is a shift in the analysis from classical statistics to machine learning in high-throughput data, because effectively analyzing all available data though directed statistical analysis is undoable.[5] However, the advanced data mining knowledge required to analyze these large amounts of data is often lacking in the toolkit of most life scientists.[6] This pushes the analysis of experimental results away from the life scientist and into the domain of the data scientist, who often lacks the expertise of the life scientist. Therefore a solution

should be sought that allows life scientists to analyze their own experiments.

HC StratoMineR, a tool for the analysis of true multivariate high content data, was recently published. It is a Web-based tool for the analysis of HCS data,[6] which allows the user to make use of the full potential of high content (HC) data. Within the workflow of HC StratoMineR, there are various opportunities for the user to visualize the data. The visualizations are carried out by R, using the ggplot2 library, generating noninteractive visualizations.[7]

Data Visualization

When data are being presented in a textual or tabular form, the amount of data that can be interpreted by a person is limited to a couple of hundred records. Beyond that, some way of visualizing the data is required to understand what information is hidden in the data.[8] Visualization of data is a very powerful way to show activity or artifacts within data. The visualization of data is not only important to achieve clear insight into what the data look like in its raw state but can also trigger new discoveries and insights.[9,10] For example, it is possible to summarize data using standard deviations, means, medians, and ranges. As interesting as these summary statistics are, they still do not tell the researcher if the distribution of data is normal or if there are any outliers in the data. In contrast, a visualization will provide the researcher with an answer to these questions.[11] Visualizations can thus give the user a better understanding of how the data are composed and provide a clear overview of its distribution.

Visual Data Mining

In data mining, statistical methods and algorithms such as, Naive Bayes, principal component analysis, square root transformation, normalization, and other methods are used to analyze large sets of data. Despite the effectiveness of these techniques, their complex nature makes the data mining process difficult to comprehend for nondata scientists. Because specific skills are required to properly configure these algorithms and interpret their results, the amount of control a researcher has over the analysis is diminished.[12]

It is important to include the flexibility and creativity of the human mind in the data exploration process to make use of the cognitive capacity of the human brain. Visual data mining (VDM) focuses on integrating the user directly into the data exploration process by presenting the data in some visual form.[1] Including VDM into the data exploration process enables users to explore large volumes of data without having to understand complex statistical or mathematical methods and algorithms. VDM can still be used with noisy and heterogeneous data through the direct involvement of the user. These aspects of VDM allow for faster data exploration, and it regularly produces better results than automated methods.[1] For example, a visualization may reveal distinct clusters in a data set, while the automated analysis of the same data set may not be able to detect these clusters due to the noisiness of the data. In addition to the detection of clusters, VDM is useful for many other data analysis tasks, such as the following: outlier detection, feature importance assessment, and the detection of patterns.[11]

Even though VDM can be performed without the use of data mining algorithms, the combination of both VDM and data mining allows for an even better data analysis solution. For instance, data mining can be used to provide visualizations with the simplification needed to make them more comprehensible, for example, reducing the number of dimensions in the data to be visualized by factor analysis. Also, VDM is useful in exploring the (intermediate) results of data mining methods or to make the process of a method more clear,[11] for example, a layer-by-layer visualization of k-means clustering. Usually, data analysis is performed by a workflow consisting of multiple data mining methods. VDM can be used to explore the results of the entire workflow or to inspect the (relative) effect of a single method in the workflow.[11,13]

However powerful, there are certain limitations to the capability of the human visual system. There is a physical limitation to a maximum of three axes in a visualization. Also, there is a limit to the number of preattentive features, such as hue, orientation, intensity, and size, that can be combined freely.[14] If three or more of such features are used in the same visualization, they can greatly reduce the comprehensibility of that visualization.

When dealing with large sets of data, misinterpretation, disorientation, and occlusion of parts of data are prone to happen.[11] Therefore, to optimally use the potential that VDM has to offer, methods are needed to communicate the main trends of the data effectively. For instance, visualizing a random subset of the data can help to increase the readability of the distribution and the variance. Simplified visualizations such as a box plot can reduce the number of elements to five when visualizing a vector, where a scatter plot can contain thousands of elements (dots) to visualize the same vector. These simplified visualization methods can support the construction of visualizations that can be interpreted efficiently by the user, even when dealing with large amounts of data.

Interactive Visualizations

VDM follows the following paradigm: overview first, zoom and filter, and then details on demand.[1] This makes the interactivity of the visualization an important aspect of VDM because it allows users to directly interact with the visualization. A theory on direct manipulation (DM) was described by Shneiderman[15]

within the context of computer applications and graphical user interfaces. DM endeavors toward an interface such that the object manipulated by the interface is part of the interface itself.[15] Typical examples of DM are zooming in on a picture using your fingertips, swiping on a tablet to the next page of a document, or uploading a file using drag and drop functionality by selecting and dragging the file to an upload form. Even though DM was devised in the early 80's, it still remains one of the standard doctrines in interface design.[16]

A good way to combine VDM with automated statistics and data mining is to provide users with an application that implements DM. Users can explore their data by direct interaction with the visualization. When interesting patterns are detected, data mining can be used to further investigate the phenomenon. Because the visualization itself is used as an interface and is constantly being displayed, users can optimally interact with their data.[11]

By equipping a visualization with DM, a user is able to directly manipulate the visualization, aiding in effective data exploration by focusing on interesting sections. Keim[1] divides the methods of interactivity of visualization into five categories:

- Dynamic projection: Through dynamic projection of a multidimensional data set, multiple dimensions can be viewed in one visualization. This method allows users to dynamically change the esthetics of a visualization, such as the ability to change the variables that are on the axis of a visualization.
- Filtering: Filtering the data to be visualized can also be very helpful in data exploration, by dividing the data into interesting subsets. This can be done by querying a data set for interesting records or by browsing the data through different subsets. The visualization will be dynamically updated according the filtered data.
- Zooming: By zooming, very large amounts of data can be condensed for an overview of the data, while interesting parts can be magnified for a more detailed inspection.
- Distortion: Interactive distortion techniques may be used to focus on a specific section of the data, while preserving the overview of the complete data set, for example, fisheye view. An example of interactive distortion is the blurring of records that are below a certain threshold, showing only the records that are above that record. An example of distortion in a heatmap is shown in the *Supplementary Figure S1* (Supplementary Data are available online at www.liebertpub.com/adt). In this example, interactive distortion techniques are used to "find" the top 10% of the data.
- Linking and brushing: Brushing over a visualization allows for the highlighting of sections, aiding in finding interesting patterns in the data. Linking visualizations to other visualizations may also provide the user with new insights. For example, linking a scatter plot to a histogram (*Supplementary Fig. S1*) combines two different outlooks over the data in one view.[1] By brushing over the histogram, the data in the scatter graph are updated.

Application of Interactive Methods

There are numerous data cleansing, preparation, and manipulation methods (algorithms) available that can be applied to (a part of) the data.[13,17] Examples include not only normalization, transformation, and scaling algorithms but also clustering and classification algorithms. Some of these are very general, while others are designed for very specific purposes, for example, the B-score normalization method, a row and column polish for the correction of plate effects in the analysis of HCS.[18] The majority of these algorithms lack a description in terms of heuristics or best practices. Moreover, the best practices that are available are not always fitting for each situation.[12] Consequently, it is not always clear which of these methods yields the best results. As previously mentioned, a visualization of the results can provide the user with information about the effect of an algorithm on the data.

In a visualization, the interactive categories of Keim, dynamic projection, filtering, zooming, distorting and linking of the data or parts thereof, can reveal patterns in the data that are not easily visible without using other methods. When using noninteractive data visualization, the application of these methods is not possible. In information processing, both speed and accuracy play an important role.[19] In the context we investigate, we require a high accuracy degree and we expect a main difference in speed. In other words, we primarily expect a difference in speed.

Hypothesis 1a: Patterns in the data can be detected faster using interactive data visualizations over noninteractive visualizations.

Hypothesis 1b: Patterns in the data can be detected more accurately using interactive data visualizations over noninteractive visualizations.

Hypothesis 2a: A faster interpretation can be made about the output of different algorithms using interactive visualizations over noninteractive visualizations.

Hypothesis 2b: A more accurate interpretation can be made about the output of different algorithms using interactive visualizations over noninteractive visualizations.

MATERIALS AND METHODS

Because we are seeking for a way to measure the effects of multiple data manipulation methods (algorithms) and simultaneously provide users with the insight of interactive data visualizations, we developed a questionnaire that includes the

ability to create interactive visualizations and noninteractive visualizations. Each question is supported with one or more visualizations.

Setup of the Experiment

The questionnaire contains 12 questions (*Supplementary Appendix S1*) with two additional test questions. The test questions are used to train the participant in the look and feel of the interface and the format of the questionnaire. The questionnaire is carried out with a control and an experimental group. In the control group, noninteractive visualizations are shown and interactive visualizations are shown in the experimental group. The same questions and visualizations are presented to both groups. See the visualizations offered with the questions in *Figure 1*. The questions are offered in a random order to take the user's attention and learning curve into account. This is to avoid the possibility that certain questions are always first or last, which might affect the representativeness of the given answers. The two measured constructs are accuracy and the time required to answer the questions. For each question, the time is measured independently and the time measured starts when the visualization is fully loaded; so the loading time and the speed of each user's computer do not affect the time to answer the question.

A classroom was prepared for participation. There were 33 students instructed to bring a laptop with an external mouse and headphones to watch the instruction video carefully. They were asked to fill in the questionnaire as accurately and as quickly as possible, with a reward of €50 for the fastest student with the least number of wrong answers. The focus is concentrated on the accuracy over speed to avoid participants clicking through the questionnaire as quickly as possible and have a few correct answers by chance. The participants were instructed not to ask any questions during the questionnaire. In total, 79 students, including computer science, information science, and bioinformatics students, from Leiden and Utrecht participated, whereof 46 people with various backgrounds participated over the Web.

Materials

The questionnaire was conducted in a Web-based environment. The interactive and noninteractive versions were performed similarly. The participants were asked to visit a Web site (http://interactiveplotting.stratominer.com/survey) to participate in the questionnaire (see screenshot in the *Supplementary Fig. S2*). Participants were randomly assigned to the experimental or control group.

The back end of the questionnaire was built using the Python Django framework, which used MySQL for data storage. The Django framework was not only responsible for the random allocation of participants to a group but also for the random ordering of questions in the questionnaire. To communicate with the frontend of the questionnaire, Django exposed a RESTful api.

The front end of the questionnaire was built using the Angular framework, which was responsible for the rendering of the interface. The Angular framework had an asynchronous connection with the Django backend through observables. The (interactive) visualizations are all generated using D3.js, with the exception of the 3D scatter graphs, which were constructed by the Vis.js library. The answering time was measured by the front end of the application. The answering time was measured from the moment that visualizations were loaded; so loading time of the visualization does not affect the answering time. However, in the noninteractive version, the switching of visualizations is artificially delayed, to simulate the time it takes to render a noninteractive visualization in ggplot2. Both the interactive and the noninteractive questionnaires are built using exactly the same engine and framework to avoid other external factors of influence.

Data Analysis Methods

The data set contains 79 records. Each record contains metadata about the participant's age, gender, and educational level. Computer literacy, Excel, data mining, and English proficiency are also measured in a Likert Scale from 1 to 5, to discover any relationship to these proficiencies or demographical properties. To measure the construct accuracy, the total number of incorrect answers per participant was measured. To measure the construct time, the sum of time for each individual question per participant was measured. To measure the construct time to compare different algorithms, the total time of the questions related to the construct was measured (*Supplementary Table S1*). To measure the construct accuracy to compare different algorithms, the total number of incorrect answers of the questions related to the construct was measured (*Supplementary Table S1*).

Students who did not finish the questionnaire completely were left out, which results in a dataset of 68 students. To measure a difference in time and accuracy between the control and experimental group, a one-tailed independent sample *t*-test is used. To measure a difference in the accuracy of detection of patterns in the data between the control and experimental group, a chi-square test was used.

Due to the complexity of the questionnaire and the fact that participants would be using the Web-based environment for the first time, introductory material was provided to the participants in advance. Each group was provided with their own relevant

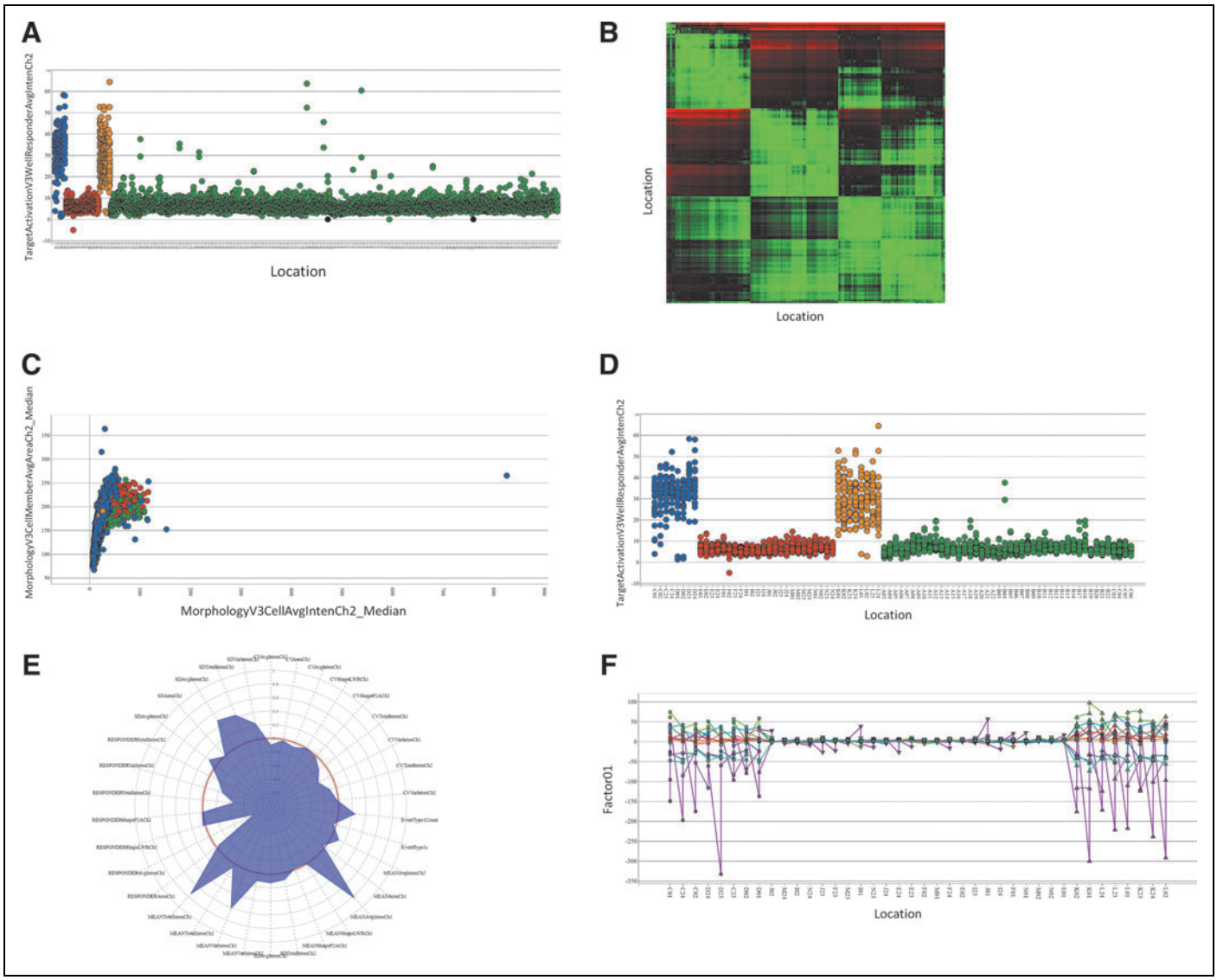

Fig. 1. Visualizations of the questionnaire. **(A)** A scatter plot with the well locations (discrete variable) on the x-axis and the variable TargetActivationV3WellRESPONDERAvgIntenCh2 on the y-axis. **Question 1:** What is the number of records where TargetActivationV3WellRESPONDERAvgIntenCh2 is 50 or higher? **(B)** A correlation matrix showing the similarity of the well locations on the x-axis and y-axis in *green*. **Question 2:** See the correlation matrix below. Select the pair that correlates between 0.999999 (99.999%) and 1 (100%). **(C)** A scatter plot showing MorphologyV3CellAvgIntenCh2_MEDIAN on the x-axis and MorphologyV3CellAvgAreaCh2_MEDIAN on the y-axis. **Question 3:** In the scatter plot below, MorphologyV3CellAvgIntenCh2_MEDIAN(X-axis) contains an outlier of 825. What is the plateName of this outlier? **(D)** A scatter plot showing the well location (discrete variable) on the x-axis and TargetActivationV3WELLRESPONDERAvgIntenCh2_MEDIAN on the y-axis. **Question 4:** What happens to the variance of the data when it is $\log_2$ transformed? **(E)** A polar plot showing the variables at the x-axis and the factor loading on the radius in a range from −1 to 1. **Question 5:** Select the variable that shows a loading of 0.82 on Factor01. **(F)** A line plot showing the well locations on the x-axis (discrete variable) and Factor 01 on the y-axis. **Question 6:** In the line plot below, 12 lines are shown. Each line represents a different microplate. Select the microplate that is most similar to microplate 2. **(G)** A scatter plot showing Factor01 on the x-axis and Factor05 on the y-axis. **Question 7:** In the dataset provided, there are two variables that result in a cross using a scatter plot. What is the combination of variables that produces a cross? **(H)** A scatter plot showing an Euclidean distance score on the x-axis and Factor01 on the y-axis. **Question 8:** In this dataset, there is a record that contains a distance value of 9039.15 and a Factor01 value of 1.33. On what wellLocation is this record located? **(I)** A scatter plot showing Factor02 on the x-axis and Factor01 on the y-axis. **Question 9:** The scatter plot below contains one *red dot*. Select the Factor01 value of the record closest to the *red dot*. **(J)** A bar plot showing an id number of the x-axis and Factor03 on the y-axis. **Question 10:** Select the IdNumber of the record with the lowest value for Factor03. **(K)** A box plot showing the plate name on the x-axis and a variable named currentVariable on the y-axis. **Question 11:** Select the two normalization methods that align the medians of the box plots on the Y-axis. **(L)** A scatter plot showing the well locations on the x-axis and a variable called currentVariable on the y-axis. **Question 12:** Select the group that contains the record with the highest value. For ease of reading, the figure can be viewed online at www.liebertpub.com/adt

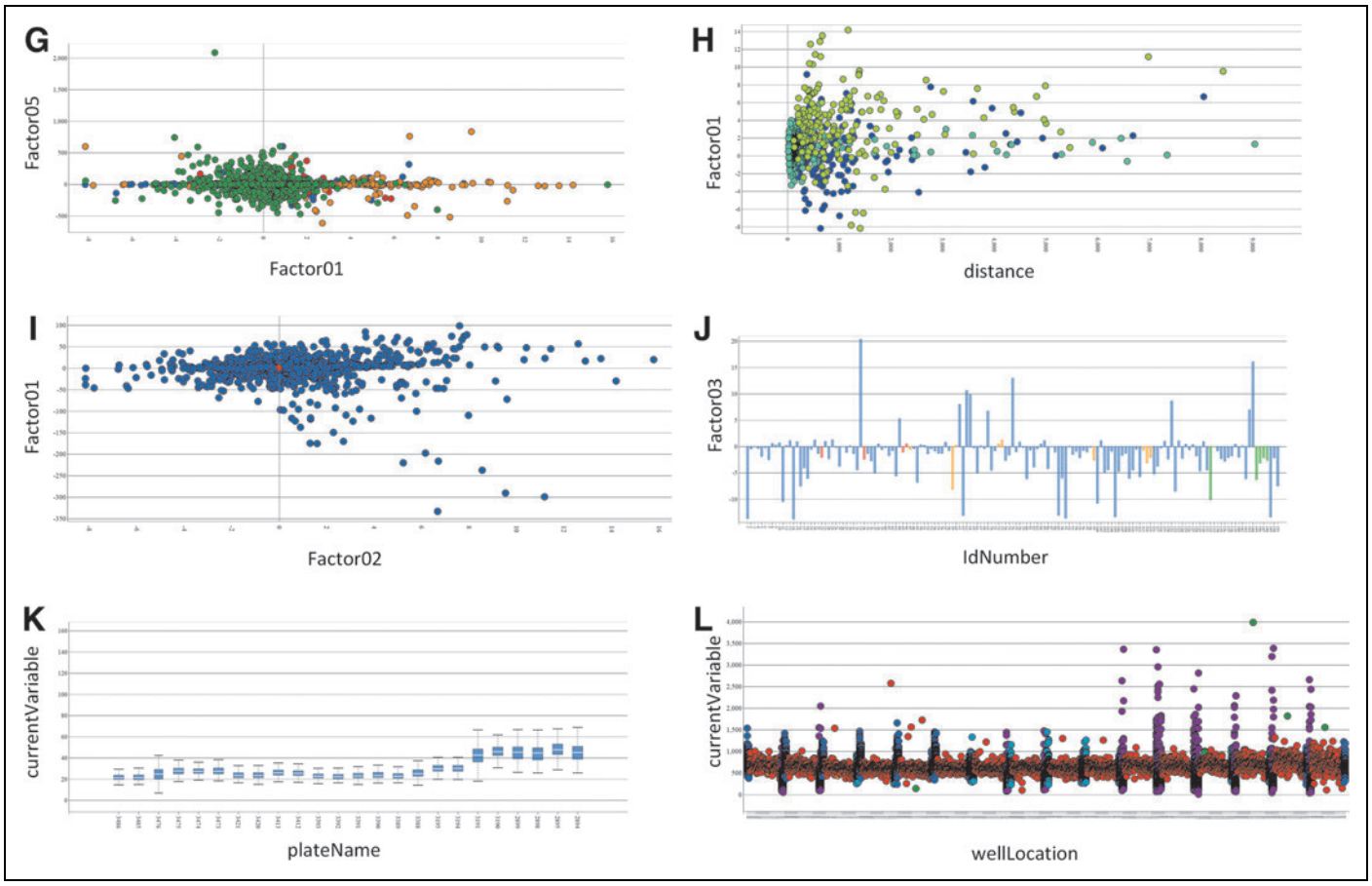

Fig. 1. (Continued).

introductory material. The introductory material consists of two test questions, each one accommodated by an explanatory video. Also, the questionnaire was provided with question-specific textual information and information to use the interface.

Table 1. The Statistical Summary of Time Comparing Methods, Total Time, and Wrong Answers in Both Conditions

Metric	Condition	*N*	Mean	SD	Standard Error	One-Tailed *P* Value
Total time (s)	Interactive	39	867.01	372.79	59.69	<0.001
	Noninteractive	29	1338.58	662.91	123.10	
Total wrong answers	Interactive	39	2.38	2.68	0.43	<0.001
	Noninteractive	29	4.41	2.24	0.42	
Time comparing algorithms (s)	Interactive	39	138.56	75.86	12.15	<0.05
	Noninteractive	29	176.57	90.95	16.89	
Wrong answers comparing algorithms	Interactive	39	0.67	0.701	0.148	>0.05
	Noninteractive	29	0.93	0.799	0.112	

Interactive and noninteractive, including the one-tailed *P* value.

RESULTS

Table 1 shows the results of the questionnaire performed in this study. The first result is the time in seconds (total time) to complete the questionnaire. On average, it took students more time to complete the questionnaire in the noninteractive group (M = 1338.58, SD = 662.91) than the interactive group (M = 867.01, SD = 372.79); $t(66) = 3.726$, $P < 0.001$ (*Fig. 2A*). A second result is the number of wrong answers given in the questionnaire (Total Wrong Answers).

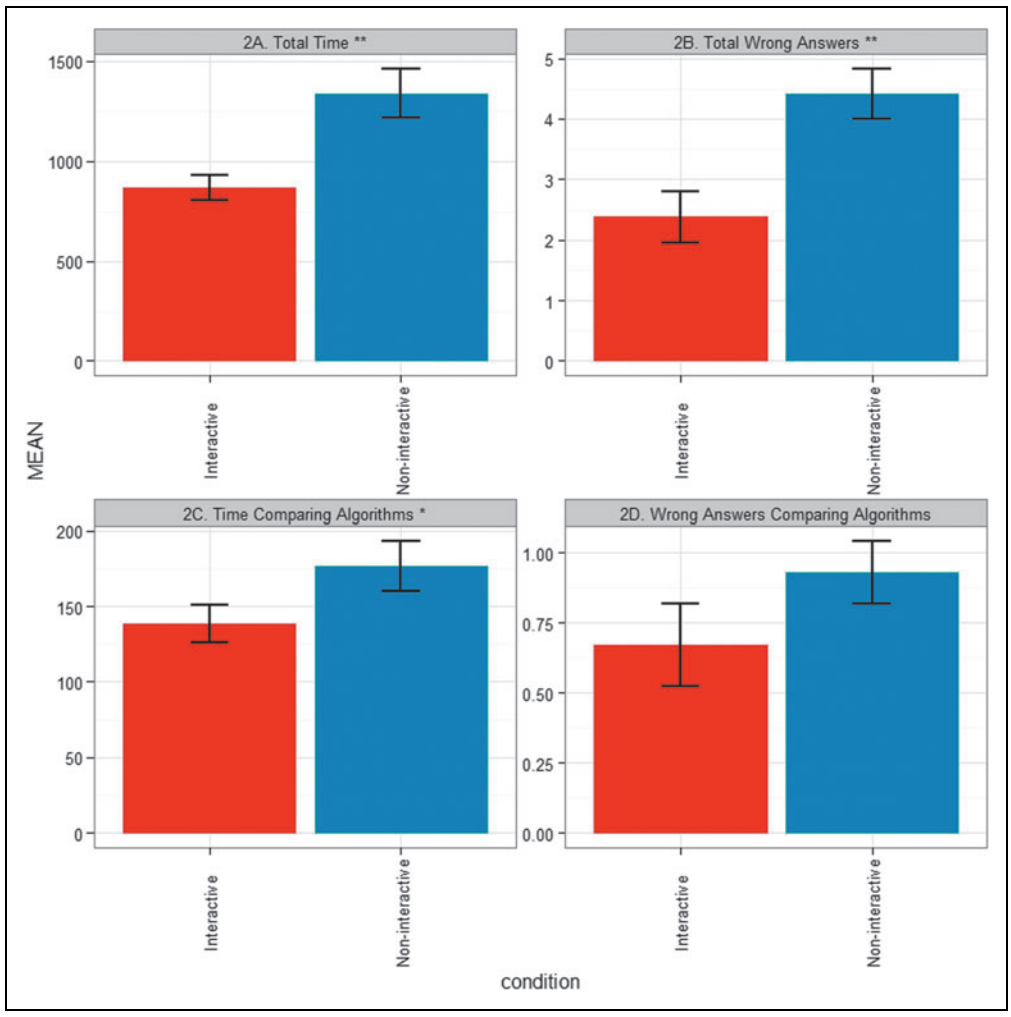

Fig. 2. Bar charts, including error bars, showing differences between interactive and noninteractive visualizations. **(A)** Total time indicates the time to complete the questionnaire, excluding the loading time of the visualizations (significant difference, one-tailed $P<0.001$**). **(B)** Total wrong answers indicate the number of wrong answers of the questionnaire (significant difference, one-tailed $P<0.001$**). **(C)** Time comparing algorithms indicates the time to compare algorithms in seconds (significant difference, one-tailed $P<0.05$*). **(D)** Wrong answers comparing algorithms (result not significant).

The questionnaire in total contains 12 questions (excluding test questions). On average, students had more wrong answers in the noninteractive group (M = 4.41, SD = 2.24) than the interactive group (M = 2.38, SD = 2.68); t(66) = 3.303, $P<0.001$ (*Fig. 2B*). A third result is the time in seconds that was required to compare multiple algorithms (time comparing algorithms). On average, it took students more time to compare algorithms in the noninteractive group (M = 176.57, SD = 90.95) than the interactive group (M = 138.56, SD = 75.86); t(66) = 1.877, $P<0.05$ (*Fig. 2C*). A final result is the number of wrong answers given in the questionnaire related to the comparison of algorithms. There was no significant association between the visualization group and the accuracy of the comparison of algorithms, $P>0.05$ (*Fig. 2D*).

The results contain relatively high standard deviations because some of the students were done extremely fast, for example, 91 s, while others took over 59 min to complete the questionnaire.

Learning Effect

The questions built in the questionnaire were presented in a randomized manner to take the user's attention and learning curve into account. This is to avoid that certain questions are always first or last, which might affect the representativeness of the given answers. Although there is a difference in time one question requires to be answered, there is still a learning curve present that can be explained by the fact that the interface is completely new and possibly a new way to visualize data. We took the questions in the real order and took the median of time across the two conditions: interactive and noninteractive. *Figure 3* shows a curve for both conditions, summarizing the median of time required to answer the first until the last question in the order the questions were offered to the participants, hence allowing to study this learning effect directly. For example, it is expected that the time to answer the questions decreases after the first few questions. As can be seen in *Figure 3*, we indeed clearly see that there is an initial peak in answering times for the first few questions, after which it seems to decrease to some plateau. We see that, despite the graph being quite noisy due to the smaller sample size and outliers, the curves both demonstrate a reduction in answering time throughout the test and indeed become more stable after about four items. Although the gap reduces, the interactive condition seems to remain somewhat faster. The last questions seem somewhat slower, which might be an indication of testing tiredness of the group of students.

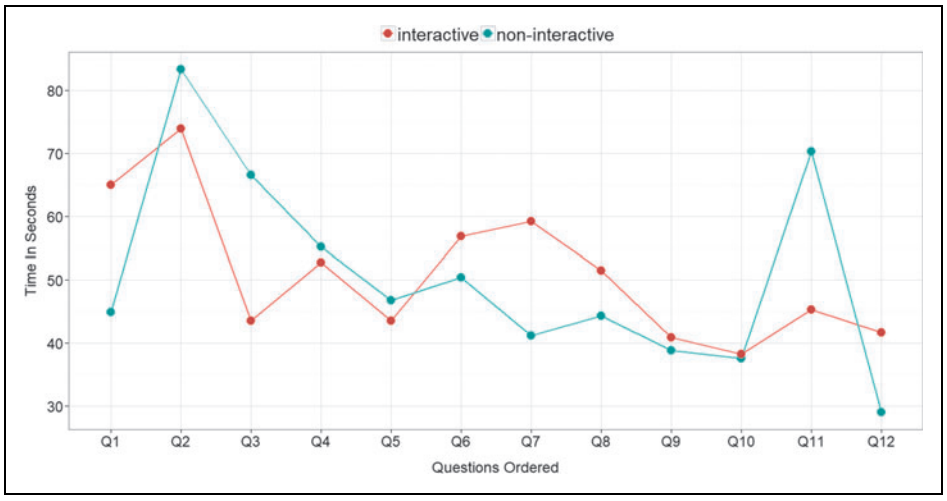

Fig. 3. The learning curve of the interactive and noninteractive versions by time taken per question. The x-axis shows the question in the order they were shown to the participants. The y-axis shows the median of the time of the i^{th} question. We regard only the ordering of questions. First question clearly takes more time, later questions less time.

Expertise

In the questionnaire, data mining expertise was measured in a Likert scale, by asking participants to rate themselves from 1 to 5 in data mining expertise, as shown in *Figure 4*. To show if there is a relationship with time/accuracy, we calculated Spearman correlations. Data mining expertise against the total time that was required to carry out the questionnaire shows a Spearman's rho coefficient of $\rho = -0.082$, NS. Data mining expertise against the total number of wrong answers shows a Spearman's rho coefficient of $\rho = -0.293$, NS. Hence, no significant monotone relationship was found between self-reported data mining expertise and response time, nor between data mining expertise and accuracy.

DISCUSSION

Visualizations with interactive methods are shown to provide better comprehensibility than noninteractive data visualizations. The overall time to perform 12 assignments was significantly decreased in the interactive group. The number of wrong answers given by the interactive group also showed a significantly lower number. The time that was required for comparing the output of algorithms was significantly decreased in the interactive group, although the number of wrong answers given by the interactive group to compare the output of algorithms did not show a significant result. Because we stressed the importance of accuracy

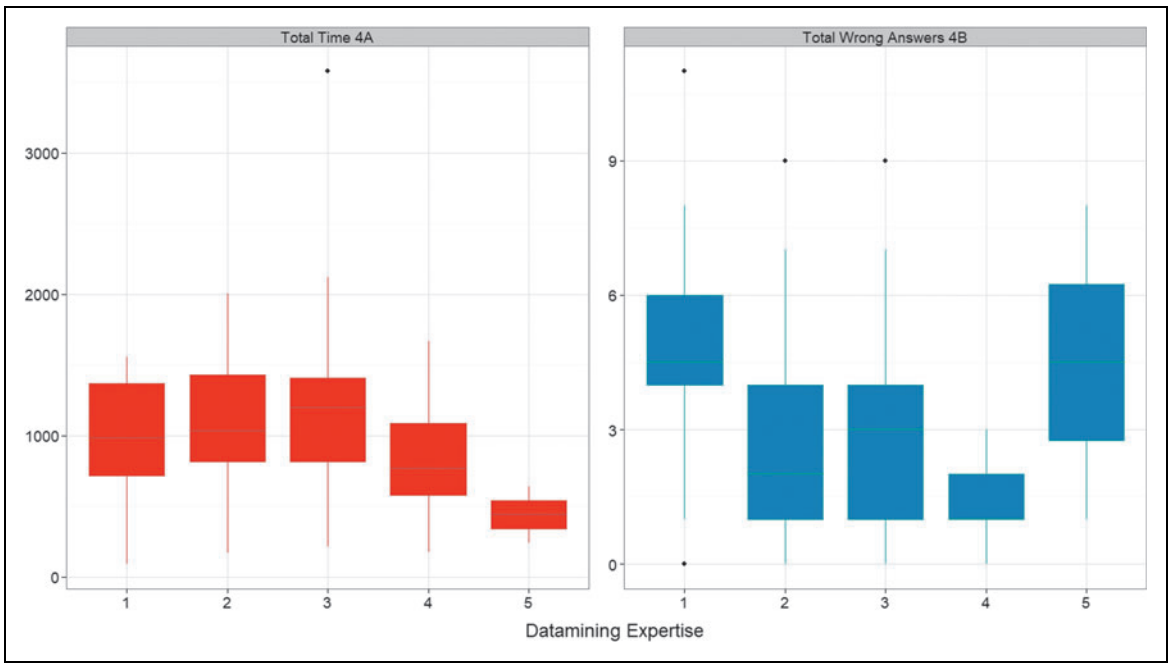

Fig. 4. Box plots showing data mining expertise in relationship to total time and total wrong answers. **(A)** The x-axis shows the data mining Likert scale from 1 to 5. The y-axis shows the box plot of the total time in seconds required to complete the questionnaire. **(B)** The x-axis shows the data mining Likert scale from 1 to 5. The y-axis shows the box plot of the total wrong answers of the questionnaire.

over time toward the participants, we influenced their speed and accuracy trade-off described by Wickelgren,[19] and hence mainly expected a reduction in time for the interactive condition. However, when using the right visualization tools, it was demonstrated that both accuracy and speed can be improved at the same time. From the 79 participants, 46 participants who performed the questionnaire over the Web did not have affinity with data analytics perse. The students who conducted the questionnaire in a classroom were students in either bioinformatics, information science, or computer science. So one might expect some diversity in knowledge of data analytics among the participants. We noticed, during the experimental setup, that random people joined and tried to fulfil the questionnaire. The majority of those people gave us feedback in that they had absolutely no idea what they were doing and were quitting the questionnaire after one or two questions. To keep our results relevant and to reduce noise, we tried to include participants in this study through a classroom session that do have more knowledge about fields related to data analytics, but have less domain knowledge as life scientists have.

The questionnaire contained 12 questions, including two test questions, to get familiar with the visualization platform. The questions that we designed cover all five categories of Keim[1] and are questions relevant to the analysis of HCS data, for example, "what is the number of data points" or "what is the plate name of this outlier." The 12 questions also include two questions covering the comparison of the output of algorithms. When we would have implemented questions that were always immediately clear, one would not perform better using interactive visualizations and there would not be a true incentive to use interactive visualizations. We believe that questions or challenges regarding data analysis can be efficiently supported by the right visualizations. Interactive visualizations can add extra value by speeding up the analysis process because of its flexible nature and decrease the user's cognitive load because of the lower burden of recalling visualization objects.

There are certain problems associated with data visualization in general. Visualizations have their limitations when dealing with large data sets, since occlusion of (parts of) the data, disorientation, and misinterpretation can occur.[11] The visualization of large data sets can also lead to the problem of overplotting, producing a visualization in where individual data points are coerced into a single solid object. Minor differences between data points are not observable in these situations, and only the trend (*e.g.*, linearity) of the data can be derived from these visualizations. With interactive data visualization, it is still possible to view these minor relationships between data points, as a result of a zoom event. The visualization of large data sets also leads to the problem that visualizations take a long time to render. Through the survey, we found an expected result that the visualization of a random sample of a large data set is informative enough to observe the trend (*e.g.*, the distribution) of the complete data set. With a sample, the rendering time of the visualization can be reduced, improving the responsiveness of the visualization process. The cognitive load can thus be reduced, when the waiting time is decreased between the inspection of different data visualizations. This leads to the next problem in data visualization: the comparison of the output of algorithms. Usually when a researcher wants to visually compare the output of multiple algorithms, there is a delay between different visualizations that represent the output of different algorithms. The data need to be manipulated by another algorithm and a new visualization needs to be created for each comparison. In this project, a platform was designed that optimizes the comparison of visualizations. Because the delay between viewing different visualizations is minimized, different visualizations can be compared faster than using regular visualization platforms. Keim[1] reported five categories of interactivity. We propose the addition of a sixth category: the interactive comparison of the output of algorithms (data manipulations). A possible side effect of this interactive method is the bias that may be introduced as researchers will set out to "find" the data algorithm that makes their data look best. At the same time, we also stress that the proposed sixth category could be part of the other methods described by Keim. For example, when linkage is used, two visualizations are interactively connected; thus they can be compared. A side note here is that, there should be a possibility to compare them side by side instead of swiping through the various visualizations that one would compare.

ACKNOWLEDGMENTS

We thank the Department of Computer Science of the Utrecht University for supporting us setting up the questionnaire and validating the results of the article. We thank the Bioinformatics department of the University of Applied Sciences Leiden for their support.

DISCLOSURE STATEMENT

W.A.O. and D.A.E. are both cofounders of Core Life Analytics B.V. For all other authors, no competing financial interests exist.

REFERENCES

1. Keim DA: Information visualization and visual data mining. *IEEE Trans Vis Comp Graph* 2002;8:1–8.
2. Marr B: Big data: 20 Mind-boggling facts everyone must read. *Forbes Magazine* 2015.

3. Singh S, Carpenter AE, Genovesio A, *et al.*: Increasing the content of high-content screening: an overview. *J Biomol Screen* 2014;19:640–650.
4. Marx V: Biology: The big challenges of big data. *Nature* 2013;498:255–260.
5. Greene CS, Tan J, Ung M, Moore JH, Cheng C: Big data bioinformatics. *J Cell Physiol* 2014;229:1896–1900.
6. Omta WA, van Heesbeen RG, Pagliero RJ, *et al.*: HC StratoMineR: A web-based tool for the rapid analysis of high-content datasets. *ASSAY Drug Dev Technol* 2016;14:439–452.
7. Wickham H: A layered grammar of graphics. *J Comput Graph Stat* 2010;19:3–28.
8. Tufte ER: The visual display of quantitative information. *J Healthc Qual* 1985;7 3–15.
9. Friendly M: A brief history of data visualization. In: *Springer Handbooks of Computational Statistics Handbook of Data Visualization* (Chen C, Härdle W, Unwin A, eds.), pp. 15–56. Springer, Berlin, 2008.
10. Menger V, Spruit MR, Hagoort K, *et al.*: Transitioning to a data driven mental health practice: Collaborative expert sessions for knowledge and hypothesis finding. *Comput Math Methods Med* 2016; DOI: 10.1155/2016/9089321.
11. Shneiderman B: Inventing discovery tools: Combining information visualization with data mining. *Inf Vis* 2002;1:5–12.
12. Costello AB, Osborne JW: Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Pract Assess Res Eval* 2005;10:1–9.
13. Fayyad U, Piatetsky-Shapiro G, Smyth P: The KDD process for extracting useful knowledge from volumes of data. *Commun ACM* 1996;39:27–34.
14. Healey CG, Booth KS, Enns JT: Visualizing real-time multivariate data using preattentive processing. *ACM Trans Model Comp Simul* 1995;5:190–221.
15. Shneiderman B. *The Future of Interactive Systems and the Emergence of Direct Manipulation*. University of Maryland, College Park, MD, 1982. Print.
16. Javed W, Elmqvist N, Yi JS: Direct manipulation through surrogate objects. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems ACM*, Vancouver, Canada, pp. 627–636, 2011.
17. Chapman P, Clinton J, Kerber R, *et al.*: *CRISP-DM 1.0 Step-by-Step Data Mining Guide*. SPSS Inc., 2000.
18. Birmingham A, Selfors LM, Forster T, *et al.*: Statistical methods for analysis of high-throughput RNA interference screens. *Nat Methods* 2009;6:569–575.
19. Wickelgren WA: Speed-accuracy tradeoff and information processing dynamics. *Acta Psychol* 1977;41:67–85.

Address correspondence to:
Wienand A. Omta, MSc
Department of Cell Biology
Center for Molecular Medicine
UMC Utrecht
Heidelberglaan 100, Room H02.313
Utrecht
3584 CX
Netherlands

E-mail: wienand@corelifeanalytics.com

Matthieu J.S. Brinkhuis, PhD
Department of Information and Computing Sciences
Utrecht University
Princetonplein 5, Room 574
Utrecht
3584 CC
Netherlands

E-mail: m.j.s.brinkhuis@uu.nl

Abbreviations Used

DM = direct manipulation
HCS = high content screening
M = mean
MySQL = my structured query language
SD = standard deviation
VDM = visual data mining