# Andromeda: Observation-Level and Parametric Interaction for Exploratory Data Analysis

**Jessica Zeitz Self, Leanna House, Scotland Leman, Chris North**
Virginia Tech
Blacksburg, USA
{jzself, lhouse, leman, north}@vt.edu

## ABSTRACT

Exploring high-dimensional data is challenging. As the number of dimensions in datasets increases, it becomes harder to discover patterns and develop insights. Dimension reduction algorithms, such as multidimensional scaling, support data explorations by reducing datasets to two dimensions for visualization. Because these algorithms rely on underlying parameterizations, they may be tweaked to assess the data from multiple perspectives. Alas, tweaking can be difficult for users without a strong knowledge base of the underlying algorithms. We present Andromeda, an interactive visual analytics tool we developed to enable non-experts of statistical models to explore domain-specific, high-dimensional data. This application implements interactive weighted multidimensional scaling (WMDS) and allows for both parametric and observation-level interaction to provide in-depth data exploration. In this paper, we present the results of a controlled usability study assessing Andromeda. We focus on the comparison of parametric interaction, observation-level interaction and a combination of the two.

### Author Keywords

Evaluation; user interface; dimension reduction; usability.

### ACM Classification Keywords

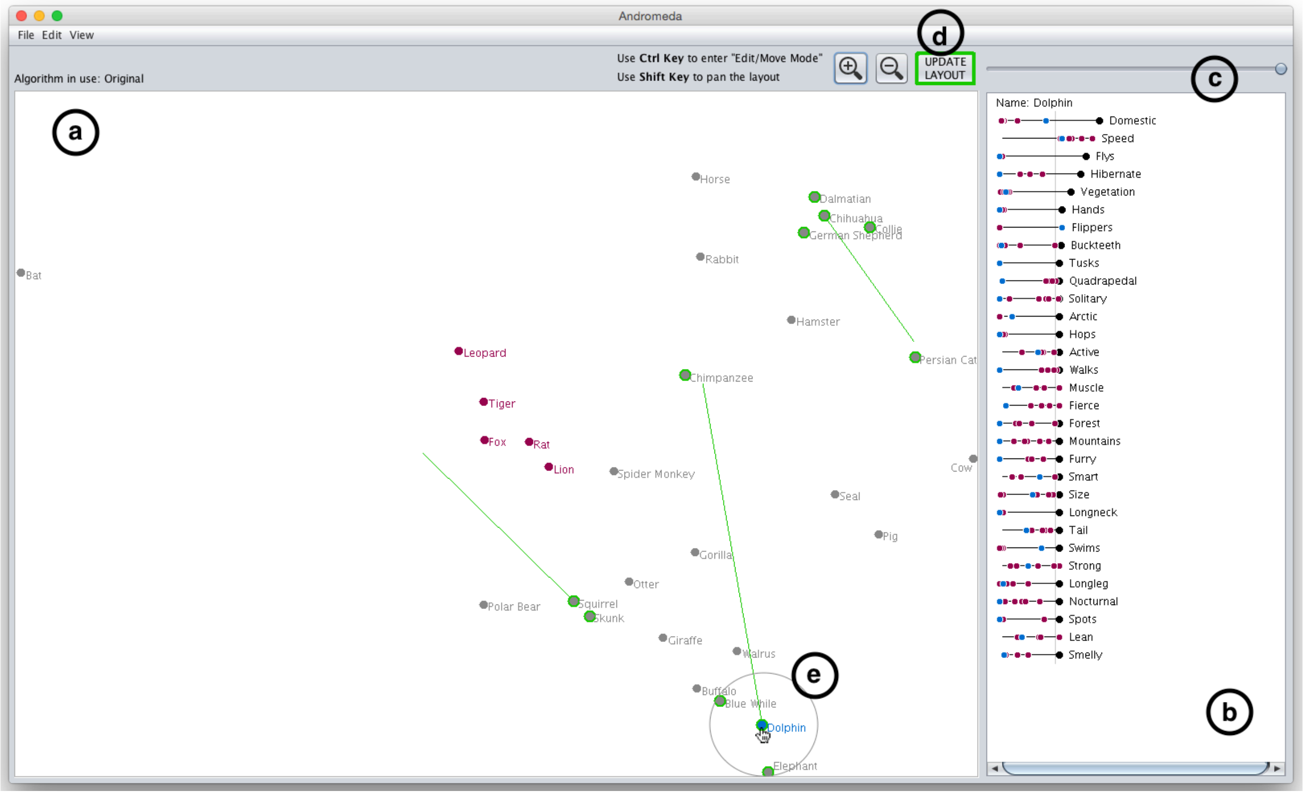
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

## INTRODUCTION

With the amount of analyzable data growing rapidly, we must develop tools to strengthen our ability to learn all that we can from data. Statistical mathematical models enable us to simplify and formalize our understanding of data. The goal of visual analytics is to design usable methods for interacting with these models to improve user data exploration techniques. For example, dimension reduction algorithms, such as Weighted Multidimensional Scaling (WMDS), project high-dimensional data onto low-dimensional (e.g. two-dimensional) space. The purpose of the algorithms is to summarize high-dimensional information in a form that is accessible to users, such as a two-dimensional graph. The visual analytics community may further improve the utility of these graphs by enhancing them with information visualization and developing tools that allow for visual interaction. In previous work, parametric and observation-level interaction (OLI) with data visualizations has been defined and shown helpful for data exploration [7,10,18]. Both forms of interaction enable users to adjust display-generating models directly and/or indirectly. In this paper, we present a tool we developed called Andromeda along with a usability study.

Andromeda is a visual analytics tool that spatializes high dimensional data in two dimensions using an algorithm called Weighted Multidimensional Scaling (WMDS) [15,16]. In the spatialization, distance reflects relative similarity; e.g., two points close to each other in the spatialization are more similar to each other in the high-dimensional space than two points far from each other. To set the spatial coordinates of the observations, WMDS relies on one parameter for each variable in the dataset. We refer to the parameters as variable weights because variables with large weights are considered more heavily in the spatialization than those with low weights. Thus, one can deepen her interpretation of a visualization in Andromeda by considering both distance and the weights; e.g., two points close to each other in a spatialization are more similar to each other *in the variables with large weights* than two points far from each other. Andromeda enables users to adjust spatializations using both parametric interaction and observation-level interaction.

Parametric interaction is available in several tools [3,12,20] in which users may specify underlying model parameters by adjusting dials and/or sliders. Although it has been shown useful, parametric interaction can challenge users who do not have a strong knowledge of the underlying model. Thus, in previous work, we developed a new way to interact with mathematical models called observation-level interaction or OLI [7,18]. With OLI, an automated


Andromeda

Use Ctrl Key to enter "Edit/Move Mode"
Use Shift Key to pan the layout

Algorithm in use: Original

Name: Dolphin

**Figure 1. Andromeda interface exploring an animal dataset: (a) the object view visualizing low-dimensional data points, (b) the parametric view displaying all dimension weights, (c) the slider tool to animate between transitions, (d) the button to update the layout, and (e) the radius highlighting near and possibly important data points comparative to the selected data point.**

procedure transforms user interactions with data visualizations (visual feedback) to parametric feedback that in turn adjusts an entire visual space. For example, in Andromeda, users may change the distance between observations by relocating them so that an automated procedure may then adjust the parameters (i.e. variable weights) in response.

With Andromeda, we seek to combine parametric interaction with observation-level interaction to provide the user with multiple ways of interacting with data. We believe tools with the multiplicity of these interactions will allow for more complete analyses than those that rely on one interaction type. We performed a controlled usability study to assess the benefits of and the drawbacks to parametric interaction, observation-level interaction and their combination. We sought to determine how the three different types of interaction (combined, OLI, and PI) support high-dimensional data analyses performed by non-experts of statistical algorithms. Our study specifically answers the following research questions:

1. Given benchmark tasks within specific categories as defined in [2], do the three different types of interaction affect the correctness of the users' answers?
2. How do the three different types of interaction (combined, OLI, PI) affect the users' insights in an open-ended analysis task?
3. How do the three interaction types affect understanding of MDS?
4. Do the three interaction types allow for effective and fluid analyses?

## RELATED WORK

Visual analytics tools aid users in exploring data. However, the tools are only useful if the design makes sense to the user, correctly portrays the underlying model, and allows the user to conduct analyses efficiently. Much research exists to guide designers during creation of interfaces for information visualization [4,19,22,25]. The visual analytics field specifically focuses on the design of interactive visualizations that provide an intuitive space that fosters insight creation and data understanding [24]. High-dimensional data is particularly difficult for users to comprehend because humans have trouble thinking about a large number of dimensions simultaneously. As discussed earlier, dimension reduction models reduce the data so that it is more manageable. Other non-traditional methods have been developed to support high-dimensional data exploration [8,13,21]. These techniques have yet to be incorporated into an interactive visual analytics tool.

IN-SPIRE's Galaxy View displays text documents as data points in topical clusters in a two-dimensional space where proximity implies relatedness [20]. Star Coordinates plots objects in high-dimensional space and then projects this space onto two-dimensions [14]. However, within both of these tools, only surface-level interactions are possible. In IN-SPIRE, the user can explore the data by selecting groups of points. The selection is cross-referenced with other types of visualizations and graphs for the user to gain more insight. Star Coordinates allows users to rotate and scale the projection. These surface-level interactions are useful, however, the user has no control over the parameters that are used to process the data.

Models have underlying parameters that can be adjusted to control how the data are reduced. Many tools exist to allow users to not only visualize high-dimensional data, but also to adjust the parameters of the model to visualize the data from multiple perspectives. Systems such as STREAMIT [1] and Dust & Magnet [26] allow parametric interaction where the user inputs feedback to update the model and in turn the model updates the visualization. STREAMIT uses a force-directed layout to visualize streaming text documents based on keyword similarity. Users can modify the numeric parameters (i.e. importance of keywords) to update the visualization. Dust & Magnet displays the parameters (i.e. dimension weights) as "magnets" within the object visualization. Users can directly interact with the magnets to modify their importance. This direct manipulation can be more intuitive for a user than increasing or decreasing a numerical value. However, both approaches still solely provide parametric interaction that could limit the depth and effectiveness of data exploration.

Some tools, such as ForceSPIRE [5,6] and Dis-Function [3], incorporate OLI for the user to physically adjust data points within a visualization. These tools utilize the relatability of OLI so the user can focus on the data rather than on learning about the statistical model.

## TOOL DESIGN

Andromeda is an interactive visual analytics tool designed to aid users in the analysis of high-dimensional data. It provides a way for users to interact with the input and output of weighted multidimensional scaling (WMDS). Andromeda supports both OLI and parametric interaction. In order to support the translation of visual interactions into transformations of the underlying parametric space, Andromeda supports OLI which allows users to interact directly with dimensionally reduced data plots. It hides the calculations of the dimensionality reduction algorithm so that the user can focus on the data using a familiar metaphor that encodes similarity with spatial proximity without requiring any knowledge of underlying statistical models. The interface is implemented using Java Swing and the MDSJ Java Library's implementation of WMDS [9]. Andromeda is composed of two main sections: the object view (Figure 1a) and the parameter view (Figure 1b). Examples throughout this paper use a modified version of an animal dataset provided by Lampert et al. that contains 49 animal objects over 72 dimensions [17]. We provided this dataset for participants to explore during the usability study.

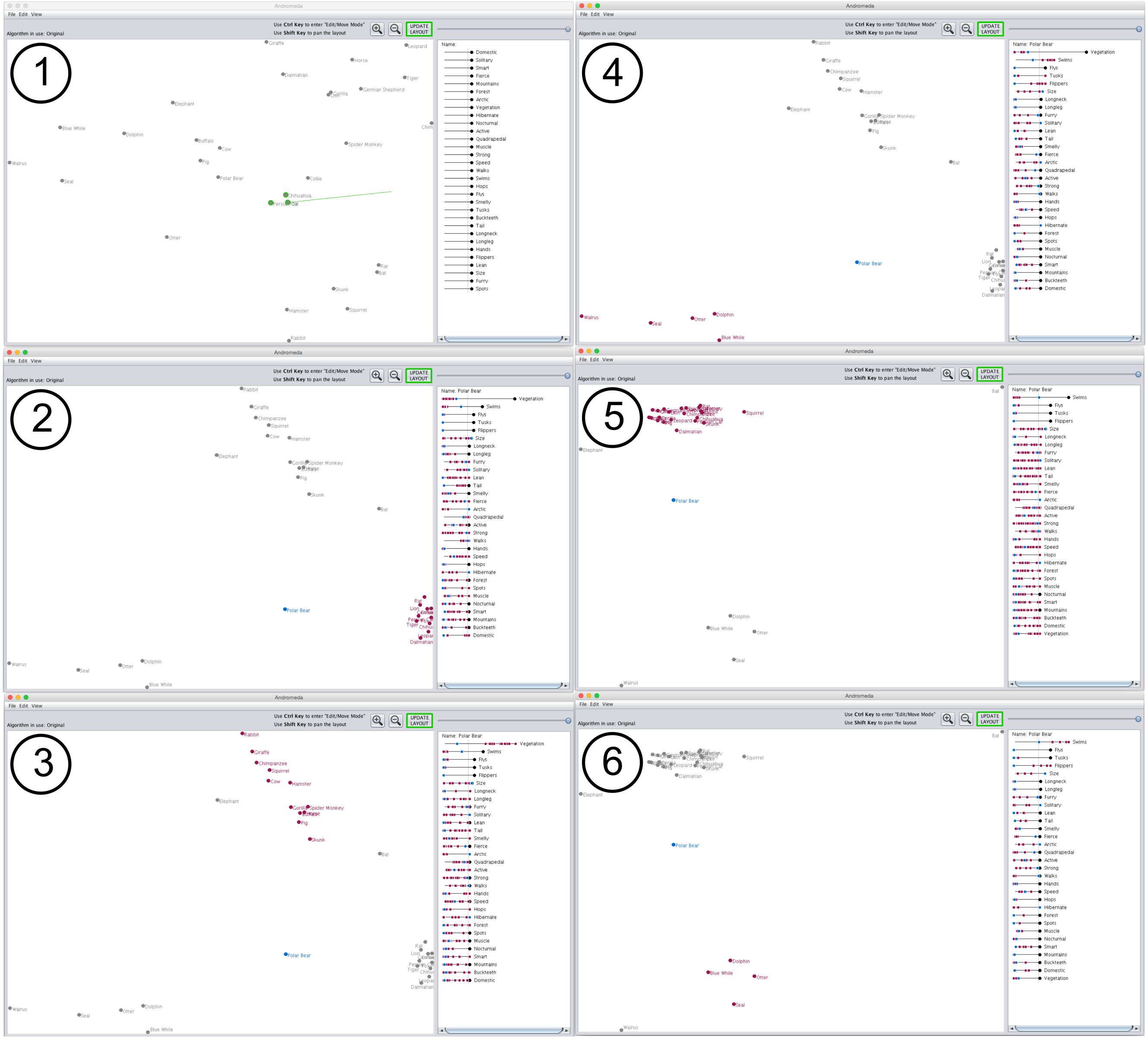
Throughout development, we had users informally assess the design of our system. We applied Andromeda in an educational setting with a graduate level visual analytics course and a graduate level information visualization course during separate semesters [23]. Each course used a different iteration of the system so that we could learn from the students' analyses. An example sequence of user interactions is shown in Figure 2. The users were prompted to reflect on their processes and explain any challenges they encountered while using the system. We analyzed the insights and processes from each class to see if users did what we expected. If not, we revamped our interactions and design choices to encourage more efficient usage and to address the challenges. The discoveries from these courses led to interface modifications and the version of Andromeda we present in this paper.

### The Object View

The resulting object layout calculated by WMDS is displayed in the object view. Each point represents one row of the high-dimensional data. This view has two modes which users can toggle between with the control modifier key. Without the modifier key, users can explore and view the data points. A user can hover over (blue point) and select points (maroon points) to view the corresponding raw data. We use color to link selected points to the parameter view where the raw data is displayed.

With the modifier key, the user enters move mode which allows for observation-level interaction. The user can manipulate points on the screen to provide input to the algorithm. When a point is moved by dragging it with the mouse, it is encoded with a green ring and a line from its original location to its new location (see Dolphin, Squirrel, and Chihuahua in Figure 1a). In this mode, points that are clicked, but not moved are considered highlighted. These points are encircled with a green ring, but do not have a line since they were not moved (see Elephant, Blue Whale, Skunk and others in Figure 1a). The green outline matches the outline of the "Update Layout" button as a visual cue that all outlined points are important to the algorithm (Figure 1d).

After points have been moved, the user can click the "Update Layout" button to recalculate the layout based on the new coordinate locations of the moved points. The algorithm only considers moved points when calculating an updated layout [11]. An optimization algorithm as described in [10,11,18] is run to find a weight vector that best represents the new coordinates of only the moved points. WMDS is then run to update the coordinates of each point given the new weight vector. Within the object view, the points animate to their new locations giving the user a



1
2
3
4
5
6
Andromeda
File Edit View
Algorithm in use: Original
Use Ctrl Key to enter "Edit/Move Mode"
Use Shift Key to pan the layout
UPDATE LAYOUT
Name: Polar Bear
Polar Bear
Vegetation

**Figure 2. This is a sequence of interactions in Andromeda. (1) Initial view with moved points. (2)-(4) Updated layout with different clusters selected. (5)-(6) Updated layout after decreasing vegetation dimension.**

visual representation of the movement of the points. The user can repeat this animation by engaging with the slider (Figure 1c). The slider allows the user to manually trace all points between the previous and current locations.

### The Parameter View

This view displays the weighted dimensions (Figure 1b). Due to the algorithm only handling continuous numerical data, the interface shows categorical or informational dimensions as static text for viewing only. Each numerical dimension is represented by an interactive line that serves as a visual representation of the relative weight compared to all other dimension lines. The user can drag the circular handle at the end of a line to adjust the weight of that dimension. Since all weights must sum to 1 (a constraint of WDMS), the interface automatically modifies all other dimensions in response to increasing or decreasing one dimension. Modifying dimensions triggers dynamic updates to the layout. Each time a user increases or decreases a dimension, WMDS recalculates the object layout based on this new weight vector in real time.

The parameter view also displays the raw data values of the high-dimensional data. All raw data values are normalized to fit a constant scale across all dimensions. This scale is used to plot the raw data onto the weight lines. When a point is selected in the object view, the corresponding raw data values are drawn onto each dimension line as a colored dot. For example, the maximum raw data value for a specific dimension will be placed on the far right of the line. A lower raw data value will appear closer to the left of the line. In Figure 1b, the selected maroon data points in the object view are animals that do not fly (third dimension line from the top). Therefore the raw data points appear toward the left of the line. As a dimension weight is increased, the plotted raw data dots are stretched to fill the line. The raw data is not changing, rather the relative distances between the values are changing based on the emphasis placed on that particular dimension.

### Interactions

Andromeda supports two types of interaction: visual observation-level interactions (OLI) and parametric interactions. Figure 3 depicts the algorithmic pipeline for each type of interaction. OLI is performed through the manipulation of objects within the object view (Figure 1a). Manipulating the objects creates a new set of low-dimensional coordinates. Andromeda calculates the optimized weight vector that best fits the low-dimensional points moved by the user. This is denoted as MDS-1 in Figure 3b. The new weight vector provides feedback as to which dimensions contribute to the new two-dimensional layout and by how much. To create a new spatialization based on this feedback, Andromeda runs WMDS again with the new weight vector and the original high-dimensional data to calculate new low-dimensional coordinates.

The parametric interaction afforded by Andromeda allows users to directly manipulate parameters of the underlying spatialization model (see Figure 3b). Andromeda allows this via manipulation of the interactive lines in the parameter view (Figure 1b). This chart displays the distribution of importance across all dimensions. By adjusting the distribution of importance, a user is providing feedback about which variables should be important while simultaneously providing the parametric feedback since this distribution is just a visual encoding of the statistical model's parameters.

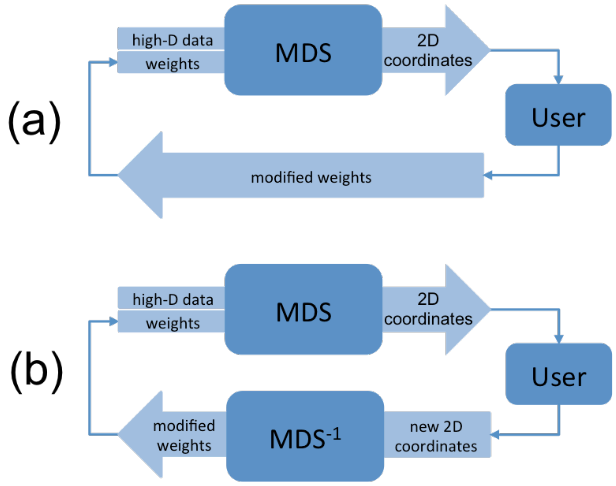


**Figure 3. Algorithmic pipeline (a) for parametric interaction and (b) for OLI or visual-to-parametric interaction.**

### Benefits of the Design

Andromeda combines parametric interaction and OLI, a combination which we argue enhances the exploratory analysis process in the following three ways:

1. With OLI, users keep their focus on the interpretation of a graph, rather than the model that created it. The objects in the dataset will presumably be familiar entities to the user, but the particular dimensions describing the objects may be different from the dimensions users have in their head. In other words, user domain knowledge may include additional dimensions about the objects that do not appear in the dataset. The user may also have meta-dimensions that she may or may not be able to be described using the dataset dimensions. For example, in the animal data, a user may be interested in a dimension she refers to as "cat-like." She may not be able to explain it, but she has a preconceived notion in her head about whether an animal fits this meta-dimension. The user can perform OLI, then the system will calculate dimensions within the dataset that define cat-like. Similarly, users may have preexisting notions of groupings, but may not have a meta-dimension to describe the grouping. Again, OLI offers the ability to instinctively group similar objects and have the mathematical model find supporting dimensions.
2. OLI provides the ability to pose what-if questions as a separate type of hypothesis testing. Users know and understand preexisting relationships between objects. For example, a user can pose that she thinks certain objects are similar. OLI can discover whether there is data to support this claim. Another question might focus on forcing an outlier into a cluster. With parametric interaction, this would require many trial and error iterations until the system converged on an appropriate parameterization. With OLI, the user can drag the outlier into the cluster and let the mathematical model do the work.
3. With OLI, users have the opportunity to manipulate the data at the object level. We claim an object level view is a more meaningful space to interact than a view consisting of a list of parameters. As the data grows, it is easier to manipulate lots of objects instead of adjusting lots of parameters. We can imagine quickly and fluidly manipulating many objects at one time. In order to provide this for parameters, we would have to considerably modify our parameter view design choices.

The above hypotheses were validated in our usability study and are discussed in the results section.

## STUDY DESIGN

The following section outlines the specific details of our study pertaining to participants, process and data collection.

### Participants

We recruited participants from both the undergraduate and graduate levels across multiple disciplines in order to obtain a diverse population in regards to experience and knowledge of data analysis. Our 30 participants spanned across five disciplines including biology, computer science, engineering, education and data analytics. No participants considered themselves experts with Multidimensional Scaling; 11 had heard of it; 16 had never used it, but had heard about it; and 3 learned about it in class. Graduate students comprised half of the participants. The remaining participants included 13 undergraduates and 2 M.S. alumni. Participants ranged from 19 to 34 years of age.

In order to test the three types of interactions, we individually disabled the OLI and PI within Andromeda to create two additional tools each with limited functionality. One third of the participants performed the study using Andromeda with all functionality. We gave another third Andromeda with PI disabled (OLI-Andromeda). The last third used Andromeda with OLI disabled (PI -Andromeda).

### Procedure and Data Collection

Participants were shown a tutorial video corresponding to one of the three tool variations (Andromeda, OLI-Andromeda or PI-Andromeda) randomly assigned to them. They were then asked to explore a high-dimensional dataset about animals [17]. It included 49 animals and 72 dimensions. The dimensions are characteristics describing the animals such as furry, speed, size, and ocean. The values ranged from 0 to 100 where 100 means high and 0 means low. For example, a grizzly bear has a furriness value of 82, whereas a blue whale has a furriness value of 0.

We asked participants to complete a survey that asked them to analyze data. The survey included 17 questions that directly related to our four research questions. The first two survey questions were biographical; survey questions 3-8 concerned low-level components (RQ1); 9-10 reflected insights (RQ2); and 11-17 reflected MDS (RQ3). Questions 3-8 and 11-17 were simple, short answer questions, whereas questions 9-10 were open ended. Additionally, we asked the participants to think aloud while they worked. The intention was to record miscellaneous thoughts and tasks performed while answering the survey questions and exploring the data (RQ4).

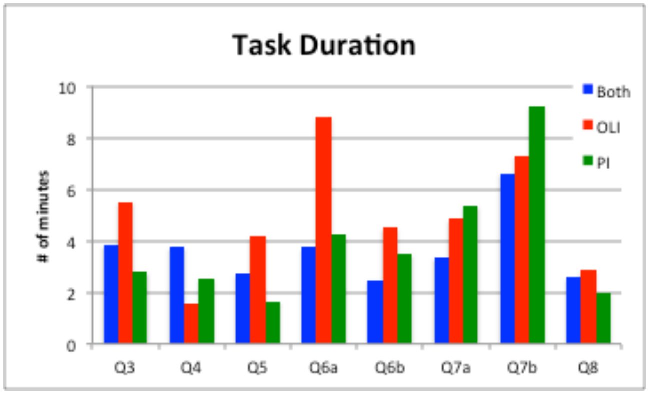


**Figure 4. The average number of minutes it took to answer each question based on type of interaction given.**

## RESULTS

Our results are divided into three sections: low-level questions, insights, and MDS concepts.

### Low-level Components: Questions 3-8

The study included six low-level questions from five different low-level components [2]. Each question corresponded to a specific component as well as an interaction (parametric or OLI) that we expected participants to use for answering the question. For example, the question, "What animal is the most timid?" was classified as a *find extremum* component and paired with parametric interaction, as we expected participants to use parametric interaction to answer the question. For questions classified as *cluster* component, e.g., "The elephant and blue whale are similar to each other, but are dissimilar to the tiger and the wolf. What other animals are like the elephant and blue whale, but not like the tiger and the wolf," we expected participants to use OLI. Refer to Table 1 for a complete list of the low-level questions, components, and paired interactions.

| # | Low-level Component | Expected Interaction |
|---|---|---|
| Q3 | Retrieve value | PI |
| Q4 | Find extremum | PI |
| Q5 | Filter | PI |
| Q6a | Characterize distribution | PI |
| Q6b | Correlate | PI |
| Q7a | Cluster | OLI |
| Q7b | Cluster | OLI |
| Q8 | Cluster | OLI |

**Table 1. All low-level component questions with corresponding component and expected interaction.**

#### *Parametric Interaction Questions*

Four of the six questions were paired with parametric interaction; i.e., we expected participants to use parametric interaction for questions Q3, Q4, Q5, Q6a, and Q6b. All 30 participants, regardless of the version of the tool, correctly answered Q3 (*retrieve value*), "How likely is it for the gorilla to live in the jungle?" On average, participants who used parametric interaction took less time to answer this question (Figure 4). Participants who did not have parametric interaction used the view mode to look at the raw data. Of the 10 participants who had both types of interactions, 7 performed parametric interaction for Q3 (Figure 5). Given both options, parametric interaction and OLI, participants were able to choose which type of interaction best fit the question at hand.

Q4 (*find extremum*) asked, "What animal is most timid?" For this question, 90% of the participants who used PI-Andromeda answered correctly, whereas 70% who used Andromeda with both interactions and 60% who used OLI-Andromeda answered correctly. Participants with parametric interaction used the tool as we expected. For example, to find the most timid animal, participants increased the weight of the timid dimension and used the updated visualization to find the most timid animal. One participant with only OLI clustered some points based on "timidness" and updated the layout. The participant refined the resulting clusters by removing some timid animals from the non-timid cluster and then settled on the correct answer. Though it is possible to answer this question using OLI, parametric interaction is more fitting for this type of dimension-specific find extremum data gathering. Of the 10 participants with both interactions available, 9 used parametric interaction to answer the question. Again, participants were able to choose the interaction that best helped them answer the question. They understood the tool's interactions and the appropriate context in which to use them.

The final two component-level parametric interaction questions resulted in similar findings. Q6 (*characterize distribution* and *correlate*) asked the participants to (a) "describe the agility characteristic" and (b) "find any other related characteristics." Participants using OLI spent a longer time answering part (a) compared to participants with parametric interaction (Figure 4). The correctness of their answers did not seem to be affected by the interactions provided. However, the group with both interactions tended to use parametric interaction to find the answers. The parameter view within the interface of Andromeda is intuitive enough for users to learn how to parametrically interact with the dimensions of the data and choose when this interaction is necessary with only a brief tutorial.

### *Observation-Level Interaction Questions*

We expected participants to use OLI for questions Q7a, Q7b, and Q8. These questions centered on rows or data points in the dataset. As with parametric interaction questions, we found that the type of interactions provided did not affect the correctness of the answers. Given the choices, participants tended to use OLI to answer these questions (Figure 5).

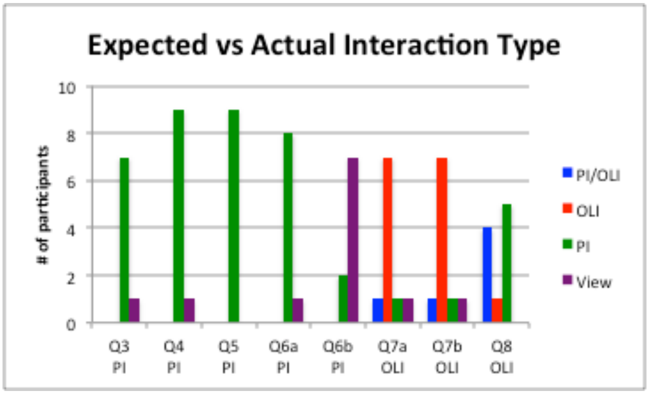


**Figure 5. Of the 10 participants provided both OLI and PI, this chart shows the type of interaction used to answer questions. The interaction labeled below the question number was the expected interaction. E.g., PI was expected to be used for question 3 and 7 of the 10 used PI.**

Q7 (*cluster*) asked participants to (a) "find animals that are like the elephant and blue whale, but not like the tiger and wolf" and (b) "find animals that are similar to the tiger and wolf, but dissimilar to the elephant and blue whale." All participants gave comprehensive answers. The following examples illustrate complex and insightful answers:

(a) *"The Humpback Whale, Rhino, Buffalo, Cow, Ox, Moose, Giant Panda, Sheep and Walrus are similar to the elephant and the blue whale but not to the tiger and the wolf. The tiger and the wolf are active, agile predators while the others are big, inactive and slow herbivores."*

(b) *"(The) Fox, Bobcat, Lion, Leopard (are similar). Some of the characteristics that distinguish the two groups are: Active, agile, meat teeth, hunter, stalker."*

Using parametric interaction tended to result in answering the question for a longer duration. Of the 10 participants with both interactions, 7 chose to use OLI.

Participants with PI-Andromeda, selected the elephant and blue whale data points to view the raw data. As they viewed the data, they would update dimensions where the elephant and blue whale were similar. These participants used a similar process to answer Q8 which asked to characterize and compare vegetarians, carnivores and omnivores. Participants either increased the weight of dimensions they believed from prior knowledge were important for explaining vegetarians, carnivores and omnivores or used prior knowledge to select animals from one of the three groups. After selecting animals, participants would look at the raw data and increase any dimensions where the selected animals were similar. This process mimics that of the OLI algorithm. The algorithm uses the data points that have been moved by the user to find a weight vector that best places the moved data points in low-dimensional space where the relative pairwise distances best match the relative pairwise distances in the user positioned layout.

For Q8 (*cluster*), "characterize and compare vegetarians, carnivores and omnivores," four participants used both interactions together giving well-formed answers. One such answer was:

> *"Vegetarians have more chew teeth, have varying degrees of activeness, and are on the lower side of agility. They have less claws, and have less meat teeth. They do not live in the ocean. Carnivores have more meat teeth and omnivores have a middle range of meat teeth."*

Participants were able to answer questions containing one single low-level component using one of the three versions of Andromeda. When given Andromeda with the best fitting interaction, participants were better able to find the answer when exploring the data. If given both parametric interaction and OLI, participants usually correctly choose which interaction was most appropriate for the type of question. We found that questions based on one or more dimensions best utilized parametric interaction, whereas questions asking specifically about data points were best suited for OLI.

### Insights: Questions 9-10

Questions 9 and 10 asked users to generate a certain number of insights to give them a chance for free-form exploration. Just as in the cognitive dimensionality study [23], we classify an insight to be more complex based on dimensionality and cardinality. Across all 30 participants, there were 96 insights. Dimensionality refers to the number of dimensions specifically mentioned within an insight. The number of dimensions per insight ranged from 0 dimensions (0D) to 10 dimensions (10D). The majority of the insights with the highest dimensionality were discovered with OLI-Andromeda. Figure 6 shows the dimensionality of the insights by tool variation.

An insight has cardinality if it specifically references one or more data points in the dataset. Insights gained from Andromeda with both types of interactions tended to reference groups of animals. Participants would meaningfully label these clusters within the insight. For example, one participant referred to "grazer animals" where "grazer" is a dimension. This insight:

> *"Grazer animals are quadrupedal and have chew teeth; they are generally inactive and more timid than not. Also, many of them are from the new world,"*

has high cardinality since it references many animals, specifically the "grazer" animals. It also has high dimensionality because it explicitly includes 5 dimensions from the dataset. We consider this a complex insight because of its high dimensionality and cardinality.

Many participants across all three groups included prior knowledge within their insights. Of the Andromeda insights, 54% included outside knowledge. Of the insights gained with PI-Andromeda and OLI-Andromeda, 49% and 23%, respectively, referenced outside knowledge. The following insight, found when using Andromeda with both types of interactions, is representative of the many well-formed and complex insights participants gained while exploring the dataset.

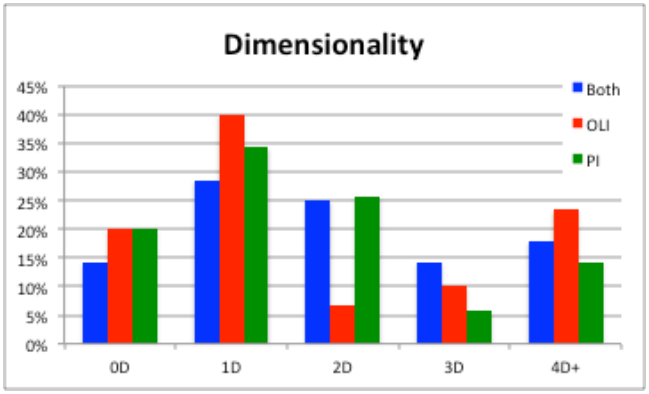


**Figure 6. Percentage of insights from each group against the number of dimensions explicitly mentioned in each insight.**

> *"I selected all of the animals that could be house pets (dogs, rabbits, cats, etc.) to see if there are characteristics that make these animals desirable to be pets compared to the other animals. I expect fierce to have a weak relationship, and domestic to have a strong relationship. I was correct in my assumptions. Other related characteristics include Timid, Ground, and Active. Active is an interesting one because these animals are grouped together in the middle showing desirable activity levels in pets - not too active that they need constant attention, but not so lazy they don't ever get up."*

This insight has high cardinality and dimensionality. The participant used prior knowledge to group animals she knows to be good pets. Her goal was to classify this group in terms of characteristics. She even included hypotheses that she was later able to confirm within the tool. Toward the end of the insight, she gives an explanation as to why the raw data backs up her claim that mid-range activity is best for pets. We see complexity in this insight based on the many facets. The participant used multiple low-level components (e.g. cluster, correlate), prior knowledge and gave a compelling argument for the insight.

This same participant went on to explore what characteristics might best explain animals that do not make good pets. A second insight states, "*Bobcats, wolves, lions, tigers, weasels, and skunks…. This new group exhibits opposite tendencies to the last in fierceness and timidness. Their speed is very high, and for the most part they are hunters*." Again, she correlates multiple dimensions and references a list of animals causing both dimensionality and cardinality to be high.

Overall, dimensionality and cardinality tend to have an inverse relationship. When an insight had high dimensionality, it had low cardinality and vice versa. However, as with the previously discussed insights, some had both high dimensionality and high cardinality. These insights represented our highest complexity insights.

If participants did not specifically mention one or more data points, they referred to a group of data points or all of the data points. For the Andromeda participants, of the 18 insights with 0 cardinality, 13 discussed a group(s) of data points. For the OLI-Andromeda participants, of the 10 insights, 9 discussed a group(s). For the PI-Andromeda participants, of the 16 insights, 14 discussed a group(s). We did not see this phenomenon as strongly with regards to dimensionality. When participants mentioned a group of dimensions, they tended to list out the dimensions by name.

On average, participants using Andromeda with both types of interactions spent more time analyzing the data. These participants took advantage of both types of interactions, OLI and parametric interaction throughout their exploration. The participants switched between the two interaction types multiple times either for a new question or to approach one question from different angles. By utilizing both parametric interaction and OLI, the participants gained a broader understanding of the data that led to a more diverse set of insights.

### MDS Concepts: Questions 11-17

The final seven questions on the questionnaire sought to explore whether participants learned basic WMDS concepts. 90% of the participants did not have any previous experience with MDS and no participants considered themselves experts. Because of this, any knowledge gained is from using Andromeda.

An important concept when using WMDS is to remember that no matter the projection, the raw data is never changed. Because of this, all plots created using WMDS are correct. Over 90% of participants regardless of the interaction type given correctly answered that there is no correct plot and that it depends on the question being explored.

Users must understand the interplay between the data point locations within the projection and the weights. With parametric interaction, increasing or decreasing a dimension weight modifies the data point locations based on the new weight vector. Data points that are more similar in reference to a dimension that is increased will now be relatively closer within the plot. We specifically asked participants to explain how and why increasing a weight would impact the relative locations of specific data points without using Andromeda. 80% of participants using Andromeda and 70% of participants using PI-Andromeda correctly explained that the fox would be positioned far from the ox, polar bear and cow, which would be clustered. Even though OLI-Andromeda participants did not have the ability to physically modify the weights, 60% of those participants correctly answered this question. They understood enough about the effect of the reverse interaction of moving the specific data points, which updates weights, to explain the hypothetical scenario of modifying a weight.

Similarly, it is important for users to understand the effect of OLI. We asked participants to explain, without using Andromeda, how and why moving the data points polar bear, deer and hamster close together and zebra far from them would impact the weight of the chewteeth dimension. Even though chewteeth would increase, 80% of all participants across all three conditions incorrectly reported that the weight would decrease. We attribute this poor performance to the participants' background knowledge of the specific data points in the question. Because participants did not look at the raw data for chewteeth for each animal data point, participants used their prior knowledge of the animal data to predict how the dimension weight would change. The second part of the question asked the participants to check their answer with the tool. Participants correctly reported an increase in the weight and attributed their mistake to incorrect knowledge about whether the animals had chewteeth.

A later question asked why adjusting the positions of the points changes the weights. Over 70% of all participants correctly explained that the weights of the most similar characteristics of the moved animals are increased which in turn adjusts all other points. This shows the students did understand how OLI affects the dimension weights, regardless of the type of interaction they were given.

Certain WMDS concepts proved more difficult for the participants to grasp given solely tasks to complete with only the use of Andromeda. Over 90% of participants across all three interaction types could not mathematically describe how Andromeda maps data points to locations in the plot. However, most participants understood that the placement depended on the amount of similarity between the more highly weighted dimensions. The difficulty lies in understanding that the algorithm is trying to minimize the stress between the low-dimensional distance and high-dimensional distance of all pairwise points. Without a more complete lesson than the video tutorial, the statistical algorithm behind Andromeda is challenging to comprehend. Regardless, we argue that a more complete understanding is not necessary to perform a useful analysis. The participants proved they understood the high-level overview of the algorithm and how the positions of the low-dimensional data points depended on the relative importance given to all dimension weights.

## DISCUSSION

During our study, we sought to discover how the three different types of interaction (combined, OLI and PI) support high-dimensional data analyses of non-experts. Our qualitative analysis of the users' analytic processes showed that the users were able to explore data within Andromeda efficiently and fluidly answering RQ4. Our participants were non-experts of WMDS and were able to easily learn to use Andromeda. While performing their analyses, participants learned simple WMDS concepts which was enough to efficiently use the tool to gain insights. Even though participants did not completely understand how the algorithm worked, it did not hinder their exploratory analyses. Our study proved that parametric interaction and OLI both provide enough context to the user to understand the algorithm at a high level answering RQ3. We even saw that participants without one of the interactions were still able to correctly answer questions specifically geared toward that missing interaction.

To answer RQ1, we found that the type of interaction accessible did not affect the correctness of the answers. Participants across all three groups were able to reasonably answer all low-level component questions. As expected, some answers were more complete than others, but we did not find significant evidence to attribute this to the type of interaction provided.

We determined that both parametric interaction and OLI are necessary for a complete and diverse analysis. All participants, no matter what type of interaction, explored the data and gained multiple insights. During the study, users with only one interaction asked if there was a way to perform the other interaction, not knowing that the interaction exists, but was disabled. One participant using PI-Andromeda actually tried to move the data points with her mouse. Users saw an opportunity to explore the data in a different way and wanted it.

In response to RQ2, even though all participants across all three interaction types (combined, OLI and PI) completed the open-ended analysis task, we found that certain analysis questions were more easily found with a particular interaction. We argue that both parametric interaction and OLI are necessary to allow for a more diverse analysis. Andromeda with both parametric interaction and OLI encouraged users to perform new tasks, beyond low-level component tasks, creating analytical gains of the system.

Through the usability study, we discovered tasks performed by participants that would not have been possible in one tool without the combination of OLI and parametric interaction. These tasks are as follows:

- **Injecting domain knowledge.** Users inject domain knowledge into their explorations. For example, users clustered certain animals they believed were similar based on previous knowledge not included within the dataset dimensions. Users included domain knowledge in their analyses no matter what interaction was provided.
- **Single dimension trends across all objects.** Users describe single dimension trends across all data points by either selecting all data points in the object view to view the raw data values in the parameter view or highly weighting one dimension to view the visualized distribution within the object view. This task was more prevalent among users with parametric interaction.
- **Comparison of clusters.** Users compare clusters and discover what dimensions best explain each cluster. OLI lends itself well to this task. Users are able to physically cluster data points to find distinguishing dimensions. This task becomes less fluid using parametric interaction.
- **Solving a subjective question.** A user wanted to determine what characteristics might describe animals that humans have as pets. She clustered pets and non-pets to determine the defining characteristics and discover any other animals that might be good pets. Subjective questions such as this that revolve around rows of the dataset are well suited to OLI. Similar questions about dimensions would require parametric interaction, proving both are essential.
- **Finding relationships between dimensions.** Users can find out if different dimensions are related to each other. One user discovered that bulbous animals have a strong tendency to swim. Dimension-based questions naturally fit with parametric interaction.
- **Forcing outliers into clusters.** Users force outliers into clusters to discover what dimensions must be emphasized to include those outliers in clusters. With OLI, users instinctively drag outliers toward clusters telling the system to find any similarities.
- **Extremes in the dataset.** Users state which two or three data points are most similar or different. Whether the extreme is in reference to a specific dimension or multiple dimensions, parametric interaction and OLI assist to find the answer.

Since we saw participants perform combinations of the above tasks, this suggests that Andromeda allows the discovery of both trivial insights through simple tasks, but also provides the opportunity for users to develop more complex insights through creative tasks. The use of OLI together with parametric interaction within an interface design delivers a well-equipped tool for visual analyses.

## CONCLUSION

In this paper, we discussed the benefits of OLI and the important role OLI plays in a well-designed visual analytics interface for exploring high-dimensional data. We stressed the importance of including both parametric interaction and OLI. With both types of interaction, a user is able to gain more complex insights and accomplish new types of tasks beyond simple low-level components of an analysis.

Providing tools such as Andromeda to domain experts, allows them to focus on exploring and learning about what most interests them: their own data. These experts can take advantage of complicated statistical algorithms that are meant to lessen the burden when analyzing complex high-dimensional data without being experts of the algorithms. Our study has shown that visual analytics tools with both parametric interaction and OLI implemented in an intuitive way have a low learning curve. Users can focus on exploring the data and not how to use the tool.

## ACKNOWLEDGMENTS

Omitted for blind review.

## REFERENCES

1. Jamal Alsakran, Yang Chen, Ye Zhao, Jing Yang, and Dongning Luo. 2011. STREAMIT: Dynamic visualization and interactive exploration of text streams. *IEEE Pacific Visualization Symposium 2011, PacificVis 2011 - Proceedings*: 131–138. http://doi.org/10.1109/PACIFICVIS.2011.5742382
2. R. Amar, J. Eagan, and J. Stasko. 2005. Low-level components of analytic activity in information

visualization. *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*: 111–117. http://doi.org/10.1109/INFVIS.2005.1532136

3. ET Brown, J Liu, C Brodley, and R Chang. 2012. Dis-Function: Learning Distance Functions Interactively. *IEEE Conference on Visual Analytics Science and Technology*: 83–92. Retrieved October 30, 2012 from http://www.cs.tufts.edu/~remco/publications/2012/VAST2012-DisFunction.pdf
4. SK Card, JD Mackinlay, and B Shneiderman. 1999. *Readings in Information Visualization: Using Vision to Think*. Academic Press. Retrieved March 30, 2015 from https://books.google.com/books?hl=en&lr=&id=wdh2gqWfQmgC&oi=fnd&pg=PR13&dq=readings+in+information+visualization+using+vision+to+think&ots=omAJaxoOQz&sig=VmDCxXUhuXrgF4aWPY-JBlpjaG4
5. A Endert, P Fiaux, and C North. 2012. Semantic Interaction for Sensemaking: Inferring Analytical Reasoning for Model Steering. *IEEE Transactions on Visualization and Computer Graphics* 18, 12: 2288–2879. Retrieved March 12, 2013 from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6327294
6. Alex Endert, Patrick Fiaux, and Chris North. 2012. Semantic interaction for visual text analytics. *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*: 473–482. http://doi.org/10.1145/2207676.2207741
7. Alex Endert, Chao Han, Dipayan Maiti, Leanna House, Scotland Leman, and Chris North. 2011. Observation-level interaction with statistical models for visual analytics. *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, IEEE, 121–130. http://doi.org/10.1109/VAST.2011.6102449
8. Michael Gleicher. 2013. Explainers: expert explorations with crafted projections. *IEEE transactions on visualization and computer graphics* 19, 12: 2042–51. http://doi.org/10.1109/TVCG.2013.157
9. Algorigthmics Group. 2009. MDSJ: Java Library for Multidimensional Scaling (Version 0.2). Retrieved from http://www.inf.unikonstanz.de/algo/software/mdsj/
10. Leanna House, Scotland Leman, and Chao Han. 2015. Bayesian Visual Analytics: BaVA. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 8, 1: 1–13. http://doi.org/10.1002/sam.11253
11. Xinran Hu, Lauren Bradel, Dipayan Maiti, Leanna House, Chris North, and Scotland Leman. 2013. Semantics of Directly Manipulating Spatializations. *IEEE Transactions on Visualization and Computer Graphics* 19, 12: 2052–2059. Retrieved October 24, 2013 from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6634115
12. Dong Hyun Jeong, Caroline Ziemkiewicz, Brian Fisher, William Ribarsky, and Remco Chang. 2009. iPCA: An Interactive System for PCA-based Visual Analytics. *Computer Graphics Forum* 28: 767–774.
13. Paulo Joia, Fernando V. Paulovich, Danilo Coimbra, José Alberto Cuminato, and Luis Gustavo Nonato. 2011. Local Affine Multidimensional Projection. *IEEE Transactions on Visualization and Computer Graphics* 17, 12: 2563–2571. http://doi.org/10.1109/TVCG.2011.220
14. Eser Kandogan. 2000. Star coordinates: A multidimensional visualization technique with uniform treatment of dimensions. *In Proceedings of the IEEE Information Visualization Symposium, Late Breaking Hot Topics* 650: 9–12. http://doi.org/10.1.1.4.8909
15. J B Kruskal and M Wish. 1978. Multidimensional Scaling. *Sage University Paper series on Quantitative Application in the Social Sciences*.
16. JB Kruskal. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29, 1: 1–27. Retrieved March 30, 2015 from http://link.springer.com/article/10.1007/BF02289565
17. Christoph H. Lampert, Hannes Nickisch, Stefan Harmeling, and Jens Weidmann. 2009. Animals with Attributes: A Dataset for Attribute Based Classification. Retrieved from http://attributes.kyb.tuebingen.mpg.de/
18. Scotland C Leman, Leanna House, Dipayan Maiti, Alex Endert, and Chris North. 2013. Visual to Parametric Interaction (V2PI). *PLoS ONE* 8, 3: e50474. http://doi.org/10.1371/journal.pone.0050474
19. T Munzner. 2014. *Visualization Analysis and Design*. CRC Press. Retrieved March 26, 2015 from https://books.google.com/books?hl=en&lr=&id=dznSBQAAQBAJ&oi=fnd&pg=PP1&dq=visualization+analysis+and+design&ots=HeNyEvM9Kp&sig=-zzLD8WjbvRF8ijvHUkXEn3uU7g
20. PNNL. 2010. IN-SPIRE Visual Document Analysis. Retrieved from http://in-spire.pnnl.gov/index.stm
21. Elisa Portes, Emilio Vital Brazil, Luis Gustavo Nonato, and Mario Costa Sousa. 2012. iLAMP : Exploring High-Dimensional Spacing through Backward Multidimensional Projection. *IEEE Conference on Visual Analytics Science and Technology*: 53–62.
22. Ben Schneiderman and Catherine Plaisant. 2005. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson Addison Wesley, Reading, MA.
23. Jessica Zeitz Self, Nathan Self, Leanna House, Scotland Leman, and Chris North. 2014. *Improving Students' Cognitive Dimensionality through Education with Object-Level Interaction*. Blacksburg. Retrieved

from http://people.cs.vt.edu/~jazeitz/vast_self_tech_report.pdf

24. J J Thomas and K a Cook. 2005. Illuminating the path: The research and development agenda for visual analytics. *IEEE Computer Society* 54: 184. http://doi.org/10.3389/fmicb.2011.00006

25. Ji Soo Yi, Youn Ah Kang, John T. Stasko, and Julie A.. Jacko. 2007. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 13, 6: 1224–1231. http://doi.org/10.1109/TVCG.2007.70515

26. Ji Soo Yi, Rachel Melton, John Stasko, and Julie a Jacko. 2005. Dust & Magnet: multivariate information visualization using a magnet metaphor. *Information Visualization* 00, April: 239–256. http://doi.org/10.1057/palgrave.ivs.9500099