Observation-Level and Parametric Interaction for High-Dimensional Data Analysis

JESSICA ZEITZ SELF, University of Mary Washington, USA
MICHELLE DOWLING, Virginia Tech, USA
JOHN WENSKOVITCH, Virginia Tech, USA
IAN CRANDELL, Virginia Tech, USA
MING WANG, Virginia Tech, USA
LEANNA HOUSE, Virginia Tech, USA
SCOTLAND LEMAN, Virginia Tech, USA
CHRIS NORTH, Virginia Tech, USA

Exploring high-dimensional data is challenging. Dimension reduction algorithms, such as weighted multidimensional scaling, support data exploration by projecting datasets to two dimensions for visualization. These projections can be explored through parametric interaction, tweaking underlying parameterizations, and observation-level interaction, directly interacting with the points within the projection. In this paper, we present the results of a controlled usability study determining the differences, advantages, and drawbacks among parametric interaction, observation-level interaction, and their combination. The study assesses both interaction technique effects on domain-specific high-dimensional data analyses performed by non-experts of statistical algorithms. This study is performed using Andromeda, a tool that enables both parametric and observation-level interaction to provide in-depth data exploration. The results indicate that the two forms of interaction serve different, but complementary, purposes in gaining insight through steerable dimension reduction algorithms.

CCS Concepts: • **Human-centered computing → Empirical studies in visualization**;

Additional Key Words and Phrases: Usability, user interface, visual analytics, dimension reduction, interaction, evaluation, data analysis

ACM Reference Format:
Jessica Zeitz Self, Michelle Dowling, John Wenskovitch, Ian Crandell, Ming Wang, Leanna House, Scotland Leman, and Chris North. 2017. Observation-Level and Parametric Interaction for High-Dimensional Data Analysis. *ACM Trans. Interact. Intell. Syst.* ?, ?, Article ? (December 2017), 36 pages. https://doi.org/0000001.0000001

Authors' addresses: Jessica Zeitz Self, University of Mary Washington, Department of Computer Science, 1301 College Avenue, Fredericksburg, VA, 22401, USA, jzeitz@umw.edu; Michelle Dowling, Virginia Tech, Department of Computer Science, 225 Stanger Street, Blacksburg, VA, 24061, USA, dowlingm@vt.edu; John Wenskovitch, Virginia Tech, Department of Computer Science, 225 Stanger Street, Blacksburg, VA, 24061, USA, jw87@vt.edu; Ian Crandell, Virginia Tech, Department of Statistics, 250 Drillfield Drive, Blacksburg, VA, 24061, USA, ian85@vt.edu; Ming Wang, Virginia Tech, Department of Computer Science, 225 Stanger Street, Blacksburg, VA, 24061, USA, mingw@vt.edu; Leanna House, Virginia Tech, Department of Statistics, 250 Drillfield Drive, Blacksburg, VA, 24061, USA, lhouse@vt.edu; Scotland Leman, Virginia Tech, Department of Statistics, 250 Drillfield Drive, Blacksburg, VA, 24061, USA, leman@vt.edu; Chris North, Virginia Tech, Department of Computer Science, 225 Stanger Street, Blacksburg, VA, 24061, USA, north@vt.edu.

2160-6455/2017/12-ART? $15.00
https://doi.org/0000001.0000001

1 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 and mathematical models enable us to explore large amounts of data and gain a better understanding of the structure, patterns, and contents within a dataset. The goal of visual analytics is to design intuitive methods for interacting with these models to improve data exploration techniques.

One method frequently utilized for understanding high-dimensional data is dimension reduction. Dimension reduction algorithms, such as Weighted Multidimensional Scaling (WMDS) [Kruskal and Wish 1978; Torgerson 1958], project high-dimensional data into low-dimensional spaces. These algorithms thus have the potential to summarize high-dimensional information in a form that is accessible to analysts, such as two-dimensional graphs [Pérez et al. 2015]. In turn, interacting with hundreds of dimensions is also accessible to analysts, as dimension reduced visualizations often avoid visual limitations that common high-dimensional visualization techniques (e.g. parallel coordinates, heatmaps) do not [Munzner 2014]. Thus, the visual analytics community may further improve the utility of dimension reduction by enhancing the output with information visualization and developing tools that allow for visual interaction.

In this paper, we consider such improvements and enhancements when using WMDS to visualize data. WMDS is easy to interpret and includes parameters that enable multiple forms of interaction. This is due to WMDS spatializations using distance between observations[1] that reflect relative similarity; two points close to each other in a WMDS low-dimensional spatialization are considered more similar to each other in the high-dimensional space than are two points far from each other. To define high-dimensional similarity, WMDS uses a vector of attribute weight parameters to compute a weighted distance function. That is, for WMDS, each attribute in a dataset is assigned a weight, and these weights reflect the importance of the associated attributes or degree to which each attributes is represented in a spatialization. Thus, the attribute weights determine the primary attributes in which observations are similar and different. For example, when all weights are equal, all attributes in a dataset are equally represented in a WMDS spatialization, and observations close to each other are more similar across all dimensions, relative to distant observations. When weights are unequal, the attributes with higher weights have more influence on a resulting spatialization than those with lower weights. Thus, observations close in proximity are similar (relative to those that are distant), primarily in the attributes with high weights.

A strong advantage of the weight parameters is that they enable WMDS to generate multiple projections of the same data. Because any one set of low-dimensional coordinates produced by WMDS cannot perfectly represent a high-dimensional dataset, it is helpful to assess multiple projections. When weights change, the directions of WMDS projections also change, and new spatializations of data are revealed. The challenge, however, is determining useful specifications of attribute weights. For this, human computer **interaction** has shown useful [Paulovich et al. 2012].

In the context of data exploration, we highlight two interaction techniques: parametric interaction and observation-level interaction. **Parametric interaction (PI)** refers to an analyst directly adjusting parameters of an algorithm or model, thereby modifying output of that model [Endert et al. 2011]. In our tool Andromeda (see Fig. 1), these parameters are weights of a distance function that is placed on the attributes of a dataset. Parametric interaction is available in several tools [Jeong

[1] In this work, we refer to items in a dataset as *observations* and the properties or dimensions of those items as *attributes*. With respect to the animal dataset used in this study, the animals themselves are the observations, and the quantitative properties of the animals (color, habitat, etc.) are the attributes. Observation-level interaction manipulates the observations directly, while parametric interaction manipulates parameters of the distance function applied to the attributes. In the examples discussed here, the parameters are the attribute weights used to bias the distance function in WMDS.

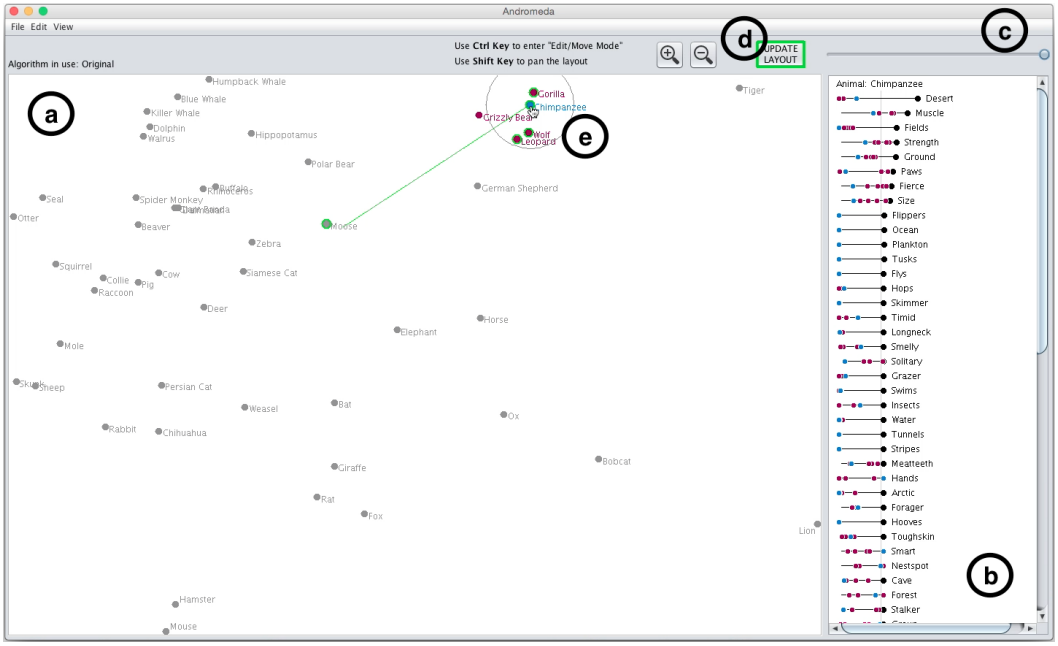

Fig. 1. Andromeda interface exploring a multi-dimensional dataset about animals: (a) the observation view projects data points onto 2D, (b) the parametric view displays all attribute weights as sliders, and selected data point values as dots on the sliders, (c) the animation slider animates transitions, (d) the button to update the layout in OLI, and (e) the green point that is being moved by the analyst to be near some selected red points.

et al. 2009; PNNL 2010] in which analysts may alter underlying model parameters by adjusting some interactive graphical component.

Although it has been shown to be useful, parametric interaction can challenge analysts who do not have a strong knowledge of the underlying model. This complexity led to the development of a new way to interact with mathematical models: observation-level interaction. **Observation-level interaction (OLI)** refers to directly interacting with individual points in a visual display, e.g., a 2D spatialization, through familiar and comfortable interactions, and having methods in place to interpret the semantic meaning of the interactions to adjust the display-generating model accordingly [Brown et al. 2012; Endert et al. 2011; House and Han 2015; Leman et al. 2013]. The natural approach for interpreting interactions relies on an inverted version of the display-generating model that uses the manipulated points to solve for new input parameters in a semi-supervised machine-learning fashion. In the case of WMDS, the inverse model takes as input a given set of moved points and learns a new weight vector that generates a new spatialization. The difference in information flow between PI and OLI is summarized in Fig. 2[2].

[2]Complementing PI and OLI interactions are **Surface-Level Interactions**. In contrast to PI and OLI, these interactions do not affect the underlying model parameters. Rather, they provide additional methods for analysts to glean insights from the data projections without modifying said projections. Examples of surface-level interactions include scaling and rotating the projection, linking data between views, and obtaining raw data values via details-on-demand [Leman et al. 2013]. As surface-level interactions do not affect the model, we focus on PI vs. OLI in this study; however, it is important to acknowledge their existence.

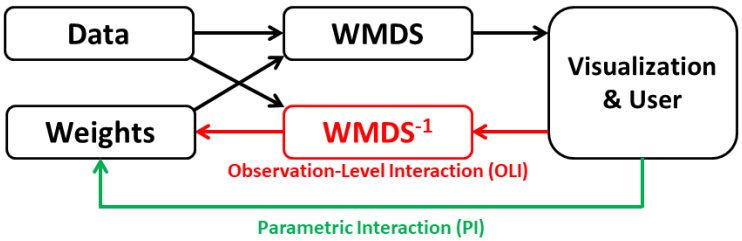

Fig. 2. A pipeline demonstrating the difference between Parametric and Observation-Level Interaction. With Parametric Interaction, an analyst directly manipulates the set of weights that control the layout algorithm. Using Observation-Level Interaction, an analyst instead manipulates the observations within the visualization, requiring the system to invert the layout algorithm and determine the appropriate weights to generate the new, user-steered visualization.

Due to the different natures of PI and OLI, we believe that these interactions afford distinct types of insights and explorations into the data. This means that PI, OLI, and their combination may afford different types of insights. Therefore, it is important for visual analytics tool designers to understand these differences in order to create appropriate tools for the task at hand.

Our contribution in this work is the design, results, and discussion of a controlled usability study to assess differences, advantages, and drawbacks to parametric interaction, observation-level interaction, and their combination. In particular, we sought to determine how the three different types of interaction (PI, OLI, and combined) support various tasks in high-dimensional data analysis when performed by non-experts of statistical algorithms. To support this usability study, we created Andromeda, a tool that combines parametric interaction with observation-level interaction to afford multiple ways of interacting with WMDS-projected data. We hypothesize that different tasks or insights will result in or inspire the use of different forms of interaction and vice versa.

Specifically, our study addresses the following research questions:

(1) **RQ1:** Given benchmark tasks within specific categories as defined in [Amar et al. 2005], do the three different types of interaction (PI, OLI, combined) affect the correctness of the analysts' answers? Does the type of interaction affect the speed at which they arrive at their answers? When afforded both types of interaction, do analysts select the type of interaction that we expect for each task?

(2) **RQ2:** How do the three different types of interaction (PI, OLI, combined) affect the analysts' insights in an open-ended analysis task?

2 RELATED WORK

2.1 Visualizing and Interacting with High-Dimensional Data

The field of visual analytics focuses on the design of interactive visualizations that may enhance the sensemaking of data, beyond what static visualizations can provide [Cook and Thomas 2005]. This is because static visualizations for complex data, no matter how intricate the method for creating them, will undoubtedly mask or hide some information in data. When this happens, it is up to analysts— users of the visualizations— to reconcile what is hidden with what is revealed to make sense of the data. Interactions may assist in this reconciliation. However, interactive tools are only useful if the design makes sense to the analyst, correctly portrays the underlying model, and allows the analyst to conduct analyses efficiently. Much research exists to guide designers during creation of interfaces for information visualization [Card et al. 1999; Munzner 2014; Shneiderman 2010; Yi et al. 2007].

High-dimensional data are particularly vulnerable to masking or hiding information in the visualized data. By the very fact that dimensions are being reduced for visualization, we know that information is lost in the process of creating visualizations. Indeed, many solutions are proposed for visualizing high-dimensional data without dimension reduction, such as parallel coordinates and heatmaps, but they cease to be effective visual abstractions as the number of dimensions enters the hundreds [Choo et al. 2010; Munzner 2014; Pérez et al. 2015]. Information is lost in the clutter and simple occlusion of important visual features.

As discussed earlier, dimension reduction models reduce the data from high-dimensional space into a more tractable low-dimensional space. These models are used in a wide variety of fields and come in many forms. Dimension reduction is used to reduce the complexity of datasets in fields such as statistics [Fukunaga 2013; Mardia et al. 1980], astronomy [Feigelson and Babu 2012; Way et al. 2012], machine learning [Jain et al. 2000; Ripley 2007], and visualization [Brehmer et al. 2014; Choo et al. 2010; Ingram et al. 2010; Jeong et al. 2009; Johansson and Johansson 2009; Kandogan 2012; Pagliosa et al. 2015]. Though we focus on WMDS in this work, the collection of dimension reduction techniques also includes Principal Component Analysis (PCA) [Jolliffe 2002], feature selection [Guyon and Elisseeff 2003], Isomap [Tenenbaum et al. 2000], and t-SNE [Maaten and Hinton 2008]. A survey of the many flavors of MDS can be found in [France and Carroll 2011]. Other non-traditional methods have been developed to support high-dimensional data exploration [dos Santos Amorim et al. 2012; Gleicher 2013; Joia et al. 2011]. These techniques have yet to be incorporated into an interactive visual analytics tool.

2.2 Alternative Interactive Tools for Multidimensional Data Exploration

A number of tools have been developed to afford analysts with the ability to interact with high-dimensional datasets. For example, IN-SPIRE's Galaxy View displays text documents as points in topical clusters in a two-dimensional space where proximity implies similarity [Pérez et al. 2015; PNNL 2010]. However, within these tools, only surface-level interactions are possible. In IN-SPIRE, the analyst can explore the data by selecting groups of points. The selection is cross-referenced with other types of visualizations and graphs for the analyst to gain more insight. These surface-level interactions are useful; however, the analyst has no control over the parameters that are used to process or display the data.

Other tools have more complex interactions that manipulate the visualization itself. Two examples of such tools include iVisClassifier [Choo et al. 2010] and Projection Inspector [Pagliosa et al. 2015]. iVisClassifier enable surface-level interactions to brush and link data displayed across parallel coordinates, heatmaps, and a scatterplot, as well as more complex interactions for outlier detection, selecting subsections of points, classifying points, and determining representative points in a cluster, the spread of a cluster, and points that are between clusters. Projection Inspector allows the analyst to interactively investigate different projections of high-dimensional data by combining different projection algorithms. In addition to these tools, other groups such as Schaefer et. al. and Johansson et. al. have found new methods for optimizing visualizations of high-dimensional data to reflect certain data structures, such as clusters or user-defined structures of interest [Johansson and Johansson 2009; Schaefer et al. 2013].

A specific class of more complex interactions enables the analyst to adjust the parameters of the dimension reduction model itself to not only visualize high-dimensional data but also to inspect the data from multiple perspectives. Systems such as STREAMIT [Alsakran et al. 2011], Dust & Magnet [Yi et al. 2005], Star Coordinates [Kandogan 2000], iPCA [Jeong et al. 2009], and DimStiller [Ingram et al. 2010] allow parametric interaction where the analyst directly manipulates the model parameters, triggering an update to the visualization. STREAMIT uses a force-directed layout to visualize streaming text documents based on keyword similarity. Analysts can modify

the numeric parameters (i.e. importance of keywords) to update the visualization. Dust & Magnet displays the parameters (i.e. attribute weights) as "magnets" within the observation visualization. Analysts can directly interact with the magnets to modify their importance. Similarly, Star Coordinates enables analysts to manipulate the axes of a biplot to increase or decrease attribute weights. This direct manipulation of the visualization can be more intuitive for an analyst 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. In addition to parametric interaction, both iPCA and DimStiller enable other forms of interaction to provide a variety of data exploration techniques. For example, iPCA allows interactions with the eigenvectors along with alternate, linked views of the data, whereas DimStiller models dimension reduction in a framework of analysis steps, chaining together operators into pipelines of expressions. Although these forms of interaction are useful, they all require the analyst to have a deep understanding of the underlying models to interact with the tool effectively.

Other tools such as User Guided MDS [Endert et al. 2011], ForceSPIRE [Endert et al. 2012a,b], and Dis-Function [Brown et al. 2012] incorporate OLI for the analyst to physically adjust points within a visualization. OLI was developed to directly manipulate observations in the plot to update the dimension reduction model parameters, instead of having to rely on manipulating parameters. Thus, these tools utilize the relatability of OLI so the analyst can focus on the data rather than on learning the details of the underlying models. However, the effects of using OLI as opposed to PI in visual analytics tasks has not yet been well studied. In this paper, we seek to address the critical need for usability studies to determine whether OLI supports analysis differently than parametric interaction.

It should be noted that some tools do enable the manipulation of observations, but not for the purpose of manipulating the underlying model parameters, and thus are considered surface-level interactions and not OLI. For example, many implementations of force-directed layouts allow analysts to drag or pin nodes in place. Similarly, tools such as GGobi allow analysts to constrain MDS layouts [Wickham et al. 2011]. Pinning introduces a form of constraint on the visual layout, constraining the underlying model to a sub-optimal solution. In essence, pinning represents a user interaction to position one or more observations in precise locations, causing the remainder of the observations to reposition in response to those constraints. In contrast, OLI interactions are an analyst request to update the underlying model in such a way that selected observations are positioned closer together or further apart without specifying precise coordinates. OLI interactions are *expressive* [Endert et al. 2011] in that they invoke model learning.

2.3 Evaluation of Insights in Multidimensional Data Exploration Tools

Under the evaluation scenario taxonomy for information visualization systems proposed by Lam et al., evaluation for multidimensional data exploration tools similar to Andromeda falls under the "Evaluating Visual Data Analysis and Reasoning" category [Lam et al. 2012]. The goal of such evaluation is to identify "if and how a visualization tool supports the generation of actionable and relevant knowledge in a domain." The output of such evaluations can include quantifiable evaluation like the number of insights obtained during analysis (for example [Saraiya et al. 2005]) or subjective evaluation such as opinions on the quality of the data analysis experience (for example [Seo and Shneiderman 2006]).

Indeed, the insight technique is a popular evaluation method for such systems in the literature. Systems such as ForceSPIRE [Endert et al. 2012b], StarSPIRE [Bradel et al. 2014], DimStiller [Ingram et al. 2010], and the analysis of representative factor generation performed in [Turkay et al. 2012] are evaluated using Use Cases or Case Studies that highlight insights generated while using each

tool. In this paper, we use this same technique to conduct a user study to evaluate the properties of insights afforded by OLI as opposed to PI.

2.4 Interaction in Human-Centered Machine Learning

While systems that learn from analysts have existed for some time, the advancement focus has undergone a recent shift from developing new machine learning algorithms to advances in studying and supporting user behavior [Amershi et al. 2014; Stumpf et al. 2009]. These human-centered machine learning systems are producing advancements in a number of domains, including but not limited to visualization [Kapoor et al. 2010; Talbot et al. 2009], text classification [Kulesza et al. 2011], image classification [Fogarty et al. 2008], and robotics [Cakmak et al. 2010; Kaochar et al. 2011].

If both parametric and observation-level interaction are generalized to their basic ideas (direct manipulation of model parameters for parametric interaction, learning new parameters based on other user interactions for observation-level interaction), then each can be found in the human-centered machine learning literature. For example, ManiMatrix [Kapoor et al. 2010] is a visualization system which allows analysts to directly refine values in a confusion matrix and see the effects of such refinements on the model, a clear example of parametric interaction. In contract, the Crayons system [Fails and Olsen 2003] represents a fully OLI system for image classification, with the analyst providing iterative feedback to the system by selecting image regions and allowing the system to classify features based on that information.

Still other projects such as EluciDebug [Kulesza et al. 2015], a system to support explanatory debugging, allow analysts to refine the model through both PI and OLI. In EluciDebug, OLI comes in the form of what Kulesza et al. call *instance-based feedback*, in which analysts can apply a label to an item, and the classifier then uses that label as part of its training set. This falls under a clustering, "this belong here" approach to OLI [Wenskovitch et al. 2018]. PI is supported by *feature-based feedback*, wherein an analyst tells the classifier *why* an item should be labeled in a certain manner due to the features that it contains or feature values that it matches. This distinction between telling the system why an item is classified in a certain way versus classifying the item and letting the system determine why, matches the difference between PI and OLI in the context of Andromeda.

3 ANDROMEDA DESIGN

Andromeda is an interactive visual analytics tool designed to demonstrate and study how interactivity aids analysts in the exploration of high-dimensional data. It provides a way for analysts to interact with the input and output of weighted multidimensional scaling (WMDS) through PI and OLI (respectively). Examples throughout this paper use a modified version of an animal dataset provided by Lampert et al. that contains 49 animal observations over 72 attributes [2009]. We provided this dataset for participants to explore during a usability study, that is described in the next section.

For development, we applied an iterative user interface design process with analysts informally assessing each iteration [Nielsen 1993]. We applied Andromeda in an educational setting in graduate and undergraduate courses related to data analysis [Self et al. 2017]. Each course used a different iteration of the system so that we could learn from the students' analyses. 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 analysts did what we expected. If not, we altered our interactions and design choices to encourage more efficient usage and to address the challenges the students encountered. The discoveries from these courses led to interface modifications and the version of Andromeda we present here. A summary of the features and visual encodings of the Andromeda system are summarized in Table 1.

Andromeda is composed of two main views: the observation view (Fig. 1a) and the parameter view (Fig. 1b). These views are discussed next.

Table 1. Andromeda Interactions and Encodings

Interaction Type	Feature	Description and Encoding
Parametric Interaction (PI)	Attribute Weight Sliders (update model weights)	Drag attribute slider handle to the right (increase weight) or the left (decrease weight). Attribute slider length will adjust. The spatialization is updated automatically after this interaction.
Observation-level Interaction (OLI)	Manipulate Observations (update model weights)	**Drag point:** Dragging a point with the mouse (with the Ctrl key down) will cause a green halo around the point and a green line tracking the point from its original position. **Highlight point(s):** Clicking a single point or drawing a box around multiple points (with the Ctrl key down) will create a green halo around the point(s). **Click "Update Layout":** Begins the Inverse WDMS computation on the points with a green halo to learn new attribute weights; the spatialization is updated automatically after this computation.
Surface-level Interaction	Details on Demand (no effect on model)	**Hover over points:** Hovering the mouse cursor over a point will display the raw data for that observation as blue points on the attribute sliders and will turn that observation blue. **Select point(s):** Clicking a single point or drawing a box around multiple points will display the raw data for the observation(s) on the attribute sliders as maroon dots and will turn those observations maroon. **View raw data:** Hovering the mouse cursor over the attribute slider handle will display a tooltip above that handle, displaying the raw attribute data values for selected points.

3.1 The Parameter View

This view displays the weighted attributes (Fig. 1b). Each attribute is represented by an attribute slider that serves as a visual representation of the relative weight compared to all other attribute sliders. Thus, this chart visualizes the model's weight vector, representing the distribution of importance across all attributes.

The **parametric interactions** afforded by Andromeda allow analysts to directly manipulate parameters of the underlying spatialization model (see Fig. 3a for a visual algorithmic pipeline). Andromeda allows this via manipulation of the attribute sliders in the parameter view. The analyst can drag the handle of the slider to adjust the weight of that attribute. Since all weights must sum to 1 (a constraint of WMDS), the interface automatically modifies the weights of all other attributes in response to increasing or decreasing the weight of one attribute. This parametric interaction provides the values for the statistical model's parameters as well as feedback about

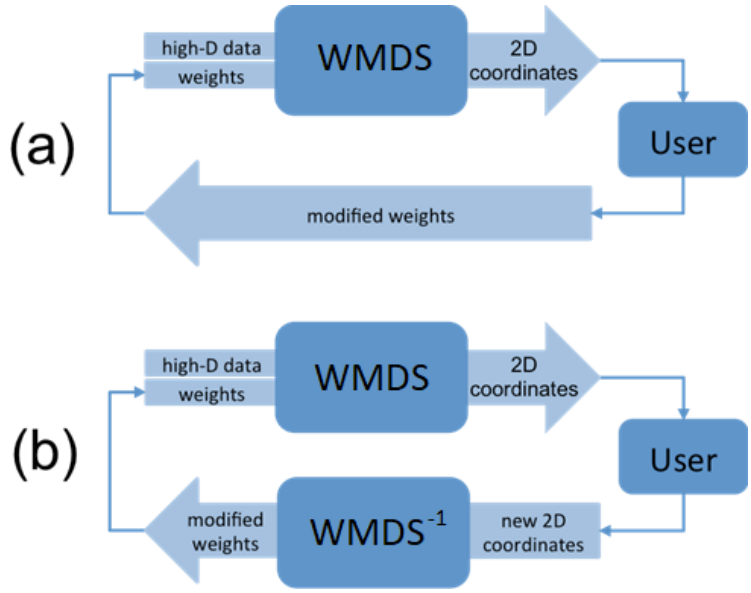

Fig. 3. Algorithmic pipeline (a) for parametric interaction and (b) for observation-level interaction.

which attributes are important to the analyst. In response to this parametric interaction, each time an analyst increases or decreases an attribute weight, WMDS is rerun to calculate the observation layout based on this new weight vector. Optionally, the weight sliders can be vertically sorted by their weight values, placing the most important attributes at the top.

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 attributes. This scale is used to plot the raw data onto the attribute sliders. When a point is selected or hovered over in the observation view, the corresponding raw data values are drawn onto each attribute slider as a colored dot. The maximum raw data value for a specific attribute will be placed on the far right of the slider, while a lower raw data value will appear closer to the left. In Fig. 1b, the selected maroon points in the observation view are animals that do not live in the desert (top attribute slider); therefore, the raw points appear toward the left of the slider. As an attribute weight is increased, the plotted raw data dots are stretched to fill the slider. The raw data is not changing, rather the relative distances between the values are changing based on the emphasis placed on that particular attribute. This visually reinforces the effect that weights have on the computed distances between points in the observation view. An example of a parametric interaction to answer Q4 from our study is shown in Fig. 4.

3.2 The Observation View

The resulting observation layout calculated by WMDS is displayed in the observation view. Each point represents one observation of the high-dimensional data. In our examples with the animal dataset, each point represents one animal. This view has two modes, View Mode and OLI Mode, which analysts can toggle between with the control (Ctrl) modifier key. Without the modifier key, analysts can explore and view the points in View Mode through the surface-level interactions of hovering over (blue point) and selecting 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.

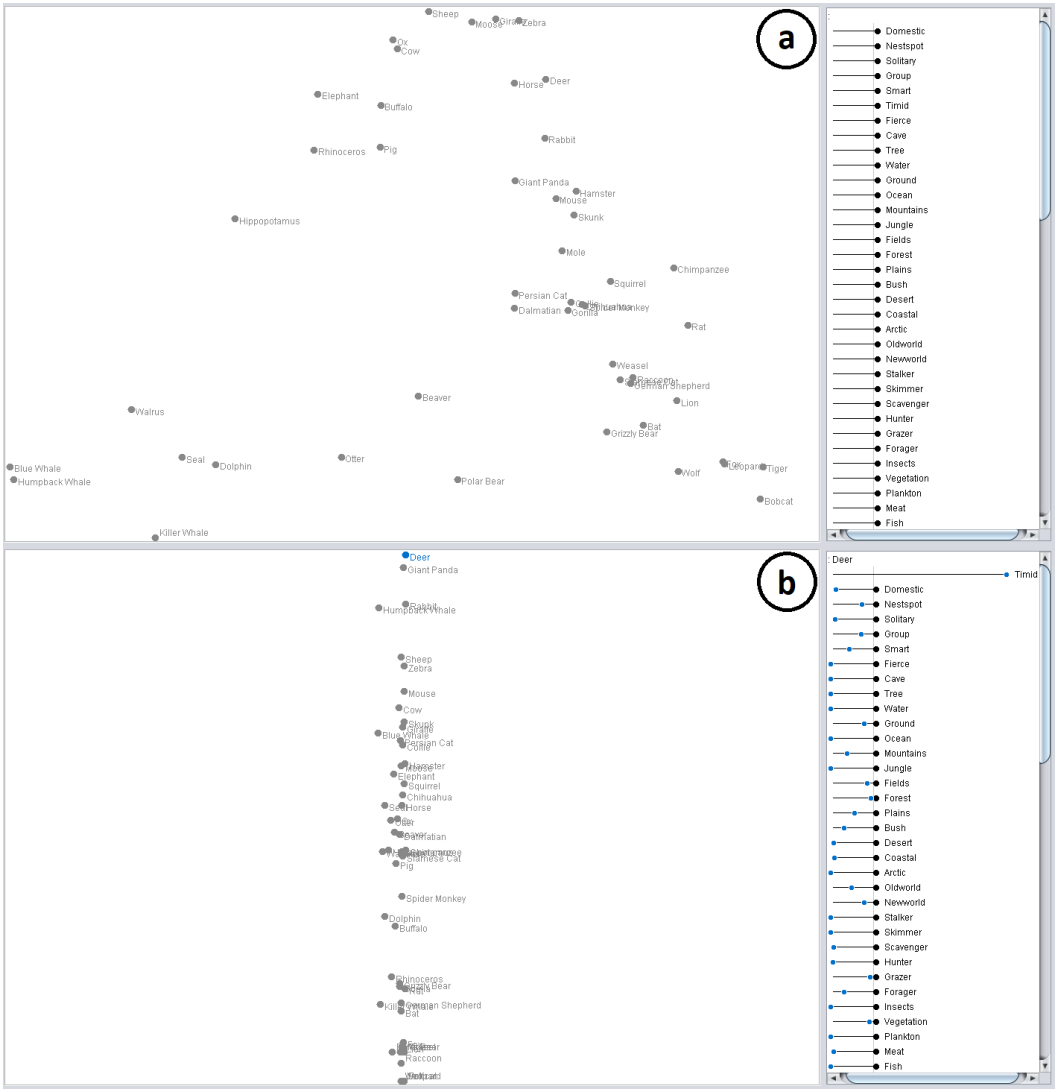

Fig. 4. A example of a parametric interaction in Andromeda, in this case answering Q4 from our study. To determine which animal is most timid, an analyst can (a) load the dataset of animals to see the initial, equally-weighted WMDS projection, and then (b) maximally increase the weight of the Timid attribute to generate a new projection. The analyst is then presented with a one-dimensional spectrum of timidity, and can use a surface-level interaction (mouseover) to check whether the animal at the top or the bottom of the spectrum represents the greatest level of timidity. The deer, at the top of the spectrum, has the highest Timid value in the dataset, as indicated by the accompanying blue glyph on the "timid" bar in (b).

With the modifier key, the analyst enters OLI Mode, which allows for **observation-level interaction**. When a point is moved in OLI Mode by dragging it with the mouse, it is encoded with a green halo and a line from its original location to its new location (see Chimpanzee in Fig. 1a). In this mode, points that are clicked (but not moved) are considered highlighted. These points are also encoded with a green halo, but do not have a line since they were not moved (see Moose,

Gorilla, Wolf and Leopard in Fig. 1a). The green halo of the moved points matches the outline of the "Update Layout" button as a visual cue that all points with green halos will be considered by the algorithm triggered by clicking "Update Layout" (Fig. 1d).

After points have been moved, the analyst can click the "Update Layout" button to perform the OLI algorithm. This algorithm recalculates the layout based on the new low-dimensional coordinates of the moved and highlighted points (points with green halos). To do so, the algorithm begins by calculating the optimized weight vector that best represents the relative pairwise distances between these low-dimensional points [Hu et al. 2013]. Thus, the moved points serve as the training feedback for machine learning the updated weights. Effectively, this inverts the WMDS algorithm to use the low-dimensional coordinates of the moved points as input to calculate the optimized weight vector as output. This part of the OLI algorithm is denoted as WMDS^{-1} in Fig. 3b. We use an optimization algorithm as described in [House and Han 2015; Hu et al. 2013; Leman et al. 2013] that minimizes a stress function with respect to the weight vector, ω, given user-specified low-dimensional coordinates, r^*,

$$\omega = \underset{\omega_1,\ldots,\omega_p}{\arg\min} \sum_{i=1}^{n} \sum_{j>i}^{n} \left(dist_L(r_i^*, r_j^*) - dist_H(w, d_i, d_j) \right)^2 \tag{1}$$

To represent each attribute weight as a proportion, the weight vector ω is constrained by the requirements that all weights must sum to 1 and that each weight must be positive. The functions $dist_L$ and $dist_H$ refer to the Euclidean distance between points i and j in low-dimensional and high-dimensional space respectively, except $dist_H$ is weighted Euclidean distance based on ω.

To complete the OLI algorithm, WMDS is then run with the new weight vector to update the coordinates of all points (i.e., not just the moved points). Low-dimensional coordinates, r, are determined based on minimizing a stress function with respect to r given the new weight vector, ω,

$$r = \underset{r_1,\ldots,r_n}{\arg\min} \sum_{i=1}^{n} \sum_{j>i}^{n} \left(dist_L(r_i, r_j) - dist_H(\omega, d_i, d_j) \right)^2 \tag{2}$$

With the new low-dimensional coordinates calculated, all points in the observation view animate to their new locations, giving the analyst a visual representation of the movement of the points. The attribute sliders are also updated to reflect the new attribute weight values computed by the optimization. The analyst can repeat this animation by interacting with the animation slider (Fig. 1c). This slider allows the analyst to manually trace all points between the previous and current locations. The end effect of the OLI interaction is that the algorithm produces a new projection that best reflects the analyst's input feedback as supported by the data. An example of observation-level interaction to answer Q7a from our study is shown in Fig. 5.

4 STUDY DESIGN

The goal of this study is to test and compare OLI and PI, and to discover how these interactions affect data analysis tasks and the types of insights gained.

4.1 Conditions and Participants

We performed a between-subjects usability study with three groups, each with a different interaction capability: an OLI group (called OLI), a PI group (called PI), and a group that had both PI and OLI (called Both). Each group had a version of the Andromeda interface with functionality to match their assigned interaction(s): one with Andromeda that only enabled OLI (referred to as OLI-Andromeda), one with Andromeda that only enabled PI (referred to as PI-Andromeda), and one with both PI and OLI enabled (referred to as Both-Andromeda). Additionally, each of these

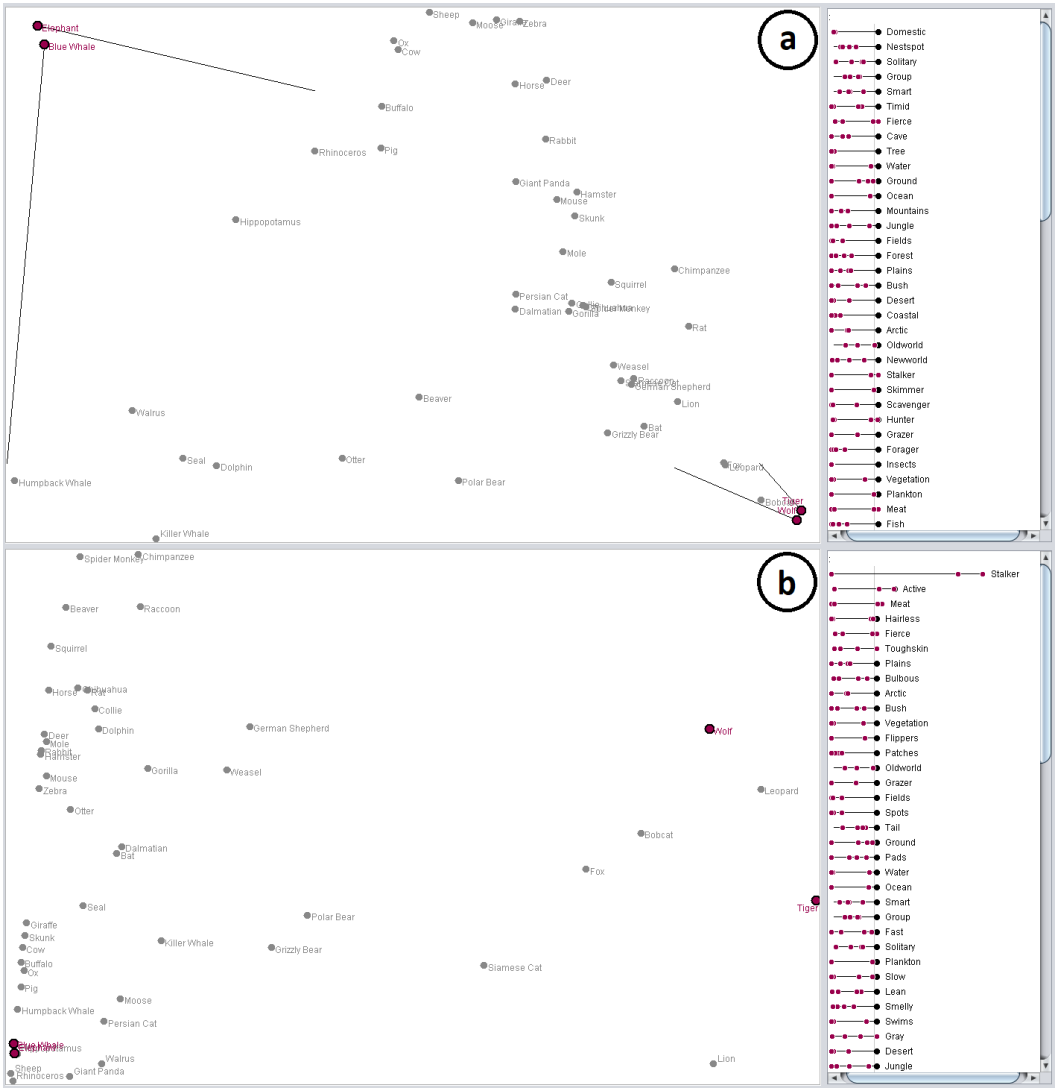

Fig. 5. A example of a observation-level interaction in Andromeda, in this case answering Q7a from our study. To determine which animals are similar to Elephants and Blue Whales, but dissimilar from Wolves and Tigers, an analyst would begin by loading the dataset (identical to Fig. 4a), and then (a) selecting and dragging the Blue Whale and Elephant to one area of the projection, and separately selecting and dragging the Wolf and Tiger into another distant area of the projection. After clicking the "Update Layout" button, the WMDS^{-1} algorithm learns new attribute weights based on only the observations that were moved or highlighted. Immediately following this computation, a new projection is displayed that repositions all observations based on the new weights (which shows that the Stalker, Active, and Meat attributes are relevant to the analyst's interaction). From the positions of the observations, the analyst can then see that animals such as the Pig and Walrus are similar to the Blue Whale and Elephant, but dissimilar from the Wolf and Tiger.

versions of Andromeda allowed surface-level interactions. Thus, the independent variable for the study is the interaction group.

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 six disciplines including biology, business, computer science, engineering, education, and data analytics. No participants considered themselves experts with Multidimensional Scaling; 11 had never heard of it; 16 had never used it, but had heard about it; and 3 had learned about it in a class. Participants ranged from 19 to 34 years of age. A summary of the participants from this study, broken down by discipline and education level, is included in Table 2. 10 participants were randomly assigned to each of the three groups: OLI, PI, and Both.

Table 2. Participants Broken down by Discipline and Level

Discipline	**# of Participants**	**Level**	**# of Participants**
Biology	6	Undergraduate	13
Business	1		
Computer Science	10	Graduate	15
Data Analytics	1		
Education	1	Professional	2
Engineering	11		
Total	30	*Total*	30

4.2 Data, Tasks, and Hypotheses

Participants used their assigned version of Andromeda to explore a high-dimensional dataset of animals [Lampert et al. 2009]. The data included 49 animals (observations) and 72 attributes. The attributes are characteristics describing the animals such as Furry, Speed, Size, and "Ocean" (referring to ocean-dwelling). The values ranged from 0 to 100, where 0 means low and 100 means high. For example, a grizzly bear has a furriness value of 82, whereas a blue whale has a furriness value of 0.

We designed two sets of data analysis tasks to perform on this data: benchmark tasks and insight tasks [North 2006]. See Appendix A.1 for the complete list of task questions from this study.

For **RQ1**, the **benchmark tasks** were designed to examine analyst performance on specific types of data analysis tasks. 8 short answer questions (Q3–Q8 on the survey including subquestions) were designed as benchmark tasks based on a subset of the low-level analysis tasks outlined by Amar et. al. that analysts pose when analyzing multi-dimensional data [2005]. We chose a variety of tasks that we believed was easiest to answer using one of the two types of interaction (OLI and PI). Table 3 lists the benchmark tasks and their hypothesized associated interactions.

To create our hypothesis on which type of interaction would enable participants to answer each question more easily, we assessed the type of task associated with each question as well as the type of data specifically related to the question. For example, Q4, "What animal is the most timid?" was classified as a *find extremum* task that centered around a specific attribute. Thus, we hypothesized this question would be easier to answer with PI rather then OLI. This is because it should be easy to find the answer by maximizing the weight of the timid attribute and finding the most timid animal at one end of the resulting nearly-1D projection (as seen in Fig. 4). In contrast, OLI would require more of a trial-and-error process of using surface-level interactions to view the timid-ness

Table 3. Benchmark Tasks

Question	Low-Level Task	Hypothesized Interaction Type
Q3 – How likely is it for the gorilla to live in the jungle?	Retrieve value	Surface
Q4 – What animal is most timid?	Find extremum	PI
Q5 – What animals are quadrupedal and slow?	Filter	PI
Q6a – Describe the distribution of the agility characteristic.	Characterize distribution	PI
Q6b – Are any characteristics strongly related to agility? How? Why?	Correlate	PI
Q7a – What other animals are like the elephant and blue whale, but not like the tiger and wolf? Why?	Cluster	OLI
Q7b – What animals are similar to the tiger and wolf, but dissimilar to the elephant and blue whale? Why?	Cluster	OLI
Q8 – Characterize and compare vegetarians, carnivores, and omnivores.	Domain Knowledge	Both

of each animal and manually search for the maximum. Similarly, tasks related to attribute filters, distributions, and correlations are all attribute-centric and therefore matched to PI.

For questions classified as *cluster* tasks, e.g., Q7a: "What other animals are like the elephant and blue whale, but not like the tiger and the wolf?" we hypothesized that OLI would be more efficient (see Fig. 5 for an example sequence of interactions) because the question focuses on data observations. With OLI, analysts can easily create these clusters and find other nearby animals. In contrast, PI would require more steps to manually find which attributes these animals share via surface-level interactions and then attempt to create the relevant clusters using PI on those attributes.

Q3 was designed as a simple retrieval task that should be easily answered using only the surface-level interaction to select the Gorilla and view its Jungle attribute value. Q8 was designed to require domain knowledge, since the data does not contain information about vegetarian, carnivore, or omnivore. We expect that both types of interaction (OLI and PI) would be useful to group animals based on their believed diet, or investigate related attributes such as "chew teeth"

Our hypothesis for RQ1 is that the participants with the appropriate interaction type (OLI or PI) should be able to answer the corresponding benchmark tasks more quickly and accurately. Furthermore, we hypothesized that the participants in the Both group, who could use both OLI and PI, would choose to use the interaction (OLI or PI) that we hypothesized was most appropriate for the given question.

For **RQ2**, the **insight tasks** were designed to examine the insight generating capabilities of the interactions. Two questions (Q9-Q10 on the survey) were open-ended insight questions. For RQ2, we hypothesize that OLI would produce observation-centric insights that emphasize cardinality, such as clusters. Additionally, we hypothesize that PI would produce attribute-centric insights that emphasize dimensionality, such as extremum and correlations.

4.3 Procedure and Data Collection

To begin, participants were shown a short tutorial video corresponding to one of the three tool variations (OLI-Andromeda or PI-Andromeda, Both-Andromeda) randomly assigned to them. Participants then completed an online survey. The first two survey questions were biographical. The next two sets of questions asked them to analyze the data according to the previously described tasks; survey questions Q3–Q8 (eight total including subquestions) concerned benchmark tasks (RQ1), and Q9–Q10 reflected insights (RQ2). Both of these sets of data analysis questions required participants to analyze the data using their assigned tool, and then type their answers into a textbox. Then, Q11–Q17 were about participants' understanding of WMDS concepts after having used Andromeda; we discuss these in Future Work in Section 6.5.

The dependent variables in the study included: (RQ1) Performance on the benchmark tasks Q3–Q8, including time, accuracy, and interaction logs; (RQ2) Written insights that were analyzed for dimensionality, cardinality, and interestingness. Additionally, we asked the participants to think aloud while they worked, with the goal of capturing miscellaneous thoughts while answering the survey questions and exploring the data. Participants each spent 1 to 1.5 hours completing the study.

5 RESULTS

Our results are divided into two sections, each mapping to the two categories of questions described in the previous section: benchmark tasks and insights.

5.1 Benchmark Tasks: Questions Q3–Q8

Per RQ1, participants answered questions containing individual benchmark tasks using one of the three versions of Andromeda (OLI-Andromeda, PI-Andromeda, Both-Andromeda). We first give an overview of the results and then discuss each benchmark task individually.

Overall, the strongest result in the benchmark tasks concerns the **expected vs actual interaction type**. Fig. 6 breaks down the 10 participants in the Both condition into the number who actually used each interaction type (OLI or PI) while performing each of the benchmark tasks, with annotations below each question displaying our hypothesized interaction type. This is good evidence that our hypothesis is supported regarding which interaction type (OLI or PI) analysts would choose when given access to both. That is, the participants in the Both condition, who had access to both OLI and PI interactions, tended to use the expected interaction type for each benchmark task, though there are several exceptions that we justify in future subsections.

We also examined the **correctness** of the participants' answers to the benchmark task questions. Table 4 shows an average correctness score for each question, with the highest average correctness for each question in bold. Correctness was measured by comparing participants' answers to a predefined correct answer using a simple 3-level scoring scheme. If a participant's answer contains all of the important criteria defined in the correct answer (e.g. list of data attributes and observations), then the score is 1. If a participant's answer contains only a portion of the correct criteria then the score is 0.5. If the answer fails to contain any of the correct criteria, the score is 0. The trends in the correctness results are not strong given our hypothesized most appropriate interactions.

We further analyzed the **time** participants took to answer each benchmark task question. Fig. 7 presents each question by average task duration for each of the interaction methods. Through video analysis, we recorded task times, not including time spent reading the task question or typing answers. While the overall pattern makes sense in terms of task difficulty , significant differences were rarely found between interaction types due to large variance and small sample sizes. Interestingly, for the Both condition, performance times tended to more closely match the

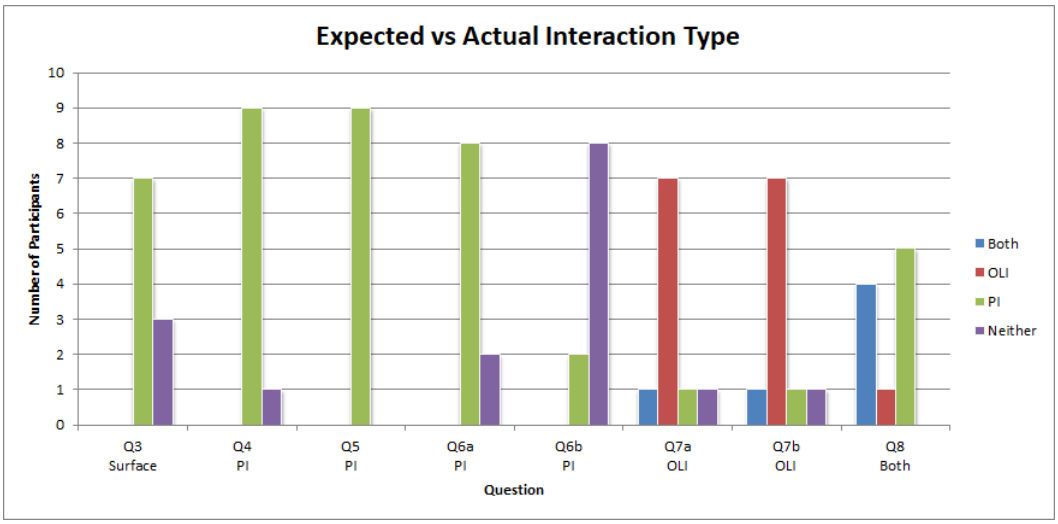

Fig. 6. Of the 10 participants provided with both OLI and PI capabilities, this chart shows the type of interaction they actually used to answer each question. The type of interaction labeled below the question number is the hypothesized interaction; e.g., PI was expected to be used for question 4, for which 9 of the 10 participants used PI. The "Neither" bars refer to participants who exclusively relied on surface-level interactions, using neither PI nor OLI to address the task. Participants performed the expected interaction a majority of the time in 5 of the 8 questions; however, questions Q3, Q6b and Q8 went against our hypotheses as discussed in the text.

Table 4. Correctness Score by Question and Interaction Type

Question	Both	OLI	PI
Q3 (Surface, retrieve)	**1.00**	**1.00**	**1.00**
Q4 (PI, extremum)	0.70	0.60	**0.90**
Q5 (PI, filter)	0.65	**0.95**	0.75
Q6a (PI, distribution)	**0.65**	0.60	0.50
Q6b (PI, correlate)	**0.65**	0.60	**0.65**
Q7a (OLI, cluster)	**0.85**	**0.85**	**0.85**
Q7b (OLI, cluster)	**0.65**	**0.65**	0.60
Q8 (Both, domain)	0.85	0.55	**0.90**

time for the hypothesized interaction type for the tasks Q4–Q7b where we correctly predicted the interaction type. Next, we discuss each task individually.

5.1.1 Surface-Level Interaction Question. As Q3 ("How likely is it for the gorilla to live in the jungle?") falls under the *retrieve value* task category, it can be addressed simply by viewing the data via surface-level interactions, and we hypothesized that participants would answer the question in this fashion. All 30 participants, regardless of the interaction type, answered this question correctly. As expected, many participants answered this question quickly, although it took some time for participants to find the Gorilla observation in the projection, suggesting the need for a search

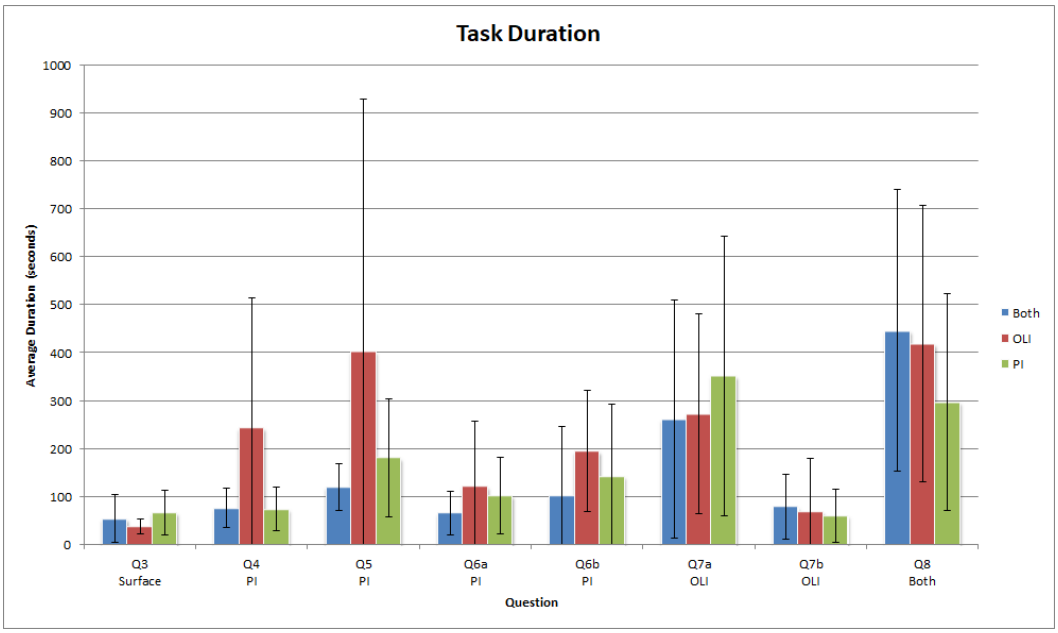

Fig. 7. The average number of seconds it took to answer each question based on type of interaction given. Each bar represents the average across all 10 participants for each interaction type. These times take into account only time spent interacting with the Andromeda, and do not consider the time that participants spent reading the questions and writing responses. Error bars represent one standard deviation. We hypothesized that the times would be shorter for interaction types that match the targeted interaction type annotated below each task. While there is some indication of this effect in Q5 and one significant result was found in Q4, overall results are not significant due to high variance and small samples.

feature. In contrast to our interaction hypothesis, 7 of the 10 participants in the Both group used PI to answer this question, while the other 3 participants simply viewed the data through surface-level interactions. Some participants thought that if they used PI to increase the weight of the Jungle attribute, it might be easier for them to find the Gorilla or to view its relative position among the other animals on the Jungle attribute. However, it is possible that this extra step actually slowed down the PI participants for this task.

5.1.2 Parametric Interaction Questions. Four of the eight questions were hypothesized to be associated with parametric interaction; i.e., we expected participants who used parametric interaction for questions Q4, Q5, Q6a, and Q6b to be able to arrive at the correct answer faster and more accurately.

Q4 (*find extremum*) asked, "What animal is most timid?" This task is clearly attribute-centric since it mentions a specific attribute name, likely putting participants in an attribute-oriented mindset at the start of the task. Participants with PI used the tool as we expected: to find the most timid animal, participants increased the weight of the Timid attribute and updated the visualization. Participants could find the Timid attribute quickly because the attributes are initially sorted in alphabetic order. Then, the participants used the new projection to find the cluster of timid animals (as in Fig. 8a). The participants then searched the cluster for the most timid animal. To illustrate an alternate approach, one participant in the OLI group clustered some points based on their domain knowledge of "timidness" and updated the layout (as in Fig. 8b). The participant refined

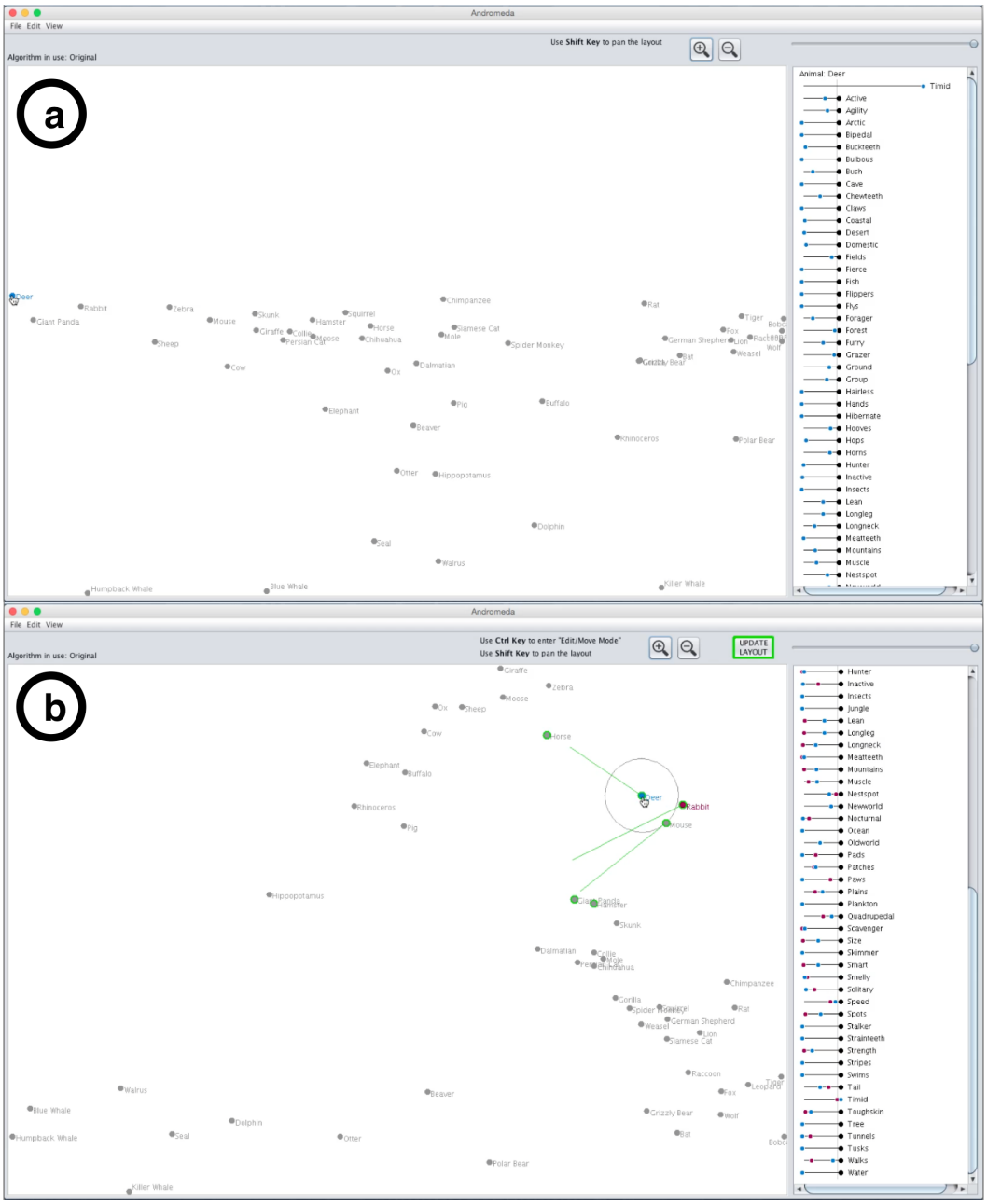

Fig. 8. Two methods of answering Q4 to locate the most timid animal. (a) This participant used PI to increase the weight of the Timid attribute. Then, the participant hovered over the points to find the cluster of animals whose raw data values for Timid were high. These animals were to the left of the projection. The Deer had the highest value for Timid making it the most timid animal. (b) This participant used OLI to group animals he knew to be timid from his domain knowledge, and by viewing the raw data of the timid attribute, localized many of the timid animals.

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, PI is a better match, which led us to hypothesize that PI would be more efficient to use. Of the 10 participants with both interactions available, 9 used PI to answer the question, supporting this hypothesis. Evidently, these participants understood the interactions supported by the tool and the appropriate context in which to use them. Participants in the PI group were also able to answer this question more accurately (9 of 10 fully correct) than those in the OLI group (6 of 10 fully correct) and those in the Both group (7 of 10 fully correct). Additionally, this task has the strongest timing results; an unequal variance, two sample t-test concludes that our PI group is faster than our OLI group ($t = -1.9585$, $df = 10, p = 0.0393$).

Q5 (*filter*) asked participants to list "What animals are quadrupedal and slow?" Here, 9 of the 10 participants in the Both group chose to use PI to answer this question (1 participant from the Both group did not answer this question), matching our hypothesis on the interaction type, though only 5 of these 10 provided a fully correct answer. Interestingly, we found that the OLI group outperformed both the PI and Both groups with respect to correctness, with 9 fully correct answers and 1 partially correct answer. Based on video analysis, it appears that some of the PI participants were confused by the fact that they needed to consider animals that were high on the Quadrupedal attribute but low on the Speed attribute, and mistakenly answered with animals that were quadrupedal and fast. Since the OLI participants were more focused on the animals, they used their own knowledge about animals to initially focus more accurately on appropriate matches. However, OLI participants did more manual searching for potential matches. Two of the OLI participants struggled with this question, and were outliers with very long performance time.

Q6a (*characterize distribution*) asked the participants to describe the distribution of the Agility attribute. This is also an attribute-centric task, and 8 of the 10 participants in the Both group appropriately used PI to answer this question, while 2 participants performed only surface-level interactions. Most PI participants increased the weight of the Agility characteristic to generate an almost 1-dimensional projection of agility (small weights on the other dimensions provide some jitter in the plot). Then they could use this plot to estimate the approximate density distribution of observations over agility. Some participants simply used surface-level interaction to select all the data points and viewed the resulting density of points on the Agility attribute slider, though the small size of the slider made it difficult to decipher.

Similarly, Q6b (*correlate*) asked the participants to list other attributes strongly related to agility. We expected participants to use PI to increase the weight of other attributes along with Agility to create a series of 2D plots. However, many participants resorted to surface-level interaction for this task by selecting agile animals (based on the previous task) and then simply viewing their distributions displayed on all the weight sliders. Thus, because this task was a follow-up to the previous task (Q6a), the participants in the Both condition had already performed PI to increase the weight of Agility. Otherwise, it is likely that they would have begun this task by performing PI on Agility. We believe this is the primary reason that the expected interaction type did not match our hypothesis for this task.

5.1.3 Observation-Level Interaction Questions. We hypothesized that OLI would be the best match for questions Q7a and Q7b since these questions centered on observations in the dataset. 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." Question Q7b was a follow-up mirror of question Q7a to encourage participants to think about multiple clusters, not just one cluster. Because the participants could build on their previous solution, the task performance time for Q7b was much faster than Q7a. In

confirmation of our hypothesis, 7 of the 10 participants in the Both condition chose to strictly use OLI to answer the questions Q7a and Q7b (Fig. 6) when given the choice of interaction types.

The following examples illustrate particularly insightful answers given by OLI participants:

> Q7a: "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."
>
> Q7b: "Fox, Bobcat, Lion, Leopard. Some of the characteristics that distinguish the two groups are: Active, agile, meat teeth, hunter, stalker."

To answer Q7a and Q7b, most OLI participants exploited the ability to manipulate observations to construct the two required clusters and then discover other similar animals placed by the algorithm in each of the two newly formed clusters as shown in Fig. 5. PI participants took a more indirect route. Typically, they first selected the elephant and blue whale points to view the raw data. As they viewed the data, they identified attributes where the elephant and blue whale were similar (such as Hairless and Speed), and then increased the weight of these attributes using PI. Interestingly, this cognitive process closely mimics that of the $WMDS^{-1}$ algorithm in which attributes that define or distinguish between clusters of points gain a higher weight. This finding confirms that the OLI process can serve to support participants' cognitive processes, by reducing cognitive load and interactive burden in completing such tasks. It also confirms the general strategy applied in the design of the OLI optimization algorithm.

5.1.4 Both-Interaction Questions. Q8 (*domain knowledge*) asked participants to "characterize and compare vegetarians, carnivores, and omnivores." Since the data did not contain information specifically about diet, participants had to exploit their domain knowledge to connect the existing data to diet in order to answer this question. We hypothesized that this task could enable both types of interactions effectively. We expected that some participants would apply their domain knowledge about individual animal's diets to organize the observations, and then learn about the relevant highly-weighted attributes. One of the Both participants did strictly use OLI in this way. Conversely, five of the Both participants instead applied their domain knowledge directly to the attributes. In this case, these participants scanned through the list of attributes and used PI to increase the weight of attributes they believed from prior knowledge were important for explaining diet. However, it seemed that participants tended to have more knowledge about the attributes than the animals in our experiment, or at least they found it easier to guess about the dietary relationships of the attributes than the animals. Because some of the OLI participants did not know the diet of specific animals, it was difficult for them to make progress with only OLI interaction available, and their guesses resulted in more incorrect answers.

Most interestingly, four of the Both participants applied their domain knowledge in both ways, using both types of interactions together in an iterative way to refine their final solution. They produced thorough, well-formed answers, such as:

> "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."

Q8 is the task where participants in the Both condition most effectively applied both types of interactions in a coordinated way. It is interesting that, when given the opportunity to apply their own domain knowledge, having both types of interaction enabled participants to apply their knowledge in a flexible manner, exploiting knowledge about both observations and attributes.

5.2 Insights: Questions 9–10

Questions 9 and 10 asked participants to list a number of insights, to give them an opportunity for free-form, open-ended exploration. Using the insight-based evaluation method, we analyze characteristics of each listed insight [Saraiya et al. 2005]. Across all 30 participants, there were a total of 93 insights listed. Each insight was examined by a single evaluator to determine the cardinality, dimensionality, and task diversity for that insight (each of these is described in more detail in the following subsections). We analyze the insights according to these three properties, plus the inclusion of domain knowledge and total time to acquire (as seen in Table 5).

Table 5. Insight Summary Statistics

Group	Insight Count	Average Dimensionality	Average Cardinality	Insights using Domain Knowledge	Average Time to Acquire (minutes)
Both	28	2.04	0.89	15 (54%)	11.43
OLI	30	2.13	1.40	7 (23%)	10.38
PI	35	1.86	1.03	17 (49%)	8.32

5.2.1 Dimensionality and Cardinality. The goal of OLI and PI are to explore large high-dimensional datasets and enable insights that are more complex in terms of dimensionality and cardinality. The hope is that such interactions can enable analysts to generate insights that are high in each of these characteristics. We hypothesized that PI will support increased dimensionality and that OLI will support increased cardinality of insights. Thus, we classify each insight based on its dimensionality and cardinality.

Dimensionality refers to the number of attributes specifically mentioned within an insight. The number of attributes per insight ranged from 0 attributes (0D) to 10 attributes (10D). Also, a few insights generically referred to "many" attributes, but we score these as 0D in the quantitative analysis of dimensionality because they lack specificity. The left side of Fig. 9 shows the dimensionality of the insights by interaction group. Interestingly, in contrast to our hypothesis, OLI participants produced many high-dimensional (higher than 3D) insights, with 23% of their insights being high-dimensional. For PI participants, 14% of their insights were high-dimensional. Participants in the Both group produced the most evenly distributed dimensionality insights, with at least 14% in each level of dimensionality.

Cardinality refers to the number of observations specifically mentioned within an insight. An insight has cardinality if it specifically references one or more observations in the dataset. The right side of Fig. 9 shows the cardinality of the insights by group. Cardinality ranged from 0 to 6. Most importantly, OLI participants made numerous insights containing a cardinality of 2 observations (50% of their insights). This is because they frequently investigated pairs of animals by dragging them together.

Also in contrast to dimensionality, there were many (over 50%) insights that generically referred to "many" animals or referenced groups of animals. But, again, unless they specifically mention some animals, we scored them as 0 cardinality for the analysis. Participants often meaningfully labeled these clusters within the insight. For example, one participant referred to "grazer animals" in 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."

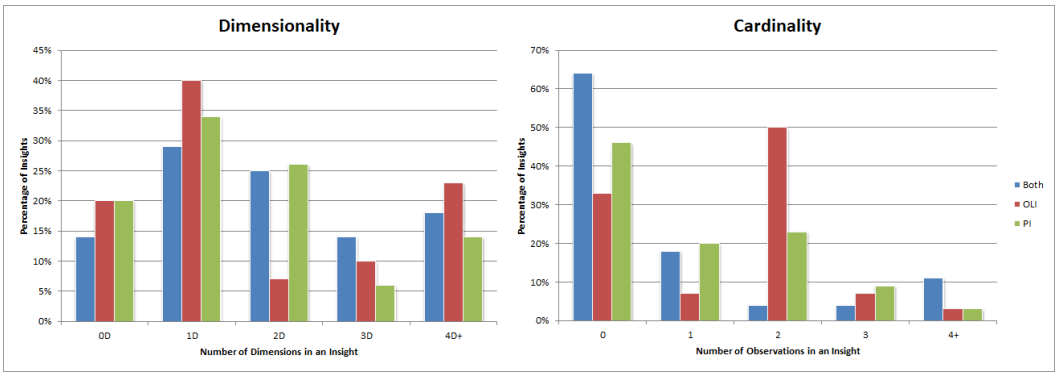

Fig. 9. Dimensionality describes the number of attributes explicitly mentioned in each insight, while Cardinality refers to the number of observations explicitly mentioned in each insight. Here, we show the frequency of both Dimensionality and Cardinality measures with respect to the participants' group. Our hypothesis that PI supports increased dimensionality of insights is not well supported, as OLI-generated insights with three or more dimensions are more frequent than PI-generated insights with three or more dimensions. Our hypothesis that higher cardinality will be seen more in OLI-generated insights than in PI-generated insights is also difficult to justify from the results. Insights with a cardinality of two were much more commonly generated by OLI, but insights with a cardinality of three or more were approximately similar between PI and OLI participants.

This insight has generically high cardinality since it references many animals, specifically the "grazer" animals (but we scored it as 0 cardinality in Figs. 9 and 10). It also has high dimensionality because it explicitly includes 5 attributes from the dataset. This is a good example of a highly complex insight because of its dimensionality and cardinality. In many similar cases, participants did not specifically mention one or more animals, but instead referred to either a group of animals or all of the animals. This especially occurred with the participants in the Both group, resulting in their low cardinality of insights.

Interestingly, with regards to dimensionality, we did not see this phenomenon of discussing all the attributes or a group of attributes collectively. In other words, when participants mentioned a group of attributes, they tended to list out the attributes by name. Considering dimensionality and cardinality together, only one insight had both dimensionality and cardinality greater than 3, as shown in Fig. 10.

5.2.2 Task Diversity. To analyze the task diversity of the insights, we categorized each insight according to the low-level tasks [Amar et al. 2005] that they contained, as shown in Fig. 11. Each insight could contain multiple low-level tasks. While the insights stemmed from a diversity of tasks, some tasks were better supported by specific interaction types. As hypothesized, OLI usage produced many *cluster* task insights. Indeed, *cluster* tasks focus on the observations, which lends itself to OLI. Likewise, participants who had PI or Both produced many *correlate* task insights, but not when using OLI alone.

5.2.3 Domain Knowledge. Many participants across all three groups exploited domain knowledge within their insights (see "Insights using Domain Knowledge" in Table 5). The following insight, provided by a participant in the Both group, is an example of a well-formed and complex insight (with how complexity of insights was determined explained in a following subsection) in which the participant exploited their domain knowledge:

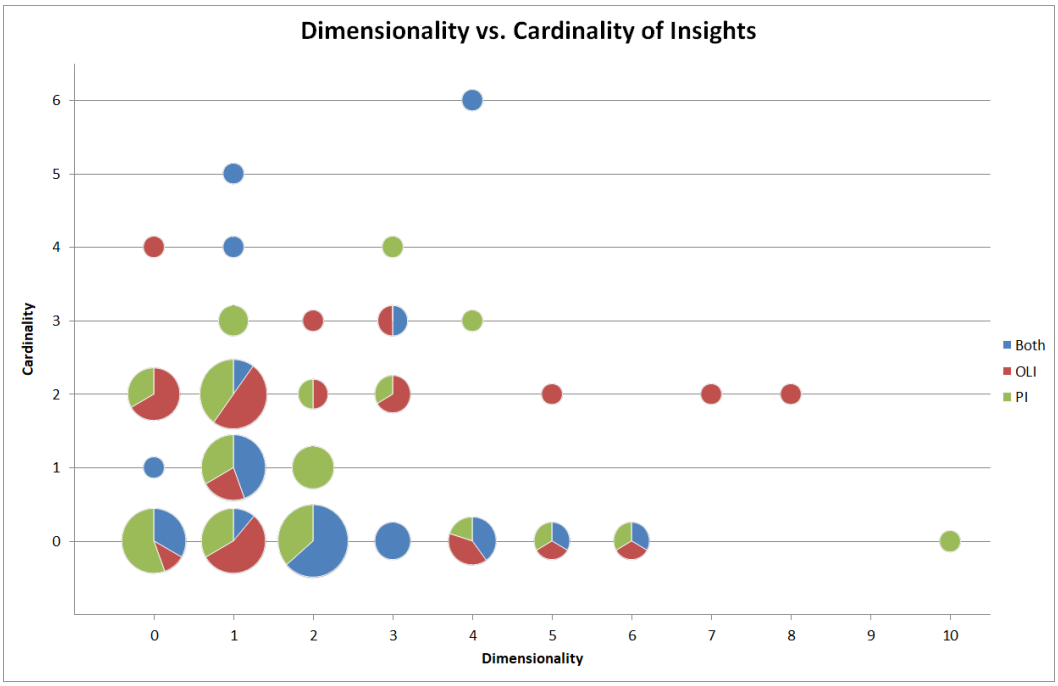

Fig. 10. A bubble plot showing the dimensionality and cardinality of each of the insights, where the bubble area represents the frequency of each (dimensionality, cardinality) pair, and each pie chart reflects the count of insights at that (dimensionality, cardinality) pair for the three groups. Many of the insights provided by participants had low dimensionality and low cardinality, though some exceptions provided either high dimensionality or high cardinality. One insight provided both high dimensionality and high cardinality. Many OLI group insights had a dimensionality of 2, as these interactions typically compare two observations.

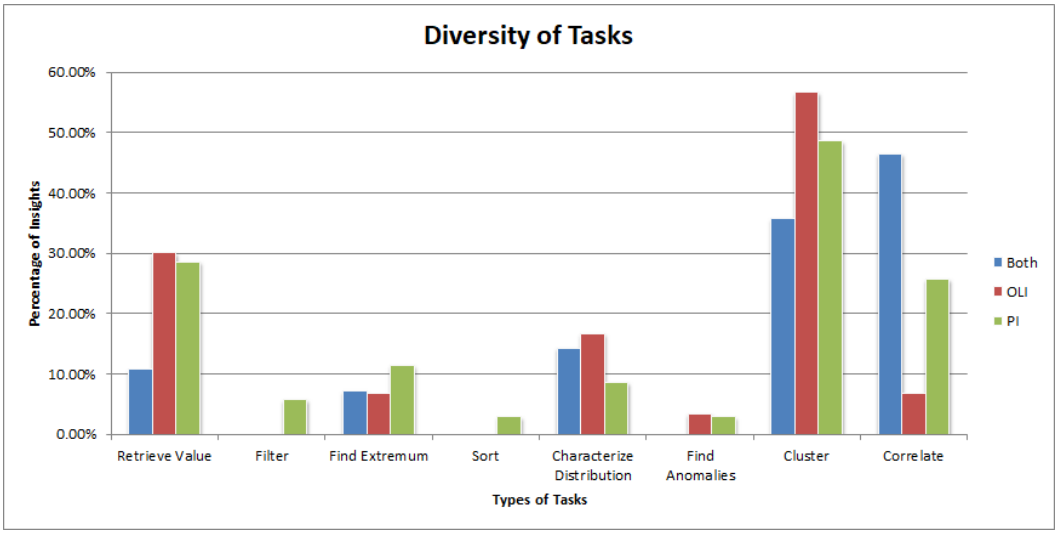

Fig. 11. Distribution of insights across low-level tasks shown as the percentage of insights that contained at least one of the tasks. For example, OLI usage concentrated on cluster task insights but not correlations.

> "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."

The participant used her domain knowledge to group animals she knew to be good pets. This produced an insight that has high general cardinality (generic groups, such as dogs and cats) and high dimensionality (5 attributes), making it a valuable complex insight. 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. The participant used multiple low-level tasks (e.g. *cluster*, *correlate*), 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:

> "From the same layout I used to find desirable pet characteristics I can compare animals that may not make good pets. Bobcats, wolves, lions, tigers, weasels, and skunks are the farthest removed from the typical pet animals so I selected them. 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 attributes and references a list of animals, creating an insight with both high dimensionality and cardinality. Her domain knowledge spun off into excellent insight.

5.2.4 Time to Acquire. Participants in the Both condition spent 3 more minutes analyzing the data than those participants with PI-Andromeda (see "Average Time to Acquire" in Table 5 for average duration). Overall, PI was the most efficient at producing insight at a rate of 2.4 minutes per insight, versus 4.1 minutes per insight for Both. Perhaps this was due to the increased familiarity with this type of interaction.

Participants in the Both group took advantage of both types of interactions, OLI and PI, throughout their exploration. On the positive side, these participants switched between the two interaction types multiple times either for a new insight or to approach one insight from different angles. On the negative side, this approach appeared to slow them down. Based on video analysis, it seems that these participants felt compelled to make use of both kinds of interactions so as to be thorough in their analysis, and therefore took longer. It is not surprising that OLI participants took longer to explore, because the OLI usage requires more interactive steps to operate (selecting and moving data points, invoking and waiting for the algorithm to update the projection), whereas PI operates in real-time as analysts drag the weight sliders.

5.2.5 External Evaluation. To understand the value and validity of the insights gained by the participants, we recruited a group of 26 graduate students (none of whom participated in the usability study) to read, rate, and comment on the insights for evaluation. Each insight was evaluated by at least 3 unique students, with an average of 5 unique students evaluating each insight.

The insights were rated based on **complexity** and **interestingness**. Complexity is a measure of how much data is involved in the insight, which provides a quantitative measure of the insight itself. To decide on complexity, raters might consider the involvement of all or large amounts of the data in a synergistic way, the uncertainty of the entire insight or a small part of the insight, or the domain knowledge included and connected to the data and the insight itself. Raters chose

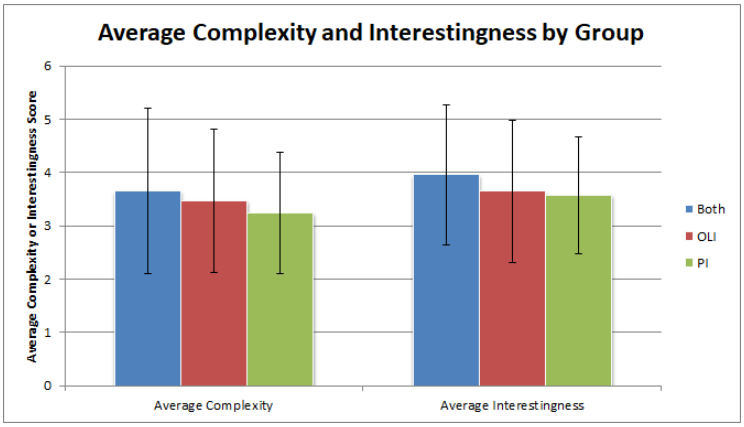

Fig. 12. The average scores provided by raters for the complexity and interestingness of the insights created by each of the three groups. Error bars represent one standard deviation. The differences between groups are not statistically significant.

complexity on a scale from simple (1) to complex (7). Raters also rated each insight on a scale from uninteresting (1) to very interesting (7). Interestingness brings in subjectivity and provides a qualitative measure of the value of the insight. It can be based on the impact it might make on the domain and the meaningfulness of the insight. Lastly, we asked the raters to explain what factors they considered and their reasoning behind the ratings for each insight. The average complexity and interestingness scores across all three groups are shown in Fig. 12.

While we did not find significant differences for complexity or interestingness across the interaction types, we did find that complexity and interestingness are quite correlated with each other (Pearson $r = 0.807$). This suggests that a more complex insight will also be more interesting than a less complex insight, and further suggests that interestingness could be measured by complexity and vice versa. Unfortunately, complexity ratings did not directly correlate with our more automated metrics of dimensionality and cardinality (individually or their sum), as shown in Fig. 13.

The raters not only rated each insight, but also explained the factors used for the ratings. Overall, highly rated insights tended to be assertions or high-level conclusions. These insights contained unexpected results, subjectivity, and uncertainty. Insights that they thought represented more high-dimensional conclusions were rated more highly. Raters preferred insights that were backed with facts from the raw data and combined with outside domain knowledge. For example, the previously mentioned insight about house pets received an average complexity score of 5.8 and an average interestingness score of 6 (both out of 7). The raters gave such explanations as:

> "This insight is very interesting as it confirms our hypothesis on pet animals as well as throws new knowledge on domestic animals which might have [an] impact in the domain. It is also complex in nature as it involves large amount of data and also good number of characteristics and its inter-relationships."
>
> "Complex because multiple categories are involved and compared. Interesting because of the real-world applications."
>
> "Takes some domain knowledge of what types of pets there are. Formed hypothesis and used the data to support it. Found interesting insight in activity level."

Insights that they considered "common knowledge" were assigned lower ratings. For example, "Pigs are domestic" received a complexity score of 1 and an interestingness score of 1. They state

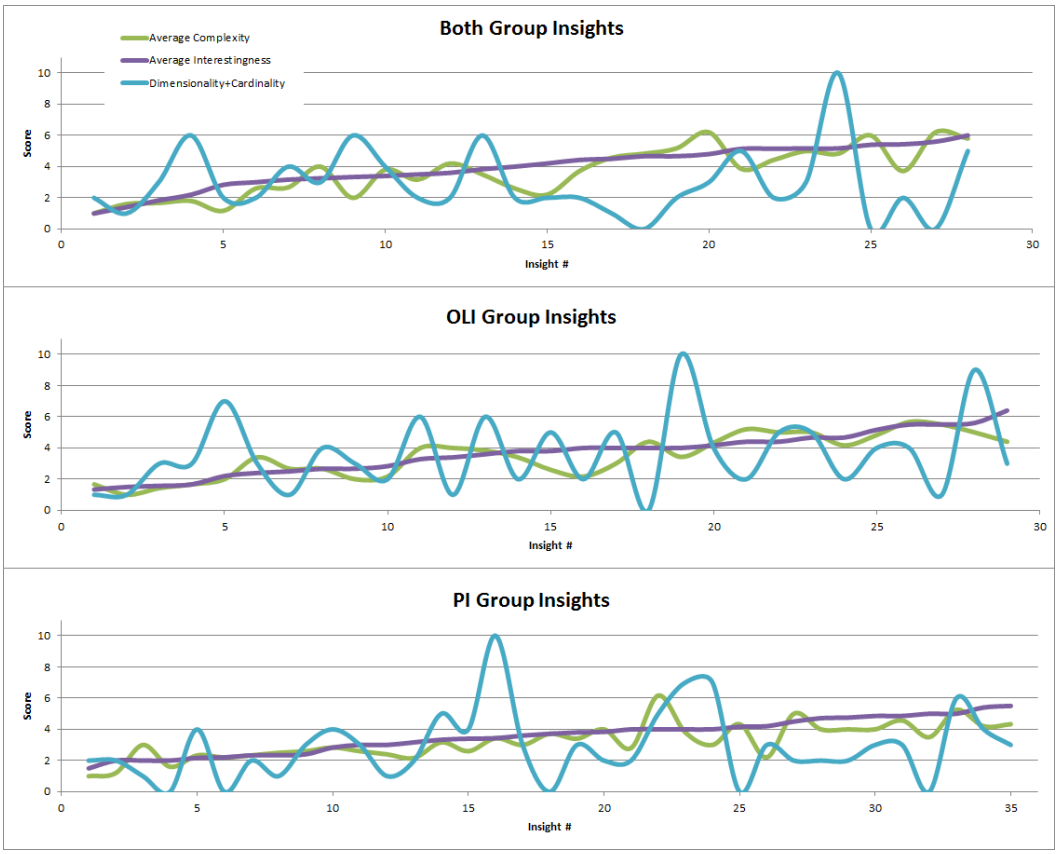

Fig. 13. The scores provided by the raters for each of the insights, showing measures of complexity, interestingness, and dimensionality+cardinality. Insights for the Both, OLI, and PI groups are separated into three charts, sorted by interestingness from low to high. The correlation between complexity and interestingness can be observed, as well as the lack of correlation between these measures and the combination of dimensionality and cardinality.

"...it's very simple (only one data point and only one feature)" and "it is common sense, thus uninteresting." Additionally, raters rated an insight low if it included incorrect knowledge. However, students rated an insight high if it identified possibly incorrect raw data, stating that finding a discrepancy in the data is important.

Even though these raters did not directly interact with Andromeda, they also considered the perceived difficulty of the interaction steps when rating the insights. If an insight seemed to require more complex interactions or multiple steps to get to, raters rated it more highly. An insight that is merely the result of a *retrieve value* task (could be found straight from the data table) was rated low.

6 DISCUSSION

During this study, we sought to discover how the three different types of interaction (Both, OLI, and PI) support high-dimensional data analyses of non-experts. First, our analysis showed that the study participants were able to explore the high-dimensional data by using these interactions. Our participants were unfamiliar with Andromeda before the study, yet were able to quickly learn

to use these interactions. While performing their analyses, participants learned simple WMDS concepts, which was enough to efficiently use the tool to answer questions and gain insights about the data. Even though participants did not completely understand how the algorithm worked, it did not hinder their exploratory analyses. We also found that many participants who lacked either one of the two interaction types were still able to correctly answer questions that we thought were specifically geared toward that missing interaction type, albeit with some additional difficulty.

Overall, participants tended to focus on the attributes when using PI, which supported the following tasks:

- **Attribute based filtering.** Participants found observations that satisfied given criteria on single or multiple attributes by using parametric interaction to increase the weight of the criteria attributes (in a fashion analogous to dynamic queries [Shneiderman 1994]) and then find the relevant cluster in the observation view.
- **Finding distributions and relationships between attributes.** By using parametric interaction, participants were able to describe single attribute trends across all points by highly weighting one attribute. Participants also correlated different attributes. One participant discovered that bulbous animals have a strong tendency to swim. Such discoveries are afforded by parametric interaction, in which the analyst can increase the weight of the desired attributes and view the distribution of data in the observation view.

Participants focused on the observations with which they were more familiar when using OLI, which supported the following tasks:

- **Comparing observations and clusters.** Participants compared observations and clusters of observations to discover what attributes best explain each cluster. Participants frequently pushed observations together or apart to identify their similarities or differences. Common operations included creating a new cluster apart from all other observations, distinguishing multiple new clusters of observations, or pushing existing clusters together. OLI lended itself well to this task, allowing participants to directly cluster points to find distinguishing attributes. This task was less fluid when using PI. A related task was **forcing outliers into or out of clusters** to discover what attributes differentiate or relate the outliers.

Some tasks were best supported by the combination of OLI and PI:

- **Using domain knowledge.** Participants exploited their pre-existing knowledge of the data attributes and observations to better connect with the data. For example, a participant wanted to determine what characteristics might describe animals that humans have as pets. The participant had meta-attributes she wanted to describe with the dataset attributes. She had a preconceived notion in mind about whether an animal fits this meta-attribute of pet. She clustered pets and non-pets to determine the defining characteristics and discover any other animals that might be good pets based on model calculations.

With the knowledge of the distinct differences of PI and OLI and the affordances each offer, the visual analytics and HCML communities can make informed decisions about the components they want to include within a tool and tailor them to the particular purposes of the tool.

6.1 RQ1: Benchmark Tasks

To answer RQ1, we did not find strong effects of interaction on the correctness of the participants' answers or their speed of performance. In general, the speed and correctness results only provide supporting evidence for the main finding concerning the expected interaction type.

Most importantly, if given both PI and OLI, participants usually correctly chose the interaction that we hypothesized to be most appropriate for the type of question, confirming our hypothesized

matching of tasks and interaction types. As shown in Fig. 6, we found that questions based on one or more attributes (Q4–Q6: filter, distribution, correlation) best utilized PI, whereas questions asking specifically about observations (Q7: cluster) were best suited for OLI. The one case where participants most exploited both types of interactions in a single task was the domain knowledge question (Q8), though participants tended to have more knowledge about attributes than observations. The seemingly refuting questions (Q3 retrieve value, and Q6b correlate) are understandable given the task context as discussed in the results section , but hint at a possible deeper connection between PI and surface-level interaction as discussed later.

These results also support both the "Match-Mismatch Hypothesis" of Gilmore and Green [1984] and the "Cognitive Fit" theory of Vessey [1991]. In both cases, these refer to the idea that when the notation or representation of the data emphasizes some aspect or type of information, a task that requires that type of information can be more easily performed. Gilmore's study focused on the properties of programming notations, while Vessey's study examined the uses of graphs and tables in decision making. Despite these studies surveying very different domains, both came to the same conclusion. Likewise, we hypothesized that tasks requiring the manipulation of dimension weights to emphasize or deemphasize attributes could be more easily performed via PI, since PI naturally afford the analyst with control over those attributes and weights. Similarly, tasks requiring a perspective on the observations could be more easily performed via OLI, since OLI affords easy comparison and manipulation of the observations.

6.2 RQ2: Insights

To answer RQ2 with the open-ended insight tasks, we found that the type of interaction used affects the types of insights gained. OLI interaction tended to produce more observation-oriented insights, such as clustering, and emphasized insights with cardinality of 2. Conversely, PI produced more correlation and 2D insights than OLI. Thus, given their differences in affordances, it might be argued that when PI and OLI are taken together they can provide for a more diverse analysis in terms of tasks types, dimensionality, and cardinality. However, this is not entirely clear since the results of the Both condition did not necessarily reflect the simple combination of the OLI and PI results. For example, the Both condition was more focused on 0 cardinality insights than either of the PI or OLI conditions (though many of these were generically high-dimensional).

6.3 Learnability and Usability Issues

Following our initial evaluation of participants' time and accuracy for the benchmark tasks, we re-examined the videos of their interactions in an attempt to discern why there was such a high variance in these results. In general, we found that there was a high variance in participants' speed of learning, desire for thoroughness of analysis, and familiarity with the data. When combined, this translated into high variance in the performance results.

Although there was a training period, multidimensional data analysis is complex and the participants learned at very different speeds. Some participants were still learning how to use Andromeda and its capabilities throughout the course of the study, trying alternative strategies to answer various questions. Thus, the performance metrics did not necessarily measure only the optimal strategies for each task, but also included some high-variance learning time. Based on participants' verbal feedback, it appeared that they found OLI more difficult to learn than PI, perhaps because PI is a more familiar concept. Clearly, there is more opportunity for future research on the learnability of these interaction techniques.

One specific learnability problem we found was that many participants ignored the green selections in OLI that indicate whether a data observation is to be considered in the OLI WMDS^{-1} algorithm's computation of the updated projection. Andromeda automatically highlights points

that the analyst moves, but also automatically highlights other nearby reference points. The automatic highlighting of reference points sometimes hindered participants rather than helped. It seemed that participants did not fully understand what the green highlighting meant, despite the pre-study tutorial. This resulted in participants unintentionally highlighting other nearby points when attempting to use OLI to answer a question. Thus, when they clicked the "Update Layout" button, undesired points were considered by the WMDS^{-1} algorithm, resulting in a projection that did not reflect the participants' intentions. The automatic highlighting was designed to solve a problem found in previous usability studies in which analysts wanted to move points with respect to reference points [Self et al. 2016]. However, because of the design of the OLI clustering tasks in this study, the participants only wanted to consider the points specifically mentioned in the tasks. New interaction designs are needed that encompass both of these types of OLI tasks, such as the use of a second view for moved points or including explicit representations of cluster boundaries [Wenskovitch and North 2017].

Yet, there is also evidence of the intuitiveness of the OLI and PI interactions. For example, during the study we noted that some participants in the PI or OLI conditions asked if there was a way to perform the other type of interaction without knowing that the interaction existed but was disabled. One participant using PI-Andromeda actually tried to move the points with her mouse. Some participants recognized these natural opportunities to explore the data in a different way and wanted them.

6.4 Asymmetry of Interaction Design

Another potential problem we discovered in the design of Andromeda is asymmetry in the design of the OLI, PI, and surface-level interactions that might cause learnability difficulties. These asymmetries might be the result of design bias due to greater familiarity and history of PI, and they indicate the need for more research on the design of effective OLI interactions and their integration with PI and surface-level interactions.

An asymmetry in the surface-level interaction makes it more closely linked to PI than OLI. Surface-level interactions can select an observation to view its attribute values on the attribute sliders. However, there is no corresponding interaction to select an attribute and view its values on the observations. Thus, surface-level interaction can be used to easily find relevant attributes, but not to easily find relevant observations. It would require tediously mousing over each observations one at a time. This encourages possible swapping of PI and surface-level interactions for some tasks, such as in Q3 (retrieve value) and Q6b (correlate). It also suggests opportunities for projecting attribute values onto observation glyphs, such as encoding the size of observations' dots based on a selected attribute.

Another asymmetry is that OLI is inherently multi-object, while PI is not. That is, for OLI to function, the analyst must manipulate multiple observations to express a set of desired distances. This means that OLI requires more steps to operate, and its greater complexity might make it more difficult to learn than PI. This suggests possible opportunities for more symmetric designs of OLI and PI interactions, such as sorting attributes based on selected observations or, more generally, projecting attributes into a weighted observation space.

6.5 Future Work on Education

The final seven questions (Q11–Q17, see Appendix A.1) on the survey sought to examine whether interactive exploration using OLI and PI supported the participants in learning basic WMDS statistical concepts. 90% of the participants did not have any previous experience with WMDS and no participants considered themselves experts. Because of this, any knowledge gained is from using Andromeda. While we did not find major differences between the interaction conditions,

the overall results indicate the potential value of these techniques for education about dimension reduction that should be further investigated.

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 valid. When directly asked (Q17), over 90% of participants, regardless of the interaction type given, correctly explained that there is no wrong projection, but rather the projection depends on the question being explored.

To understand WMDS, analysts must also understand the interplay between the attribute weights and the point locations within the projection. For example, Q11 specifically asked participants to predict how and why increasing the weight of the "smartness" attribute (PI) would impact the relative locations of a specific set of points. 80% of participants using Both-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. Surprisingly, even though OLI participants had not experienced the ability to directly modify the weights (PI), 60% of those participants correctly answered this question. They understood enough about the effect of the reverse interaction (OLI) of moving the specific points, which updates weights, to explain the hypothetical scenario of directly modifying a weight.

Certain lower-level WMDS concepts proved more difficult for the participants to grasp. Over 90% of participants across all three interaction types could not mathematically describe how Andromeda maps observations to locations in the plot. However, most participants understood conceptually that the placement depended on the amount of similarity between the more highly weighted attributes. The difficulty lies in understanding that the algorithm is trying to minimize the stress between the low-dimensional projected distances and high-dimensional distances 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 answers the participants provided to this question proved they understood the *high-level* conceptual overview of the model and how the positions of the low-dimensional points depended on the relative importance given to all attribute weights. Using this knowledge, most participants understood how to analyze the data, get answers to the questions, and gain novel insight.

This preliminary data suggests that interactive exploration using OLI and PI supports students in learning the high-level concepts of dimension reduction. Indications from our results here, as well as from related studies [Chen et al. 2016; Self et al. 2017], show that this is a promising open research area for education.

6.6 Limitations

This study also had several limitations that, if addressed, could produce more detailed results. First, we did not pre-screen for prior domain knowledge. For example, we noticed that some of the international student participants were not very familiar with the English names of some of the more unusual animals. This is a confound that may have caused them to focus on the attributes rather than the observations for some of the tasks such Q8, which required the most domain knowledge.

Second, we could have worded some of the tasks more specifically, such as asking for a specific number of attributes or observations in the benchmark task questions, to make the scoring system simpler and more definitive. Also, we noted that some statistical terms, such as "distribution," may not have been familiar to some of the participants, or they may have had difficulty describing the distribution without multiple-choice prompts.

Third, while we attempted to carefully produce the introductory instructional video, it is possible that the video may have biased the participants towards certain types of interactions or insights.

7 CONCLUSIONS

In this paper, we examined the roles of OLI and PI in high-dimensional data analysis. Key findings are:

- When performing benchmark tasks, our hypotheses about which interactions participants would choose to accomplish certain tasks were mostly confirmed. The interaction type did not have clear effects on task performance time and accuracy.
- The interaction type emphasizes insights on either the observations (OLI), including cluster and 2-cardinality tasks, or the attributes (PI), including correlation and 2D tasks. Given both types of interaction, analysts apply their domain knowledge to the analysis.
- The type of interaction did not affect the analysts' *high-level* understanding of WMDS.

We stress the importance of understanding the distinct differences between PI and OLI. This study particularly clarifies the *distinction* of OLI for interacting with dimension reduction models. OLI is fundamentally different than PI given its observation-centric nature, and it supports analysis from a different angle than that of PI. In particular, OLI emphasizes clustering and cardinality. By understanding the differences, advantages, and drawbacks of PI, OLI, and their combination, the visual analytics and HCML communities can better advance the use of dimension reduction models.

APPENDIX

A.1 Study Questions

In this appendix, we provide the questions asked to the participants for our study.

Participant Information:

(1) Please enter the provided participant number and the date.
   - Participant #:
   - Group #:
   - Date:
   - Age:
   - School level (i.e. freshman, sophomore):
   - Major:

(2) How familiar are you with Multidimensional Scaling?
   - Never heard of it (1)
   - Have heard about it, but have never used it (2)
   - Learned about it in a class (3)
   - I am an expert (4)

Benchmark Tasks (see Section 5.1):

Use Andromeda to analyze and provided animal dataset and answer the following questions. The dataset includes 49 animals and 75 characteristics describing the animals. For each characteristic, 100 means high and 0 means low. For example a grizzly bear is very furry, therefore furriness would be close to 100, but a fish would be 0.

(3) How likely is it for the gorilla to live in the jungle?
(4) What animal is most timid?
(5) What animals are quadrupedal and slow?
(6) (a) Describe the distribution of the agility characteristic. (b) Are any characteristics strongly related to agility? How? Why?

(7) (a) The elephant and blue whale are similar to each other, but are dissimilar to the tiger and wolf. What other animals are like the elephant and blue whale, but not like the tiger and wolf? Why? (b) What animals are similar to the tiger and wolf, but dissimilar to the elephant and blue whale? Why?
(8) Characterize and compare vegetarians, carnivores and omnivores.

Insights (see Section 5.2):

(9) Use Andromeda to explore and learn about the data. Write a list of 4 or more interesting insights you gain from the data and justify each with appropriate rationale including evidence to back up your claims.
(10) Write your reflections on how you conducted your analysis.

WMDS Concepts (see Section 6.5):

(11) (a) Without Andromeda, briefly explain how and why increasing the weight for smartness would impact the relative locations of the following data points: ox, polar bear, fox, and cow. (b) Using Andromeda, check your explanation. Did the answer surprise you?
(12) (a) Without Andromeda, briefly explain how and why moving data points polar bear, deer, and hamster close together and zebra far from them would impact the weight of chewteeth. (b) Using Andromeda, check your explanation. Did the answer surprise you?
(13) Within Andromeda's data point view, what do the distances between points mean?
(14) Briefly describe how Andromeda maps data points to locations in the plot.
(15) Why does adjusting the weights change the plot?
(16) Why does adjusting the positions of the points in the plot change the weights?
(17) Since Andromeda can create multiple plots from the same high-dimensional dataset, which plot is right? Please explain your answer.

Wrap Up:

(18) List any significant usability problems you encountered when using the Andromeda.

A.2 Glossary

In this appendix, we provide a concise list of terms used throughout this work.

(1) **Andromeda** – An interactive visual analytics tool designed to demonstrate and study how interactivity aids analysts in the exploration of high-dimensional data.
  (a) **Observation View** – One view of Andromeda, displaying the layout calculated by WMDS. Each point in the view represents one observation of the high-dimensional data.
    (i) **View Mode** – A mode of Andromeda's Observation View, which allows analysts to explore and view the observations in the WMDS projection. Supported interactions include hovering over and selecting points view view the corresponding raw data.
    (ii) **OLI Mode** – Another mode of Andromeda's Observation View, which allows analysts to manipulate the observations, thereby creating a new set of low-dimensional coordinates. Supported interactions include clicking and dragging observations. These moved points are what are used by the WMDS^{-1} inverse algorithm triggered by clicking the "Update Layout" button above the observation view.
  (b) **Parameter View** – A second view of Andromeda, displaying the weighted attributes in the input dataset. Categorical data is displayed as static text, while numerical attributes are displayed by attribute sliders that represent the relative weight of the associated attribute.

(i) **PI Mode** – The interaction mode associated with Andromeda's Parameter View, allowing an analyst to drag the handle on the attribute sliders to the right to increase or left to decrease the weight of that attribute.

(c) **OLI-Andromeda** – A version of Andromeda provided to study participants that only allowed for observation-level interactions.

(d) **PI-Andromeda** – A version of Andromeda provided to study participants that only allowed for parametric interactions.

(2) **Attribute** – A property or dimension of an item in a dataset. With respect to the animals dataset used in this study, the attributes are the quantitative properties of the animals such as color, diet, and habitat.

(3) **Complexity** – A measure of how much data is involved in an insight; used for evaluating insights in Section 5.2.

(4) **Interestingness** – A more subjective evaluation of insights that is based on creativity and synthesis, providing a qualitative measure to the insights described in Section 5.2.

(5) **Observation** – The items in a dataset. With respect to the animals dataset used in this study, the observations are the animals themselves.

(6) **Observation-Level Interaction (OLI)** – An interaction technique allowing an analyst to interact with individual points in a visual display through familiar and comfortable interactions, requiring the model to interpret the semantic meaning of each interaction to adjust the model parameters through the use of an inverted visualization algorithm.

(7) **Parametric Interaction (PI)** – An interaction technique allowing an analyst to directly adjust the underlying parameters of an algorithm or model, thereby modifying the output of that model.

(8) **Stress Function** – A function that measures the difference between the pairwise high-dimensional distances between observations and the low-dimensional distances between observations in the projection.

(9) **Surface-Level Interaction** – An interaction in a static projection that views the raw data, links data between views, or scales or rotates the projection, and thus does not alter the underlying model parameters.

(10) **Weight Vector** – The collection of weights that are associated with the attributes in the dataset.

(11) **Weighted Multidimensional Scaling (WMDS)** – A frequently-used dimension reduction algorithm that projects high-dimensional data onto low-dimensional space. The algorithm computes pairwise distances between observations in high-dimensional space, and then attempts to preserve those distances in the low-dimensional projection.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under grants IIS-1447416, IIS-1218346, and DUE-1141096.

REFERENCES

Jamal Alsakran, Yang Chen, Ye Zhao, Jing Yang, and Dongning Luo. 2011. STREAMIT: Dynamic Visualization and Interactive Exploration of Text Streams. In *2011 IEEE Pacific Visualization Symposium*. IEEE, 131–138.

R. Amar, J. Eagan, and J. Stasko. 2005. Low-level Components of Analytic Activity in Information Visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* 111–117. https://doi.org/10.1109/INFVIS.2005.1532136

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (2014), 105–120.

L. Bradel, C. North, L. House, and S. Leman. 2014. Multi-model semantic interaction for text analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST).* 163–172. https://doi.org/10.1109/VAST.2014.7042492

Matthew Brehmer, Michael Sedlmair, Stephen Ingram, and Tamara Munzner. 2014. Visualizing Dimensionally-Reduced Data: Interviews with Analysts and a Characterization of Task Sequences. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*. ACM, 1–8.

E. T. Brown, J. Liu, C. E. Brodley, and R. Chang. 2012. Dis-function: Learning Distance Functions Interactively. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 83–92. https://doi.org/10.1109/VAST.2012.6400486

M. Cakmak, C. Chao, and A. L. Thomaz. 2010. Designing Interactions for Robot Active Learners. *IEEE Transactions on Autonomous Mental Development* 2, 2 (June 2010), 108–118. https://doi.org/10.1109/TAMD.2010.2051030

Stuart K Card, Jock D Mackinlay, and Ben Shneiderman. 1999. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann.

Xin Chen, Leanna House, Jessica Zeitz Self, Scotland Leman, Jane Robertson Evia, James Thomas Fry, and Chris North. 2016. Be the Data: An Embodied Experience for Data Analytics. In *2016 Annual Meeting of the American Educational Research Association (AERA)*. 20.

J. Choo, H. Lee, J. Kihm, and H. Park. 2010. iVisClassifier: An Interactive Visual Analytics System for Classification Based on Supervised Dimension Reduction. In *2010 IEEE Symposium on Visual Analytics Science and Technology*. 27–34. https://doi.org/10.1109/VAST.2010.5652443

Kristin A Cook and James J Thomas. 2005. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. Technical Report. Pacific Northwest National Laboratory (PNNL), Richland, WA.

E. P. dos Santos Amorim, E. V. Brazil, J. Daniels, P. Joia, L. G. Nonato, and M. C. Sousa. 2012. iLAMP: Exploring High-Dimensional Spacing through Backward Multidimensional Projection. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 53–62. https://doi.org/10.1109/VAST.2012.6400489

Alex Endert, Patrick Fiaux, and Chris North. 2012a. Semantic Interaction for Sensemaking: Inferring Analytical Reasoning for Model Steering. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2879–2888.

Alex Endert, Patrick Fiaux, and Chris North. 2012b. Semantic Interaction for Visual Text Analytics. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 473–482.

Alex Endert, 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)*. 121–130.

Jerry Alan Fails and Dan R. Olsen, Jr. 2003. Interactive Machine Learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI '03)*. ACM, New York, NY, USA, 39–45. https://doi.org/10.1145/604045.604056

Eric D Feigelson and G Jogesh Babu. 2012. *Modern Statistical Methods for Astronomy: With R Applications*. Cambridge University Press.

James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlik: Interactive Concept Learning in Image Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 29–38. https://doi.org/10.1145/1357054.1357061

Stephen L France and J Douglas Carroll. 2011. Two-Way Multidimensional Scaling: A Review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 41, 5 (2011), 644–661.

Keinosuke Fukunaga. 2013. *Introduction to Statistical Pattern Recognition*. Academic press.

D.J. Gilmore and T.R.G. Green. 1984. Comprehension and Recall of Miniature Programs. *International Journal of Man-Machine Studies* 21, 1 (1984), 31–48. https://doi.org/10.1016/S0020-7373(84)80037-1

Michael Gleicher. 2013. Explainers: Expert Explorations with Crafted Projections. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2042–2051.

Isabelle Guyon and André Elisseeff. 2003. An Introduction to Variable and Feature Selection. *Journal of machine learning research* 3, Mar (2003), 1157–1182.

Leanna House and Chao Han. 2015. Bayesian Visual Analytics: BaVA. *Statistical Analysis and Data Mining* 8, 1 (2015), 1–13.

Xinran Hu, Lauren Bradel, Dipayan Maiti, Leanna House, and Chris North. 2013. Semantics of Directly Manipulating Spatializations. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2052–2059.

S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. MÃűller. 2010. DimStiller: Workflows for Dimensional Analysis and Reduction. In *2010 IEEE Symposium on Visual Analytics Science and Technology*. 3–10. https://doi.org/10.1109/VAST.2010.5652392

Anil K Jain, Robert P. W. Duin, and Jianchang Mao. 2000. Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 1 (2000), 4–37.

Dong Hyun Jeong, Caroline Ziemkiewicz, Brian Fisher, William Ribarsky, and Remco Chang. 2009. iPCA: An Interactive System for PCA-based Visual Analytics. In *Computer Graphics Forum*, Vol. 28. Wiley Online Library, 767–774.

Sara Johansson and Jimmy Johansson. 2009. Interactive Dimensionality Reduction through User-Defined Combinations of Quality Metrics. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 993–1000.

Paulo Joia, Danilo Coimbra, Jose A Cuminato, Fernando V Paulovich, and Luis G Nonato. 2011. Local Affine Multidimensional Projection. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2563–2571.

Ian Jolliffe. 2002. *Principal component analysis*. Wiley Online Library.

Eser Kandogan. 2000. Star Coordinates: A Multi-Dimensional Visualization Technique with Uniform Treatment of Dimensions. In *Proceedings of the IEEE Information Visualization Symposium*, Vol. 650. 22.

E. Kandogan. 2012. Just-in-Time Annotation of Clusters, Outliers, and Trends in Point-Based Data Visualizations. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 73–82. https://doi.org/10.1109/VAST.2012.6400487

Tasneem Kaochar, Raquel Peralta, Clayton Morrison, Ian Fasel, Thomas Walsh, and Paul Cohen. 2011. Towards Understanding How Humans Teach Robots. *User modeling, adaption and personalization* (2011), 347–352.

Ashish Kapoor, Bongshin Lee, Desney Tan, and Eric Horvitz. 2010. Interactive Optimization for Steering Machine Classification. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 1343–1352. https://doi.org/10.1145/1753326.1753529

Joseph B Kruskal and Myron Wish. 1978. *Multidimensional Scaling*. Vol. 11. Sage.

Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, NY, USA, 126–137. https://doi.org/10.1145/2678025.2701399

Todd Kulesza, Simone Stumpf, Weng-Keen Wong, Margaret M. Burnett, Stephen Perona, Andrew Ko, and Ian Oberst. 2011. Why-oriented End-user Debugging of Naïve Bayes Text Classification. *ACM Transactions on Interactive Intelligent Systems* 1, 1, Article 2 (Oct. 2011), 31 pages. https://doi.org/10.1145/2030365.2030367

Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. 2012. Empirical Studies in Information Visualization: Seven Scenarios. *IEEE Transactions on Visualization and Computer Graphics* 18, 9 (2012), 1520–1536.

Christoph H Lampert, Hannes Nickisch, Stefan Harmeling, and Jens Weidmann. 2009. Animals with Attributes: A Dataset for Attribute Based Classification. (2009).

Scotland C. Leman, Leanna House, Dipayan Maiti, Alex Endert, and Chris North. 2013. Visual to Parametric Interaction (V2PI). *PloS one* 8, 3 (2013), e50474.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

Kantilal Varichand Mardia, John T Kent, and John M Bibby. 1980. Multivariate Analysis. (1980).

Tamara Munzner. 2014. *Visualization Analysis and Design*. CRC Press.

Jakob Nielsen. 1993. Iterative User-Interface Design. *Computer* 26, 11 (1993), 32–41.

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

Paulo Pagliosa, Fernando V Paulovich, Rosane Minghim, Haim Levkowitz, and Luis Gustavo Nonato. 2015. Projection Inspector: Assessment and Synthesis of Multidimensional Projections. *Neurocomputing* 150 (2015), 599–610.

Fernando V Paulovich, Cláudio T Silva, and Luis Gustavo Nonato. 2012. User-Centered Multidimensional Projection Techniques. *Computing in Science & Engineering* 14, 4 (2012), 74–81.

Daniel Pérez, Leishi Zhang, Matthias Schaefer, Tobias Schreck, Daniel Keim, and Ignacio Díaz. 2015. Interactive Feature Space Extension for Multidimensional Data Projection. *Neurocomputing* 150 (2015), 611–626.

PNNL. 2010. IN-SPIRE Visual Document Analysis. (2010).

Brian D Ripley. 2007. *Pattern Recognition and Neural Networks*. Cambridge university press.

Purvi Saraiya, Chris North, and Karen Duca. 2005. An Insight-Based Methodology for Evaluating Bioinformatics Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 11, 4 (2005), 443–456.

Matthias Schaefer, Leishi Zhang, Tobias Schreck, Andrada Tatu, John A Lee, Michel Verleysen, and Daniel A Keim. 2013. Improving Projection-Based Data Analysis by Feature Space Transformations. In *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 86540H–86540H.

Jessica Zeitz Self, Nathan Self, Leanna House, Jane Robertson Evia, Scotland Leman, and Chris North. 2017. Bringing Interactive Visual Analytics to the Classroom for Developing EDA Skills. In *Consortium for Computing Sciences in Colleges, Eastern Region (CCSC-ER)*. 10.

Jessica Zeitz Self, R.K. Vinayagam, James Thomas Fry, and Chris North. 2016. Bridging the Gap between User Intention and Model Parameters for Data Analytics. In *SIGMOD 2016 Workshop on Human-In-the-Loop Data Analytics (HILDA 2016)*. 6.

Jinwook Seo and Ben Shneiderman. 2006. Knowledge Discovery in High-Dimensional Data: Case Studies and a User Survey for the Rank-by-Feature Framework. *IEEE Transactions on Visualization and Computer Graphics* 12, 3 (2006), 311–322.

Ben Shneiderman. 1994. Dynamic Queries for Visual Information Seeking. *IEEE Software* 11, 6 (Nov. 1994), 70–77.

Ben Shneiderman. 2010. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson Education India.

Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting Meaningfully with Machine Learning Systems: Three Experiments. *International Journal of Human-Computer Studies* 67, 8 (2009), 639–662. https://doi.org/10.1016/j.ijhcs.2009.03.004

Justin Talbot, Bongshin Lee, Ashish Kapoor, and Desney S. Tan. 2009. EnsembleMatrix: Interactive Visualization to Support Machine Learning with Multiple Classifiers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 1283–1292. https://doi.org/10.1145/1518701.1518895

Joshua B Tenenbaum, Vin De Silva, and John C Langford. 2000. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *science* 290, 5500 (2000), 2319–2323.

Warren S Torgerson. 1958. Theory and Methods of Scaling. (1958).

Cagatay Turkay, Arvid Lundervold, Astri Johansen Lundervold, and Helwig Hauser. 2012. Representative Factor Generation for the Interactive Visual Analysis of High-Dimensional Data. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2621–2630.

Iris Vessey. 1991. Cognitive Fit: A Theory-Based Analysis of the Graphs Versus Tables Literature*. *Decision Sciences* 22, 2 (1991), 219–240. https://doi.org/10.1111/j.1540-5915.1991.tb00344.x

Michael J Way, Jeffrey D Scargle, Kamal M Ali, and Ashok N Srivastava. 2012. *Advances in Machine Learning and Data Mining for Astronomy*. CRC Press.

J. Wenskovitch, I. Crandell, N. Ramakrishnan, L. House, S. Leman, and C. North. 2018. Towards a Systematic Combination of Dimension Reduction and Clustering in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (Jan 2018), 131–141. https://doi.org/10.1109/TVCG.2017.2745258

John Wenskovitch and Chris North. 2017. Observation-Level Interaction with Clustering and Dimension Reduction Algorithms. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics (HILDA'17)*. ACM, New York, NY, USA, Article 14, 6 pages. https://doi.org/10.1145/3077257.3077259

Hadley Wickham, Dianne Cook, Heike Hofmann, Andreas Buja, et al. 2011. tourr: An R Package for Exploring Multivariate Data with Projections. *Journal of Statistical Software* 40, 2 (2011), 1–18.

Ji Soo Yi, Youn ah Kang, John Stasko, and Julie Jacko. 2007. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (2007), 1224–1231.

Ji Soo Yi, Rachel Melton, John Stasko, and Julie A Jacko. 2005. Dust & Magnet: Multivariate Information Visualization using a Magnet Metaphor. *Information Visualization* 4, 4 (2005), 239–256.