Information Visualization
http://ivi.sagepub.com/

A comparison of benchmark task and insight evaluation methods for information visualization

Chris North, Purvi Saraiya and Karen Duca

Information Visualization 2011 10: 162 originally published online 8 August 2011
DOI: 10.1177/1473871611415989

The online version of this article can be found at:
http://ivi.sagepub.com/content/10/3/162

Published by:
SAGE
http://www.sagepublications.com

Info Vis

Research Paper

A comparison of benchmark task and insight evaluation methods for information visualization

Information Visualization
10(3) 162–181

DOI: 10.1177/1473871611415989
ivi.sagepub.com
SAGE

Chris North[1], Purvi Saraiya[1] and Karen Duca[2]

Abstract

This study compares two different empirical research methods for evaluating information visualizations: the traditional benchmark-task method and the insight method. The methods are compared using criteria such as the conclusions about the visualization designs provided by each method, the time participants spent during the study, the time and effort required to analyse the resulting empirical data, and the effect of individual differences between participants on the results. The study compares three graph visualization alternatives that associate bioinformatics microarray time series data to pathway graph vertices in order to investigate the effect of different visual grouping structures in visualization designs that integrate multiple data types. It is confirmed that visual grouping should match task structure, but interactive grouping proves to be a well-rounded alternative. Overall, the results validate the insight method's ability to confirm results of the task method, but also show advantages of the insight method to illuminate additional types of tasks. Efficiency and insight frequently correlate, but important distinctions are found.

Categories: H.5.2

[Information Interfaces and Presentation]: User Interfaces – evaluation/methodology.

Keywords

empirical evaluation, graph visualization, time series data analysis, benchmark tasks, insight

Introduction

Visualization tools are often evaluated in controlled studies that use benchmark tasks.[1,2] Participants are usually given a variety of predefined tasks to perform on pre-selected data during the course of the study. The performance time and accuracy of the participants' responses for the tasks are recorded and later analysed to evaluate the visualization tools. However, such studies often fail to represent the real-world data analysis scenario, which is less guided and much more in depth.[3]

An attempt to capture the real world exploratory data analysis scenario in a short-term study used an insight-based method.[4] The method used an unguided protocol requiring the participants to think aloud about the insights they glean from the data. The visualization tools were then analysed based on the quantifiable characteristics of the insights that can be measured uniformly across participants. Thus, in contrast to the controlled studies that use benchmark tasks, the insight method does not use predefined tasks and instead treats tasks as dependent variables in the experiment.

While the insight method appeared useful, there are open questions about how the method compares with the traditional benchmark task method and whether the method should be used instead of the benchmark task method to provide meaningful statistical analyses between visualizations or as a complementary approach. Thus, the goal of this paper is to compare both methods: the task-based and insight-based methods. Such studies to compare empirical research methods are more common to the field of usability engineering, but less frequent in the information visualization domain. Thus, a broader research goal is to

[1]Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA.
[2]Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA.

Corresponding author:
Chris North, Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA
Email: north@vt.edu; http://infovis.cs.vt.edu/

investigate if such a comparison between methods can be done for information visualization.

A secondary goal is to evaluate the design choices for visualizations that integrate two types of data; in particular, visualization of graphs with associated time series data for the bioinformatics domain. In bioinformatics, graphs or 'pathways' with a node-link representation are typically used to represent interactions between biomolecules (genes, proteins, etc.). Multidimensional time series data from high-throughput experiments such as gene expression microarrays[5] are often analysed in context of these biological pathway graphs. The graphs provide important biological context to otherwise raw time series numerical data analysis.[6]

Figure 1 shows the overlay of time series data in the context of a graph. Each vertex in the graph corresponds to a row in the time series data set, and each experiment treatment or time point is a column. Three common visualization methods used by current bioinformatics software tools to overlay multidimensional time series data on graphs[6] include (a) overlaying data on graph vertices for one time point at a time (Figure 2) by encoding a visual property (e.g. colour) of the vertex and using sliders or similar interaction to animate the graph to other time points; (b) data from all the time points can be overlaid simultaneously by using complex node glyphs (Figure 3); or (c) small multiples can be used to simultaneously display a miniature graph for each time point (Figure 4). Although many user studies have been conducted to evaluate graph visualization, few studies have evaluated alternatives for graphs with associated multidimensional data.

Literature review

The literature related to the design space for visualizing graphs with associated time series is summarized by Saraiya et al.[7] While that study examined the use of multiple views including parallel-coordinates plots, this paper focuses on the primary graph representation itself.

Comparison of information visualization studies

Different types of studies have been used to evaluate visualization tools, as summarized by various authors.[1,2,4,8] The shortcomings of typical controlled experiments based on benchmark tasks has sparked discussion about the need to develop new evaluation methods for visualization tools that better represent real-world data analysis scenarios and provide better feedback about the usability of the data representation method.[2,3,4,9] To fill this need, the biennial BELIV (beyond time and errors: novel evaluation methods for information visualization) workshop focuses on research of novel evaluation methods in information visualization, and a variety of new methods can be found in its proceedings.[10] Methods related to insight-based evaluation are one such growing class of new methods.[4,11–13] Rigorous comparison of methods is still needed to guide evaluators.

The literature for comparisons of empirical research methods used to evaluate information visualization tools is sparse, and mostly anecdotal. Golovchinsky et al.[14] discuss general guidelines for better tasks and methods to evaluate visualizations. Chen et al.[15] present recommendations for more consistent and comparable user studies based on their meta-analysis. Kosara et al.[16] discuss user studies for information visualization and the lessons learned from these studies and how these were used to design more effective visualization tools and evaluation studies. House et al.[17] presented a panel discussion summarizing research for visualization evaluation using human subjects, including suggestions and guidelines for conducting such studies based on their experiences. Tory et al.[18] suggest expert reviews as an alternative in certain contexts where designing and conducting user studies can be difficult. Carpendale[8] argues for the need to apply a greater variety of evaluation methodologies in information

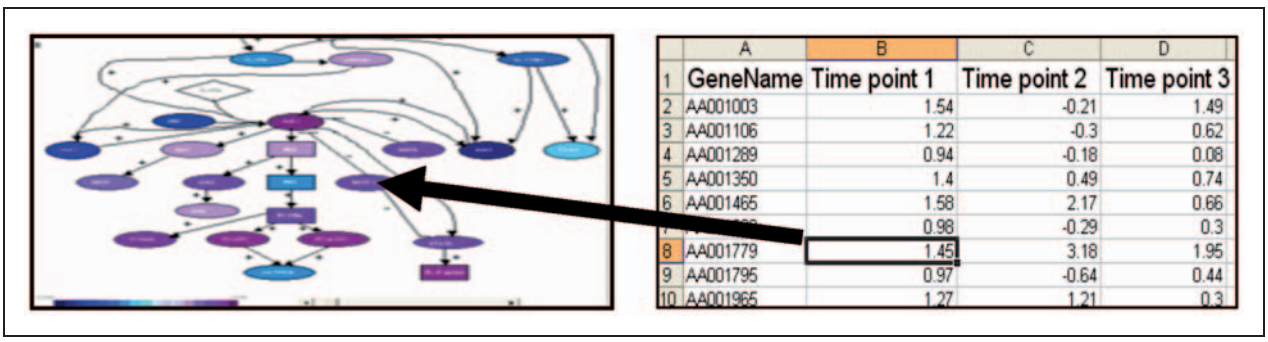

Figure 1. An example of linking time series data to graphs.

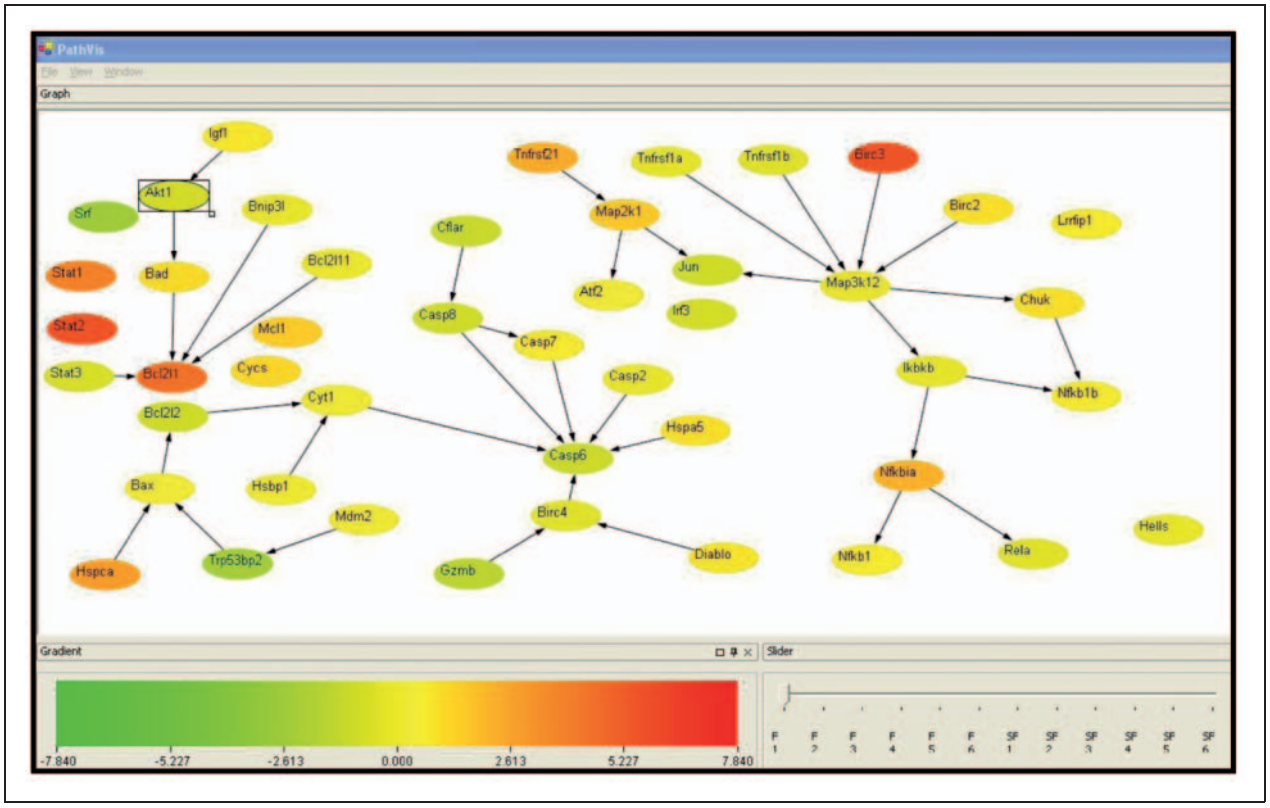

Figure 2. Single time point – overlay a single time point on all graph vertices, using vertex colour. A slider is used to navigate between all 12 time points.

visualization, especially methods that enable increased qualitative analysis.

Comparisons of studies in usability engineering

Several studies have been conducted to analyse and compare methods typically used for evaluating user interfaces. Examples include Steves et al.,[19] who compared usage-based evaluation techniques and inspection methods for groupware systems; Brush et al.,[20] who compared the effectiveness of local versus remote usability studies; Bekker et al.,[21] who compared two methods for evaluating children's computer games; Jeffries et al.,[22] who compared usability testing methods with multiple participants to heuristic evaluation; Hartson et al.,[23] who provided a list of criteria that can be used to compare usability evaluation methods; John et al.,[24] who reported a detailed case study of six usability methods that evaluate each method's usability error predictive power to actual user tests; and Doubleday et al.,[25] who compared different usability testing methods for information retrieval tasks.

Although studies have been conducted to evaluate usability methods that analyse user interfaces with respect to each other, studies to evaluate empirical research methods for evaluating visualization tools are rare. Most of the usability methods are compared based on the number of usability errors found, severity of these errors, and participants' and facilitators' experience in the study. Since the dependent variables for the usability methods are usually the same (usability errors), such direct comparisons between the evaluation methods are possible. However, the dependent variables for the benchmark task-based method (performance time, accuracy) and the insight method (data insights) are different. Also, the evaluation for visualization tools investigates a wider range of options (e.g. data representation method, interaction mechanisms used, etc.) as compared with typical user interface evaluation. Hence, higher level measures, such as task taxonomies, conclusions about the visualizations, time spent by the participants in the study, effort spent to analyse the resulting empirical data, and so on, need to be used for meaningful comparisons between these two evaluation methods.

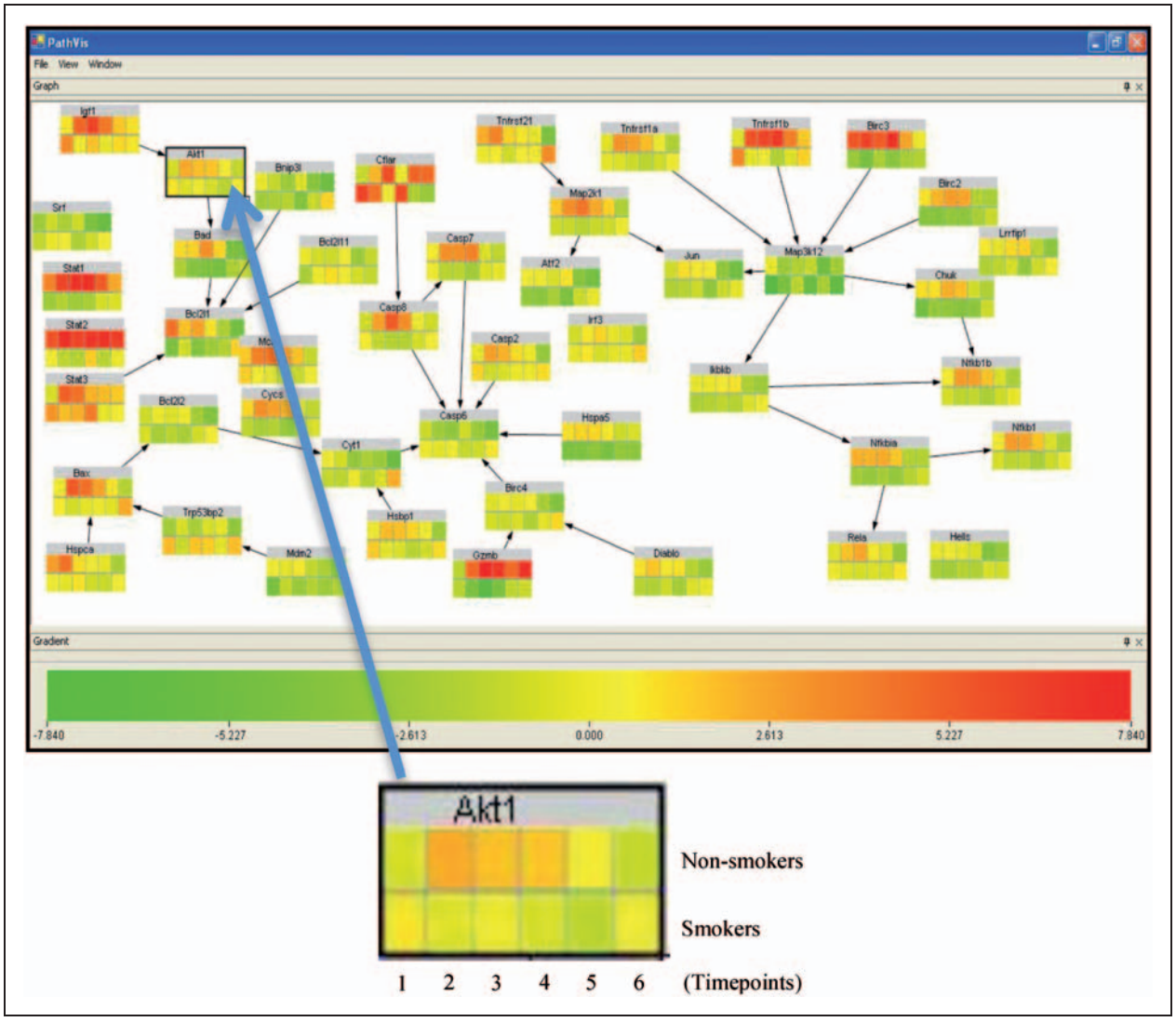

Figure 3. Multiple time points – overlay all 12 time points as a miniature heatmap onto each graph vertex.

Experiment design

Design and motivation

The aim of this study is to analyse and compare the results of two empirical evaluation methods, using three different visualization alternatives that support the analysis of time series data within the context of network graphs. Conceptually, a 2×3 between-subjects design examines the following two independent variables:

1. two empirical evaluation methods: benchmark tasks method and the insight method; and
2. three graph visualization design alternatives.

Technically, this is executed in the form of a pair of matched and synchronized studies, one study for each of the two evaluation methods (benchmark tasks method or insight method), appropriately drawing on a shared participant pool with random assignment to each study. The goal is to compare the experiences of the two studies, with particular attention to comparing the results generated by each study.

The *benchmark task evaluation method* tests users on a predefined set of benchmark tasks and measures their performance time and accuracy. In general, this method seeks to identify how well the tested visualization supports a specific set of analytical tasks. Conversely, the *insight evaluation method* employs an open-ended protocol in which users independently

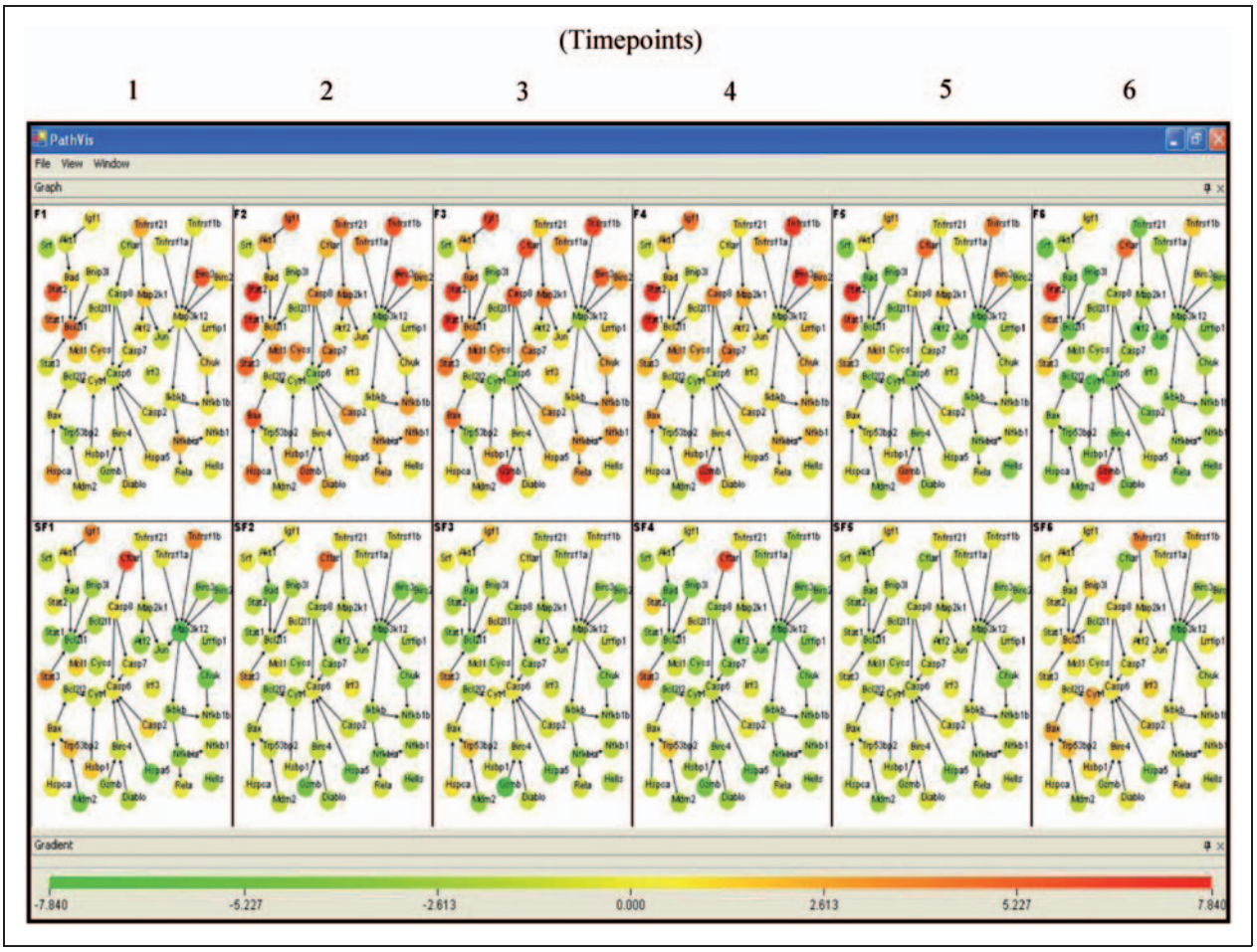

Figure 4. Multiple graphs – multiple (12) small graphs, each displays one of the time points using vertex colour. Top row is non-smokers, bottom row is smokers.

identify and report insights they find in the data, which are then categorized and counted. This method seeks to identify what analytical insights the tested visualization causes users to recognize. A key question is how the reported insights will relate to the benchmark tasks, and how their corresponding performance rates will compare.

Data and scenario

The biologists we collaborated with conducted a mouse gene expression microarray experiment to analyse the impacts of tobacco smoking on immune response to flu infection. The actual data was 45,001 rows (genes) $\times$ 72 columns (time points and conditions). The biological significance of the data and the actual analysis process for this data by biologists are presented in Gualano et al,[32] and Saraiya et al.[29]

A directed graph, having 46 vertices (genes) and 36 edges (gene interactions) representing an actual immune response pathway, was linked to a time series data set representing gene expression for 12 time points (Table 1). Thus, the participants in the experiment were working with a small subset of the actual data. However, the graph size was based on the typical size of the pathways used by the biologists at the time of the experiment, and was corroborated by the average size of publicly available pathways found in the STKE Database of Cell Signalling.[26]

This is an increasingly common data analysis scenario in biology research.[6] Essentially, gene expression experiments enable biologists to determine the activity level of individual genes in an organism by measuring their RNA or protein outputs. Thus, biologists can begin to understand the complex workings of a biological system by subjecting it to some external stimulus (such as infection or tobacco) and observing how it responds at the molecular level over time in comparison with the control condition. Because the biological systems are extremely complex and the data contain a

Table 1. Data used for the study

Data type	Description
Graph	A directed graph having 46 vertices and 36 edges. Each node had an out degree of 0 to 3
Time series data	Gene expression values for 12 time points for each vertex. Of these, six time points measured expression values for flu infection for non-smokers, and the remaining six time points corresponded to flu infection for smokers

large amount of noise, it is increasingly important to analyse the resulting time series data with relation to other biological information.

Such other biological information frequently occurs in the form of network graphs. For example, pathways represent known or hypothesized cellular workflows, similar to computational data flow diagrams. Interaction networks represent known compatibilities between biomolecules, such as chemical reactions. Gene ontologies represent hierarchical categorical relationships, such as gene functions or spatial compartments of proteins within the cellular structure.

Biologists' analysis task is to make sense of the observed results within the context of the other known information, and thus hopefully contribute new information such as recognizing a potential new pathway link or a new gene function for a previously unknown gene and thereby updating the known networks. Hence, the overall analytical task is often exploratory in nature. In general, biologists want to discover how patterns in the time series relate to patterns in the graph. For example, in the case of pathways, it is expected that temporal relationships will cascade and follow the directed edges in the pathway graph. If not, this might indicate that some other biological process is intervening that needs to be uncovered.

Scalability is an issue in biology data sets. Gene expression measurement methods are continuously improving, enabling the capture of tens of thousands of data points. Computationally generated relationship networks can result in massive 'hair balls'. Biologists take great care to validate, reduce, and curate data to more manageable sizes, for example by conducting statistical significance tests to generate shorter 'gene lists' that contain only genes that were affected by the stimuli with high probability. While we currently test modestly sized data, this would normally occur within the context of a broader analysis process that honed the data down to such genes of interest. Furthermore, actual visualization applications for biologists will need to combine the basic representation techniques studied here with other information visualization methods to support much greater scalability (see, for example, Barsky et al).[28]

Graph visualization design alternatives

Three graph visualization design alternatives that overlay the time series data onto the graph nodes were used in the study. The visualizations are drawn from Saraiya et al.[7] and are based on a simple *taxonomy of visual grouping*: (a) time series by interaction, (b) time series within graph, and (c) graph within time series. The visual encoding of the time series data on each graph vertex was based on a common colour scheme used in bioinformatics, i.e. the colour scale from yellow to green was used to display negative data values, and yellow to red was used to display positive data values.[7] The colour encoding was preserved on all three visualizations so as to eliminate encoding as a factor and focus the comparison on the visual grouping factor. In each visualization, mousing over a vertex displays in a tooltip the name of the vertex, the time point position, and the numerical time series data value. The visualization alternatives are:

- single time point (1 Tpt): This visualization overlays values for one time point on a vertex at a time (Figure 2). A slider lets users iterate over all the time points in the data. This is the most common method used in existing bioinformatics software, such as Cytoscape.[27]
- multiple time points (M Tpts): This visualization overlays data from all the time points onto each vertex using a miniature heat map (Figure 3). This method is occasionally used in biology presentations and technical reports to display data results.
- multiple graphs (M Graphs): This visualization uses the small-multiples technique to display a miniature graph for each of the time points in the data (Figure 4). This method is used in some more recent tools, such as Cerebral.[28]

As a proposed design guideline, our *visual grouping hypothesis* states that each visualization alternative will perform best with tasks that have a query grouping

Table 2. Benchmark tasks for the task-based method

No.	Task
T1	Which of the following genes shows a positive value for all non-smoking time points but negative for all smoking time points?
T2	What is the overall expression pattern for non-smoking time points vs. smoking time points?
T3	Which of the following genes is negative for all 12 time points?
T4	Which of the following time points has the maximum number of positive genes?
T5	Which of the following time points has the maximum number of negative genes?
T6	At which of the following time points, for both conditions, do most genes change their values from previous time points?
T7	How many genes are between Map3k12 and Rela?

structure that match the visualization's visual grouping structure. For example, understanding how the graph changes over time will best match multiple graphs (graph within time series), but understanding how vertices' time series pattern changes over the graph will best match multiple time points (time series within graph).

Experiment protocol

A total of 60 participants, 10 for each visualization alternative for each evaluation method, participated in the study. As the data had a biological background, all the participants in the study were undergraduate students who had taken courses in molecular biology. Hence, the participants were all familiar with the basic concepts of gene expression data and experiments, although they were not familiar with this specific data set.

Before beginning, the participants were given a brief introduction to the visualization alternative that they were randomly assigned and the data background used in the study. Then, the protocols were different depending on the randomly assigned evaluation method.

Benchmark task method protocol. Participants were required to perform the seven tasks listed in Table 2 in order. All the tasks were multiple-choice questions, with five possible choices. The tasks were based on the observed analysis tasks of the biologists who designed the biology experiment and analysed the actual data.[29] The tasks involve a variety of different types of questions about the data, focus on different aspects of the data, and vary in complexity. Owing to the scoring requirements of the benchmark method,[3] these tasks represent low-level components[30] of the biologists' high-level analyses, and thus necessarily simplify the actual analytic process. Time and accuracy were automatically measured for each task (Table 3). At the end, participants were given the opportunity to verbally comment on any feedback about the visualization.

Table 3. Dependent variables for the task-based method

Time to answer each question
Accuracy of answers
Overall time spent in the study
Feedback about the visualization alternative

Table 4. Dependent variables for the insight method

Data insights reported
Time at which each insight was reported
Overall time spent in the study
Feedback about the visualization alternative

Insight method protocol. Participants were asked to analyse the data in a think-aloud fashion, verbally reporting any findings they thought were interesting, until they felt that they had learned all they could from the data. Participants were asked to distinctly report each finding, which we then recorded as individual insight occurrences. The experimenter sat next to the participants during the study, silently observing the participants' data analysis process and also recording the data insights and the times at which these were made since beginning the study (Table 4). Although participants tended to comment on the tools

throughout, they were also given an opportunity at the end to provide feedback about the visualization.

Results

In this section, we report the results of each study separately. In the next section we compare the results between the studies.

Benchmark task method results

Overall performance. On performing analysis of variance (ANOVA) of the task-based study, we found that there were no significant differences between the participants on the total time spent in the study or overall differences on the accuracy for the tasks for all the three visualization alternatives (Figure 5). However, the participants using single time point visualization were somewhat more accurate ($p = 0.06$) than multiple time point visualization.

Performance for individual tasks. Significant results from ANOVA tests on tasks between the three visualization alternatives for time and accuracy (Figure 6) are summarized in Table 5. Although tasks 4 and 5 were equivalent, task 5 required more careful analysis than task 4, as the time point at which most nodes were positive was more obvious than the time point at which there were most negative nodes. The results clearly indicate that differences in the visual representation of the data significantly affected users' ability to perform specific tasks. In particular, the visual grouping structure of the visualization helped or hindered users in performing certain tasks. It seems that visual grouping according to the task structure increases the perceptual salience of the task data pattern.

In general, the single time point visualization resulted in the most consistently good performance, whereas the other two varied more depending on the task. Multiple time points did not perform very well. It performed well on task 1, which involves examining the time series pattern for a single gene, because it visually represents a node's time series pattern

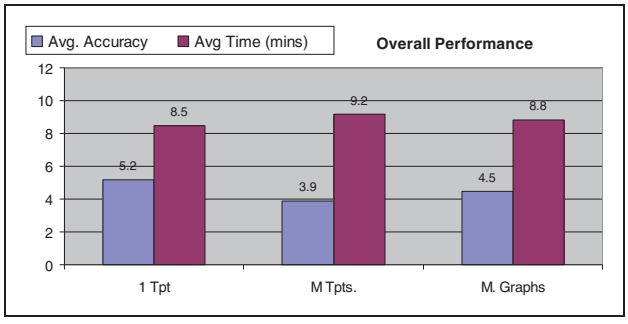

Figure 5. Average time that participants spent in the benchmark task study (minutes), and the average number of correct responses (out of seven tasks), for all three visualization types.

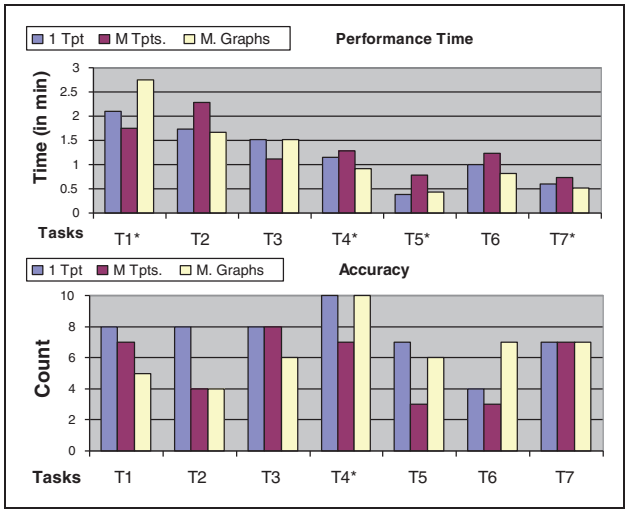

Figure 6. Average time (minutes), and total count of correct responses (out of 10 participants) for each task, for all three visualization types.
*indicates significant differences.

Table 5. Summary of comparing the task-based performance of the three visualization alternatives for each benchmark task

Tasks	1 Tpt	M Tpts	M Graphs
T1	–	–	Slowest ($p = 0.04$)
T2	–	–	–
T3	–	–	–
T4	–	Least accurate ($p = 0.03$)	Faster than M Tpts ($p = 0.01$)
T5	–	Slowest ($p = 0.04$)	–
T6	–	–	Weakly faster ($p = 0.1$) than M Tpts
T7	–	–	Faster than M Tpts ($p = 0.04$)

succinctly as a visual group. Task 3 has a similar trend but not significantly so. Accordingly, multiple graphs performed poorly on task 1. It had an advantage in tasks 4, 5, and 6, which involve finding time points with interesting overall expression patterns. Task 6 in particular requires comparing adjacent time points. This is also expected according to the visual grouping hypothesis, because multiple graphs visually groups first by time point, which matches the task structure. Unexpectedly, multiple graphs also performed well on task 7, a graph topology task.

Task-based conclusions of visualization alternatives. The analysis of individual task performances and summary in Table 5 leads to the conclusions about the visualization alternatives summarized in Table 6. Most of the conclusions about the single time point and multiple time points visualization alternatives are similar to the findings from a previous task-based study.[7] This confirms that the single time point is advantageous over multiple time points for tasks that involve analysing or searching for individual time points of interest.

Insight method results

Overall performance. On performing ANOVA of the insight-based study, we found that participants using the multiple graphs visualization spent significantly less time in the study than other participants ($p = 0.02$). The single time point visualization produced a greater number of distinct data insights than multiple time points ($p = 0.07$), and multiple time points produced a greater number of insights than multiple graphs ($p = 0.06$). These results are summarized in Figure 7. We eliminated duplicate insights so that each insight is distinct within a participant. However, the insights may be repeated across participants when more than one participant reported the same data insight. Normalizing the results to compute the average rate of *insights per minute* gives an indication of insight efficiency. Single time point produced the fastest insight rate (0.9 insights/min) compared with the other two (0.7 insights/min).

Interestingly, in contrast to the task method, with the insight method it is better when participants spend more time. As the protocol was open-ended, this meant that they believed they could continue to make more findings, and in fact they did make more findings because the number of insights correlates with time spent. Single time point users spent the most time and gained the most insights. Since the single time point visualization involves the most interactivity (the time point slider), this might confirm the hypothesis that interactivity plays in important role in engaging users in data analysis,[4] thus generating more insight. It is possible that the interactivity enables flexibility, which encourages users to explore many possibilities and gain more different types of insight. Given that the instructions were to continue until they learned all that could be gained from the data, it is likely that they continued because they recognized the *opportunity* for more insight.

For multiple graphs, quitting early because one believes that no more learning can be done does not seem like a valuable quality for a visualization, especially because higher insight totals from the other visualizations indicate that there was more information

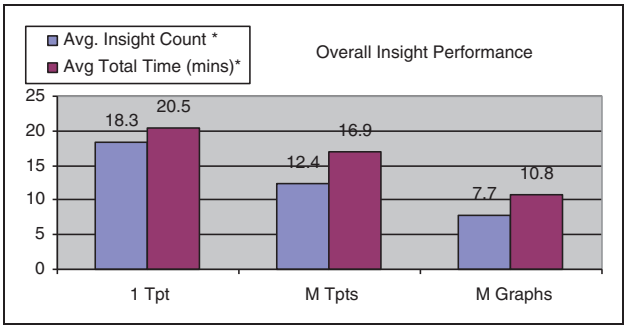

Figure 7. Average amount of time (minutes) participants spent in the insight study, and average count of data insights reported per participant, for all three visualization types. *indicates significant differences.

Table 6. Conclusions about visualization alternatives from the task-based study

1 Tpt	M Tpts	M Graphs
Consistent good performance for all tasks	Fast for single gene analysis (T1)	Slowest for single gene analysis (T1)
Fast for single gene analysis (T1)	Slowest and least accurate for time point analysis (T4, T5, T6)	Fast and accurate for time point analysis (T4, T5, T6)
Fast and accurate for time point analysis (T4, T5)	Slow for graph topology (T7)	Fast for graph topology (T7)

available to learn from the data. Why did they stop? Users simply believed that the multiple graphs visualization could not help them anymore, and this probably indicates its limitations (or perceived limitations) to certain types of insights. No more interesting patterns were clearly recognizable to the users. A critical problem with the multiple graphs visualization is that it does not produce many insights related to the expression patterns of individual genes, which are apparently an important class of numerous insights in this domain scenario, as seen in the next section.

Another potential explanation for the variation in total insight is that the multiple graphs visualization caused users to think that some insights found were simply not worth verbally reporting. But this explanation seems least likely owing to the correlation between total insight and total time. It was not the case that users were sitting quietly while exploring and not reporting insights as they found them; rather, they actually stopped and claimed to be done.

Performance based on insight category. Abstracting individual insights into categories provides a useful way to analyse how visualization alternatives affect the type of insights users gain.[4] Upon analysing the participants' insights using an open coding strategy, we found that all of them could be grouped into seven distinct categories, described below, based on the aspects of the data that the insights involve. Each insight belongs to only one of these categories. Figure 8 summarizes the participants' performance based on these insight categories for all three visualization alternatives. The seven categories are:

1. Gene expression: Most frequent data insights reported time series expression patterns for a single gene, e.g. 'Gene *Gzmb* displays positive values for all the non-smoking time points except the first time point, but is negative for all the smoking time points'.
2. Topology: Some of the insights involved only the graph topology. This did not include any information about the associated time series data, e.g. 'The *map3k12*, *casp6*, and *bcl2ll* genes seem to be major focal points in the graph as they have a lot of arrows pointing towards them.' None of the participants using M Graphs reported such insights.
3. Topology and expression: Some of the insights reported by the participants investigated gene expression based on graph topology or effects of genes on each other connected directly or indirectly through other genes, e.g. 'All the genes towards the outside, i.e. *Trnf2*, *birc3*, etc. are more positive for almost all the time points than the inside ones that they are supposed to affect.'

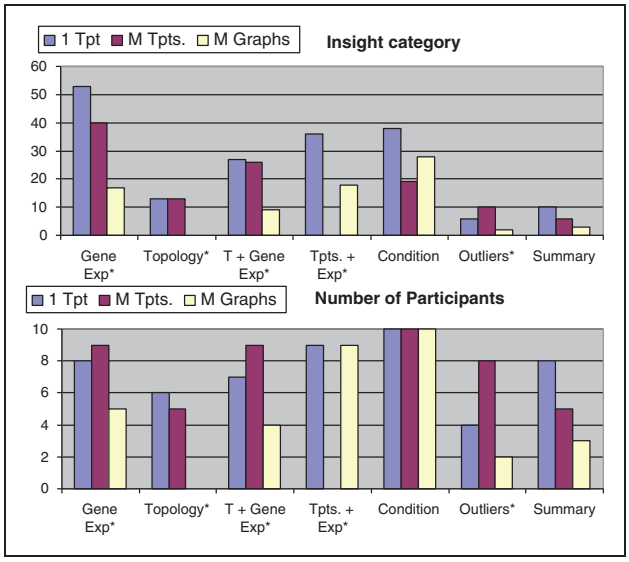

Figure 8. Total number of insights, and the number of participants (out of 10) who reported these, for each insight category.
*indicates significant differences.

4. Time point analysis: Some participants reported insights that investigated overall graph expression at a particular time point, e.g. 'A lot of genes are negatively expressed at time point 5 for smokers as compared to all other time points'. None of the participants using multiple time points reported such insights.
5. Experiment conditions: All participants in the study evaluated the differences in the gene expression between smokers and non-smokers, which appropriately reflects the goals of the biology experiment from which the data set originated, e.g. 'Overall, non-smokers have more positively expressed genes than smokers.'
6. Outliers: Some participants identified a few genes that displayed different expression patterns than other genes in the graph, e.g. '*Stat1* gene is different than other genes, as it upregulates with time for non-smokers, whereas most other genes downregulate.'
7. Summary: Some participants tried to summarize their high-level findings about the data or suggested future research based on their data analysis. These insights are most similar to the *hypothesis* insight characteristic that was ranked very important in the original insight study,[4] e.g. 'Smokers don't have many highly expressed genes, and a lot of them may reduce the gene expression of the subsequent genes. This may eventually lead to less expression for the overall immune system against the flu for smokers.'

Table 7. Summary of comparing the number of insights generated by the three visualization alternatives for each insight category

Category	1 Tpt	M Tpts	M Graphs
Gene expression	–	–	Weak least ($p = 0.1$)
Topology	–	–	Least ($p = 0.03$)
Topology and expression	–	–	Least ($p = 0.05$)
Time point analysis	Most ($p = 0.03$)	Least ($p = 0.01$)	–
Condition	–	–	–
Outliers	–	More than M Graphs ($p = 0.01$)	–
Summary	More than M Graphs ($p = 0.02$)	–	–

Table 7 lists the results from the ANOVA between participants on the number of distinct insights for each category reported by each participant using each visualization alternative. These results clearly indicate that differences in the visual representation of the data caused users to gain significantly different kinds of insight, even though they had complete freedom to explore indefinitely. In particular, the visual grouping structure of the visualization encouraged or discouraged users from gaining certain types of insight. It seems that visual grouping according to the task structure of a certain insight type increases the perceptual salience of the insight data pattern, making it more likely to be noticed and reported by the user as an insight.

In general, single time points offered the most consistently good insight production across all seven insight categories and across all users. It was advantageous over the other visualizations for analysing specific time points, owing to its ability to focus on each time point, and generating important summary insights, perhaps indicative of its interactive engagement.

The Multiple time points visualization was advantageous for spotting outlier genes, but was extremely poor at individual time point analysis, generating a total of zero such insights. This conforms to the visual grouping hypothesis because the visual grouping of a single gene's time series pattern makes it salient to visually compare with all other genes, but makes it difficult to visually focus on one time point across the graph. Given its low insight rate, multiple time points produced an unusually large number of topology and expression insights.

The multiple graphs visualization was disadvantageous for gene expression, topology, and relating gene expression to topology. The fact that multiple graphs had less overall time and insight led naturally to it also having fewer insights in many of the insight categories. One possibility could be to normalize the results in terms of total time, thus focusing on insight rates for each category. In that case, given multiple graphs' overall low total insight amount, it was particularly efficient at generating insights about the conditions.

Insight conclusions about visualization alternatives. Participants' performance on the insight categories and the summary of data analysis results in Table 7 lead to the conclusions about the visualization alternatives listed in Table 8. Measurement of other characteristics of insights[4] was also attempted, but did not produce useful results in this study. For example, when quantifying the domain value of individual insights, the domain experts found only the 'summary' category of insights to be more valuable than the others, probably owing to the relatively novice experience level of the subject pool. Also, for correctness, no obviously incorrect insights were found.

Comparison between methods

Total time spent

On performing overall ANOVA, participants in the insight method spent significantly more total time in the study than those in the task-based method ($p = 0.01$) (Figure 9). Participants using single time point and multiple time points visualizations spent significantly more time ($p = 0.01$) in the insight method than in the task method. The total time measure only includes the time specific to either the task or insight portion of the experiments, not the introduction and feedback portions that were common to both study methods.

The task method produced consistent study lengths across all three visualization conditions, whereas the insight method produced a high variability in study length depending on condition, with single time point users taking nearly twice as long as multiple graphs users. In terms of administration, this variability can make it more difficult to schedule and conduct

Table 8. Conclusions about the visualization alternatives from the insight-based study

1 Tpt	M Tpts	M Graphs
Encouraged users to work longer and find more total insights	More insights than M Graphs on single genes, topology, expression, and *outlier nodes*	Shortest total time and fewest total insights
Consistent good performance for all insight categories	No analysis of time points	Fewest insights about single genes, and expression topology
Most insights for time point analysis	Emphasized insights relating expression to topology	No analysis of graph topology
More insights than M Graphs on single genes, topology, expression, and *summary findings*		Fewer insights on outliers and summary
		Emphasized insights comparing conditions and time points

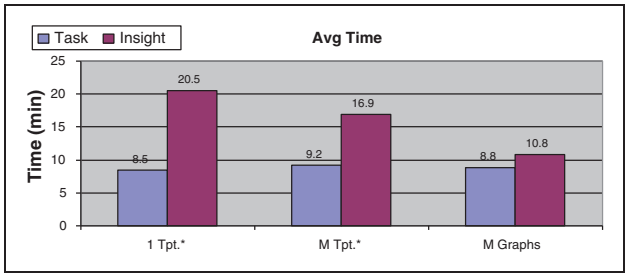

Figure 9. Average time participants spent in the study for each visualization type, for task and insight methods. *indicates significant performance differences.

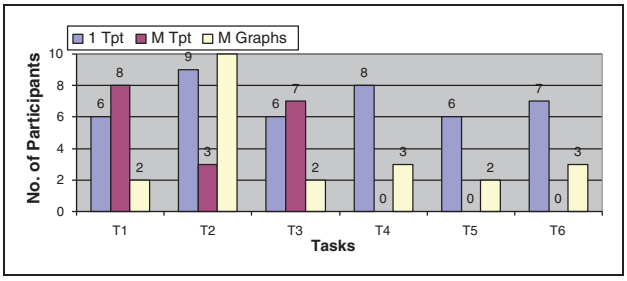

Figure 10. Number of participants for each visualization who reported insights equivalent to the tasks in the task-based method.

the study. It also means that enforcing a fixed time on an insight study could greatly bias the results. However, through this open-ended length, the insight method uncovers a new type of metric related to user engagement, or perceived value of the visualization in terms of the expected amount of available insight. The fact that single time point users took longer than the others in insight, yet took the same amount of time as the others in the task method, confirms the insight rate finding. Its overall task speed is about the same, yet it offers more opportunity for total learning.

Performance comparison on tasks and categories

Benchmark tasks with equivalent insights. It is possible to directly relate tasks and insights. Some participants in the insight method reported insights very similar to the benchmark tasks that were used in the task-based method. For example, similar to task 4 in the task method, some participants in the insight method reported that time point 4 for non-smokers has the maximum number of positively expressed genes. Figure 10 summarizes the number of participants that made insights comparable to a particular benchmark task. Since task 7 was very specific, it had no exact insight matches and we omitted it from the analysis. This chart is an interesting hybrid of the insight and task methods, because it shows how well each visualization enables a specific predefined set of insights.

Interestingly, the insight results in Figure 10 closely mirror the task-based performance results in Figure 6. There is a correspondence between task performance time and amount of insight. Visualizations that resulted in slower user performance time on a task also produced fewer insights equivalent to that task. Likewise, visualizations that resulted in a faster user performance time on a task also produced more insights equivalent to that task. This represents a confirmation between the two study methods. More generally, this represents an important discovery about how insight can occur: *certain types of insights are generated when a visualization makes those types of insight quick to acquire*.

None of the participants using multiple time points reported insights involving time point analysis (tasks 4, 5, and 6). The participants using multiple time points were significantly slower or less accurate on these tasks in the task-based method. The effect was confirmed and found even more significant in the insight method when analysing the insight category: time point analysis. This indicates that providing a set of predefined tasks potentially forces participants to perform a type of analysis that they would not otherwise perform and that the visualization would not otherwise encourage. This brings into question the *ecological validity* of prescribed task-based studies; that is, the prescribed tasks do not well represent what will actually occur in the real-life analysis.

Benchmark tasks versus insight categories. For a broader comparison, we can abstract the benchmark tasks into the insight categories and compare the results based on category. Table 9 groups the benchmark tasks according to their corresponding insight categories. Corresponding results about the visualization alternatives for each study method are also listed. Table 10 then directly compares the findings in terms of whether simple better/worse differences were detected, not detected, or not tested.

Confirmation: Interestingly, for most of the tasks, each of the two evaluation methods confirmed or partially confirmed the other's results. Again, there is some correspondence between task performance and insight performance. Tasks 1 and 3 required the participants to analyse expression values for individual genes. The task-based method found that multiple graphs was slowest at one of these tasks. Similarly, the insight method found that multiple graphs produced the least of that type of insight. These confirm the visual grouping hypothesis, since multiple graphs splits apart each node representation. For task 2, which examined differences between conditions, neither method produced significant differences, although their trends correlated.

More confirmatory results are for tasks 4–6. These tasks required analysis of the graph to find interesting time points. Both evaluation methods found significant advantages of the single time point visualization and disadvantages of the multiple time points visualization in these types of tasks. Both methods also found some advantages of multiple graphs for some of these tasks. These results doubly confirm the visual grouping hypothesis, since time point analysis is best supported by visualizations that group first by time point.

Table 10. Findings of both methods related in terms of better/worse differences detected in visualization performance

Differences detected by both methods that confirm each other
M Graphs is the worst for single gene expression analysis M Tpts is the worst for time point analysis
Differences detected by both methods that refute each other
M Graphs is better (task)/worse (insight) than M Tpts for topology
Differences detected only in insight method
1 Tpt is the best for time point analysis 1 Tpt is better than M Graphs for topology 1 Tpt is the best overall, M Graphs worst overall
Differences detected only in task method
None
Differences found in insight method, not tested in task method
M Graphs is the worst for expression topology M Tpts is better than M Graphs for outliers 1 Tpt is better than M Graphs for summary

Table 9. Comparison of visualization performance for both methods on the tasks according to insight category

Task	Task study result	Insight category	Insight study result
T1	M Graphs slowest.	Gene expression	M Graphs least insights
T3	No differences		
T2	No differences	Condition	No differences
T4	M Tpts least accurate	Time point analysis	1 Tpt most
	M Graphs faster		M Tpts least
T5	M Tpts slowest		
T6	M Graphs faster		
T7*	M Graphs faster	Topology	M Graphs least

*indicates opposite conclusions about the visualizations between the two methods.

The conclusions of the two methods suggest identical visualization design recommendations.

These confirmatory results reinforce an important link between how efficiently and accurately a visualization supports a particular task and how much of that type of insight a user gains from using the visualization.

Refutation: However, task 7 shows opposite results between the two evaluation methods. For task 7 we found that the participants performed this task faster using multiple graphs than multiple time points. However, in the insight method participants using multiple graphs produced significantly less topology insight (no insight, in fact) than the other tools. This result refutes the previously mentioned link and highlights the difference between imposed tasks and self-discovered tasks. It is possible for a visualization to support a given task more efficiently than other visualizations, but *not* to promote or encourage the use of that task to gain insight about it as much as the other visualizations.

Some possible explanations for this phenomenon are: (1) the visualization steered users towards other types of insights that were made more perceptually salient; (2) the visualization made insights of that type appear as uninteresting or irrelevant to the problem; or (3) the methods are measuring effects at different levels – the task-based method is more perceptually orientated, capturing perceptual efficiency, whereas the insight method is more cognitively orientated, probing at the user's thought processes.

Detection: The insight method results are a superset of the task method results. All of the performance advantages and disadvantages detected by the task method are accounted for by differences detected with the insight method. The task method did not detect any additional differences. The insight method also detected all of the task categories used in the task method. As shown in Figure 10, only task 7, which was a very specific topology task, did not have nearly identical matching insights, although there were other topology insights generated.

On the other hand, the insight method detected additional performance differences that the task method did not, even for task categories that were specifically tested by the task method. These detected differences are further refinements of previously mentioned findings. For example, while both methods found that multiple time points is the worst at time point analysis, the insight method further found that single time point is the best for that type of task, even better than multiple graphs. It is somewhat surprising that the insight method enabled further statistical differences to be found in comparison to the task method.

Extension: Furthermore, the insight method found additional important insight categories that had not been considered for the task-based method, including the topology and expression, outlier, and summary categories. Interestingly, however, the insight result for outlier confirms a similar result from a previous task-based study[7] where it was also found that a multiple time points visualization was faster and more accurate to search for outlier genes, i.e. genes that display different behaviour than most other genes.

Hence, the insight method extended the findings and revealed additional differences between the visualization alternatives for those categories. This provides new useful information that was not provided by the task method. In particular, the single time point and multiple time points visualizations produced more of these types of insight than multiple graphs.

This reflects how designing benchmark tasks for a task-based study can overprescribe and bias the results. This is a form of experimenters' bias that is frequently overlooked in task-based studies, especially when overall task performance is computed in a way that does not take task profiling into account. The insight method overcomes this bias by simply measuring what the users, interacting with the visualization, identify as the important insights or tasks. It lets the users and visualization tools determine what is relevant.

The insight method enabled the simultaneous discovery and evaluation of new task types. At the lowest level, tasks and insights are similar concepts. They are both identifiable units of discovery, such as those identified in various low-level analytic tasks taxonomies.[30] However, insights also allow for more higher level constructs, such as the summary insights found here, which can be more complex, deeper, domain-relevant, uncertain, or vague. These insights do not make good benchmark tasks because they are too difficult to measure and score.[3] Thus, the task method does not adequately evaluate such tasks.

Conclusions about visualization design

Table 11 summarizes conclusions for the visualization alternatives using both methods. As the dependent variables for both methods are different, they provided different conclusions about the visualizations. The task-based method provided feedback in terms of accuracy and performance time. The insight method provided feedback based on the types of data insights the visualization generated. As the tasks were preselected in the task-based study, they provided feedback that allowed designers to judge accurately whether or not a visualization design *supports* a particular task. The unguided insight method provided feedback at a

Table 11. Comparison of the conclusions about the visualization alternatives from both evaluation methods

Visualization	Task-based method	Insight-based method
1 Tpt	Consistently good performance Good single gene, and time point analyses	More time and insights overall Consistently good performance Good single gene, time point, topology, expression, and summary insights
M Tpts	Good single gene analysis Poor time point, and graph topology analyses	Good single gene, topology, expression, and outlier node insights Poor time point insight
M Graphs	Good time point, and graph topology analyses Poor single gene analysis	Less time and insights overall Good conditions, and time point insights Poor single gene, expression, graph topology, outlier, and summary insights

higher level, suggesting what kinds of data analysis a particular visualization method *motivate* and thus how fruitful the visualization was. The fact that users did not perform certain analysis tasks with a visualization does not necessarily mean that the task is not supported, but that the visualization encouraged users to focus on other data analysis aspects.

Overall, in terms of the three visualizations tested, the results of the task-based method tended to favour single time point and multiple graphs visualizations over multiple time points, except for single node analysis tasks, whereas the insight method results strongly favoured the single time point and multiple time points visualizations over multiple graphs, and revealed multiple graphs as the least successful except in time point and conditions analysis tasks.

In terms of general visualization design principles, both methods resulted in strong support for the *visual grouping hypothesis*, which suggests that the visual grouping structure of the visualization should match the task conceptual structure. This was particularly evident in single node gene expression and time point analyses. Time point analysis (which involved analysing and comparing graphs at individual time points) was most successful in single time point and multiple graphs visualizations, which visually group first by time point. All values at a single time point are visually grouped and separated from other time points. This is *column-centric* from the data point of view (Figure 1). On the other hand, single gene and outlier analyses (which involved analysing and comparing the time series of individual genes across the graph) were most successful in multiple time points, which visually groups first by gene. All time points for a single gene are visually grouped together within the node, and separated from other gene nodes. This is *row-centric* from the data point of view (Figure 1).

Interestingly, single time point visualization did well at both types of tasks, suggesting that interaction was able to overcome limitations of visual grouping. Single time point visually grouped primarily by time point (column-centric, with column selected by the slider), and did very well with the corresponding time point analysis tasks. Yet, at the same time, it *interactively grouped* by gene (row-centric) such that all the time points of a node would visually appear in the same location over time as the user dragged the slider, and did very well with the corresponding single gene expression analysis tasks. This design concept of interactive grouping can provide a useful middle ground between the visual grouping decision alternatives.

Empirical data analysis process

The data analysis process for the task-based method was more straightforward than the insight method. It required the use of standard statistical analysis methods including ANOVA and paired *t*-tests. It took about 1 day to finish the entire process, as the investigators had previous experience of analysing such data. The use of autograded multiple-choice questions for the benchmark tasks helped make this process simple and efficient.

The data analysis process for the insight method was more complex. The amount of empirical data collected for the insight method supported richer analysis options. The participants' insights were analysed first to find suitable categories to group the insights. The choice of categories can be dependent on the investigators' preferences and data understanding. A discussion was required between the investigators and domain biology experts to finally agree to a list of categories. With meetings involved it took about 3–4 days

to finish the data analysis. Thus, in contrast to the task-based method, data analysis for the insight method is more complicated and subjective. It is possible that other analysts may have grouped the insights differently. For future work, a more generalized insight categorization[30] can be attempted.

This effort is partly offset during the design phase of the task-based method by the need to design the benchmark task set to test. This requires time and subjectivity by the investigator to interview domain experts and decide on the most import task set, and then to prepare the benchmark task materials.

User feedback about the visualizations

Much more valuable user feedback was collected from the insight method, even though we did not require it.

Usability issues. Although both methods were conducted to evaluate visualization alternatives, the insight method required more interaction with the participants. The experiment protocol for the insight method required a closer observation of the participants' data analysis procedure and one-to-one interaction. This made it easier to notice if the participants were having any difficulties with the user interface. Also, while performing data analysis in the insight method, participants verbally commented about the visualization interfaces such as 'the choice of colour is weird', 'the time point labels are difficult to understand', etc. This was natural because of the think-aloud protocol. Such valuable information was missed in the task-based method. For example, we also observed in the insight method that participants using the single time point visualization enjoyed the study because the visualization was more interactive than the other visualization alternatives. This enjoyment may have prompted these participants to spend more time in the study than the other participants.

Visual representations. Participants in the insight method provided more feedback about the visual representations of the graph. While analysing the data, participants frequently commented on their difficulties and suggested other data representation methods that they thought would better support some of their data analysis tasks. For example, participants using the multiple graphs visualization commented that it was difficult for them to focus on the time series of a single gene only. Participants using the multiple time points visualization commented that they were having trouble focusing on a single time point. They said that somehow the visualization was prompting them to focus on the overall node expressions, and they suggested that various interactions could be added to enable them to highlight genes or drill down. These comments qualitatively validate the rationale behind the *visual grouping hypothesis*.

Effect of individual differences

The task-based method produced less overall variance, since it provided all the participants with an equivalent set of tasks. The list of tasks provided very specific direction to the participants throughout the study. This prevented the participants from getting confused about what to do next. Also, it made the study experience similar for all these participants.

In contrast, the insight method was open-ended and it was important for the study that the participants think aloud. It is possible that some participants were more communicative than others, and thus reported more insights than other participants who may have actually had similar data insights but chose not to verbalize them all. Sometimes participants, depending on the type of visualization alternative they were using, felt that some insights were so noticeable that they may be too trivial and not worth reporting. Thus, findings from the insight study were more likely to be affected by the individual differences between the participants.

The participants in the insight method were suspicious of our intentions, and some asked if the data insights they were reporting made sense, or if they could be provided with more details as to what they should be reporting so that they can be more helpful. When some participants in the insight method became confused about the purpose of the study, sometimes they needed to be encouraged to report insights. We just answered 'Yes, that makes sense'. Some users required more prompting than the others. It may be helpful in future studies to decide if the participants should be provided with such encouragement to make the study more uniform. A few participants reported that the entire study felt as if there was some catch involved to it. They thought there was either something that they were supposed to definitely notice, or that we wanted them to completely miss. At the end of the study, when participants were ready to leave, they wanted to know if they behaved as we expected them to and what was the purpose of the entire study.

Participant motivation for data analysis

Unmotivated subjects were easier to recognize in the insight method. All the participants in the study were undergraduate biology students. To encourage participation in the study, they received some course credit.

It is likely that some participants came only for the credit and were not motivated to perform data analysis. For the task-based method, it could be either lack of motivation or usability of the visual representation that adversely affected participants' performance (especially accuracy). For the insight method it was more obvious to identify which issue (motivation or usability) was the cause of problems. It was easier to notice unmotivated participants because there was more communication with the investigator. The participants would often comment that they were tired or say 'I just came from class, my mind is blank, please give me a minute to rest'. We also noticed that participants who came during the weekend were more relaxed and interactive in the study, whereas participants who came during the weekdays were less inclined to spend as much time in the study. Potentially, as such unmotivated subjects can be recognized in the insight method, they could be filtered from the study so as to focus on a more realistic scenario. Motivational rewards could also be offered.

Discussion and limitations

The insight method presented previously[4] recognized several characteristics of an *insight* such as hypothesis generation, breadth versus depth, directed versus undirected, correctness, and domain value. For the data analysis discussed here, we focused on insight categorization. Grouping insights by categories provided us with sufficient basis to compare the studies for the present discussion.

Also, the data and tools used here were more simplistic to reduce the learning time and allow users to complete the analysis in limited time. For real-world data analysis scenarios, an analyst spends much more time analysing the data. The original data set from which the data for this study was selected was 45,001 rows by 72 time points and required about 3 months of data analysis by the biologists. The most important subgraph found after a few months of data analysis, and the associated time series data set which was just 46 rows by 12 time points, was used in this study. Thus, although the short-term studies provide important feedback and enable rigorous comparison of alternatives, they miss the amount of feedback provided by a longitudinal study[9,29,31] for visualization tool usage. However, an advantage of the insight method is that it can be applied in a longitudinal study.[29]

The participants in this study were undergraduate biology students. For the insight method, at the end of the data analysis some participants were confident about their data analysis and could summarize the data or make hypothesis about the biological phenomenon suggested by the data. Such comments were ranked very highly in the original insight study.[4] The single time point visualization produced the highest number of this type of insight. However, although the participants had a background in biology, they did not have enough familiarity with the specific immunity phenomena examined by this data set. Any such hypotheses were merely speculations. They would not be able to judge the actual value of such findings. Although the purpose of overlaying gene expression data onto pathway graphs is to provide meaningful biological context for the data,[6] in this case the participants did not have adequate background with this specific pathway and genes to fully exploit that meaning. Thus, apart from the summary insights, an attempt at ranking the insights in this study by the experts resulted in most insights being rated at a similar value and considered 'dry' data analysis without much biological inference, so was not a useful measure in this case. While the insight method worked well to capture and measure insight counts in this simple scenario, more complex scenarios can offer richer insight analysis. Of course, these issues were irrelevant to the task-based evaluation method, which forced the use of 'dry' data analysis tasks in order to support a straightforward scoring scheme for time and accuracy.[3]

Conclusion

This study was conducted to compare two empirical research methods for evaluating visualization alternatives, the benchmark task-based method and the insight-based method. Table 12 summarizes the two methods, and Table 13 summarizes their empirical results. As the dependent variables for both methods are different, the studies were compared based on abstractions of their results and on higher level criteria most relevant to evaluating visualization tools.

The fundamental difference between the methods is in their treatment of user tasks, which impacts how subjective bias enters the process. The task method forces experimenters to design benchmark tasks, which prescribes the results and threatens ecological validity. The insight method lets users determine the tasks, in the form of insights that they identify, requiring qualitative analysis to produce quantitative results, which threatens repeatability. Through this, though, the insight-based method provided a way to capture a realistic data analysis scenario and a wider range of comparison factors. A higher level analysis such as grouping insights into task categories enabled indirect comparison of the task and insight method

Table 12. General comparison of the benchmark task and insight evaluation methods

Comparison factor	Task-based method	Insight-based method
Purpose	Evaluate specific research question about task performance	Evaluate insight generated in realistic analytic scenario
Design prep	Prepare benchmark tasks and scoring scheme	Prepare problem scenario
	Better with simple data, tools, tasks	Better with complex data and tools
Experiment	Benchmark task protocol	Open-ended protocol
design	Form based	Think aloud
	Time and accuracy	Capture insights
	Can be multiplexed	Interaction with user
	Short-term study only	Can be longitudinal[2]
	Longer preparation time	Variable procedure time
User tasks	Determined by experimenter	Determined by user (user identifies insights)
Participants	Any users	Expert, motivated users
	Many users	Motivation is detectable
		Train without biasing
Empirical data analysis	Processing scores data	Coding rich insight and usability data
	Quantitative statistical analysis	Statistical analysis
		Higher variance
		Longer analysis time
Primary outputs	Identify tasks *supported* by a visualization	Identify tasks *promoted* by a visualization
	Perceptual, mechanical task efficiency	Cognitive, interactive learning efficiency (amount of insight)
	(time, accuracy)	Statistical differences
	Statistical differences	Detects new tasks, ignores unneeded tasks
	Feedback on selected tasks only, ensures coverage of those tasks	Higher level tasks, user hypotheses, Summary
	Low-level tasks	Qualitative feedback and analytic process
Subjective bias	Choice of benchmark tasks and scoring scheme	Coding of insights and categories
Bias threat	Ecological validity	Repeatability

results. There are several key findings in correlating the results between the evaluation methods, in terms of comparing the three visualization design alternatives:

- Insight confirms task method: Many of the findings in the task method were confirmed, or even amplified, in the insight method. Overall, both methods found advantages of the single time point visualization. For example, both methods showed that single time point is the most successful and multiple time points is the least successful at time point analysis. This may provide some *validation* of the insight method, that it detects effects found by the task method.
- Insight refutes task method: However, some findings were counter, indicating that users *behave differently* when not in the forced direction of a task-based method. The task method tended to favour the multiple graphs visualization, while the insight method favoured the multiple time points visualization. As a specific example, even though participants performed the graph topology task fastest using multiple graphs in the task-based method, they gained the least insight about topology using multiple graphs in the insight method. In fact, none

Table 13. Comparison of the benchmark task and insight evaluation empirical results

Benchmark task method result	Insight method result	
	More insight	Less insight
Fast and accurate	*Confirm*	*Refute*
	Time point tasks on single time point and multiple graphs	Topology tasks on multiple graphs
	Single gene tasks on single time point and multiple time points	
Slow and inaccurate	*Refute*	*Confirm*
	Topology tasks on multiple time points	Time point tasks on multiple time points
		Single gene tasks on multiple graphs
No difference detected	*Expand*	*Expand*
	Single time point overall	Multiple graphs overall
	Time point tasks on single time point (uniquely)	
	Topology tasks on single time point	
Not tested	*Extend*	*Extend*
	Outlier tasks on multiple time points	Topology and expression on multiple graphs
	Summary tasks on single time point	

of the multiple graphs users made any topology insights. Thus, because of its unguided protocol, the insight method may allow participants to miss certain type of tasks. The fact that participants did not gain topology insight does not mean that the task is not supported by the visualization, but indicates that the visualization does not provoke the participants to look for it.

- Insight expands task method: The insight method found further statistical differences for which the task method did not detect corresponding differences. Most surprisingly, the task method did not find any further differences that went undetected by the insight method. For example, the insight method found that single time point was the best visualization for time point analysis tasks. Thus, the insight method expanded on the above, confirming and refuting differences with further refinements. The insight method was surprisingly *more powerful* than the task method in this study.
- Insight extends task method: Although the task-based method is more uniform, it provides feedback only on the preselected tasks. Designing proper benchmark tasks is non-trivial[3] and requires deep domain knowledge. For example, the insight study found that single time point promoted more summary tasks and multiple time points performed well at finding outlier nodes. We did not get this information from the task-based method because we did not have benchmark tasks to reflect those potential insight categories. The summary insight category represents an important *higher-level task*, in which users hypothesize about biological meaning, which is difficult to capture in a simple benchmark task method that scores time and accuracy. The insight method extended to new results by offering the opportunity to discover important new task types from users.

In terms of visualization design guidelines, both methods confirmed the *visual grouping hypothesis*, along with further support from qualitative feedback from the insight method participants. Visualizations should be designed so that group data according to the query structure of the desired tasks. The insight results also highlighted the importance of interaction, as in the single time point visualization, to motivating insight generation. *Interactive grouping* by visual animation can provide an effective middle ground between the visual grouping alternatives.

Ultimately, the fundamental difference between the two methods creates a subtle but important distinction between what they measure. The task method measures how efficiently a visualization *supports* a given task, whereas the insight method measures how much a visualization *promotes* a given task to users.

Overall, the many confirmatory results suggest how insight functions. Insight typically occurs when tasks are efficiently supported by a visualization. However, the subtle adverse effects of the distinction are revealed in the few refuting results, indicating that support and promote do not necessarily go hand in hand in actual visualization usage. This distinction also suggests that attempting to combine the methods might produce unexpected interactions or biases. Then, which of these measures is more valuable? Both are certainly useful. In the end, if we want to know what users will actually *do* with a visualization, then this is best measured by the latter.

Acknowledgements

We thank Peter Lee for developing the visualization tools, Vy Lam for providing data, and Dr Joe Cowles and Dr Carla Finkielstein for encouraging student participation in the study.

References

1. Chen C and Czerwinski M. Empirical evaluation of information visualizations: an introduction. *IJHCS* 2006; 53: 631–635.
2. Plaisant C. The challenge of information visualization evaluation. *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI '04)*. New York, NY: ACM, 2004, pp.109–116.
3. North C. Towards measuring visualization insight. *IEEE Comput Graphics App* 2006; 26(3): 6–9.
4. Saraiya P, North C and Duca K. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Trans Vis Comput Graphics* 2005; 11(4): 443–456.
5. Duggan D, Bittner B, Chen Y, Meltzer P and Trent J. Expression profiling using cDNA microarrays. *Nat Genet* 1999; 21: 11–14; .
6. Saraiya P, North C and Duca K. Visualizing biological pathways: requirements analysis, systems evaluation and research agenda. *Info Vis* 2005; 4(3): 191–205.
7. Saraiya P, Lee P and North C. Visualization of graphs with associated time series data. *Proceedings of IEEE InfoVis 2005. IEEE*. 2005, pp.225–232.
8. Carpendale S. Evaluating information visualizations. In: Kerren A, Stasko JT, Fekete J-D and North C (eds) *Information Visualization, Lecture Notes In Computer Science*. Vol. 4950, Berlin, Heidelberg: Springer-Verlag, 2008, pp.19–45.
9. Shneiderman B and Plaisant C. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. *Proceedings of the 2006 AVI workshop on Beyond time and errors: novel evaluation methods for information visualization (BELIV '06)*. New York, NY: ACM, 2006, pp.1–7.
10. Bertini E, Perer A and Lam H. Beyond time and errors: Novel evaluation methods for information visualization (BELIV Workshop). Available at: http://www.beliv.org/ (accessed 01 February 2011).
11. Chang R, Ziemkiewicz C, Green TM and Ribarsky W. Defining insight for visual analytics. *IEEE Comput Graphics App* 2009; 29(2): 14–17.
12. Plaisant C, Fekete JD and Grinstein G. Promoting insight-based evaluation of visualizations: From contest to benchmark repository. *IEEE Trans Vis Comput Graphics* 2008; 14(1): 120–134.
13. Smuc M, Mayr E, Lammarsch T, Aigner W, Miksch S and Gartner J. To score or not to score? Tripling insights for participatory design. *IEEE Comput Graphics App* 2009; 29(3): 29–38.
14. Golovchinsky G and Belkin NJ. Innovation and evaluation of information exploration interfaces: A CHI98 Workshop. *SIGCHI Bull* 1999; 31: 22–25.
15. Chen C and Yu Y. Empirical studies of information visualization: a meta-analysis. *IJHCS* 2000; 53: 851–866.
16. Kosara R, Healey C, Interrante V, Laidlaw D and Ware C. Thoughts on user studies: why, how, and when. *IEEE CG&A* 2003; 23(4): 20–25.
17. House D, Interrante V, Laidlaw D, Taylor R and Ware C. Panel: Design and evaluation in visualization research. *IEEE Visualization Conference '05*. IEEE 2005, p.117.
18. Tory M and Möller T. Evaluating visualizations: Do expert reviews work? *IEEE CG&A* 2005; 25(5): 8–11.
19. Steves MP, Morse E, Gutwin C and Greenberg S. A comparison of usage evaluation and inspection methods for assessing groupware usability. *Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work (GROUP '01)*, New York, NY: ACM, 2001, pp.125–134.
20. Brush AJ, Ames M and Davis J. A comparison of synchronous remote and local usability studies for an expert interface. *CHI '04 extended abstracts on Human factors in computing systems (CHI EA '04)*. New York, NY: ACM, 2004, pp.1179–1182.
21. Bekker M, Baauw E and Barendregt W. A comparison of two analytical evaluation methods for educational computer games for young children. *Cognition Tech Work* 2008; 10(2): 129–140.
22. Jeffries R and Desurvire H. Usability testing vs heuristic evaluation: Was there a contest? *SIGCHI Bull* 1992; 24(4): 39–41.
23. Hartson HR and Andre TS. Criteria for evaluating usability methods. *IJHCI* 2001; 13(4): 373–410.
24. John B and Marks S. Tracking the effectiveness of usability evaluation methods. *BIT* 1997; 16(4/5): 188–202.
25. Doubleday A, Ryan M, Springett M and Sutcliffe A. A comparison of usability techniques for evaluating design. *Proceedings of the 2nd conference on Designing interactive systems: processes, practices, methods, and techniques (DIS '97)*. ACM: New York, NY, 1997, pp.101–110.
26. STKE, The Signal Transduction Knowledge Environment, Database of Cell Signalling. *Science Signalling*. Available at: http://stke.sciencemag.org/cm/ (accessed 01 February 2011).
27. Cytoscape: An Open Source Platform for Complex Network Analysis and Visualization. Available at: http://www.cytoscape.org/ (accessed 01 February 2011).
28. Barsky A, Munzner T, Gardy J and Kincaid R. Cerebral: visualizing multiple experimental conditions on a graph with biological context. *IEEE Trans Vis Comput Graphics* 2008; 14: 1253–1260.
29. Saraiya P, North C, Lam V and Duca K. An insight-based longitudinal study of visual analytics. *IEEE TVCG* 2006; 12(6): 1511–1522.
30. Amar R, Eagan J and Stasko J. Low-level components of analytic activity in information visualization. *Proceedings of IEEE InfoVis '05* (Minneapolis, MN) 2005, pp.111–117.
31. Seo J and Shneiderman B. Knowledge discovery in high-dimensional data: case studies and a user survey for the rank-by-feature framework. *IEEE Trans Vis Comput Graphics* 2006; 12(3): 311–322.
32. Gualano RC, Hansen MJ, Vlahos R, et al. Cigarette smoke worsens lung inflammation and impairs resolution of influenza infection in mice. *Respiratory Res* 2008; 9(1): 53.