Bringing Interactive Visual Analytics to the Classroom for Developing EDA Skills

Jessica Zeitz Self
Nathan Self
Leanna House
Jane Robertson Evia
Scotland Leman
Chris North
Virginia Tech

Key Words: dimension reduction, multidimensional scaling, visual analytics, multivariate analysis, education

Abstract

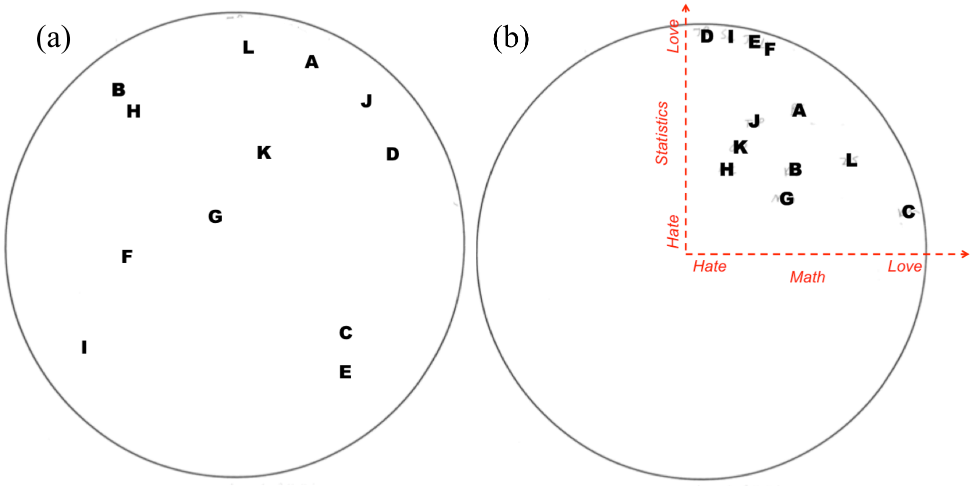
This paper addresses the use of visual analytics in education for teaching what we call *cognitive dimensionality* (CD) and other EDA skills. We present the concept of CD to characterize students' capacity for making dimensionally complex insights from data. Using this concept, we build a vocabulary and methodology to support a student's progression in terms of growth from *low cognitive dimensionality* (LCD) to *high cognitive dimensionality* (HCD). Crucially, students do not need high-level math skills to develop HCD. Rather, 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 high-dimensional insights. In this paper, 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' CD begin low and improve.

1 INTRODUCTION

The phrase “Big Data" is practically redundant. Today's datasets are big; at low cost, advanced technology has enabled almost every industry, scientific field, branch of government, etc., to collect new and more data than ever. However, datasets are just tables of numbers without humans to discover, process, reflect, and communicate information in the data (Thomas and Cook 2005; Endert et al. 2014). This means that, to learn from large datasets, humans are called upon to integrate what they know with tens, hundreds, even thousands of observations and variables at once. In practice, because of the number of dimensions they consider and the methods they use to learn from data, students tend to limit what they can learn from data. In this paper, we present evidence of this limitation and how the use of a new analytics tool may foster improved exploratory data analysis skills influencing what students learn from data. Improved data analysis methods influence EDA skills, which subsequently influences what students learn when performing an analysis. Therefore, if we improve the methods, we in turn improve EDA skills and increase what students' learn from data.

Our work was initially motivated by an observation we made while teaching an introductory lesson on data visualization to first year college students. In the lesson, students were given a dataset about themselves; i.e., the rows in the dataset represented the students and the columns were student descriptors such as Age, Average Number of Study Hours Per Week, Average Number of Texts Sent Per Week, Love of Statistics (on a scale of 0 to 100), etc. that were compiled from a survey. With these data, the students were asked to use their intuition and map the class in a provided circle in which distance between students reflected relative difference between them. For example, students mapped close to each other in the circle were to be considered more similar to each other than students mapped farther apart, according to the map maker (i.e., the student who created the map). Every student created his/her own map and the purpose was for each student to intuit the meaning of distance for himself/herself while considering the variables before technical methods for visualization were introduced, such as Principal Component Analysis (PCA) and Multidimensional Scaling (MDS) (Kruskal and Wish 1978; Jolliffe 2002). Results from this mapping exercise were surprising. The highest number of variables used by any one student was three and almost all students used only two. For example, one student pretended that there were two axes and plotted Love of Math versus Love of Statistics (see Figure 1b). To provide a comparison, an instructor completed the same exercise and considered seven variables in his map: Age, Exercise, Study Habits, Alcohol Consumption, Politics, Hours Sleep, and Television Watching (see Figure 1a).

Figure 1. Example Maps



An instructor's (a) and student's (b) map of the class participants; each letter represents either a student or professor and the red axes depict what the student stated he was thinking when creating the map.

We describe what we observed by coining the phrase *cognitive dimensionality* (CD). Even though the students had access to 25 variables, the students consistently only considered two; they had *low cognitive dimensionality* (LCD). There are several reasons that may explain the observed LCD that range from the exercise itself (e.g., unclear and subjective) to the students' cognitive abilities. However, based on experience and feedback from students, we made clarifications to the exercise and re-implemented the task in other settings, including an undergraduate class of approximately 60 students. We observed the same LCD phenomenon.

In the face of LCD, it stands to reason that the exploratory data analysis (EDA) skills of students may also suffer; LCD hinders students' approaches for discovering complex insights (e.g., insights which rely on more than two variables) in high-dimensional data. If so, teaching comprehensive EDA methods may boost CD and/or visa versa. Alas, the process of data exploration is hard to teach in its own right. It relies on problem solving skills that can be application specific, personal, and creative in that they often rely on selecting and merging several analytical methods (computational, statistical, and visual) together to discover different types of structure and features in the data. Furthermore, some might argue that students require upper-level mathematical skills to apply multi-dimensional EDA methods, including PCA and MDS projection methods. However, supported by education theory (Section 3), we assert that technical, mathematical skills are not required to learn from data, to develop new EDA skills, and to increase CD (Leman and House 2012). Rather, with the right tools and opportunities students may construct their own forms of EDA to learn from and appreciate high-dimensional data.

In this paper, we performed a three-part observational study (Section 4) to assess the impact of an interactive analytics tool called Andromeda on EDA skills including the ability to think high-dimensionally. Andromeda is a tool that enables users (e.g., students) to explore data visually, based on their own conjectures. That is, Andromeda responds to user feedback and presents multiple two-dimensional data projections that are solved from sequential, user (e.g., student) interactions. We were motivated to bring Andromeda to the classroom because constructivist learning strategies (Piaget and Inhelder 1969) are grounded in offering experiences for students to repeatedly assimilate and/or accommodate new information. Students assimilate new information when it fits with their existing knowledge and develop ways to accommodate new information otherwise. When students use Andromeda, data explorations become student-centric experiences during which students are called upon to reconcile – either assimilate or accommodate – repeatedly the relationship between two-dimensional visualizations and a high-dimensional dataset. Insights from data result when reconciliations are successful.

In the three-part observational study, we evaluate similarities and differences in written insights made by students when a) they do not use an analytical tool, b) they use a statistical package, such as R or Matlab and c) they use Andromeda. We use standard methods in visual analytics (Amar, Eagan, and Stasko 2005; North 2006) to define and measure insights, as well as to declare the tasks the students used to make them. Variation in the tasks informs us about the EDA skills of our study subjects. For example, results from the study suggest that students do, indeed, start with LCD and limited EDA skills, but have the ability to improve when using software. Student CD and diversity in EDA increases after using Andromeda.

The remainder of the paper proceeds as follows. In Sections 2 and 3 we highlight background research, including Andromeda, and educational theories that inspired us to consider that Andromeda may influence EDA skills such as CD. In Sections 4 and 5 we describe, respectively, the observational study, including the three-part assignment, and a summary of the results from our study. We conclude with a discussion in Section 6.

2 CURRENT WORK IN VISUAL ANALYTICS: ANDROMEDA

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 and improve CD.

2.1 Visual Analytics

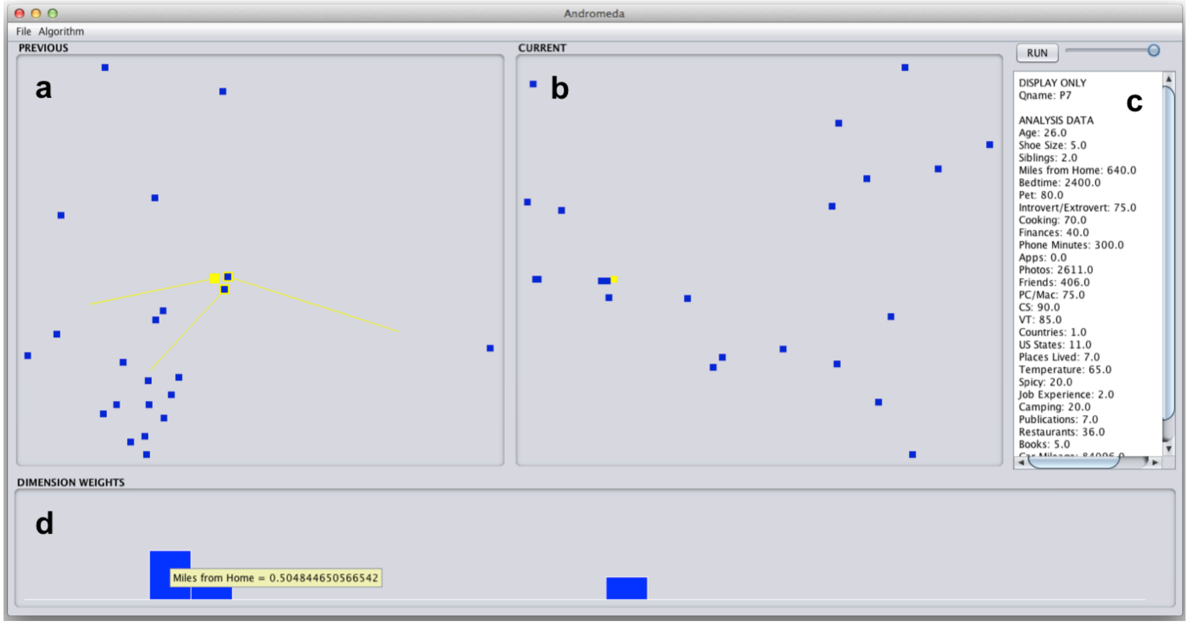
Visual Analytics is the "science of analytical reasoning facilitated by interactive visual interfaces" (Thomas and Cook 2005). Thus research is devoted to developing methods by which humans may visualize and interact with data in ways that make sense to them; humans are a central component in the process of making sense and learning from data. There are many ways in which humans may interact with data (Saraiya, North, and Duca 2005; North 2006). For example, when provided two-dimensional visual projections of data, three types of interaction are defined in Leman et al. (2013): Surface Level, Parametric, and Visual to Parametric. Surface level interactions are comparable to read-only actions include highlighting, zooming, and filtering observations in scatterplots that do not update the model. Parametric interactions enable analysts to adjust specifications for parameters in models that create the visualizations. Visual to Parametric Interaction (V2PI) allows analysts to indirectly adjust parameters of the model by adjusting the visualization.

V2PI (Leman et al. 2013) is an example of Object Level Interaction (Endert et al. 2011) and is a deterministic version of Bayesian Visual Analytics (BaVA) (House, Leman, and Han 2010). Using a dynamic 5-step process, V2PI quantifies surface level interactions to adjust tunable parameters in the display-generating model to create new visualizations. We refer to the surface level interactions as cognitive feedback (about the model) and a parameterized version as parametric feedback. With V2PI, users communicate their ideas and judgments about the data through an easy-to-interpret, low-dimensional visualization. These judgments are quantified in a form to update an underlying model, and thus, visualization. Users never have to leave the visual domain of a dataset to explore its high dimensions. When V2PI is committed multiple times, users explore high-dimensional data in a sequence that parallels their mental processing of the data. For an example of V2PI (as well as the other forms of interaction), consider the use of Andromeda.

2.2 Andromeda

Andromeda enables all three forms of interaction while providing visualizations based on Weighted Multidimensional Scaling (WMDS) (Carroll and Chang 1970; Schiffman, Reynolds, and Young 1981). WMDS is a linear projection method that includes one parameter, i.e., weight, per data dimension to reflect its relative importance in a visualization. In turn, pairwise relative distances between observations reflect relative similarities/differences between observations, particularly in the dimensions with the largest weight. Figure 2 provides a screenshot of the Andromeda interface which includes two versions of a WMDS scatterplot (labeled a and b in Figure 2), a list of the variables (labeled c in Figure 2), and a bar graph of designated weights for each variable in the dataset (labeled d in Figure 2). There are two data visualization panels (labeled a and b in Figure 2) of WMDS plots, to which we refer as the previous and current views, respectively. 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. In this section, we describe how to use each component of the Andromeda interface and conclude with how Andromeda supports student-centric data explorations.

Figure 2. Andromeda Interface



DISPLAY ONLY
Qname: P7

ANALYSIS DATA
Age: 26.0
Shoe Size: 5.0
Siblings: 2.0
Miles from Home: 640.0
Bedtime: 2400.0
Pet: 80.0
Introvert/Extrovert: 75.0
Cooking: 70.0
Finances: 40.0
Phone Minutes: 300.0
Apps: 0.0
Photos: 2611.0
Friends: 406.0
PC/Mac: 75.0
CS: 90.0
VT: 85.0
Countries: 1.0
US States: 11.0
Places Lived: 7.0
Temperature: 65.0
Spicy: 20.0
Job Experience: 2.0
Camping: 20.0
Publications: 7.0
Restaurants: 36.0
Books: 5.0
Car Mileage: 84006.0

Miles from Home = 0.504844650566542

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.

2.2.1 Surface Level Interaction in Andromeda

To deepen their understanding of what students see, as well as decide whether to interact, users may hover over data points. Hovering is a surface level interaction. When a user hovers over a data point (in either view), Andromeda responds in two ways. One, the values for each variable that describe the hovered point appear in the side Data Panel (labeled c in Figure 2). Two, because the views are linked, hovering over a data point in one view will highlight the same data point in the other view. This provides a means for users to make sense and develop their own interpretation of the difference between the two visualizations.

2.2.2 Parametric Interaction in Andromeda

Andromeda enables users to adjust the weights of the model. To do so, users change a weight parameter of WMDS by clicking on the top of a bar and dragging it up or down to increase or decrease the variable's relative weight to other variables (MacKenzie 1992). Recall, the larger the weight the more influence the variable has in a visualization than others with lower weights. 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.

2.2.3 Visual to Parametric Interaction in Andromeda

The advantage of V2PI is that users who have knowledge or a conjecture about a subset of data points may use it to inform WMDS without making global statements, such as adjusting weights directly. In the current view (labeled b in Figure 2), users may apply V2PI by dragging observations to change pairwise distances between them. Recall that, in WMDS visualizations, relative pairwise distances between points convey relative pairwise similarities/differences. For example, two points close together are more similar to each other than two points far apart. The two-dimensional coordinates of the points are solved by WMDS, given assigned weights for each dimension. V2PI inverts WMDS in that, given new coordinates for three or more observations, Andromeda solves for new weights; cognitive feedback is communicated in the form of changing some pairwise distances between observations by dragging and parametric feedback includes new weights solved from the coordinates of the dragged points. Andromeda uses the new weights to re-implement WMDS for the entire dataset and display a new visualization; i.e., a new "current" view.

Andromeda communicates the updated values of the weights by the bar graph in the bottom panel (labeled d in Figure 2). The maximum height of any bar is one; the taller the bar for a variable, the more weight it has in the visualization. For example, in Figure 2, the previous view (labeled a) includes yellow lines to show from where the points were dragged to create both the current view in panel b and the bar graph in panel d. Notice all observations relocated in panel b and the variable with the largest weight is Miles From Home and has the weight 0.505.

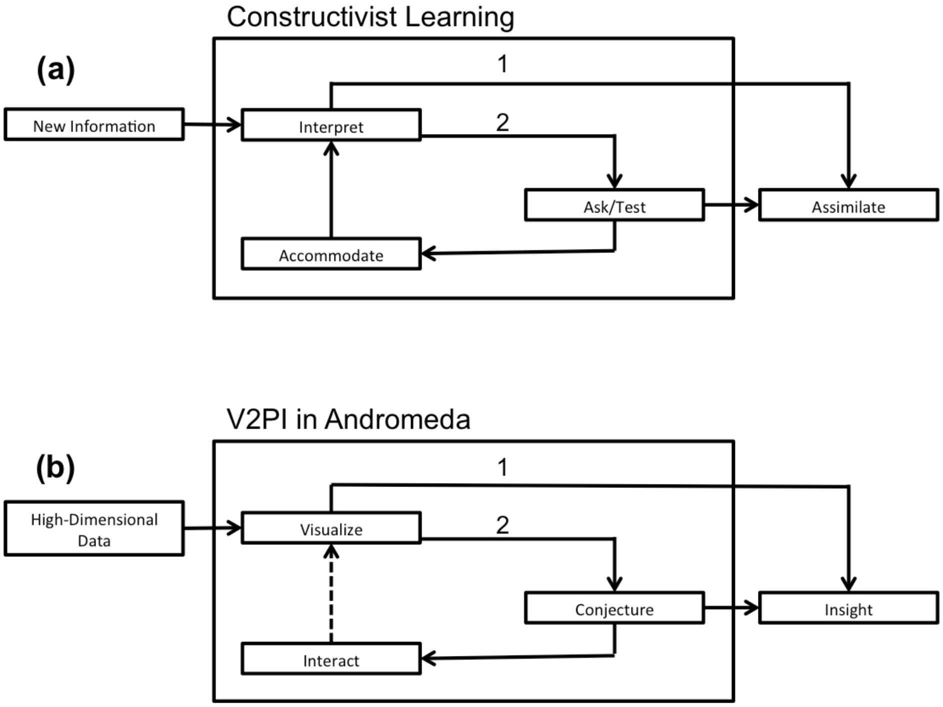
3 CONSTRUCTIVIST AND EXPERIENTIAL LEARNING STRATEGIES WITH ANDROMEDA

The parallels between how students naturally learn and the V2PI process make Andromeda ideal for teaching EDA skills and related concepts (e.g., dimensionality) using a constructivist approach. With a subtle, but important difference, students assess and learn from data using Andromeda in the same way that they assess and learn from the world around them.

Jean Piaget believed that children construct knowledge by exploring the world and employing psychological process to make sense of their environment (Piaget and Inhelder 1969). There are two key processes in the journey to constructing knowledge and understanding: assimilation and accommodation. Whether one assimilates or accommodates new information is dependent upon personal existing schema, or prior knowledge, and the generation of new ideas. Figure 3a diagrams Piaget's learning process with assimilation and accommodation central to constructivist learning. It starts with students' interpretations of new information. In the face of new information, students decide whether it exists within current schema or not. When it does (choice 1 in Figure 3a), students assimilate. However, when it does not (choice 2), students often make and evaluate inquiries to ultimately accommodate. In turn, there is a new interpretation of the information that follows – either an adjusted schema or new schema that formed from accommodation. Students may either assimilate the new interpretation or continue the loop for accommodation.

In the context of data exploration, users of V2PI in Andromeda follow the same learning process, shown in Figure 3b. Data are summarized by a visualization that may or may not make sense to the user. When it (or aspects of it) does make sense (labeled (1) in Figure 3b), the user makes insights. However, when it does not make sense or the user wants to learn more, the user interacts by adjusting the location of observations. In turn, Andromeda, not the user, translates the adjustments in a parametric form to update the visualization (i.e., interpretation of the data) and the user may either make insights or repeat the process to learn more.

Figure 3. Diagrams of Learning



Diagrams of learning (a) and learning with V2PI with Andromeda (b): Given either an interpretation of new information or a Visualization of high-dimensional data, students have two choices, labeled 1 and 2 in each diagram. The first choice results in the discovery of something new or reinforcement of what is known, whereas the second choice continues the students in the learning process.

The translation made by Andromeda in Figure 3b is represented by a dotted arrow and highlights the important difference between typical learning and learning with Andromeda. By parameterizing interactions, Andromeda communicates back to the user the meaning or impact of their interactions, relative to *all* of the data when new visualizations are created. This is highly specific, automated, just-in-time guidance for users. In other forums, such guidance has shown useful in experiential learning activities (Edelson 2001) . Also, students are called upon iteratively to not only make decisions about the data, but also their own translated feedback and what it means in both low- and high-dimensional spaces. The importance of decision making with data and the role of technology, in general, is echoed by Garfield and Ben-Zvi. Fundamental to EDA is the presence of creative, problem solving or reasoning skills that are fostered in 'Statistical Reasoning Learning Environments' – particularly those environments that use technology to lighten students' cognitive burden and allow them to "focus [on] the more important task of learning how to choose appropriate analytic methods,… interpret results,…[and] visualize concepts" (Garfield and Ben-Zvi 2009).

4 METHODS

We implemented a set of three iterative assignments to assess students' cognitive dimensionality (CD) and other 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 the students themselves and colleagues. The final dataset included 23 observations and 27 variables.

Similar to what was described in the introduction, the dataset provided for analysis was collected from the students by issuing a survey with 27 personal questions that had numeric answers. Exemplary questions include:

> 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*.

4.1 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.

4.1.1 Manual

The first assignment was open-ended and intended for us to establish a baseline for students' current CD and 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.

4.1.2 Statistical Computing Environments

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) (Jolliffe 2002) and either unweighted or weighted multidimensional scaling plots (MDS or WMDS, respectively) (Carroll and Chang 1970; Kruskal and Wish 1978; Schiffman et al. 1981). 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.

4.1.3 Andromeda

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.

4.2 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. Hence, we refer to previous research to translate the assignments into a structured form that identifies and quantifies the quality of insights. We hypothesize that the methods used for analysis influences EDA skills which in turn influences 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.

Saraiya et al. (2005) define insight as a unit of discovery such that an insight is a distinct observation by a user. Though their study was in the bioinformatics domain they stress that their definition and methodology can be applied in any domain. Plaisant et al. offer that an insight may simply be a "nontrivial discovery about the data" (Plaisant, Fekete, and Grinstein 2008). In this paper we adopt the definition of insight as *an observation by a student about the data*. Most students denoted separate insights by a bulleted list within each assignment. Multiple sentences might comprise one single insight with a single conclusion. Some students included a description of the techniques and processes they used to discover their insights. For our analysis, we aggregated all insights provided by the 18 students across the three assignments. Across all students and assignments, 257 insights were made.

Several researchers have worked on defining characteristics of insights, such as Saraiya et al. (2005), North (2006) and Chang and Ziemkiewicz (2009). Collectively, they 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. Arguably, subjective characteristics, such as relevance or usefulness in a particular domain are hard to measure systematically. We focus our attention on those that are quantifiable and apply the following metrics for each insight:

- *Dimensionality*: Each dimension that was explicitly listed in an insight is tallied. This way, dimensions that a student voluntarily decided to name are treated as dimensions that were important in the generation of that insight. Insights that mentioned no dimensions were given a zero for this measure. The more dimensions mentioned, the more complex the insight.
- *Cardinality*: Each data point that is explicitly listed in an insight is counted. Insights that do not mention any particular data points count as a zero. The higher the cardinality, the more complex the insight.
- *Relationship cardinality*: Most insights in our study involved comparisons of points. We categorize the nature of the relationship such as one-to-many, one-to-all, one-to-one, etc. Insights that are merely one-to-one are less complex than others which are one-to-many and one-to-all.

To further evaluate insights, as well as EDA skills, we also consider the tasks completed to make insights (Amar et al. 2005). 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 or EDA endeavors taken to make the insight. To do so, we use the analytic task taxonomy of low-level components outlined in (Amar et al. 2005) and identify the following tasks taken for each insight:

- *Retrieve value*: We consider each explicit listing of a numerical value, either raw or derived, as a retrieve value task. For example, in the manual assignment, several insights listed computed similarity scores or data values from raw data dimensions. Each unique value that appears in an insight is tallied as one retrieve value task.
- *Filter*: As described in (Amar et al. 2005), filter tasks involve finding all data that satisfy a given condition. For example, insights that listed students that are older than *x* or have 0 siblings, for example, would have contained filter tasks, but no insights did such things.
- *Compute derived value*: Compute derived value tasks were tallied for any insights that involved a derived value, regardless of whether derived by hand, code, or output (e.g., by Andromeda). For insights that report more than one derived value, such as comparisons between MDS plots, multiple compute derived value tasks were tallied.
- *Find extremum*: Find extremum tasks were counted when an insight dealt with some number of the top or bottom values of any single dimension. For example, in many cases, insights were of the form person P is *most* similar to person Q; the extremum of person Q's similarity score is reported. Cases such as persons P, Q, and S are most similar to person T are also of this task.

- *Determine range*: Student conclusions that involved describing the range of values in a dimension are counted as determine range tasks. These only occurred in the manual assignment.
- *Characterize distribution*: Insights that describe the general pattern of all data points over a dimension or space are counted as characterize distribution tasks. For example, insights from the manual assignments often described skewness in histogram plots, whereas others tended to describe the spread of observations in data projections.
- *Find anomalies*: These tasks were tallied for insights that describe unexpected observations, e.g., statistical outliers in the raw data or in derived values.
- *Cluster*: Insights that identified members of clusters or relationships between clusters were counted as cluster tasks.
- *Correlate*: Correlate tasks were assigned to insights when a correlation between two dimensions was discovered.

Insights may be the result of one or more than one analytical task. To exemplify how we quantify insights and determine tasks, consider the following two examples.

> 1) "[Name redacted] has an exorbitant amount of camping trips." (manual assignment)
>
> 2) "When the class of four outliers consisting of nodes of [four names redacted] are brought closer to the main cluster to integrate them all, [two names redacted] move even farther from the main cluster. It can be seen that attributes such as Q14 (PC v/s Mac), Q21 (Degree of spicy food liked) and Q24 (Number of publications as an author) gain more weights and hence become more prominent in order to accommodate the new cluster." (Andromeda assignment)

The first insight results from one task, *find anomalies*. The fact that the student mentioned has taken many camping trips makes him an anomaly or outlier in the dataset to the student who wrote this insight. We classify this insight as dimensionality one and cardinality one with relationship cardinality being one. The second insight contains three tasks: *compute derived value*, *find anomalies*, and *cluster*. This insight was based on the two-dimensional projection visualized in Andromeda. Since the visualization is created from the raw data, we classify the projection as derived data. Therefore the insight includes a *compute derived value* task. The student specifically mentions outliers and clusters of points which constitutes the *find anomalies* and *cluster* tasks, respectively. This insight has dimensionality of three and cardinality of six with relationship cardinality being many-to-many. The second insight is of higher complexity than the first insight. We infer that the student who gained the second insight has more developed EDA skills than the student who gained the first insight.

In this observational study, we also assess the EDA skills of students by documenting general techniques students used to inspire tasks (which resulted in specific insights). Defining techniques was a challenge, but resolved to two main techniques that reflect directly how students assimilate and accommodate new information. These techniques include:

- Comparing before/after interaction
- Using outside knowledge

We documented these techniques directly from the insights as well as from written narratives from each student reflecting on her process analysis. The second example insight above denotes a before/after interaction. The student explains that clustering one set of points causes another set of points to become outliers. Although, before/after interaction is arguably easier for assignments that rely on Andromeda, it is feasible for all three assignments. For example, students may manually or use statistical software to delete variables or filter observations to assess changes in data summaries.

5 RESULTS

To assess potential gains over the course of the three assignments, we summarize the insights, tasks and techniques. We describe the implications of our results in Section 6.

5.1 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 and we discuss each row of the table in detail.

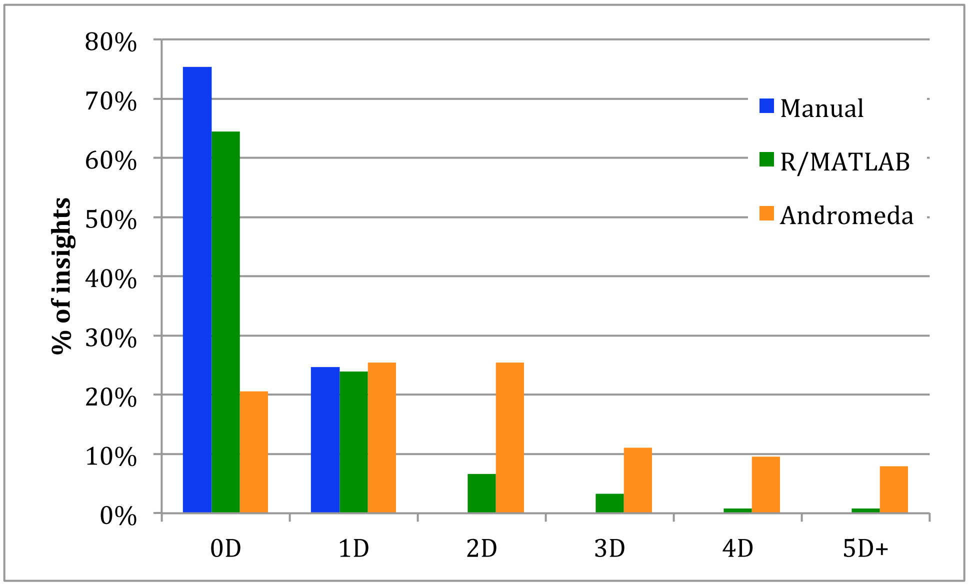
Table 1. Insight Trends

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

5.1.1 Dimensionality

Figure 4 graphs dimensionality by assignment. Of the manual insights, 75% did not refer to any dimension. Most of these zero-dimensional insights included finding extremums 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 to a dimension was made. The remaining 25% of manual insights only considered one dimension. These insights focused on the anomalies and extremums of a single dimension. For example, an insight considering one dimension stated who in the dataset was the youngest (age dimension) or who wore the largest shoe (shoe size dimension).

Figure 4. Dimensionality of Insights



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

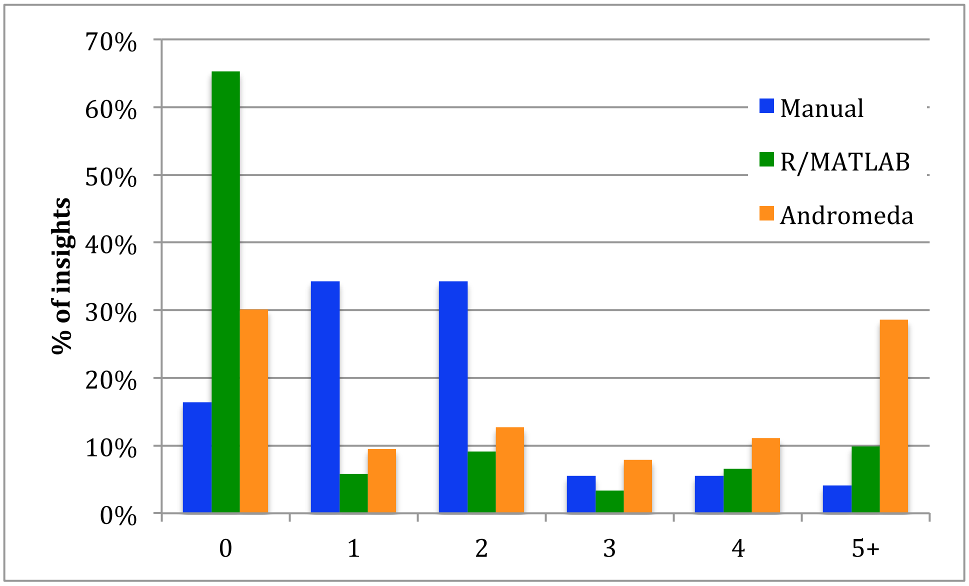
Similar to manual insights, 64% and 24% of statistical environment insights did not reference any dimensions or referenced one dimension, respectively. Of the zero-dimensional insights, many focused on characterizing the distribution of derived values from MDS, whereas, most of these one-dimensional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

Relative to the manual and statistical environment assignments, 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, as depicted Figure 4. Even though the percentages are small, we see a shift in the complexity of insights when using Andromeda. 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.

5.1.2 Cardinality and Relationship Cardinality

We categorized the insights based on cardinality and relationship cardinality. Our categories for relationship cardinality consisted of no cardinality (meaning no reference to any particular data point or group of points), one-to-one, one-to-many, one-to-all, many, many-to-many, many-to-all, and all. Figure 5 and Figure 6 list the percentages per assignment.

Figure 5. Cardinality of Insights

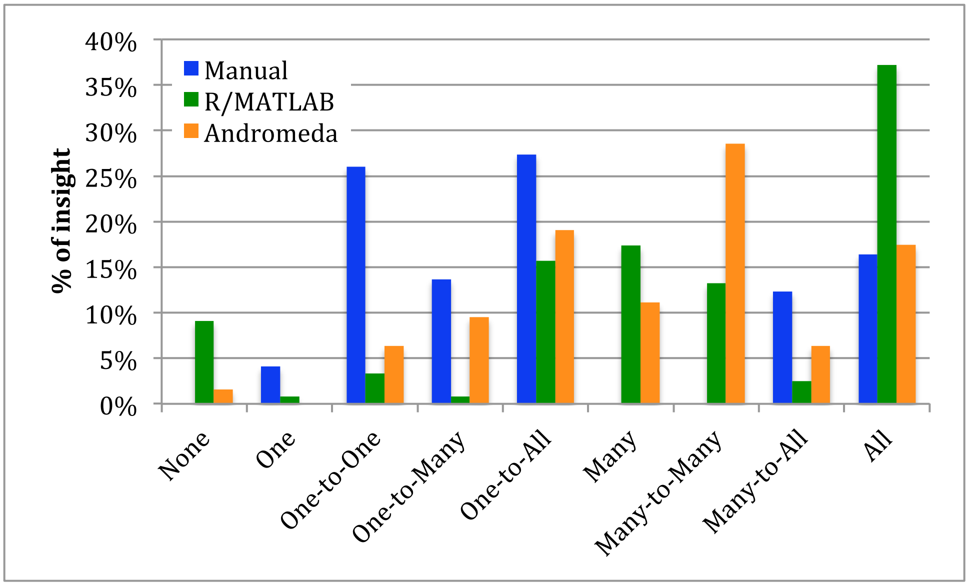


Percentage of insights from each assignment against the number of observations explicitly mentioned in each insight.

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. For example, a student stated the he is fairly compatible with most of his classmates. This insight was based on the student having overall high similarity values with all pairs.

As seen in Figure 6, 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 5, 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.

Figure 6. Relationship Cardinality of Insights



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

5.1.3 Task Diversity

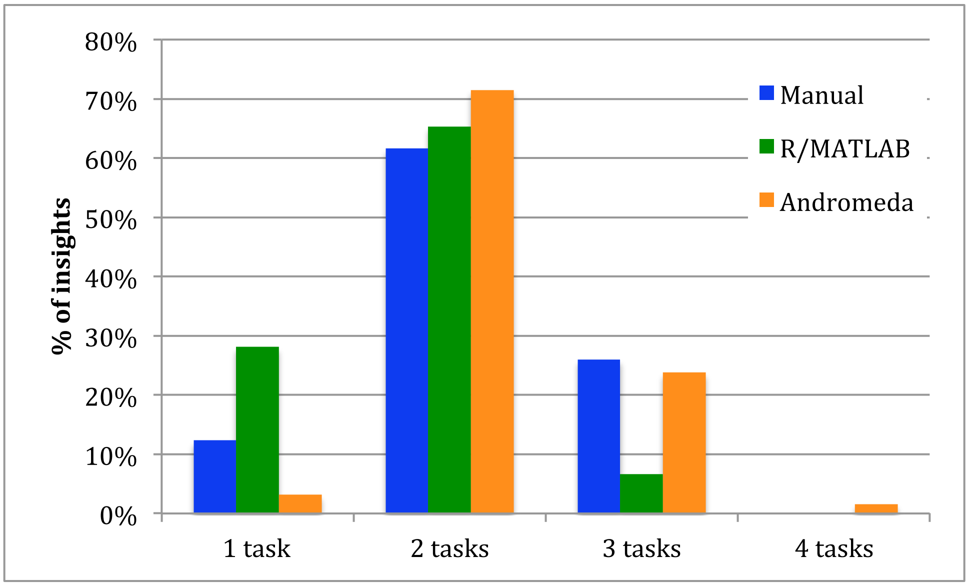
Insights made using different tasks reflect both the complexity of insights and the EDA skills of the students. Again we summarize differences in the use of components for the manual, statistical environment, and Andromeda assignments.

The 73 manual insights, 121 statistical environment insights, and 63 Andromeda insights contained 156, 216, and 141 individual tasks respectively. Summaries of the tasks are displayed in Figure 7 and Figure 8. In Figure 7, 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 8 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 (see Figure 8); 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 this particular dimension." These insights also included the highest or lowest top two or three persons based on a single dimension.

Figure 7. Number of Distinct Tasks Per Insight



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

For the statistical environment assignment, the second most prevalent task within this context was *characterize distribution*. Of the 121 insights, 28% contained a characterize distribution task (Figure 8). 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. They explained that the data points within each group are similar, however, they would not provide evidence as to how the data points are similar.

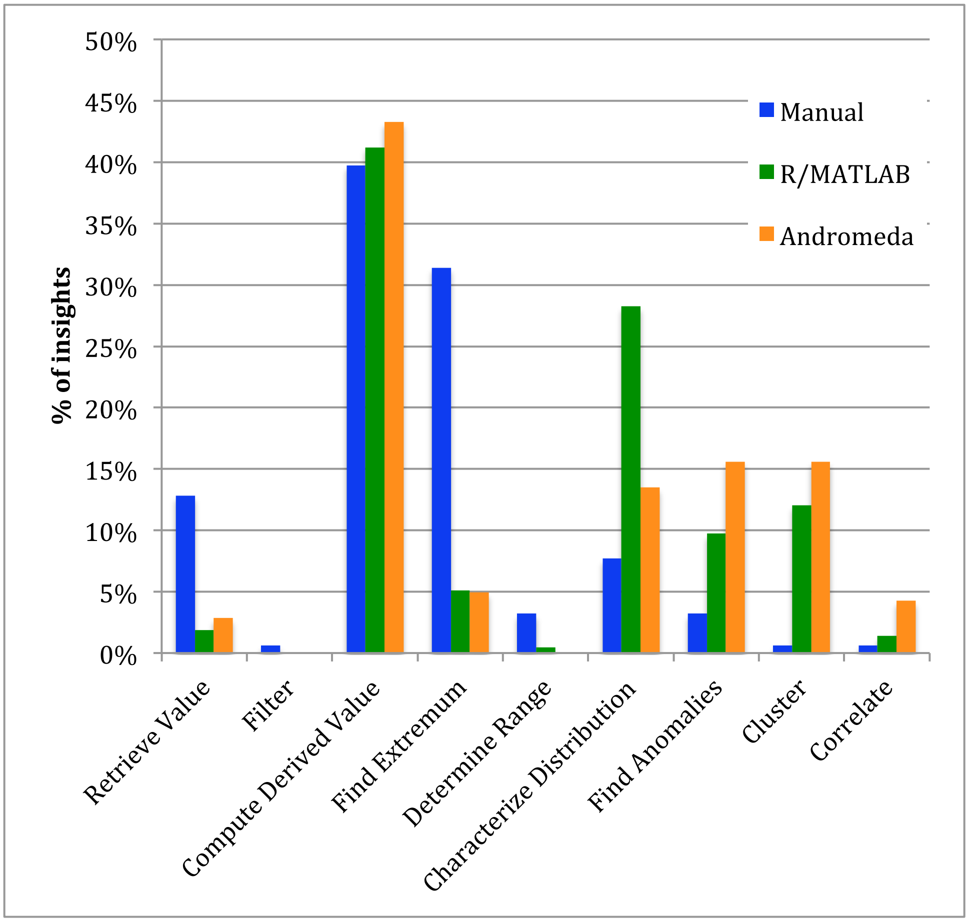
The distribution of tasks for Andromeda insights is comparable to that of statistical environment. However, *finding anomalies* (16%) and *clustering* (16%) tied for being the second most useful task with the use of Andromeda when making insights. In particular, we note clustering as a common task, as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

5.2 Techniques

We took note of two techniques (*before/after interaction* and *outside knowledge*) students used for analysis. As expected, students tended to use techniques that were easiest for a given tool. In Andromeda, *before/after interaction* would define the comparison of two spatial layouts with interactions between the two. Of the Andromeda insights, 44% took advantage of the before/after technique as opposed to 3% of manual insights and 12% of statistical environment insights (Figure 9). Only one student utilized a before/after interaction for the manual assignment. Two of those insights explained what happened to the similarity values when one dimension was removed from the dataset.

As for the use of *outside knowledge*, students often supported their insights with outside knowledge not included in the dataset: 3% of manual insights, 6% of statistical environment insights, and 8% of Andromeda insights included outside knowledge (Figure 9). Such knowledge might explain the difference between two clusters being international students and American students when nationality of each student was not included in the dataset. One student described a large split between two clusters to be based on non-student versus student roles in the dataset. Another student based the similarity of two students on them participating in the same research group in academia. By bringing in this outside knowledge, the students were connecting their analysis to the real world.

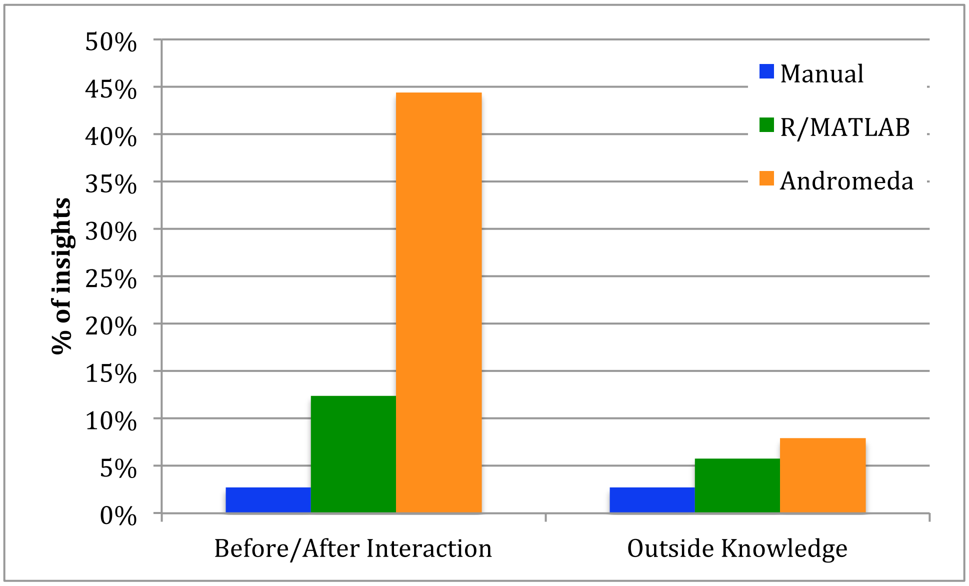
Figure 8. Diversity of Tasks



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

Also, often students grouped dimensions based on an overarching description. This occurred across all assignments, but predominantly in the Andromeda assignment. For example, most students grouped dimensions such as age, shoe size, miles from home and number of siblings because they considered them unchangeable attributes, whereas they grouped dimensions such as preferred outdoor temperature, love of computer science, and PC versus Mac since these dimensions are opinion based. With these types of dimension categorizations, student would make overarching insights about the dataset. For example, multiple students combined all dimensions having to do with travel (number of US states visited, number of countries visited, and number of places lived) and concluded that most students in the class were well travelled.

Figure 9. Analysis Techniques



The bar chart depicts the percentage of insights that were developed using a before/after interaction or outside knowledge.

6 DISCUSSION

We hypothesized that, relative to other assignments, insights derived from the use of Andromeda would be more complex and reflect both an increase in CD and improved EDA skills. Findings from the observational study support our hypothesis. Students successively gained higher CD as they progressed from manual computation to static visual encodings and from statistical environments to interactive, Andromeda spatializations. Additionally, the tasks and techniques employed with Andromeda reflected more advanced EDA skills than those with the manual and statistical environment assignments.

With the manual approach, students tended to discuss the relationships between two individuals. Students, specifically 10 of the 13 who provided written insights, adopted an egocentric perspective where they focused on the similarities and differences between themselves and a single other individual. Of the 13, 12 students did break the egocentric approach and branched out to discuss who was most similar or dissimilar from everyone or from some third person. However, insights still solely focused on the similarity and dissimilarity of data points. Insights in the manual assignment rarely identified clusters of students with similarities and usually did not compound on themselves towards deeper insight. Most students were concerned with the extremes in a single dimension such as oldest or most similar with respect to one dimension. When using Andromeda, students focused less on themselves and more on clusters of data points. If students did reference themselves, it tended to be within a cluster of people and identified multiple dimensions in support of their insights, in accordance with the dimension weights reported by Andromeda.

In the statistical environment assignment, students progressed from reporting extremes to describing overall trends for specific dimensions, however still focusing on one dimension. All 17 students that wrote down insights within their assignment continued to ask basic comparison questions and describe the histograms and dimensionally reduced plots they made. However, a few students began to increase the weights of a subset of dimensions in the weight vector for the PCA and MDS plots. Andromeda insights tended to characterize the entire distribution in reference to many dimensions. While describing static MDS plots generated by statistical environments, students identified members of clusters without offering suggestions as to why the data points might be clustered. In contrast, insights from the Andromeda assignment tended to offer explanations for which dimensions caused a given clustering.

Andromeda insights show a better understanding of the data. The tools interactions provoked more exploratory analyses that focused on testing hypotheses rather than on simplistic visualization summarization. Students did not follow one line of inquiry, but pursued alternative viewpoints which helped to thwart the tunneling of their thought processes. Students tended to focus on clusters of points instead of single data points. Of the 16 who recorded insights within their assignment, 8 students clustered data points of students they thought to be similar for validation. Andromeda insights about clusters have a deeper understanding about why those clusters formed. For example, students would inspect the weight distribution to make conclusions about which dimensions were important for this clustering. Several times students moved two to three points on top of each other and the resulting visualization placed the points farther apart. Students deduced that in the ways the selected points are similar, the data points are actually more similar to completely different points. Students had internalized that by moving a subset of points. Andromeda is actually arranging all the data points based on the similarity of the subset even if the subset is not similar. In one case, a student discovered two data points that were only similar in one dimension. The student claimed this insight based on the resulting spatialization showing all data points in a line with one high contributing dimension.

Coercing outliers into main clusters was a unique process used in the Andromeda assignment. During their analysis, 3 out of the 16 students who recorded insights when using Andromeda forced outliers into a bigger cluster to see what dimension weights it took for the outliers to be similar to another set of data points. For example, a student moved four outlying data points closer to a cluster and discovered these four students were similar to the main cluster based on PC versus Mac, food spiciness and number of publications. This process is difficult to replicate with the manual or statistical environment tools.

Students also performed PI (increasing weights in the bar chart) to see which groups of students were most similar along a subset of dimensions. Of the 16 students who recorded insights during the Andromeda assignment, 5 students grouped dimensions based on user-defined categories. Two students increased the weights of all dimensions having to do with technology (number of apps, number of phone minutes, PC versus Mac, number of social network friends, and love of computer science) to discover trends within the class. As stated above, students did perform PI during the statistical environment assignment as well, but the interaction was more natural in Andromeda.

These findings suggest that students' mental models of the dimension reduction technique and the interaction associated with V2PI are fairly accurate. The interactive methods introduced in Andromeda (e.g., V2PI) bolster students' EDA skills. These skills include higher cognitive dimensionality, higher cardinality, a mix of relationship cardinality, and the ability to take advantage of a diversity of combined tasks. By utilizing these new skills students generate more complex insights.

We acknowledge that our study has limitations compared to a formal controlled experiment. CD is quantifiable and because of our baseline assignment(s) we identified a potential problem in EDA education. Students, even at the graduate-level in computer science, tend to think in low dimensions. Our observational study further supports the role technology can play in the classroom to foster experiential learning and improve how we teach EDA. The assignments did not specifically ask for more complex conclusions compared to previous assignments, yet student insights did increase in complexity paralleling the complexity of the tools. Albeit 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 assignments; but, 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.

7 CONCLUSION

This paper introduced the concept of cognitive dimensionality (CD) relative to data dimensionality and evaluated CD, an exploratory data analysis (EDA) skill. The analysis process of our interactive tool, Andromeda, mimics that of the learning process strengthening its usability in the classroom. We conjectured that a student's CD and other EDA skills would increase in complexity provided a tool such as this. To support this conjecture, we presented a classroom study using a series of assignments to assess changes in student analyses provided varying analytical tools. The contributions of our study are as follows:

- Students by default demonstrated low cognitive dimensionality in the baseline assignment.
- When provided better tools, students increased their cognitive dimensionality and other EDA skills.
- Students found more complex and higher dimensional insights with these tools.
- The Andromeda tool that supports visual-to-parametric interaction (V2PI) helped

students find novel high-dimensional insights.

Our contributions will lead future studies in furthering the research of cognitive dimensionality and of education in data analytics, as well as, the building of better tools for these purposes.

Acknowledgements

This work was supported by NSF grants DUE-1141096, IIS-1218346 and IIS-1447416.

References

Amar, R., J. Eagan, and J. Stasko (2005), "Low-Level Components of Analytic Activity in Information Visualization," in *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, 111–17, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1532136.

Carroll, J D and J J Chang (1970), "Analysis of Individual Differences in Multidimensional Scaling via an N-Way Generalization of Eckart-Young Decomposition," in *Psychometrika*, 35, 238–319.

Chang, R and C Ziemkiewicz (2009), "Defining Insight for Visual Analytics," in *IEEE Computer Graphics and Applications*, 29(2), 14–17, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4797511.

Edelson, DC (2001), "Learning‐for‐use: A Framework for the Design of Technology‐supported Inquiry Activities," in *Journal of Research in Science Teaching*, 38(3), 355–85, http://onlinelibrary.wiley.com/doi/10.1002/1098-2736(200103)38:3%3C355::AID-TEA1010%3E3.0.CO;2-M/pdf.

Endert, Alex, Chao Han, Dipayan Maiti, Leanna House, Scotland Leman, and Chris North (2011), "Observation-Level Interaction with Statistical Models for Visual Analytics," in *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, IEEE, 121–30, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6102449.

Endert, Alex, M. Shahriar Hossain, Naren Ramakrishnan, Chris North, Patrick Fiaux, and Christopher Andrews (2014), "The Human Is the Loop: New Directions for Visual Analytics," in *Journal of Intelligent Information Systems*, http://link.springer.com/10.1007/s10844-014-0304-9.

Garfield, Joan and Dani Ben-Zvi (2009), "Helping Students Develop Statistical Reasoning: Implementing a Statistical Reasoning Learning Environment," in *Teaching Statistics*, 31(3), 72–77, http://dx.doi.org/10.1111/j.1467-9639.2009.00363.x.

House, L, S Leman, and C Han (2010), *Bayesian Visual Analytics (BaVA)*, \url{http://fodava.gatech.edu/node/34}.

Jolliffe, Ian (2002), *Principal Component Analysis*, 2nd ed., John Wiley and Sons, Ltd, http://onlinelibrary.wiley.com/doi/10.1002/9781118445112.stat06472/full.

Kruskal, J B and M Wish (1978), “Multidimensional Scaling,” in *Sage University Paper series on Quantitative Application in the Social Sciences*.

Leman, SC and L House (2012), “Improving Mr. Miyagi’s Coaching Style: Teaching Data Analytics with Interactive Data Visualizations,” in *Chance*, 25(2), 4–12, http://www.tandfonline.com/doi/full/10.1080/09332480.2012.685362.

Leman, Scotland C, Leanna House, Dipayan Maiti, Alex Endert, and Chris North (2013), “Visual to Parametric Interaction (V2PI),” in *PLoS ONE*, 8(3), e50474.

MacKenzie, I Scott (1992), “Fitts’ Law as a Research and Design Tool in Human-Computer Interaction,” in *Human-computer interaction*, 7(1), 91–139.

North, Chris (2006), “Toward Measuring Visualization Insight,” in *IEEE Computer Graphics and Applications*, 26(3), 6–9.

Piaget, J and B Inhelder (1969), *The Psychology of the Child*, Basic Books.

Plaisant, Catherine, Jean-Daniel Fekete, and Georges Grinstein (2008), “Promoting Insight-Based Evaluation of Visualizations: From Contest to Benchmark Repository.,” in *IEEE transactions on visualization and computer graphics*, 14(1), 120–34, http://www.ncbi.nlm.nih.gov/pubmed/17993707.

Saraiya, Purvi, Chris North, and Karen Duca (2005), “An Insight-Based Methodology for Evaluating Bioinformatics Visualizations,” in *IEEE Transactions on Visualization and Computer Graphics*, 11(4), 443–56, http://www.ncbi.nlm.nih.gov/pubmed/16138554.

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

Thomas, J J and K a Cook (2005), “Illuminating the Path: The Research and Development Agenda for Visual Analytics,” in *IEEE Computer Society*, 54, 184, http://vis.pnnl.gov/pdf/RD_Agenda_VisualAnalytics.pdf.

Jessica Zeitz Self
Department of Computer Science
Virginia Tech
114 McBryde Hall (0106)
225 Stanger Street
Blacksburg, VA 24061
jzself@vt.edu

Nathan Self
Department of Computer Science
Virginia Tech
114 McBryde Hall (0106)
225 Stanger Street
Blacksburg, VA 24061
nwself@vt.edu

Leanna House
Department of Statistics
Virginia Tech
406-A Hutcheson Hall (0439)
250 Drillfield Drive
Blacksburg, VA 24061
lhouse@vt.edu

Jane Robertson Evia
Department of Statistics
Virginia Tech
406-A Hutcheson Hall (0439)
250 Drillfield Drive
Blacksburg, VA 24061
robj@vt.edu

Scotland Leman
Department of Statistics
Virginia Tech
406-A Hutcheson Hall (0439)
250 Drillfield Drive
Blacksburg, VA 24061
leman@vt.edu

Chris North
Department of Computer Science
Virginia Tech
114 McBryde Hall (0106)
225 Stanger Street
Blacksburg, VA 24061
north@cs.vt.edu