BRINGING INTERACTIVE VISUAL ANALYTICS TO THE CLASSROOM FOR DEVELOPING EDA SKILLS

Jessica Zeitz
Computer Science Department
University of Mary Washington
Fredericksburg, VA 22401
540 654-5996
jzeitz@umw.edu

Nathan Self, Leanna House, Jane Robertson Evia, Scotland Leman, Chris North
Discovery Analytics Center
Virginia Tech
Blacksburg, VA 24061
{nwself; lhouse; robj; leman; north} @vt.edu

ABSTRACT

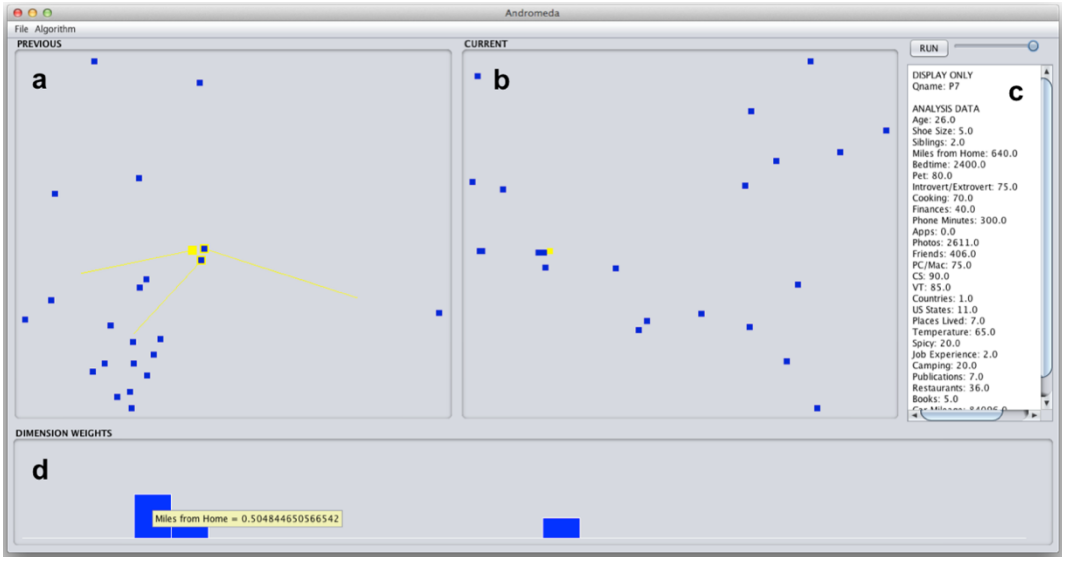
This paper addresses the use of visual analytics in education for teaching Exploratory Data Analysis (EDA) skills. EDA is inherently a creative, knowledge discovery process that often takes place before formal technical statistical analyses. A challenge in teaching EDA is that there is often no right nor wrong way to conduct EDA, yet, given a dataset, some EDA can be more comprehensive or insightful than others, based on the kinds of insights made. Also, in the face of high-dimensional data, students are often limited by how they relate to the data and their technical skills for EDA. How can students make complex insights from high-dimensional data, if they do not have the technical skills to explore the data from multiple, high-dimensional perspectives? In this paper, we use our own tool called Andromeda that enables human-computer interaction with a common, easy to interpret visualization method called Weighted Multidimensional Scaling (WMDS) to promote the idea of making complex insights. We present Andromeda and report findings from a series of classroom assignments to 18 graduate students. These assignments progress from spreadsheet manipulations to statistical software such as R and finally to the use of Andromeda. In parallel with the assignments, we saw students' cognitive dimensionality (CD) begin low and improve.

INTRODUCTION

Today's datasets are big. Advances in technology have enabled almost every organization to collect more data than ever. To explore and learn from these large datasets, humans are called upon to assimilate what they know with tens, hundreds, even thousands of observations and variables at once. However, assimilation is stifled by both limitations of the human brain (e.g., conceptualizing more than three dimensions) and practical methods for summarizing datasets. This is particularly true for students. Due to limited experience with data and analytical methods, students are challenged to explore data. In this paper, we present evidence that a new analytics tool may foster improved Exploratory Data Analysis (EDA) skills.

In this paper, we performed a three-part observational study (Section 4) to assess the impact of an interactive analytics tool called Andromeda on pre-defined EDA skills. Andromeda is a tool that enables users to explore data visually, based on multiple linear projections of data *and* personal conjectures about the data.

Figure 1. Andromeda Interface



This is a screenshot of Andromeda during an analysis: (a) the previous view panel depicting the previous spatialization, (b) the current view panel depicting the most recent spatialization, (c) the detail panel displaying the raw data, and (d) the dimension weights bar chart visualizing the dimensional reduction weight vector of the model in the current view panel.

Effectively, when students use Andromeda, data explorations become student-centric experiences during which students are called upon to reconcile – either assimilate or accommodate [1] – repeatedly the relationship between two-dimensional visualizations and a high-dimensional dataset. Insights from data result when reconciliations are successful. Results from the study suggest that students start with limited EDA skills, and improve when using software and Andromeda.

CURRENT WORK

This paper brings research in statistics and visual analytics to the classroom. In this section, we highlight relevant components of our research that provide the necessary, technical background for us to assert that Andromeda enables students to construct EDA skills.

Visual Analytics

Visual Analytics (VA) is the "science of analytical reasoning facilitated by interactive visual interfaces" [2]. VA research is devoted to developing methods by which humans may visualize and interact with data in ways that make sense to them; humans are a central component in the process of making sense of data. There are many ways in which humans may interact with data [3], [4]. For two-dimensional visual projections of data, three types of interaction are defined in [5]: Surface Level, Parametric, and Visual to Parametric. Surface level interactions are

comparable to read-only actions include highlighting, zooming, and filtering observations that do not update the visualization. Parametric interactions enable analysts to adjust specifications for parameters in models that create visualizations. Visual to Parametric Interaction (V2PI) allows analysts to indirectly adjust parameters of the model by adjusting the visualization [5] [6] [7]. With V2PI, users communicate their ideas and judgments about the data through an easy-to-interpret, low-dimensional visualization. These judgments are quantified and used to update an underlying model, and from there, the visualization.

Andromeda

Andromeda enables all three forms of interaction and provides visualizations based on Weighted Multidimensional Scaling (WMDS) [8], [9]. WMDS is a linear projection method that includes one parameter (weight) per dimension which reflects its relative importance in a visualization. With WMDS, pairwise relative distances between observations reflect relative similarities/differences between observations, particularly in the dimensions with the largest weight. Figure 1 provides a screenshot of the Andromeda interface which includes two WMDS scatterplots, a list of the variables, and a bar graph of designated weights for each variable in the dataset. The two WMDS plots represent previous and current views. The current view is the result of performing either V2PI or Parametric Interaction. We provide both views to enable users to compare and assess the impact of their interactions. Andromeda supports all three forms of interaction.

***Surface Level Interaction in Andromeda*:** When a user hovers over a data point (in either view), Andromeda displays that point's values for each variable in theData Panel (labeled c in Figure 1) and highlights the same data point in the other view. Hovering provides a means for users to make sense and develop their own interpretation of the difference between the two visualizations.

***Parametric Interaction in Andromeda*:** Users may click and drag the top of a bar in the dimension weights bar chart to increase or decrease that variable's weight relative to other variables [10]. The plot in the current view panel is then redrawn with the updated WMDS spatialization. The old spatialization in the current view panel then transitions to the previous view panel.

***Visual to Parametric Interaction in Andromeda*:** In the current view (labeled b in Figure 1), users may apply V2PI by dragging points in the spatialization. Andromeda then solves the inverse of WMDS, using the new pairwise distances between points to compute new variable weights. Andromeda uses the new weights to re-compute WMDS for the entire dataset and update the current view. Andromeda communicates the updated values of the weights in the bar graph in the bottom panel (labeled d in Figure 1).

METHODS

We implemented a set of three iterative assignments to assess exploratory data analysis (EDA) skills. The assignments were given over a three-week period in a graduate visual analytics course with 18 students enrolled. The assignments involved analyzing data collected from a survey given to students and colleagues. The survey included 27 personal questions that had numeric answers, such as, *What*

is your age? On a scale from 0-100, how much do you like cooking? How many apps do you have on your smartphone? Students had the option of releasing their name on the survey or completing the survey anonymously. The final dataset included 23 observations and 27 variables. We refer to this final dataset as *survey data*.

Assignments

The three assignments required the students (i.e., participants in the study) to analyze the survey data using three separate tools with increasing complexity: first, by hand; second, with R or MATLAB; and, third with Andromeda. For each assignment, students were asked to analyze the survey data and develop insights about their classmates; e.g., find patterns or relationships among students. Visualizations were encouraged, particularly those that relied on proximity to encode similarity.

Manual Assignment: The first assignment was open-ended and was intended for us to establish a baseline for students' current EDA skills. It required students to calculate a 23 x 23 similarity matrix using a metric, such as cosine similarity. Then, without other mathematical techniques or algorithms, the students created a hand-drawn 2-dimensional map of the class and listed insights they discovered about the data. Deriving the similarity matrix was our way to open the discussion for the meaning of distance and to encourage the students to use spatial proximity to convey relative similarity among the data points. To make insights, the students could use simply summary statistics, their visualization, the cosine similarity matrix, and/or a similarity matrix of their own.

Statistical Computing Environments Assignment: The second assignment built upon the first by adding computational and visual representations, with limited interaction. Students used typical analytical tools, including R or MATLAB. The assignment suggested that students create standard data plots such as histograms, scatterplots, scatterplot matrices, and parallel coordinate plots, as well as projections from principal component analysis (PCA) [11] and either unweighted or weighted multidimensional scaling plots (MDS or WMDS, respectively) [8], [9], [12]. Note that insightful students manually interacted with WMDS by filtering the data and/or adjusting visualization parameters directly; i.e., visualizing subsets of the data and/or adjusting variable weights in WMDS. To complete the assignment, students were asked again to list their insights.

Andromeda Assignment: The third and final assignment provided Andromeda and asked the students one more time to list insights from the data. To complete this assignment, students received a short tutorial on the basic functionality of Andromeda and were taught how to take screen shots so that they could provide images of their work to support their claims.

Data Collection

Crucial to the evaluation of this observational study is the definition and measure of insights, as well as the methods applied for making insights. We assume that the methods used for analysis influence EDA skills which in turn influence the quality of insights. By measuring the insights, a representation of what the students learned, across multiple methods of analysis, we can infer the development of

students' EDA skills.

Inspired by [3] and [13], we adopt the definition of insight as *an observation by a student about the data*. Thus, for 18 students across the three assignments, 257 insights were made. To analyze these insights, we characterize their complexity and depth. Collectively, [3], [4], [14] list several properties for insights that reflect complexity and depth, including: time taken to reach the insight, the domain-specific significance for the insight, whether the insight leads to a new hypothesis, and whether the insight is qualitative, unexpected, correct, or broad. We focus our attention on those that are quantifiable and document the following for each insight:

- *Dimensionality*: Each dimension that was explicitly listed in an insight is tallied.
- *Cardinality*: Each data point that is explicitly listed in an insight is counted.
- *Relationship cardinality*: We categorize the comparisons of points as a relationship such as one-to-many, one-to-all, one-to-one, etc.

To further characterize insights, we also consider the tasks completed to make insights [15]. The idea is that deep insights are those that accumulate and build over time and upon other insights. We measure accumulation by counting the tasks completed to make the insight. To do so, we use the analytic task taxonomy of low-level components outlined in [15] and identify the following tasks taken for each insight: *retrieve value*, *filter, compute derived value*, *find extremum, determine range, characterize distribution, find anomalies,* identifiying members of *clusters, correlate* between dimesions. Note that insights may be the result of one or more than one analytical task.

RESULTS

To assess potential gains over the course of the three assignments, we summarize the insights and tasks. We describe the implications of our results in the Discussion section.

Table 1. Insight Trends

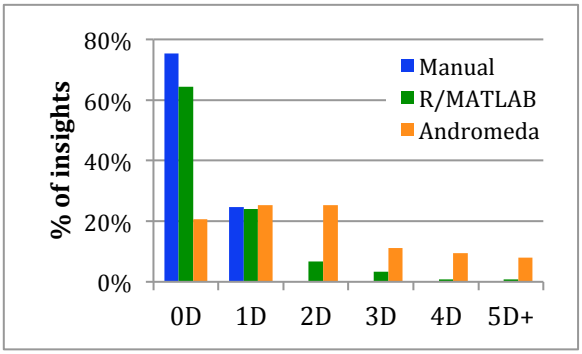
Data	Assignment		
	Manual	***Statistical Environment***	***Andromeda***
Total insights	73	121	63
Number of students	13	17	16
Avg. dimensionality	0.25	0.60	2.25
Avg. cardinality	1.66	1.22	3.11
Avg. number of tasks	2.13	1.79	2.24

Insight Complexity

Across all 18 students, there were 73 insights for the manual assignment, 121 insights for the statistical computing environment assignment, and 63 insights for the Andromeda assignment. We summarize differences and similarities in the insight complexity across the assignments according to dimensionality, cardinality, relationship cardinality and diversity of tasks. Insight trends are described in Table 1.

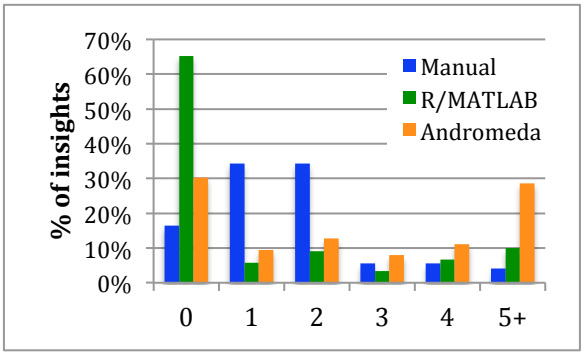
***Dimensionality*:** Of the manual insights, 75% did not refer to any dimension (Figure 2). Most of these zero-dimensional insights included finding extremums

Figure 2. Dimensionality of Insights



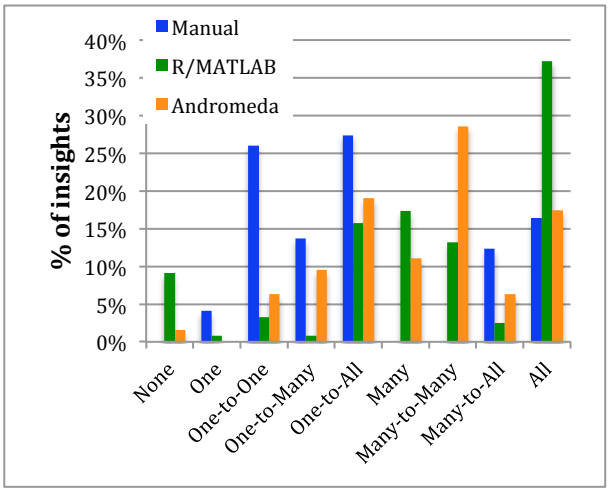
Percentage of insights from each assignment against the number of dimensions explicitly mentioned in each insight.

Figure 3. Cardinality of Insights



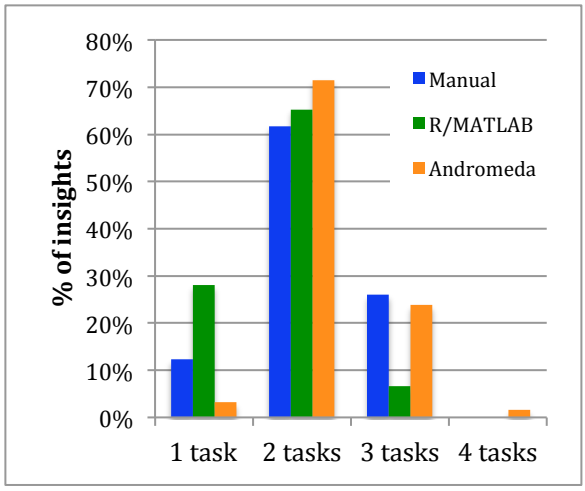
Percentage of insights from each assignment against the number of observations explicitly mentioned in each insight.

Figure 4. Relationship Cardinality of Insights



Percentage of insights from each assignment against the relationship cardinality based on observations mentioned in each insight.

Figure 5. Number of Distinct Tasks Per Insight



Distribution of insights across tasks; percentage of insights that contained at least one of the tasks.

based on the computed similarity values, comparing two individuals, or characterizing the distribution of the data based on similarity or dissimilarity. Such insights seem natural to include at least one dimension, yet no reference was made. The remaining 25% of manual insights only considered one dimension. These insights focused on the anomalies and extremums of a single dimension.

Similar to manual insights, 64% and 24% of statistical environment insights did not reference any dimensions or referenced one dimension, respectively. Of the zero-dimensional insights, many focused on characterizing the distribution of derived values from MDS, whereas, most of these one-dimensional insights stemmed from characterizing the histogram of that particular dimension.

For example, one such insight stated the students in the dataset have lived relatively few places (number of places lived dimension). A small percentage of insights did refer to two to five dimensions, which is a step up from manual insights. When two dimensions were listed within an insight, a correlation stemming from a

scatterplot was mentioned. The insights referring to three to five dimensions were gleaned from a PCA plot which explained that certain dimensions contributed most to a particular component. Also, students clustered dimensions based on a self-imposed category, e.g., travel behavior consisting of number of countries visited, number of US states visited, and number of places lived.

The spread across the number of dimensions per insight increased for those made with Andromeda. Albeit, 45% of Andromeda insights are zero- or one-dimensional, but almost 30% reference three or more dimensions (Figure 2). Even though the percentages are small, we see a shift in the complexity of insights when using Andromeda. Students produced insights consisting of up to ten dimensions and greatly increased the number of insights using two, three or four dimensions.

***Cardinality and Relationship Cardinality*:** We categorized the insights based on cardinality and relationship cardinality. Figure 3 and Figure 4 list the percentages per assignment. Notice that manual insights tended to focus on either zero to three people in the dataset or the entire dataset of people. Often, these insights were egocentric in that the student compared himself/herself to another person or the entire to who was most similar or different and/or to learn how he/she was similar or different from the entire class.

As seen in Figure 4, insights for the statistical environment assignment consisted mainly of all (37%), many (17%), and one-to-all (16%). The static MDS and PCA plots lend themselves to observing the entire layout of the points. Whereas, Andromeda seemed to inspire comparisons for subsets of the data, 29% made many-to-many comparisons. These comparisons had high cardinality given the students mostly compared clusters of data points. As seen in Figure 3, 29% of Andromeda insights included 5 or more data points. Half of these insights referenced clusters of points within the visualization. The remaining insights with cardinality of 5 or more referenced outliers and how they compared to one or more clusters within the visualization.

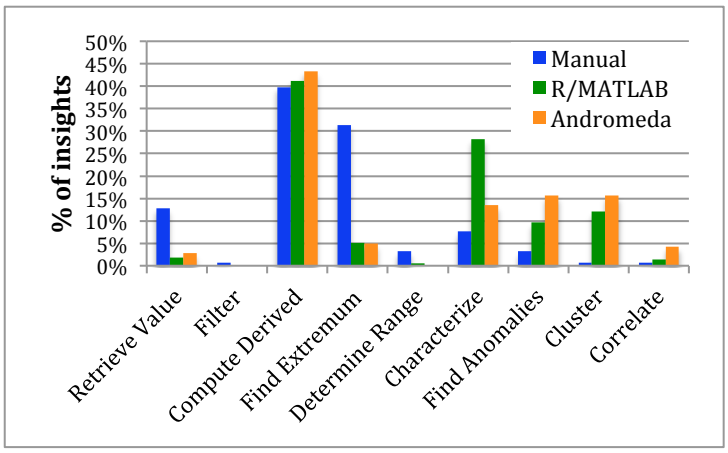
Task Diversity: Insights made using different tasks reflect both the complexity of insights and the EDA skills of the students. The 73 manual insights, 121 statistical environment insights, and 63 Andromeda insights contained 156, 216, and 141 individual tasks respectively (see Figure 5 and Figure 6). In Figure 4, we notice that the most number of tasks used to make an insight was four and resulted when using Andromeda. Surprisingly, however, the number of tasks employed (on average) with the manual assignment was higher than those used for the statistical environment.

Figure 6 shows that a high percentage of all tasks were *compute derived value*. This observation makes sense because all three assignments manipulated the raw data in some way. For the manual assignment, the derived data included similarity values and matrices, whereas, the statistical environment and Andromeda assignments produced derived data from histograms and/or dimension reduction algorithms (e.g., PCA, MDS and WMDS). *Find extremum* was the most prevalent task within the manual insights (Figure 6); 31% of the 73 manual insights. Most insights from the *find extremum* component were of the form "Person *X* had the highest/lowest raw data value for single dimension." These insights also included the highest or lowest top two or three persons based on a single dimension.

For the statistical environment assignment, the second most prevalent task within this context was *characterize distribution* (Figure 6). Students would describe unique histogram distributions of single dimensions. For the static MDS and PCA plots, students would describe the general location of data points based on proximity and visible groups. For example, many students stated that the data points formed *n* number of groups.

The distribution of tasks for Andromeda insights is comparable to that of the statistical environment. However, *finding anomalies* (16%) and *clustering* (16%) tied for being the second most useful task with the use of Andromeda when making insights. In particular, we note clustering as a common task, as this correlates with the use of dimensions in that many students described clusters that formed by learned dimensions after interacting.

Figure 6. Diversity of Tasks



Percentage of insights against the number of tasks included in the insights.

DISCUSSION

Relative to other assignments, insights derived from the use of Andromeda were more complex and reflect an improvement in EDA skills. Findings from the observational study support our hypothesis.

- Students increased the number of dimensions considered as they progressed from manual computation to static visual encodings and from statistical environments to interactive, Andromeda spatializations.
- When using Andromeda, students focused less on their own data point and more on clusters of data points. If students did reference their own data point, it tended to be within a cluster of data points and identified multiple dimensions in support of their insights, in accordance with the dimension weights reported by Andromeda.
- Insights from the Andromeda assignment tended to offer explanations for which dimensions caused a given clustering, whereas insights from the other assignments did not.
- Students did not follow one line of inquiry, but pursued alternative viewpoints which helped to thwart the tunneling of their thought processes.

We acknowledge that our study has limitations compared to formal controlled experiments. For example, the order of assignments may have confounded these results in that students may have felt compelled to generate new or more complex insights in sequence or built insights upon knowledge gained in previous assignment. However, it is clear that the insights gained through Andromeda would have been difficult to gain using the other two approaches. Thus, students without additional lessons in EDA constructed on their own how to make complex insights.

CONCLUSION

In this paper, we conclude that students' EDA skills may improve naturally with interactive analytics tools, such as Andromeda. To make this conclusion, we collected data from an observational study within which we measured insights from three EDAs resulting from three student assignments. This research may lead future work in how to assess EDA skills and measure quality of insights, as well as building better tools for data exploration.

ACKNOWLEDGEMENTS

This research was supported in part by National Science Foundation grants IIS-1447416 and DUE-1141096.

REFERENCES

[1] Piaget, J., Inhelder, B., *The psychology of the child*. Basic Books, 1969.

[2] J. J. Thomas, J.J., Cook, K., "Illuminating the path: The research and development agenda for visual analytics," *IEEE Computer Society*, vol. 54, pp. 184, 2005.

[3] Saraiya, P., North, C., Duca, K., "An Insight-based Methodology for Evaluating Bioinformatics Visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, no. 4, pp. 443–456, 2005.

[4] North, C., "Toward Measuring Visualization Insight," *IEEE Computer Graphics and Applications*, vol. 26, no. 3, pp. 6–9, 2006.

[5] Leman, S.C., House, L., Maiti, D., Endert, A., North, C., "Visual to Parametric Interaction (V2PI)," *PLoS One*, vol. 8, no. 3, pp. e50474, 2013.

[6] Endert, A., Han, C., Maiti, D., House, L., Leman, S., North, C., "Observation-level interaction with statistical models for visual analytics," in *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 121–130, 2011.

[7] House, L., Leman, S., Han, C., "Bayesian Visual Analytics: BaVA," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 8, no. 1, pp. 1–13, 2015.

[8] Carroll, J. D., Chang, J. J., "Analysis of Individual Differences in Multidimensional Scaling via an N-way Generalization of Eckart-Young Decomposition," *Psychometrika*, vol. 35, pp. 238–319, 1970.

[9] Schiffman, S. S., Reynolds, M. L., Young, F. W., *Introduction to Multidimensional Scaling: Theory, Methods, and Applications*. New York: Academic Press, 1981.

[10] MacKenzie, I. S., “Fitts’ law as a research and design tool in human-computer interaction,” *Human-computer Interaction*, vol. 7, no. 1, pp. 91–139, 1992.
[11] Jolliffe, I., *Principal Component Analysis*, 2nd ed. John Wiley and Sons, Ltd, 2002.
[12] Kruskal, J. B., Wish, M., “Multidimensional Scaling,” *Sage University Paper Series on Quantitative Application in the Social Sciences*, 1978.
[13] Plaisant, C., Fekete, J.-D., Grinstein, G., “Promoting insight-based evaluation of visualizations: from contest to benchmark repository.,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 1, pp. 120–34, 2008.
[14] Chang, R., Ziemkiewicz, C., “Defining insight for visual analytics,” *IEEE Computer Graphics and Applications*, vol. 29, no. 2, pp. 14–17, 2009.
[15] Amar, R., Eagan, J., Stasko, J., “Low-level components of analytic activity in information visualization,” *IEEE Symposium on Information Visualization (INFOVIS)*, pp. 111–117, 2005.