Be the Data: Embodied Visual Analytics

Xin Chen, Jessica Zeitz Self, Leanna House, John Wenskovitch, Maoyuan Sun, Nathan Wycoff, Jane Robertson Evia, Scotland Leman, and Chris North

Abstract—With the rise of big data, it is becoming increasingly important to educate groups of students at many educational levels about data analytics. In particular, students without a strong mathematical background may have an unenthusiastic attitude towards high-dimensional data and find it challenging to understand relevant complex analytical methods, such as dimension reduction. In this paper, we present an embodied approach for visual analytics designed to teach students about exploring alternative 2D projections of high-dimensional data points using weighted multidimensional scaling. We propose a novel concept, *Be the Data*, to explore the possibilities of using human's embodied resources to learn from high-dimensional data. In our implemented system, each student embodies a data point, and the position of students in a physical space represents a 2D projection of the high-dimensional data. Students physically move within the room with respect to each other to collaboratively construct alternative projections and receive visual feedback about relevant data dimensions. In this way, students can pose hypotheses about the data to discover the statistical support as well as learn about complex concepts such as high-dimensional distance. We conducted educational workshops with students in various age groups inexperienced in complex data analytical methods. Our findings indicate that *Be the Data* provided the necessary engagement to enable students to quickly learn about high-dimensional data and analysis processes despite their minimal prior knowledge.

Index Terms—Embodied interaction, visual analytics, high-dimensional data, visualization in education.

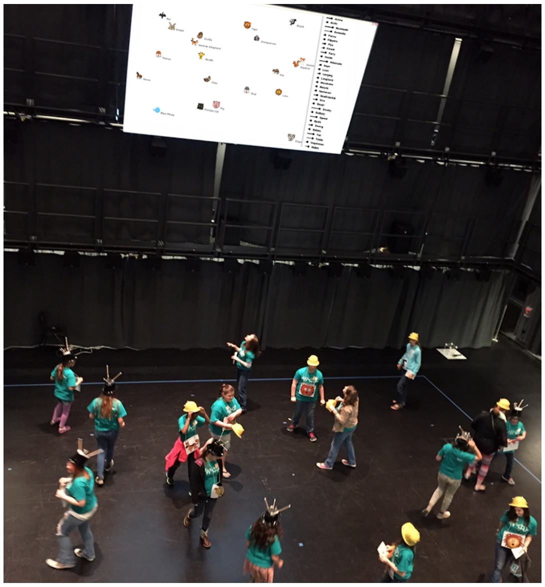


Fig. 1: In *Be the Data*, each students represents an individual data point. A bird-eye view of student locations in the physical space is shown on a large display above them.

1 Introduction

Data is becoming increasingly complex, resulting in a clear need to advance research and education in knowledge discovery from large data. Educators are called upon to teach students data analytical techniques, which rely on complex mathematical models to formalize interpretations from data. For example, analysts use dimension reduction algorithms (e.g., multidimensional scaling and principal components analysis) to project high-dimensional data onto lower (e.g., two or three) dimensions to help them explore and understand the data.

It is difficult for students to explore and understand high-dimensional data, especially for those without data analytics experience. They lack prior mathematical knowledge to understand the complexity of reducing many dimensions to few dimensions [1], [2]. This lack of knowledge may further prevent students from learning about these methods due to the cost of significant cognitive effort. Data analytics coursework is typically not approachable by students until they have mastered necessary quantitative theory and methods. In addition, students who find statistical techniques difficult may confront non-cognitive factors (e.g., unenthusiastic attitude) and try to avoid such subjects [3], [4]. Moreover, learning from high dimensional data requires comprehensive critical thinking skills that extend beyond the application of mathematical methods, such as formalizing alternative hypotheses, communicating personal judgment, exploring multiple solutions, and assessing implications of discoveries. Educators are called to create innovative and effective instructional methods to generate interest in data analytics, to make learning high-dimensional data analysis more approachable to the untrained population (e.g., students with low proficiency in mathematics or statistics), and to provide hands-on experience for students to engage in the analysis process [2], [3].

Specifically, we seek to address the problem of helping novice students to learn about high-dimensional data projections using Weighted Multi-dimensional Scaling (WMDS) [5]. Methods like WMDS provide a mechanism to convert a large (and daunting for novice students) multi-dimensional table of data into a more intuitive scatterplot, in which the spatial distance between points

- *Xin Chen is with Bloomberg. Email: chenxin@vt.edu*
- *Jessica Zeitz Self is with the University of Mary Washington Department of Cmoputer Science. Email: jzeitz@umw.edu*
- *Leanna House, Nathan Wycoff, and Scotland Leman are with the Virginia Tech Department of Statistics. Emails: {lhouse | nathw95 | leman}@vt.edu*
- *John Wenskovitch and Chris North are with the Virginia Tech Department of Computer Science. Emails: {jw87 | north}@cs.vt.edu*
- *Maoyuan Sun is with the University of Massachusetts Dartmouth Department of Computer and Information Science. Email: smaoyuan@umassd.edu*
- *Jane Robertson Evia is with the Virginia Tech Department of Practice. Email: robj@vt.edu*

Manuscript received ...

approximately represents their high-dimensional dissimilarity. Alternative projections can be produced by adjusting weight parameters on the data dimensions. But what do these plots mean and how are they produced from the data? Simply showing students the mathematical definition of WMDS is not likely to help (Fig. 2).

Students who have not previously studied high-dimensional data need to grasp several key concepts, including the notions of data dimensions, dimension weights, weighted high-dimensional distance, and 2D distance. Creative new pedagogical approaches that engage students with these concepts are needed to broaden the reach and appeal of educational outreach for big data analytics.

To address the above needs, we developed a novel approach, *Be the Data*, which applies embodied interaction [6] to visual analytics for education and outreach [7]. Specifically, we employed an interactive system (shown in Fig. 1) to teach students analytical concepts for understanding high-dimensional data projection. *Be the Data* involves five core principles:

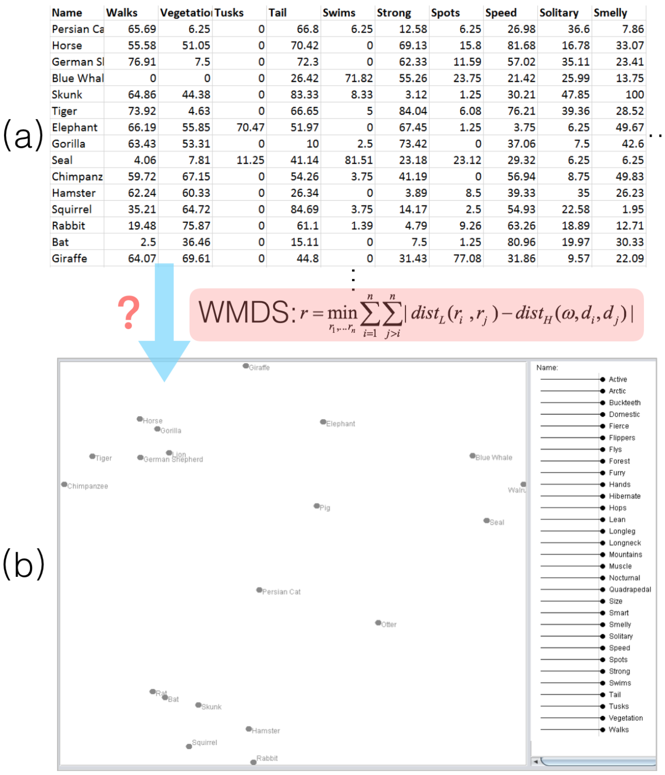
1) Each student **embodies a unique data point** in a high-dimensional dataset.
2) Students are **immersed in a 2D projection** represented by their physical coordinates in a shared physical space.
3) Students construct projections by **physically moving** themselves with respect to each other within the space.
4) Students receive real-time **visual feedback about projection parameters**, in the form of dimension weights that explain their current projection.
5) Students work together to **collaboratively explore alternative projections**, pose hypotheses based on their data domain knowledge, and learn from the data.

In this paper, we aim to explore how students exploit this novel embodied approach to understand high-dimensional data and to learn data analytics processes. The key contributions of this paper are as follows:

- We present the system, *Be the Data*, which exploits embodiment in visual analytics to invoke embodied learning.
- We describe multiple *Be the Data* workshops that we conducted with students at several age groups to assess its value.
- We identify students' data analytical strategies that employ this form of embodiment.
- We find both qualitative and quantitative evidence of student improvement in understanding high-dimensional data.

In previous work, we presented an overview of the system description and a case study [7], initial results of a workshop with one age group focusing on high-level learning strategies [8], and an alternative application of the concept to social settings [9]. Here, we expand on the concept, system, and workshop descriptions, and present detailed results of five sets of workshops covering several different age groups, including concepts learned, collaborative strategies employed, analytical data structures explored, and engagement.

We structure the remainder of this paper as follows: In Sect. 2, we provide an overview of related work to place the contributions of *Be the Data* in context. In Sect. 3, we discuss the conceptual overview and instantiation of the *Be the Data* system, including the visualization, tracking system, and dynamic clustering components. Sect. 4 describes the workshops that used *Be the Data* as well as our data collection and analysis methodology. This leads into the results of our evaluation, presented in Sect. 5. We discuss how well our studies and analysis met our high-level goals for *Be the Data* in Sect. 6, addressing the limitations of our studies as well as future research directions, and finally conclude in Sect. 7.



(a)

Name	Walks	Vegetatio	Tusks	Tail	Swims	Strong	Spots	Speed	Solitary	Smelly
Persian Ca	65.69	6.25	0	66.8	6.25	12.58	6.25	26.98	36.6	7.86
Horse	55.58	51.05	0	70.42	0	69.13	15.8	81.68	16.78	33.07
German Sl	76.91	7.5	0	72.3	0	62.33	11.59	57.02	35.11	23.41
Blue Whal	0	0	0	26.42	71.82	55.26	23.75	21.42	25.99	13.75
Skunk	64.86	44.38	0	83.33	8.33	3.12	1.25	30.21	47.85	100
Tiger	73.92	4.63	0	66.65	5	84.04	6.08	76.21	39.36	28.52
Elephant	66.19	55.85	70.47	51.97	0	67.45	1.25	3.75	6.25	49.67
Gorilla	63.43	53.31	0	10	2.5	73.42	0	37.06	7.5	42.6
Seal	4.06	7.81	11.25	41.14	81.51	23.18	23.12	29.32	6.25	6.25
Chimpanz	59.72	67.15	0	54.26	3.75	41.19	0	56.94	8.75	49.83
Hamster	62.24	60.33	0	26.34	0	3.89	8.5	39.33	35	26.23
Squirrel	35.21	64.72	0	84.69	3.75	14.17	2.5	54.93	22.58	1.95
Rabbit	19.48	75.87	0	61.1	1.39	4.79	9.26	63.26	18.89	12.71
Bat	2.5	36.46	0	15.11	0	7.5	1.25	80.96	19.97	30.33
Giraffe	64.07	69.61	0	44.8	0	31.43	77.08	31.86	9.57	22.09

. . .

?

WMDS: $r = \min_{r_1,\ldots r_n} \sum_{i=1}^{n} \sum_{j>i}^{n} | dist_L(r_i , r_j) - dist_H(\omega, d_i, d_j) |$

Fig. 2: Students find it difficult to understand high-dimensional data projection. (a) A portion of a high-dimensional dataset about animals. This dataset, in total, includes 30 animals with 30 variables. The dataset is obtained from [10]. (b) A WMDS plot of the same dataset. Without a strong understanding of the WMDS algorithm, it is difficult to understand the spatial mapping from data table to data projection.

2 Related Work

This section discusses the novelty of *Be the Data* in the context of existing work. We begin with a discussion of physical interaction with data in immersive and high-resolution environments, noting that these systems allow analysts to closely interact and even be surrounded by data, but does not allow an analyst to *become* a data point. We follow with a discussion of co-located collaborative visual analytics, highlighting the benefits of collaborative exploration of data that were found with other research projects. Next, we discuss the benefits of interactive visualization in the context of teaching the analysis and exploration of high-dimensional data. Finally, we summarize the concept of Observation-Level Interaction and inverse WMDS, which drives the weight updates through user interaction with the projection.

2.1 Physically-Embodied Interaction with Data

With technological breakthroughs beyond traditional desktop settings, we have witnessed a growing number of visualization applications that extend interaction into the physical world for more embodied data exploration and collaboration [11], [12]. Similarly, the Immersive Analytics agenda [13] seeks to more deeply embed users in their data through advanced display and interaction techniques. For example, interactive surfaces enable users to directly interact with more degrees of freedom when manipulating data [14]. High-resolution or stereoscopic 3D display technologies allow users to physically navigate data [15], [16]. Attempts have been made to physicalize virtual data into manipulative artifacts, with physical attributes (e.g., size, shape, materials, etc.) encoding data [17].

Although these contributions demonstrate different physically-embodied ways to interact with data, they are similar in the way that they place users in, or at least closer to, their data. We call this perspective "*Be In the Data*". Our work differs from such approaches by taking the embodiment to the extreme. Instead of placing users in the data, users actually become data points. Therefore, we call this new perspective of visual analytics "*Be the Data*", differing from "*Be In the Data*" in the perspective that the user takes during analysis. Rather than looking into the data points as an observer, users themselves take on the perspective of a data point. This may give users a more egocentric perception to conjecture various relationships with other data points. It has the potential to render an engaging experience, enabling students to personally identify with the data and how analytical methods affect the data, which could further lead to deep insights.

2.2 Co-located Collaborative Visual Analytics

Co-located collaboration can benefit visual analytics tasks [18], [19], [20], [21], [22]. Researchers found that closely-collaborating teams shared their findings more frequently, reported more correct facts, required fewer hints, and gained a higher task score than loosely coupled teams. Ownership and awareness of collaborator actions are two important factors to coordinate group efforts [23]. Leveraging spatial affordances of the co-located space, our work enables both ownership and awareness in collaborative visual analytics. Being a data point, each individual is responsible for and fully controls the position of their own data point. Meanwhile, everyone is aware of others' positions since they are located in the same room and also mirrored on a shared display. Students can determine their own positions and negotiate with each other in a coordinated manner, requiring them to learn about others' data and discuss relationships between their data points. Data exploration naturally evolves through social interactions to construct alternative projections of the data. No one is left out.

2.3 Teaching High Dimensional Data with Visualization

Teaching novice students about high-dimensional data analytics is challenging. A common problem is that students lack interest and mathematical background [3], [4]. Interactive visualizations have been employed because they enable students to learn and apply data analytical techniques in a visually interactive manner [24]. However, it may be challenging for learners to conceptualize complex analytical operations on abstract data via simplistic interaction mechanics suggested by a mouse and a keyboard [25]. Some have argued that learning new mathematical concepts is metaphorically structured with physical interaction [26] and fully embedded in body actions [27]. For example, Howison et al. [28] provides empirical evidence that body movement is able to evoke basic arithmetic operations to understand proportional equivalence (e.g., $\frac{2}{3} = \frac{4}{6}$). However, few educational applications for data analytics use technology other than standard desktop or laptop computers. This indicates a need for novel approaches to teach high-dimensional data analytics. In this work, we take a proactive approach to employ human bodies as movable data points. With this, we can potentially map data analysis and exploration to integrated sensory activities rather than just visual perceptions.

2.4 Observation-Level Interaction

Dimension reduction techniques like WMDS are often used to encode data similarity as spatial proximity. The "*proximity equals similarity*" visual metaphor is cognitively effective because people naturally group similar information together and move different information apart. This phenomenon has been found in both physical settings (e.g., organizing notes on a white board) [29] and virtual environments (e.g., organizing digital documents on large displays) [30]. This metaphor shields analysts from the complexity of the high-dimensional data by providing a simple 2D (or 3D) plot. However, since loss occurs in the projection, analysts can interact with parameters of the algorithm to explore many possible projections, thus looking at the data from different perspectives. For example, WMDS applies weight parameters to each data dimension, enabling the analyst to put more or less emphasis on each dimension in the computation of the high-dimensional similarities (or distances) between data points.

In contrast, dimension reduction tools like Andromeda [31] also implement *Observation-Level Interaction* (OLI) [32], in which analysts can directly manipulate the data points (”observations” in statistics terminology) in the 2D projection. The inverse WMDS algorithm then computationally back-solves the WMDS algorithm to identify the optimal data dimension weight parameters that would most closely produce the desired projection. OLI is especially useful when analysts have domain knowledge about the data points. They can organize the data points accordingly to learn how the data dimensions relate to their domain knowledge. Thus, users are able to explore various projections of high-dimensional data based on their conjectures of relationships among the data. Learners communicate their judgment that data points are similar by pulling them closer and data points are different by pushing them apart. In turn, Andromeda identifies data dimensions that potentially explain the user's judgments and provides this as visual feedback. We exploit OLI in *Be the Data* to enable young students to interact with data points with which they have significant domain knowledge, a dataset about animals. OLI can encourage students to think specifically about the distances between points, a key concept in dimension reduction.

In previous work [7], [8], we briefly described the system implementation and initial results. Here we provide a complete and detailed presentation of the concept, system, workshops, and evaluation results. We have also applied this system in a different setting, a social meeting scenario in which users embody data about themselves [9].

3 Be the Data

3.1 Conceptual Overview

Be the Data exploits a large interactive room called the *Cube*, and includes a large overhead display, a vision-based motion tracking system, and a software system for direct manipulation of high-dimensional data. Figure 1 shows an overview of the system.

To use the system, a group of students enter the Cube and embody virtual data points by wearing trackable hats. With the hats, their positions within the Cube are detected in real-time. Students manipulate the layout of the data points by walking around in the Cube. There is a large display overhead, where the students' current positions and the corresponding impact on data dimension weights are displayed as visualizations (Fig. 3). For example, if we consider a high-dimensional dataset about animals, shown in

Fig. 2a, each student represents an animal (data point) and their position in the Cube is visually shown on the display (Fig. 3).

The underlying algorithm of the visualization relies on an inversion of Weighted Multi-Dimensional Scaling (WMDS) [5]. WMDS visually plots data points in a 2D Euclidean space to reveal the relative pair-wise distances of the data points in the high-dimensional space. Conversely, *Be the Data* applies an inverse WMDS machine learning algorithm based on Leman et al. [32] that was introduced in Sect. 2.4 and is described in more detail in Sect. 3.2.1. Using the OLI technique, the system maps the changes in students' data point positions to adjustments on the dimension weight parameters. Students change their coordinates in the 2D space by walking in the Cube. In turn, they are provided with real-time feedback (i.e., a new set of dimension weights) that best explains their current layout. When students move several times to adjust the projection, they are effectively exploring the same dataset from multiple perspectives.

The goal of this approach is to help students learn to think about data in a high-dimensional manner, getting them beyond thinking about just one dimension at a time. By embodying an animal data point, they become intimately familiar with all the multi-dimensional data values associated with that animal. By comparing with other students, they understand the multi-dimensional differences between their animals. Positioning themselves in the room requires that they think about the concept of distance and how that relates to the multi-dimensional differences between animals. The visual feedback helps them recognize how the distances they chose emphasized some dimensions over others, perhaps as a result of their own external knowledge (dimensions not contained in the dataset) that they applied. Their acting out the projection helps them understand the behavior of the WMDS analytical process. The changes over time as they move help them to realize that there are many possible interpretations of high-dimensional data, and that they can explore many different kinds of questions about the data.

3.2 System Description

Be the Data blends both physical and virtual user interfaces through the concept of the data projection. The physical interface includes three parts: the physical space of the Cube, the tracking system, and the large display. The virtual interface includes interactive visualizations. Different from traditional input devices (e.g., mouse and keyboard), students interact with the system by walking around in the Cube. Such physical movement in the Cube is captured in real time by tracking the location of the students' hats with the vision-based tracking system [7]. Each hat is specifically designed to contain a unique visual identifier that is automatically mapped to one of the animal data points. The hat movement is then transformed into 2D coordinates according to their x-y position in the room, which are then used as input for the inverse WMDS and dynamic clustering algorithms. The layout and the calculated weights are presented as interactive visualizations on the large display. The students are also each given a lanyard that contains the name and picture of their assigned datapoint animal so that other students can see what they are, and a card containing a print out of their data row of the table so they know their own data values.

3.2.1 Interactive WMDS Visualization

The interactive visualization includes two essential views: a WMDS plot and a dimension chart, organized left and right respectively on the large display (Fig. 3).The WMDS plot reflects the current physical layout of the students in the Cube (from a bird's eye view). With the animal dataset, the data points are visualized as animal icons and associated animal name labels. The dimension chart lists the data dimensions in alphabetical order and reveals the current values of their weight parameters as a bar chart.

Conceptually, the weight parameters in the dimension chart, when input into the WMDS algorithm, produce the projection of the data as shown in the WMDS plot. This can be interpreted such that data points close to each other in the plot are relatively similar while those far from each other are relatively different in the data dimensions that are emphasized (i.e., dimensions that are weighted more).

However, in *Be the Data*, the process is actually reversed. The students control the projection, which is input into an inverse WMDS algorithm to compute and display the weight parameters in the dimension chart. The parameters are computed such that if they were then input into WMDS, it would produce a projection plot as close as possible to the students' current layout. In WMDS plots, the precise location of the observations are unimportant. Rather, the pairwise distances between the observations define the projection. Note that it might be impossible to find parameters that precisely reproduce the students' projection, if there is inadequate statistical support in the data for that exact projection. Note also that, since the inverse WMDS algorithm is a complex optimization problem, it necessarily computes an approximate solution.

Initially, all weights are equal (Fig. 3a), and students must start in the actual corresponding WMDS data point positions (marked on the floor). As students change the layout by rearranging themselves in the room, the weights get updated to explain the students' choice of positions (Fig. 3b). The length of the dimension bar reflects its relative weight as compared to other bars: longer means a higher weight. For example, as demonstrated in Fig. 3a and Fig. 3b, the Tiger moves closer to the Pig, thus the Tiger is more similar to the Pig than the remaining animals in the dimensions with higher weights, such as Flippers, Hibernate, and Size.

The underlying algorithm of *Be the Data* is based on WMDS, which maps three or more dimensions to two dimensions. WMDS visually plots the data in 2D Euclidean space to represent the pair-wise data point distances in the high-dimensional space. The 2D layout is determined by weight parameters $\omega = [\omega_1, \omega_2, ..., \omega_p]$, which reflects the relative importance of each dimension, where p is the number of dimensions. The 2D coordinates r of the initial plot of high-dimensional data d is determined by minimizing a stress function that computes the total error between the high-dimensional and low-dimensional distances between points:

$$r = \underset{r_1,\dots r_n}{\arg\min} \sum_{i=1}^{n} \sum_{j>i}^{n} |dist_L(r_i, r_j) - dist_H(\omega, d_i, d_j)| \qquad (1)$$

where n is the number of data points, $dist_L(r_i, r_j)$ is a L2 Euclidean distance function between 2D points r_i and r_j, and $dist_H(\omega, d_i, d_j)$ is a dimensionally-weighted L1 (Manhattan) distance function between high-dimensional points d_i and d_j.

Be the Data takes advantage of the inverse WMDS algorithm as described in [32] so that we can map the layout changes to the weight adjustments. The inverse algorithm solves for the weights ω given updated low-dimensional coordinates $r*$ determined by the students' positions in the room:

$$\omega = \underset{\omega_1,\dots \omega_p}{\arg\min} \sum_{i=1}^{n} \sum_{j>i}^{n} |dist_L(r_i^*, r_j^*) - dist_H(\omega, d_i, d_j)| \qquad (2)$$

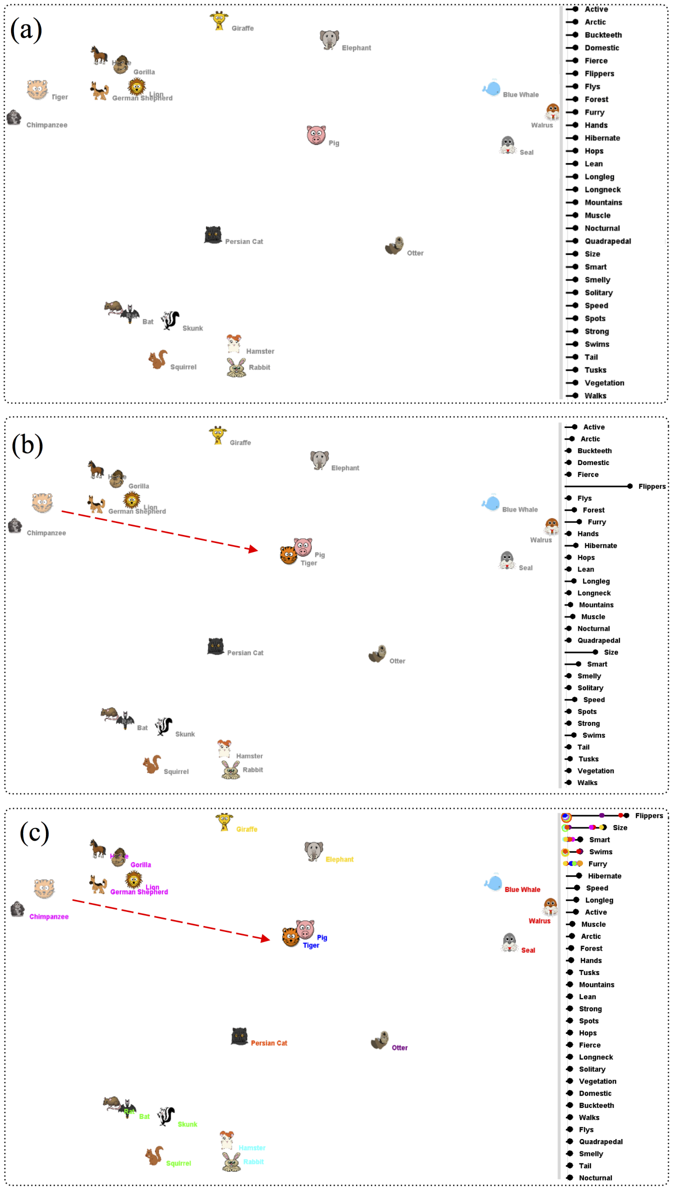


Fig. 3: The plot on the overhead large display to visualize students' locations in the room. (a) The students' initial locations. (b) When students move in the room, they change the two-dimensional coordinates in the WMDS plot, which changes the relative weights of dimensions. (c) Data points are grouped into colored clusters. Dimension chart is ordered by weights.

3.2.2 Tracking System

To track the locations of participants during each workshop, *Be the Data* uses an OptiTrack motion tracking system to locate the spatial position of unique hats. The OptiTrack system consists of 24 Oqus cameras positioned around the system, and uses Qualisys Track Manager (QTM) software to collect and process data from the cameras in real-time. Each of the trackable hats is a rigid body with 4-6 uniquely-positioned reflective markers placed upon it. The QTM software identifies the (x,z) position of each hat, and streams those positions to *Be the Data*. Our implementation allows for tracking and differentiation of more than 50 objects. We estimate that 24 cameras provides centimeter accuracy in spatial precision with 4 millisecond latency [33]. Further details are provided in Chen et al [7].

3.2.3 Dynamic Clustering

Although the WMDS plot and dimension chart visualize the dimension weight changes that explains the physical layout, they cannot explicitly reveal all of the data values of all the data points. This is problematic, because while the weights explain the distances, they do not adequately explain the actual values to understand the semantics of the layout. For example, in Fig. 3b, the pig and tiger are similar in Flippers, but is that because they both have high or low values for Flippers? We know that animals with similar Flippers are grouped, but where are the animals with high Flippers?

To remedy this, we visualize aggregations of only the highly-weighted dimensions according to clusters in the plot. This adequately reduces the number of values visualized to an easily perceptible amount. To do this, we incorporate a dynamic clustering algorithm to visualize 2D spatial clusters in the WMDS plot (Fig. 3c). The system automatically and dynamically identifies low-dimensional clusters of data points as students move around the Cube.

Clusters are determined in real-time by an optimized method of *k-means*, described in detail in Chen et al [7]. Briefly, the number of clusters (k) is determined using the heuristic elbow method to identify an optimal number of clusters for the current projection [34]. This method computes an error measure for increasing values of k until the improvement in the error falls below a specified threshold. The value of k at this "elbow" is then used for classification. Each cluster is assigned a unique color and used for the data point labels in the WMDS plot. As the students move around and the clusters update, the algorithm attempts to preserve the cluster coloring by tracking cluster centroids over time. The data values of the cluster centroids (i.e., the average values of all data points in that cluster) are visualized in the dimension chart for the highly-weighted dimensions as colored points on the dimension bars, with color corresponding to cluster color and position on the bars corresponding to those aggregate values. The data values are scaled according to the dimension bar lengths (weights). This scaling helps students understand the effect of the weight on determining distance between points on the bars.

For example, in Fig. 3c, the red cluster ranks highest on the Flippers dimension, which indicates that the red cluster had the highest average value in the Flippers dimension. This makes sense, since the red cluster contains the Blue Whale, Walrus, and Seal. It also explains why this cluster is so far away from many other clusters which have 0 for Flippers. When the dynamic clustering feature is enabled, the dimension chart is sorted based on their weights. This potentially helps students to quickly identify features that differentiate the clusters.

We include the WMDS plot in the visualization for three purposes. First, it enables the students to see all data points simultaneously. Awareness of each other's locations and movements is important for collaborators to coordinate their activities [20]. We make each individual's input salient so that the consequences of everyone's decision is visible and counts for the group result. Meanwhile, it raises individual responsibility for decisions about the data being embodied. Such ownership is helpful to decrease student inclinations to simply follow others. Second, the graphic representation combined with physical layouts are coherent with students' mental model of the *proximity* $\approx$ *similarity* metaphor. Imagery activity bridges the abstract with the concrete [35], combining physical and virtual. Third, the interactive visualization explicitly documents students' thinking process, which can be replayed for instructional and research purposes.

Workshop	Grade Level	Number of Participants
ICAT	3rd grade	80
AWC	6th and 7th grade	62
CEED	10th or 11th grade	50
STEP	Pre-college	33
DA	Undergraduate	49

TABLE 1: Participant educational level.

4 Evaluation

We conducted educational and outreach workshops to explore how novice students (i.e., students without mathematical background in high-dimensional data) employ embodied interaction with *Be the Data* to learn and analyze high-dimensional data. Specifically, we sought to answer the following two questions:

1) Did students learn key concepts about high-dimensional data, including weight variables, relative distance, dimension reduction, and data exploration? Did this increase their confidence?
2) How did students exploit *Be the Data* to learn about data and data analytics processes?

4.1 Participants

Over the course of a year, we conducted five different workshops at STEM outreach activities at our institution, involving students from a range of age groups. The workshops were designed as educational activities embedded within broader outreach events. Thus, the workshops were not designed as controlled studies, but more like ecologically valid field studies. We recruited 62 participants in 6th and 7th grade at an Association for Women in Computing (AWC) workshop, 50 participants in 10th and 11th grade at Center for the Enhancement of Engineering Diversity (CEED) workshops, 33 recent high-school graduates entering college in the Student Transition Engineering Program (STEP) workshop, 49 undergraduate participants in a data analytics (DA) introduction workshop (Table 1), and 80 3rd graders at an Institute for Creative Arts and Technology (ICAT) outreach event.

Participants were new to high-dimensional data analytics and the WMDS method. For ICAT and AWC participants, none of the school curriculum had yet covered WMDS related concepts according to the students' teachers. CEED, STEP and DA participants were asked on a survey about their familiarity with WMDS. Of 123 returned responses, 75 students checked "never heard of it," 45 students checked "heard of it but never used it," 3 student checked "learned about it," and 0 students checked "expert on it."

4.2 Workshop and Procedure

In groups of 20-30, students were asked to analyze a high-dimensional dataset of 20-30 animals with 30 dimensions (Figure 2a) using *Be the Data*. Each dimension reflects the degree (on a scale of 0–100) to which animals could be described by that characteristic (e.g, *skunks* were rated 100 in the dimension *smelly*, whereas *horses* were rated 33 in the same dimension).

Before the workshop started, students completed a pre-survey. After the workshop, they completed a post-survey. Due to event scheduling constraints, ICAT participants did not take the pre-survey or post-survey. The dynamic clustering feature was implemented later and was only available for the DA workshop.

The workshop lasted approximately 30 minutes. It started with a short introduction on high-dimensional data. The instructor explained the concept of high-dimensional data using the animal data shown in a table (Fig. 2a), and identified dimensions as columns in the table. Then, the instructor explained how to use the system and gave examples. Specifically, she explained that the visualization on the screen is a 2D projection of the given high-dimensional data, and that positions of animal icons on the visualization represented students' coordinates in the room. The instructor asked students to move randomly in the cube and let them look at the visualization and weight changes. The instructor explained the *proximity* ≈ *similarity* metaphor to interpret the WMDS visualization with highly-weighted dimensions. The instructor picked a few of the students, two near and two far, to share their data and explain the reason for their distances from each other. The instructor completed the explanation by suggesting that the system could be used to explore data randomly or intentionally when addressing research questions. For example, the instructor asked, "What makes some animals good or bad pets?" With guided discussion, the students would group and/or align themselves according to animals they liked to disliked as pets.

After the introduction, students as a group were asked by the instructor to suggest questions that they would like to investigate collaboratively with the tool. Some students raised their hands and proposed questions. If necessary, the instructor selected and refined a proposed question to ensure that it allowed for the exploration of the entire dataset as a group, rather than drawing a conclusion about a single animal or subset of animals. Although the students did not suggest questions that could not be answered easily with *Be the Data*, the instructor was ready to nudge students away from questions that were not about the observations. That is, questions referring to, say, correlation between variables, are hard to answer with *Be the Data*, but questions for which the observations (i.e., animals) are the subjects of interest comply with the nature of OLI [36]. If the group initially had difficulty thinking of questions, the instructor gave them more sample questions to solve, such as, "What may describe herbivores, omnivores, and carnivores?" One creative question asked by the students was, "What makes animals good to eat?"

After selecting a question, the students as a group then collaboratively used the system to answer the question as they collectively saw fit. Questions required whole-group collaboration in that each student needed to move according to their own intuition but also in concert with the other students. Further collaborations resulted when some students did not know where to move or disagreed on where to move (refer to Table 3). Once the group decided that they had reached their desired positions and were done moving, the system updated the visualization to display the new dimension weights. The students then reflected on their findings by discussing the results as a group. In some cases, the students were given the opportunity to further adjust their positions and the instructor provoked some additional discussion about what was learned about the animal data.

4.3 Data Collection and Analysis

To answer the research questions listed at the beginning of this section, we collected qualitative and quantitative data from recorded video, pre-surveys, and post-surveys. The overhead video recordings preserved anonymity of participants while still allowing researchers to investigate the workshop execution. The surveys included multiple choice and open-ended questions that reflected students' understanding of technical concepts, as well as attitudes towards the workshop and data analytics. Differences in pre- and post-survey answers were used to measure potential gains from the workshop. Unfortunately, due to errors on the pre-survey of the AWC event, we do not have results for AWC's pre-survey. We learned from the AWC workshop that changes in the survey questions and data management were needed to warrant analytical comparisons between the pre- and post-surveys. We further improved survey questions in the DA and CEED workshops to measure students' learning more thoroughly.

To analyze the quantitative data from the surveys, we include a Bayesian approach. There has been some debate over the use of classical methods which rely on p-values to make inference [37]. The p-values can easily be miscalculated and/or misunderstood when comparing them to a type I error (i.e., a significance level α). For this paper, we employ Bayesian models to analyze the quantitative data. In many (if not all) cases, a classical results in the same, final inferences as those that we report in Sect. 5. To support readers unfamiliar with Bayes, we report p-values when applicable.

The quantitative data from the surveys were primarily from two types of questions: (a) questions with right or wrong answers, and (b) questions that request students to rate their attitude. We analyze right or wrong answers per question and survey with Beta-Binomial Bayesian models [38]. From these models, we learn the posterior distribution for the probability of a correct answer, and then we estimate the distribution for the difference in probabilities between the pre- and post-surveys. We use the distribution for the difference to assess and infer changes in the student responses before and after *Be the Data*. For example, consider the question "identify dimensions on DR plot" (for the STEP workshop). The *maximum a posteriori* (MAP) difference is 0.48 with a credible interval $(0.22, 0.67)$ (in Table 2). This means that the most probable difference between the pre- and post-probabilities of being correct is 0.48. Because the credible interval does not overlap 0, we can infer that students' understanding did change during the workshop. When we report p-values for these questions, small p-values may reject the null hypothesis that the pre- and post-probability of being correct are equal, as it is in this example (0.0008).

We analyze questions about attitude using a very similar approach. Rather than learning the posterior distribution for the probability of being correct, we learn the posterior distribution for the mean attitude response. To do so, we use a Gaussian model with reference priors [38] to analyze responses per question and per survey. In turn, we estimate the distribution for the difference in means. For example, the MAP estimate for the difference in mean attitude toward the statement "Analyzing data is boring" in CEED is -2.09 with a credible interval $(-2.96, -1.21)$. This means that most probable difference in mean is -2.09. This difference is notable because 0 is not included in the credible interval. When we report p-values for these questions, small p-values may reject the null hypothesis that the pre- and post-attitude means are equal.

To analyze qualitative data from the surveys' open-ended questions, we had two researchers grade participant responses independently and compare their grades to assure accuracy. That is, the open ended questions on the survey have 1) several right and wrong answers and 2) countless ways to write such answers. The two graders determine which answers are right or wrong, and do not offer partial credit. For example, if asked to explain why there are discrepancies in two WMDS plots of the same data but with different dimension weight specifications, there are countless answers. Two correct answers include "The weights changed" and "The smelly and walks variables played a greater part." Whereas, two wrong answers include, "The points changed locations" and "I don't know." These correctness scores were then evaluated quantitatively as above.

The last form of data that we analyzed were collected via videos. From the videos, we observed when and how students collaborate to solve problems. For example, to evaluate Research Question 2, we observe how or when students collectively communicate, strategize, and respond to updated layouts. From these data, we do not analyze the accuracy of students' insights about the data, but rather the student actions when using *Be the Data*. In Section 5, we summarize findings we make from the observations.

5 RESULTS

In this section, we discuss outcomes and lessons learned after conducting the five workshops. We begin with a detailed discussion of some of the key concepts learned from the workshops, discussing the participants' improvement in interpreting high-dimensional data projections and how they are generated. These results are discussed using results from pre- and post-surveys taken by participants. We follow this discussion with examples of students learning about a dataset using *Be the Data*, as well as providing a taxonomy of layouts generated by the workshop participants and their relationship to the way the participants interpreted the dataset. Finally, we discuss user engagement with *Be the Data* and comment on potential improvement for future iterations of the system.

5.1 Key Concepts Learned

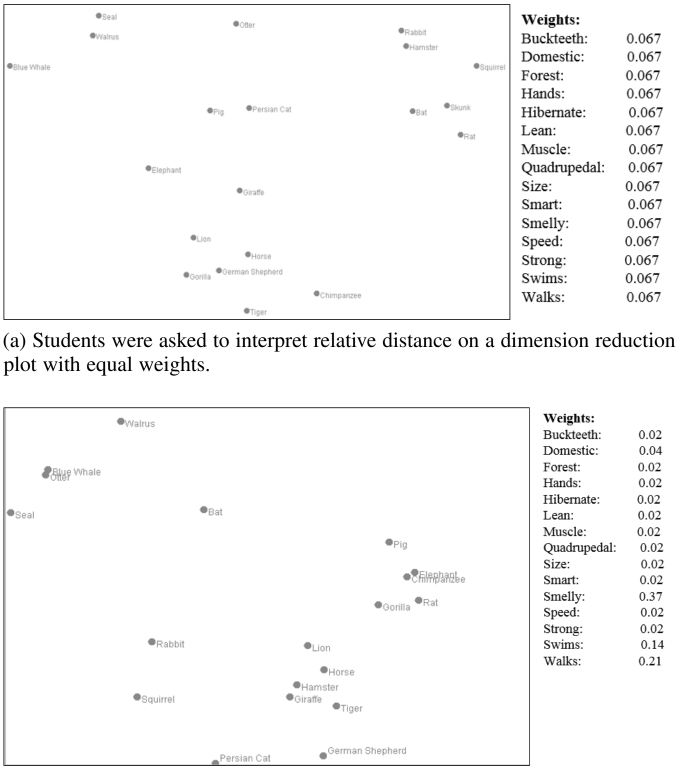
After the study, 60 out of 62 students returned post-surveys in AWC, 47 out of 50 students returned both pre- and post-surveys in CEED, 28 out of 33 students returned both pre- and post-surveys in STEP, 49 out of 49 students returned pre-surveys and 48 returned post-surveys in DA. The results of correctness proportion in the pre- and post-surveys, expected differences, credible intervals, and p-values are shown in Table 2. It suggests that students gained knowledge about the following concepts: *relative distance*, *dimension reduction*, and *data exploration*. It also indicates that students showed more interest and confidence in learning about high-dimensional data.

Students learned concepts about relative distance in dimension reduction plots. To evaluate their understanding, students were asked to interpret similarities of data points in a dimension reduction visualization. In all the workshops, students were asked questions on a dimension reduction (DR) plot with all weights equal (Fig. 4a). In the AWC workshop, 92% of the students answered correctly afterwards. In the STEP workshop, there was strong evidence of improvement (credible interval $(0.01, 0.35)$). In the CEED and DA workshops, almost all the students (96% for both) answered correctly afterwards. There was no improvement, due to the high rate of correctness in the pre-survey (also 96% for both). The relatively high percentage of correctness in all workshops indicate that the *proximity* $\approx$ *similarity* visual metaphor is an intuitive idea that can be exploited for usability in data analytics.

In both CEED and DA workshops, we asked three additional questions to further study student understanding of relative distance. First, students were asked to interpret similarity of data points on a dimension reduction plot with unequal weights (Fig. 4b). There was no significant improvement due to the high correctness rate (96% in CEED and 94% in DA) in the pre-survey.

Workshop	Question	Pre Correctness	Post Correctness	Expected Difference	Credible Interval	p-value
Key concept: relative distance						
AWC	Relative distance on a DR plot with all weights equal	—	0.92	—	—	—
STEP	Relative distance on a DR plot with all weights equal	0.79	0.96	0.15	$(0.01, 0.35)$*	0.0495*
CEED	Relative distance on a DR plot with all weights equal	0.96	0.96	0	$(-0.09, 0.09)$	0.999
CEED	Relative distance on a DR plot with weights not equal	0.96	0.98	0.02	$(-0.06, 0.11)$	0.557
CEED	Explain changes in two DR plots	0.19	0.60	0.40	$(0.21, 0.57)$*	<0.001*
CEED	Predict the change if you want a data point to be closer to a cluster	0.32	0.85	0.52	$(0.35, 0.68)$*	<0.001*
DA	Relative distance on a DR plot with all weights equal	0.96	0.96	0	$(-0.09, 0.09)$	0.9834
DA	Relative distance on a DR plot with weights not equal	0.94	0.98	0.03	$(-0.05, 0.13)$	0.3204
DA	Explain changes in two DR plots	0.38	0.60	0.21	$(0.03, 0.41)$*	0.0265*
DA	Predict the change if you want a data point to be closer to a cluster	0.46	0.79	0.33	$(0.15, 0.50)$*	0.0005*
Key concept: dimension reduction						
AWC	Identify dimensions on a 2D plot	—	0.57	—	—	—
AWC	Identify dimensions on a 3D plot	—	0.50	—	—	—
AWC	Identify dimensions on DR plot	—	0.78	—	—	—
STEP	Identify dimensions on a 2D plot	0.93	0.93	0.01	$(-0.15, 0.15)$	0.8346
STEP	Identify dimensions on a 3D plot	0.75	0.79	0.04	$(-0.18, 0.25)$	0.8115
STEP	Identify dimensions on DR plot	0.29	0.75	0.48	$(0.22, 0.67)$*	0.0008*
CEED	Identify dimensions on DR plot	0.66	0.94	0.27	$(0.12, 0.42)$*	0.001*
DA	Identify dimensions on DR plot	0.73	0.96	0.22	$(0.09, 0.36)$*	0.0022*
Key concept: data exploration						
STEP	Exploratory nature of data	—	0.86	—	—	—
CEED	Exploratory nature of data	—	0.96	—	—	—
DA	Exploratory nature of data	—	0.90	—	—	—
Interests and Confidence						
		Pre Average	**Post Average**	**Expected Difference**	**Credible Interval**	**p-value**
CEED	Analyzing data is boring	4.63	2.54	−2.09	$(-2.96, -1.21)$*	<0.001*
CEED	The lack of mathematical background prevents me from analyzing high-dimensional data	3.37	1.33	−2.04	$(-2.97, -1.12)$*	<0.001*
CEED	I know what is meant by the term, high-dimensional data	2.61	9.00	6.39	$(5.42, 7.36)$*	<0.001*
DA	Analyzing data is boring	2.63	1.88	−0.75	$(-1.48, -0.02)$*	0.0519
DA	The lack of mathematical background prevents me from analyzing high-dimensional data	3.49	2.24	−1.25	$(-2.23, -0.27)$*	0.0191*
DA	I know what is meant by the term, high-dimensional data	4.97	8.53	3.56	$(2.53, 4.58)$*	<0.0001*

TABLE 2: A quantitative summary of students' learning of key concepts (i.e., relative distance, dimension reduction, data exploration), and interest and confidence towards learning about high-dimensional data before and after the workshop. DR stands for dimension reduction. Columns 3 and 4 are observed proportion of correct answers for the pre- and post-surveys. Columns 5 and 6 show the expected difference and the credible interval for the difference in proportions. Column 7 contains the p-values from a two-tailed two sample t-test. The * in columns 6 and 7 flags questions with a significant difference in pre- and post-surveys.



Weights:

Buckteeth:	0.067
Domestic:	0.067
Forest:	0.067
Hands:	0.067
Hibernate:	0.067
Lean:	0.067
Muscle:	0.067
Quadrupedal:	0.067
Size:	0.067
Smart:	0.067
Smelly:	0.067
Speed:	0.067
Strong:	0.067
Swims:	0.067
Walks:	0.067

(a) Students were asked to interpret relative distance on a dimension reduction plot with equal weights.

Weights:

Buckteeth:	0.02
Domestic:	0.04
Forest:	0.02
Hands:	0.02
Hibernate:	0.02
Lean:	0.02
Muscle:	0.02
Quadrupedal:	0.02
Size:	0.02
Smart:	0.02
Smelly:	0.37
Speed:	0.02
Strong:	0.02
Swims:	0.14
Walks:	0.21

(b) Students were asked to interpret relative distance on a dimension reduction plot with unequal weights.

Fig. 4: Students interpreted relative distance on dimension reduction plots.

Second, they were asked to explain if and why their answer changed between the equal weight plot to the unequal weight plot. With a qualitative analysis of their written responses, we found strong evidence of improvement (credible interval $(0.21, 0.57)$ in CEED, credible interval $(0.03, 0.41)$ in DA). For example, some students answered correctly afterwards as "The weights changed" and "The smelly and walks variables played a greater part."

16 students who answered incorrectly in the pre-survey by simply restating the distance changes at the graphical presentation level (e.g., the points got closer on this plot) changed their answers in the post-survey to consider weighted variables. For example: "Because we are comparing the [points] based on different weighted categories than before. Before, all categories were equal, and now they have weighted values" and "The weight of the variables changed, maybe because of a different research question that was asked."

Third, students were asked which dimension weight(s) needed to change if they want one particular data point (e.g., *seal*) to be closer with a cluster (e.g., *blue whale* and *otter*). Based on their written responses, we found strong evidence of improvement (the credible interval $(0.35, 0.68)$ in CEED, $(0.15, 0.50)$ in DA). Six students who did not answer the question correctly in the pre-survey answered correctly in the post-survey, such as "much higher weight on "swims" and "strength and swims variable are weighted heavier."

Seven students who answered incorrectly in the pre-survey (e.g., "It has to go higher on both the axes") answered correctly in the post-survey (e.g., "Change the weight again and base it according to what lives in water"). Two students who answered the question too generally (e.g., "The weights of the variables") answered it more specifically in the post-survey (e.g., "The weight of some variables such as *swims*").

Students learned to interpret multiple dimensions on dimension reduction plots (henceforth, *DR* plot). To evaluate their knowledge about DR plots compared with 2D and 3D scatterplots, students were asked which dimensions are used to plot data in a given 2D scatterplot, 3D scatterplot and *DR* plot. Students needed to identify the correct two, three, or all of the dimensions, respectively. The workshops did not specifically teach about the 2D and 3D scatterplots; we assume students already understood those concepts and we used them as a baseline for comparison. In the AWC workshop, 57%, 50%, and 78% of the students respectively answered these questions correctly afterwards. Students in the STEP workshop showed a significant improvement in understanding the *DR* plot (credible interval $(0.22, 0.67)$). Before the instruction, the students understood the 2D scatterplot and 3D scatterplot, but not the *DR* plot. After the instruction, their understanding of the *DR* plot approximately reached the same level as those of the other plots. In the DA and CEED workshops, we only asked the question on a *DR* plot. There was strong evidence of improvement (credible interval $(0.12, 0.42)$ in CEED, $(0.09, 0.36)$ in DA).

Students understood the exploratory nature of high-dimensional data. To evaluate their understanding, we asked students, on the STEP and DA post-survey, to explain whether it is possible to create many dimension reduction plots from the same high-dimensional data. This question was not asked in the pre-survey because the question was meaningless before students had the opportunity to explore the data. Based on their written responses, we found that 86%, 96%, and 90% of students in the STEP, CEED, and DA workshops, respectively, understood that many different plots could be created to investigate answers to different questions. For example, some students answered: "Depend[ing] on the research question asked, different variables are taken into consideration more heavily than others" and "There are different interpretations of the same data."

Two students further elaborated that the visualization helped them to see the data in an easily perceptible way. Participants tend to focus on a few variables, although there are often many more in play. The visualization clearly presented these impacting variables to the students.

In the CEED and DA workshops, we studied students' attitudes in learning about high-dimensional data before and after the workshop. Students answered three questions using a 0-10 scale, with 0 = strongly disagree and 10 = strongly agree with the question statement. The results of average scores in the pre- and post-surveys, p-values, and effect size are shown in Table 2. For the statement that "Analyzing data is boring," there is a significant difference in CEED and marginal difference in in DA (DA students' responses indicate some prior interest in data analytics).

Results for the statement that "The lack of mathematical background prevents me from analyzing high-dimensional data" showed strong evidence that students, after attending the workshop, were more confident in analyzing high-dimensional data with relatively weak mathematical background.

Results from the question that "I know what is meant by the term, high-dimensional data" showed strong evidence that students believed that they understood the topic better after attending the workshop. In the follow-up question asking what they had learned, students appreciated the relevance of high-dimensional data analysis in their daily lives:

- "Anybody is capable of analyzing high-dimensional data and applies this more often than one would initially think."
- "The system is a neat way to gather sentiment and it could be used for other projects like political sentiment, which would be cool."
- Some students mentioned other high-dimensional data (e.g., sports, countries) that could be analyzed in this manner.

5.2 Collaborative Learning with *Be the Data*

Students learned about data through embodied interaction. They *progressively formalized data relationships through embodied social interactions*. We demonstrate these behaviors with one group who answered the question, *what make some animals good to eat*. Five key steps of their analysis process are shown in Fig. 5.

In the initial exploratory phase, one student who represented the *skunk* separated herself from others as an obviously inedible animal. This led to the system identifying *smelly* as an important variable (Fig. 5a). The *skunk* is an outlier in this dimension with a value of 100, while other animals range from 1 to 51. The system increased the weight on the dimension *smelly* to reveal the reason for this layout. Then, students discussed with their neighbors about whether their animal was edible, based on their own knowledge about their embodied animals. This indicates that students took the ownership of their data point. Some students gradually took a more dominant role in directing others to move. For example, one student directed the crowd and pointed with her hands (Fig. 5b), *"less edible animals move here, more edible animals move over there..."*. Instead of data dimensions, students focused on their embodied animals, the data points. Soon they formed two groups: *non-edible* animals on the upper right corner and *edible* animals on the lower left corner (Fig. 5c). This led the system to increase the weight on dimensions *buckteeth*, *domestic*, and *hops*.

Then, the student (as noted in Fig. 5d) who embodied the *rat* realized that she belonged to neither group, since she felt that some cultures do eat rats. She moved out of the inedible group and stood directly in between the two groups to indicate partial edibility, and to see what made her unique: *buckteeth*. This prompted other students to reconsider and refine their positions. The *non-edible* group spread themselves out by discussing with their neighbors (Fig. 5e). Additionally, the girl representing the *Persian cat* in the edible group thought that she should not be as edible as the *rabbit* and *pig*. This led the students to construct a layout representing a spectrum of edibility rather than a simple binary edibleness. The system then revealed that for this layout, the dimensions *buckteeth*, *domestic*, *forest*, *smelly*, and *tusks* had their weights significantly increased compared with other dimensions.

In this scenario, students progressively formalized data relationships through social interactions, and the system *expressively transformed their domain knowledge* about animals into changes to the visualized dimension weights. In the analysis process, students focused on their embodied data objects and moved physically to explore data relationships. The visualizations evolved as students continuously interpreted and changed their judgments about the data. The *layout evolution* – from one outlier (Fig. 5a), to a binary classification (Fig. 5c), and finally to a distribution spectrum (Fig. 5e) – indicates that students progressively refined a deeper understanding of the data. Such progressive deep interpretation of the data is also revealed in the increasing number of high-weighted dimensions (comparing the dimensions in the three figures) for explaining the student generated projections, representing their growth in high-dimensional thinking.

Learning with *Be the Data* involves embodiment through a dynamic process of human-computer and human-human interaction (Table 3). This unique physical interactive setting enables various human-computer interactions where human behaviors can directly steer computation (e.g., walking in the room or joining or leaving a group). In turn, computational results, presented as visualizations, contribute to progressive interpretation of the data. Students looked at the visualization to evaluate their positions in some group, gained feedback about dimensions, and then determined the next step (e.g., leaving a group). Simultaneously, the embodied setting of *Be the Data* also helped to promote human-human interactions for data exploration. Students took ownership of their embodied data points,

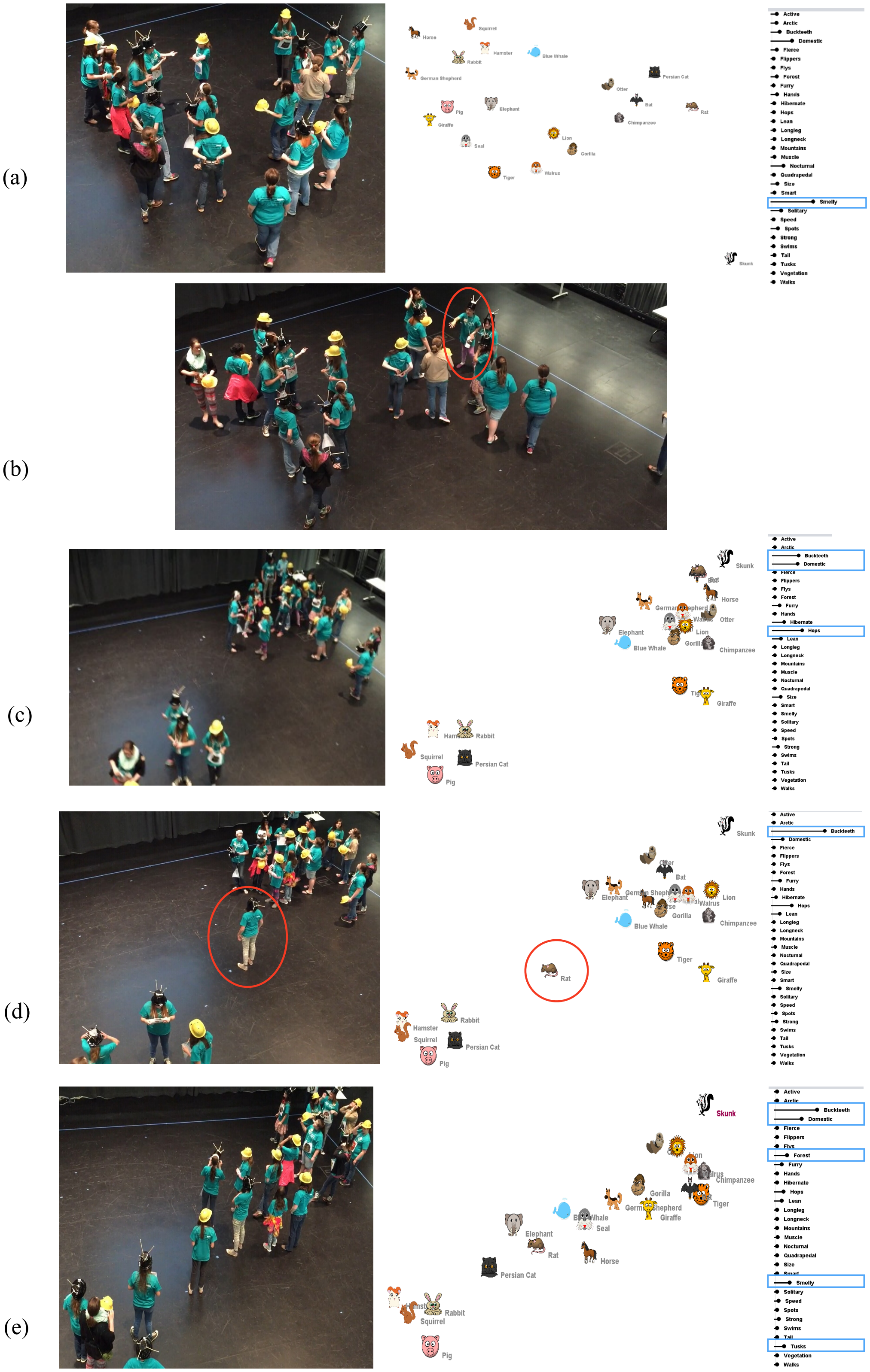
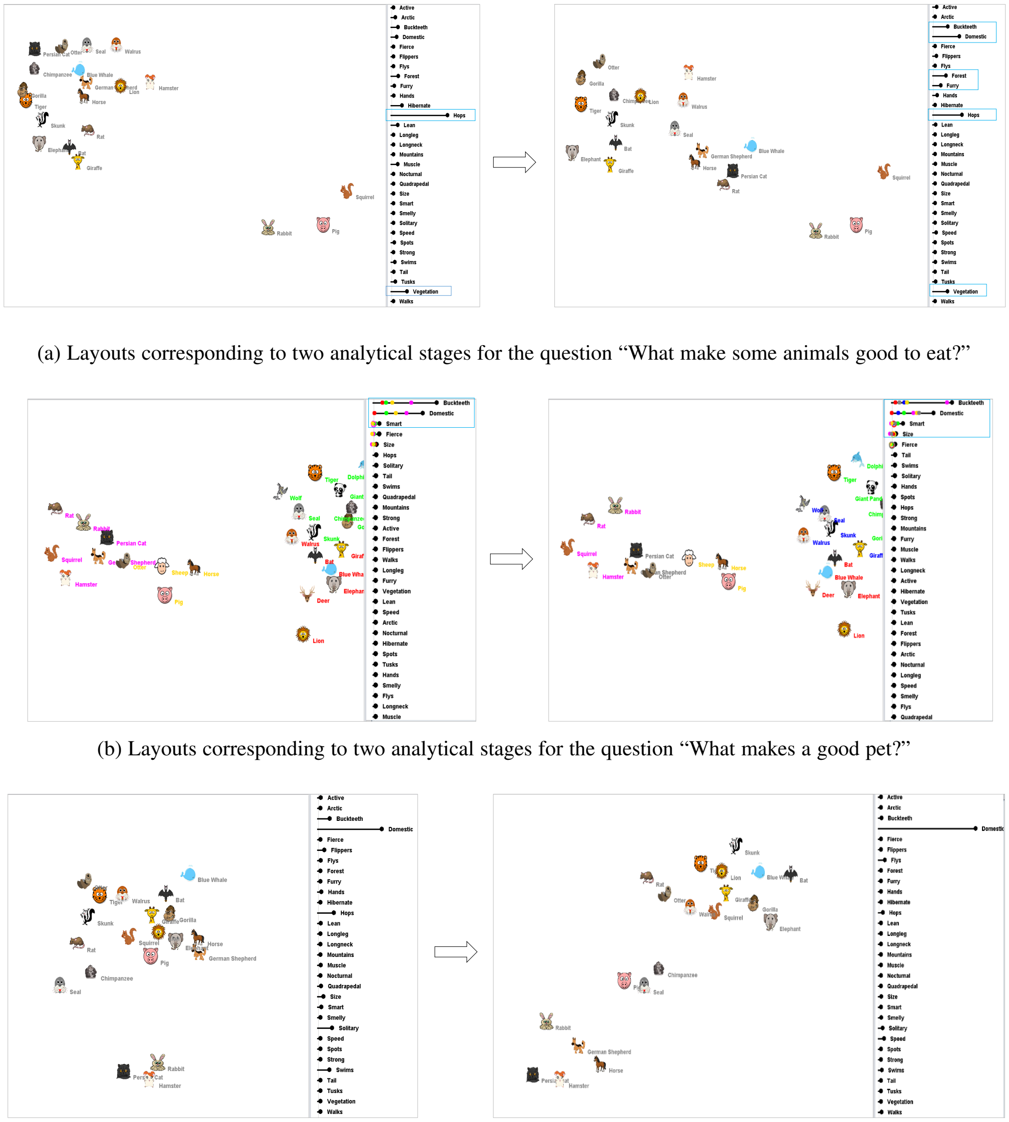


Fig. 5: Students exploit embodiment to understand the data. (a) The student who represented *skunk* moved far from other students to the bottom-right corner (not shown)). (b) Students discussed alternative positioning and one student took leadership. (c) Students formed two separate groups, indicating a binary hypothesis about the data: *edible* versus *non-edible*. (d) The student representing *rat* moved to between the groups. (e) Other students adjusted positions, which revealed a final hypothesis as a spectrum of edibility.



(a) Layouts corresponding to two analytical stages for the question "What make some animals good to eat?"

(b) Layouts corresponding to two analytical stages for the question "What makes a good pet?"

(c) Layouts corresponding to two analytical stages for the question "What differentiates wild and domestic animals?"

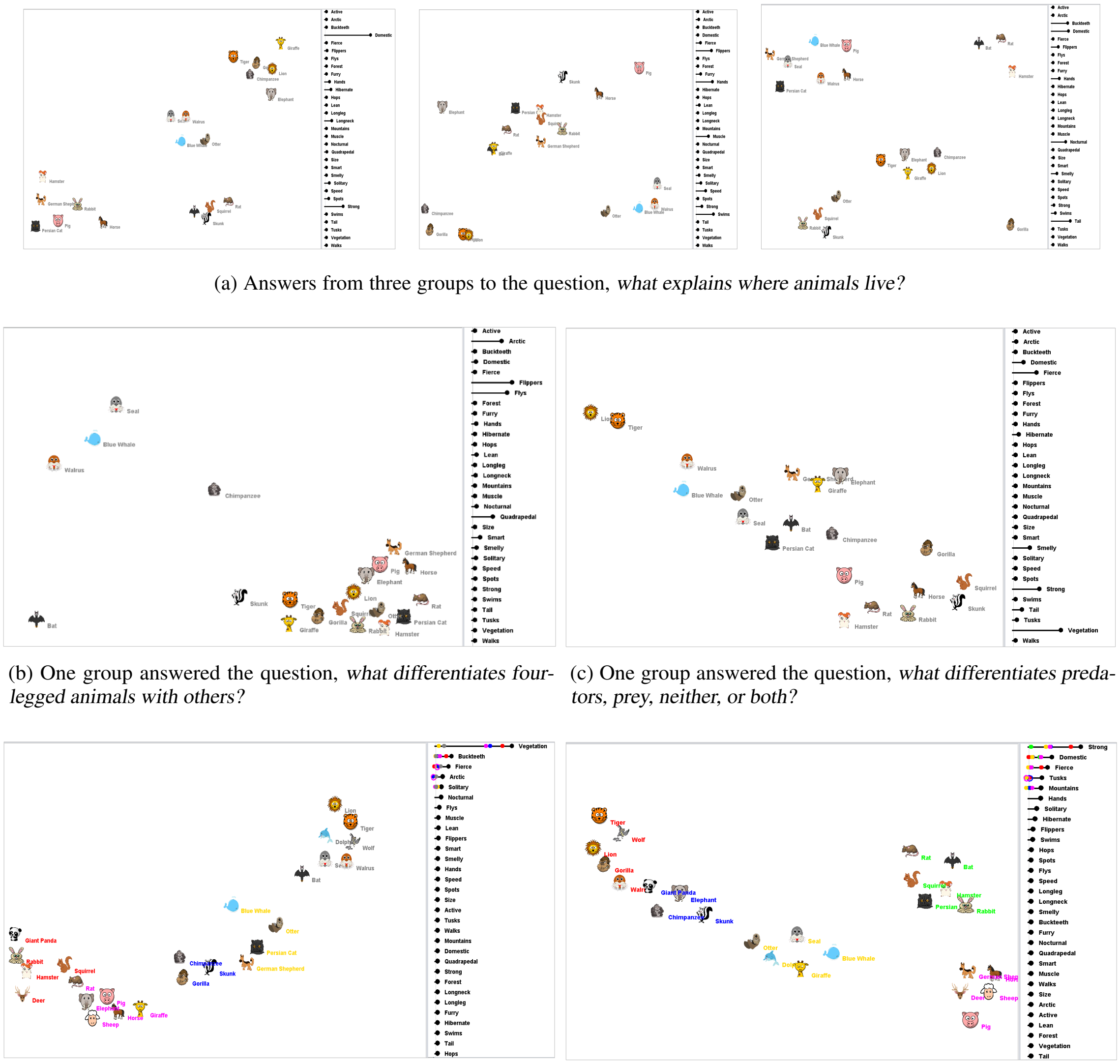
Fig. 6: Three examples of layout changes corresponding to two analytical stages for a given question.

which helped to elicit discussions. Some students took on leadership roles and directed others to move. The free movement capability gave students many opportunities to interact with others. This made learning with *Be the Data* more collaborative, as students were more engaged with others during the learning process. The coupling of physical and virtual enabled both of these forms of interaction to relate and intertwine. Physical behaviors served both human-computer and human-human interactions.

5.3 Analytic Strategies of Students

We observed four typical structures of student generated layouts with *Be the Data*. These layouts are exemplified in Figs. 5, 6, and 7, and display 1) *outliers*, 2) *binary groups* indicating a boolean understanding of data, 3) *multiple clusters*, and 4) *linear spectra*. Different structures reflect different interpretations of the data, and, more importantly, reflect differences in exploratory analytic strategies. In total, we identified three analytic strategies applied by students when using *Be the Data*: *broadening* analysis, *in-depth* analysis, and *multi-perspective* analysis. That is, it seems that with flexible social interactions, groups of students are able to *broaden* their analysis by considering more dimensions, perform *in-depth* analysis by reinforcing previously identified dimension(s), and hold *multiple perspectives* to interpret a dataset. We use the narratives portrayed by Figs. 5, 6, and 7 to support our discovery of these analytic strategies.

Broadening Analysis. We found students progressively considered more dimensions of the data during their exploratory analysis process, revealed as layout structure changes in Figs. 5 and Fig. 6. With instant system and peer feedback, students continuously explored assumptions and changed their judgments of the data. In Fig. 5, the layout evolution from *one outlier* (Fig. 5a) to *binary groups* (Fig. 5c), to a *linear spectrum* (Fig. 5e) demonstrates that the potentially considered dimensions (shown with blue boxes in these figures) increased from one to three to five. Similarly, in Fig. 6a and Fig. 6b, in early stages, students tended to focus on two dimensions in the data. In later stages they expanded their focus to four or five dimensions of the data. The increase in considered dimensions represents a *broadening analysis* strategy. Students progressively understand the data by considering more dimensions, and these dimensions are incrementally added to their analysis result. This



(a) Answers from three groups to the question, *what explains where animals live?*

(b) One group answered the question, *what differentiates four-legged animals with others?*

(c) One group answered the question, *what differentiates predators, prey, neither, or both?*

(d) One group answered the question, *how vegetarians, carnivores, and omnivores differ?*

(e) One group answered the question, *what animal you want to encounter on a deserted island?*

Fig. 7: Students interpret the same dataset from different perspectives, revealed as different visual layouts.

is revealed as the layout changes during their analysis. We only observed increased dimensions in the questions that seemed more subjective than others. For example, the interpretation of whether an animal is good to eat or is a good pet is subjective, while the interpretation of whether an animal is domestic is objective. In contrast to this strategy used to answer subjective questions, we found another strategy for the objective questions.

In-depth Analysis. We found that students confirmed key dimensions in early exploration stages and reinforced their findings at later stages, as reflected in Fig. 6c. In this figure, students are exploring an answer to the question, *What makes wild animals and domestic animals different?* At first, students formed two groups (Fig. 6c left). Under further investigation however, the students formed three groups (Fig. 6c right) and discovered that the dimension *domestic* increased in weight while others decreased. With the structural change students first confirmed the role of a key dimension, *domestic*, and then reinforced their confirmation.

Multi-perspective Analysis. Considering the exploratory nature of high-dimensional data, it is useful for students to consider different perspectives and thus produce different visualizations to gain new insights from the same data. As represented by layouts in Fig. 7, perspective changes when having either 1) multiple groups of students answering one common question and/or 2) the same or different sets of students answering multiple questions. For example, three groups of students answered the same question, “What explains where animals live?” differently (Fig. 7a). By bringing different external domain knowledge to bear, different visualizations were produced. Also, four groups asked and answered different questions to explore this dataset from different perspectives (Figs. 7b-7e). Different student groups had a variety of expertise and experiences and they were motivated by difference questions, so the students’ exploration perspectives were different.

5.4 Engagement and Usability Issues

Be the Data provided an enjoyable and engaging user experience. Students enjoyed the workshop as indicated in their responses to the open-ended question about their workshop experience. Among 183 responses (60 returned from AWC, 28 from STEP, 48 from DA, 47 from CEED), “interactive,” “moving around,” and “interesting” (or “fun,” “engaged,” and “cool”) were the most frequently mentioned positive factors, being mentioned 53, 68, and 69 times respectively. “Collaboration” and “teamwork” were

Human-Computer Interaction
• Moving around in the room
• Joining or leaving a group
• Looking at the visualization to determine positions
Human-Human (Social) Interaction
• Discussing with others, while identifying with their embodied animal: *"I am rabbit, so where is squirrel?"* *"Panda is too close to me [dolphin]."* *"Lion is over there. So I [tiger] need to move."* *"You [seal] and I [walrus] are the same, let's move together."*
• Directing others to move: *"Good to eat on the left, not good to eat on the right."* *"You guys want to go over there to be separate from carnivores."* *"We need to spread out."* *"Whale needs to move that way."*
• Preserving the authority and initiative to determine their own position: *"I don't think I [gorilla] am a good pet."* *"I (rat) might not belong to the non-edible group [and wandered away]."* *"I am pretty sure I [chimpanzee] am an omnivore."*
• Exploiting their shared spatial capabilities: *"They gotta be in that corner."* *"Cat, come over here."*
• Moving around to discuss with different groups of people: Large group collaboration spontaneously occurred to form clusters in the space. Small group collaboration spontaneously occurred to refine positions with neighbors.

TABLE 3: A summary of human-computer and human-human interaction in *Be the Data*. Both forms of interaction feed into each other and steer the underlying model, which further updates the visualizations. Embodied interactions of moving in the space simultaneously serve both HCI and HHI purposes.

mentioned 26 times. Students described *Be the Data* as "extremely interesting," "way more fun than class," and felt that they were more "engaged in the activity (as opposed to a lecture)."

Therefore, students felt that "data can be fun, not always tedious and boring." An elementary school teacher, while observing his students' activities in *Be the Data*, commented that "I have never seen my students so engaged for so long."

From those 183 responses, students also mentioned that they like the new technology (15 occurrences) and visualization (22 occurrences), and believed it was an intuitive way (6 occurrences) to learn data. Students commented that it was the "most unique data organization tool I've seen" and remarked on "how intuitive it was being able to see what we were doing on the screen."

While some students mentioned that the workshop was educational and informative (22 occurrences), two students were concerned about not learning the algorithm in a way that enabled them to analyze data mathematically. One student specifically noted the intuitiveness of performing multi-dimensional analysis and the difficulties of describing and quantifying data details. One student disliked the workshop.

6 Discussion

Our goal was to explore this new extreme form of embodiment to support student learning about high-dimensional data and analytical processes. *Be the Data* is an attempt to unify physical world experience and computational analytics experience in a natural way. We focused on an exploratory qualitative and quantitative analysis of how the students used this embodied approach to learn. Our results suggest that students gained knowledge about relevant concepts, produced various inferences from the data, and were engaged in the collaborative data exploration.

Be the Data offered an intuitive medium for students to reason about abstract data. The familiarity of the embodied "proximity $\approx$ similarity" metaphor seemed to support an efficient conceptualization of the underlying mathematical model. This could be due to exploiting students' spatial awareness capabilities in the real world [39], [40] to interpreting the spatial organization of abstract data. As students moved, they received immediate visual feedback of dimension changes. They built intuitions about relationships between their movement and the dimension changes. They were able to explore and test hypotheses to increase their understanding of the data in a high-dimensional way. Gigerenzer describes such learning as "Gut Feelings" [41]. He gave an example that although very few people would be able to calculate the parabolic curve that the ball takes and solve the problem mathematically, they are able to run towards the location where the ball will come down and catch the ball. This idea is similar to *Be the Data*: students gained understanding about dimension reduction concepts (as indicated by the improvement in the post-survey) and were able to draw multivariate insights that were not stemming directly from the mathematical formula. Such learning experience might lead to a better understanding later when they are confronted with the mathematical underpinnings. This is an interesting potential future study.

Yet, the embodiment in *Be the Data* is more than just the physical interaction with the space [6]. Based on the qualitative results, the social interaction and the students' close identification with the animal they embodied also seemed to play major roles. For example, when discussing with each other, students frequently used first person pronouns and possessives when referring to their animal's data values (Table 3). These characteristics seemed to contribute to the high level of engagement of the students in the data analysis tasks.

Be the Data has the potential to promote STEM education and outreach in data analytics. We conjectured that students, especially younger ones, would not pay attention to a data analysis task on the screen for 30 minutes. But with *Be the Data*, students as young as elementary and middle school age were engaged in data exploration throughout the workshop. The complexity of the data model formulation often scares students away from learning data analytics [3], but students might enjoy learning data analytics in a more natural way as doorway into future mathematics [4], [42]. Results from our user study suggested that *Be the Data* made the complex analytical method approachable to novices, made the data analysis tasks appealing, sparked interest, and encouraged further exploration of the subject. Therefore, we expect that *Be the Data* could be applied to reach a broad population of learners who are not necessarily knowledgeable about multivariate analysis algorithms.

6.1 Limitations

Our study has several limitations. It did not have a control group to compare against. More work is needed to understand how learners would perform differently, for example, when given a desktop application or attending a lecture. It is unknown if embodied physical interaction improved the collaborative understanding of information over purely virtual interactions. It would be difficult to tease apart all the dimensions of embodiment and social interaction that are so closely intertwined in the *Be the Data* approach, and to factorially experiment on their individual contributions to learning. However, our research is valuable in the identification of learners' analytical and collaborative strategies (qualitative) that employ this form of embodiment, accompanied by the evidence of learning (quantitative) and their reactions. A comparison of different applications and educational methods is beyond the scope of the current exploration.

6.2 Future Work

There are many ways *Be the Data* could be improved and extended. For the visualization, viewing the cluster labels in the graph and cluster values on the dimension chart is not ideal, since students have to switch their focus from left to right to connect the information from both. To reduce such effort, theme labels near clusters of dots could offer a potential solution [43]. Also, saving analytical artifacts during usage might benefit data exploration processes and educational activities. For example, if the system could identify and save multiple layouts during the exploration process, students could compare various explorations, and return to their key analytical steps if they are stuck in the current exploration. This would have helped in our workshops to enable the students to compare the results from multiple student groups.

In addition, students could gain access to data value details and even parametric interactions on the weight parameters through linked hand-held devices. Students could visualize details about the distances to their neighbors, and interactively tune weights or modify choices of distance metrics to see how it affects the projection. In such cases, it would be very interesting to visually project the data points onto the floor of the room, and students could chase their data points around the room as the mathematical WMDS model updates. This could increase the game-like fun of *Be the Data*. Handheld devices with WiFi or GPS tracking could also potentially be used to eliminate the need for the expensive vision-based tracking system and hats, making *Be the Data* more broadly portable to classrooms and other spaces. This could exploit student interest in games such as Pokémon GO, but applied to data analytics education.

It would also be interesting to explore applications of the *Be the Data* concept to other types of statistical models for data analytics, beyond WMDS. For example, students could examine differences between multiple types of dimension reduction models that have been parameterized for OLI interaction, such as MDS, PCA, and GTM [36], or learn about clustering algorithms that use OLI techniques [44] and even topic modeling for text analytics [45]. Many other types of statistical models could be parameterized and inverted to support OLI interaction [32], and then applied using the *Be the Data* form of embodied interaction.

7 Conclusion

We created the novel *Be the Data* concept, system, and associated educational workshops, in which students embody unique data points in a high-dimensional dataset and physically explore alternative projections as a collaborative group. We demonstrated its effectiveness in educational and outreach scenarios with students ranging from elementary to undergraduate ages and analyzed their embodied analytical strategies. *Be the Data* is able to empower students, who have relatively low sophistication in the underlying data analytics model, to learn and draw high-dimensional inferences from the data. As a result, we believe that *Be the Data*, through maximizing embodiment, enables engagement that may benefit education in high-dimensional data analytics.

Acknowledgments

This research was supported in part by NSF grants IIS-1447416, IIS-1218346, and DUE-1141096, and a seed grant from the Virginia Tech Institute for Creativity, Arts, and Technology.

References

[1] M. Verleysen *et al.*, "Learning high-dimensional data," *Nato Science Series Sub Series III Computer And Systems Sciences*, vol. 186, pp. 141–162, 2003.

[2] J. S. Yi, R. Melton, J. Stasko, and J. A. Jacko, "Dust & magnet: multivariate information visualization using a magnet metaphor," *Information Visualization*, vol. 4, no. 4, pp. 239–256, 2005.

[3] N. S. Ashaari, H. M. Judi, H. Mohamed, Tengku, and M. T. Wook, "Student's attitude towards statistics course," *Procedia - Social and Behavioral Sciences*, vol. 18, pp. 287–294, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877042811011542

[4] P. M. Valero-Mora and R. D. Ledesma, "Using interactive graphics to teach multivariate data analysis to psychology students," *Journal of Statistics Education*, vol. 19, no. 1, 2011. [Online]. Available: https://www.learntechlib.org/p/109366

[5] J. B. Kruskal and M. Wish, *Multidimensional scaling*. Sage, 1978, vol. 11.

[6] P. Dourish, *Where the action is: the foundations of embodied interaction*. MIT press, 2004.

[7] X. Chen, J. Z. Self, L. House, and C. North, "Be the data: A new approach for immersive analytics," in *IEEE Virtual Reality 2016 Workshop on Immersive Analytics*, 03/2016.

[8] X. Chen, L. House, J. Z. Self, S. Leman, J. R. Evia, J. T. Fry, and C. North, "Be the data: An embodied experience for data analytics," in *2016 Annual Meeting of the American Educational Research Association (AERA)*, 04/2016.

[9] X. Chen, J. Z. Self, M. Sun, L. House, and C. North, "Be the data: Social meetings with visual analytics," in *International Workshop on Visualization and Collaboration (VisualCol 2016)*, 11/2016.

[10] C. H. Lampert, H. Nickisch, S. Harmeling, and J. Weidmann, "Animals with attributes: A dataset for attribute based classification," 2009.

[11] B. Lee, P. Isenberg, N. H. Riche, and S. Carpendale, "Beyond mouse and keyboard: Expanding design considerations for information visualization interactions," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2689–2698, Dec 2012.

[12] J. C. Roberts, P. D. Ritsos, S. K. Badam, D. Brodbeck, J. Kennedy, and N. Elmqvist, "Visualization beyond the desktop–the next big thing," *Computer Graphics and Applications, IEEE*, vol. 34, no. 6, pp. 26–34, 2014.

[13] T. Chandler, M. Cordeil, T. Czauderna, T. Dwyer, J. Glowacki, C. Goncu, M. Klapperstueck, K. Klein, K. Marriott, F. Schreiber *et al.*, "Immersive analytics," in *Big Data Visual Analytics (BDVA), 2015*. IEEE, 2015, pp. 1–8.

[14] P. Isenberg, T. Isenberg, T. Hesselmann, B. Lee, U. von Zadow, and A. Tang, "Data visualization on interactive surfaces: A research agenda," *IEEE Computer Graphics and Applications*, vol. 33, no. 2, pp. 16–24, March 2013.

[15] C. Andrews, A. Endert, B. Yost, and C. North, "Information visualization on large, high-resolution displays: Issues, challenges, and opportunities," *Information Visualization*, pp. 341–355, 2011.

[16] A. Febretti, A. Nishimoto, T. Thigpen, J. Talandis, L. Long, J. D. Pirtle, T. Peterka, A. Verlo, M. Brown, D. Plepys, D. Sandin, L. Renambot, A. Johnson, and J. Leigh, "Cave2: a hybrid reality environment for immersive simulation and information analysis," 2013. [Online]. Available: http://dx.doi.org/10.1117/12.2005484

[17] Y. Jansen, P. Dragicevic, P. Isenberg, J. Alexander, A. Karnik, J. Kildal, S. Subramanian, and K. Hornbæk, "Opportunities and challenges for data physicalization," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI '15. New York, NY, USA: ACM, 2015, pp. 3227–3236. [Online]. Available: http://doi.acm.org/10.1145/2702123.2702180

[18] L. Bradel, A. Endert, K. Koch, C. Andrews, and C. North, "Large high resolution displays for co-located collaborative sensemaking: Display usage and territoriality," *International Journal of Human-Computer Studies*, vol. 71, no. 11, pp. 1078–1088, 2013.

[19] C. Forlines and C. Shen, "Dtlens: Multi-user tabletop spatial data exploration," in *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '05. New York, NY, USA: ACM, 2005, pp. 119–122. [Online]. Available: http://doi.acm.org/10.1145/1095034.1095055

[20] J. Heer and M. Agrawala, "Design considerations for collaborative visual analytics," *Information visualization*, vol. 7, no. 1, pp. 49–62, 2008.

[21] P. Isenberg, D. Fisher, S. A. Paul, M. R. Morris, K. Inkpen, and M. Czerwinski, "Co-located collaborative visual analytics around a tabletop display," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 5, pp. 689–702, May 2012.

[22] M. Tobiasz, P. Isenberg, and S. Carpendale, "Lark: Coordinating co-located collaboration with information visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 10 650–1072, Nov 2009.

[23] P. Isenberg and S. Carpendale, "Interactive tree comparison for co-located collaborative information visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1232–1239, Nov 2007.

[24] S. Leman, L. House, and A. Hoegh, "Developing a new interdisciplinary computational analytics undergraduate program: A qualitative-quantitative-qualitative approach," *The American Statistician*, vol. 69, no. 4, pp. 397–408, 2015.

[25] V. Gallese and G. Lakoff, "The brain's concepts: The role of the sensory-motor system in conceptual knowledge," *Cognitive neuropsychology*, vol. 22, no. 3-4, pp. 455–479, 2005.

[26] G. Lakoff and M. Johnson, *Metaphors we live by*. University of Chicago press, 2008.

[27] G. Lakoff and R. E. Núñez, *Where mathematics comes from: How the embodied mind brings mathematics into being*. Basic books, 2000.

[28] M. Howison, D. Trninic, D. Reinholz, and D. Abrahamson, "The mathematical imagery trainer: From embodied interaction to conceptual learning," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. New York, NY, USA: ACM, 2011, pp. 1989–1998. [Online]. Available: http://doi.acm.org/10.1145/1978942.1979230

[29] A. C. Robinson, "Collaborative synthesis of visual analytic results," in *2008 IEEE Symposium on Visual Analytics Science and Technology*, Oct 2008, pp. 67–74.

[30] C. Andrews, A. Endert, and C. North, "Space to think: Large high-resolution displays for sensemaking," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10. New York, NY, USA: ACM, 2010, pp. 55–64. [Online]. Available: http://doi.acm.org/10.1145/1753326.1753336

[31] J. Z. Self, N. Self, L. House, S. Leman, and C. North, "Improving students' cognitive dimensionality through education with object-level interaction," 2014.

[32] S. C. Leman, L. House, D. Maiti, A. Endert, C. North *et al.*, "Visual to Parametric Interaction (V2PI)," *PloS One*, vol. 8, no. 3, 2013.

[33] Qualisys, "Qualisys Track Manager QTM: Motion Capture software for tracking all kind of movements," http://content.qualisys.com/2015/10/PI_QTM.pdf.

[34] R. L. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953. [Online]. Available: http://dx.doi.org/10.1007/BF02289263

[35] M. Johnson, *The body in the mind: The bodily basis of meaning, imagination, and reason*. University of Chicago Press, 2013.

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

[37] R. L. Wasserstein and N. A. Lazar, "The ASA's Statement on p-values: Context, Process, and Purpose," *The American Statistician*, vol. 70, no. 2, pp. 129–133, 2016.

[38] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*. Taylor & Francis, 2014, vol. 2.

[39] S. R. Klemmer, B. Hartmann, and L. Takayama, "How bodies matter: Five themes for interaction design," in *Proceedings of the 6th Conference on Designing Interactive Systems*, ser. DIS '06. New York, NY, USA: ACM, 2006, pp. 140–149. [Online]. Available: http://doi.acm.org/10.1145/1142405.1142429

[40] R. J. Jacob, A. Girouard, L. M. Hirshfield, M. S. Horn, O. Shaer, E. T. Solovey, and J. Zigelbaum, "Reality-based interaction: A framework for post-wimp interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '08. New York, NY, USA: ACM, 2008, pp. 201–210. [Online]. Available: http://doi.acm.org/10.1145/1357054.1357089

[41] G. Gigerenzer, *Gut feelings: The intelligence of the unconscious*. Penguin, 2007.

[42] J. Al-Aziz, N. Christou, and I. D. Dinov, "Socr "motion charts": An efficient, open-source, interactive and dynamic applet for visualizing longitudinal multivariate data," *Journal of Statistics Education*, vol. 18, no. 3, 2010. [Online]. Available: https://www.learntechlib.org/p/109359

[43] E. Kandogan, "Just-in-time annotation of clusters, outliers, and trends in point-based data visualizations," in *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, ser. VAST '12, 2012, pp. 73–82.

[44] J. Wenskovitch and C. North, "Observation-level interaction with clustering and dimension reduction algorithms," in *Proceedings of the 2Nd Workshop on Human-In-the-Loop Data Analytics*, ser. HILDA'17. ACM, 2017, pp. 14:1–14:6.

[45] A. Endert, P. Fiaux, and C. North, "Semantic interaction for sensemaking: Inferring analytical reasoning for model steering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2879–2888, Dec 2012.

Xin Chen is currently a research software engineer at Bloomberg. She received her Ph.D. in Instructional Design and Technology and M.S. in Computer Science from Virginia Tech in 2016.

Jessica Zeitz Self is an Assistant Professor of Computer Science at the University of Mary Washington. She received her PhD in Computer Science from Virginia Tech in 2016.

Leanna House is an Associate Professor of Statistics at Virginia Tech, and Deputy Director of the Computation Modeling and Data Analytics degree program. She earned her Ph.D. in Statistics from Duke University.

John Wenskovitch is a graduate student in Computer Science at Virginia Tech. He received a Master's Degree in Computer Science from the University of Pittsburgh and a Bachelor's Degree in Software Engineering from Gannon University.

Maoyuan Sun is an Assistant Professor in the Computer and Information Science Department at the University of Massachusetts Dartmouth. He received his Ph.D. in Computer Science from Virginia Tech in 2016.

Nathan Wycoff is a graduate student in Statistics at Virginia Tech. He received a Bachelors of Science in Statistics from the same department in 2016.

Jane Robertson Evia is an Assistant Professor of Practice in Statistics at Virginia Tech. She earned her Ph.D. in Educational Psychology from the University of North Carolina at Chapel Hill.

Scotland Leman is an Associate Professor of Statistics at Virginia Tech. He received his Ph.D. from Duke University in the Department of Statistical Science, and a M.S. degree in Computational Engineering from Stanford University.

Chris North is a Professor of Computer Science at Virginia Tech, and Associate Director of the Discovery Analytics Center. He earned his Ph.D. in Computer Science from the University of Maryland, College Park.