Be the Data:
Embodied Visual Analytics

Xin Chen

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Science and Applications

Christopher L. North, Chair

Leanna L. House, Co-Chair

Nicholas F. Polys

July 15, 2016

Blacksburg, Virginia

Keywords: Visual Analytics, Embodied Interaction, Collaboration

Be the Data: Embodied Visual Analytics

Xin Chen

(ABSTRACT)

With the rise of big data, it is becoming increasingly important to educate students about data analytics. In particular, students without a strong mathematical background usually have an unenthusiastic attitude towards high-dimensional data and find it challenging to understand relevant complex analytical methods, such as dimension reduction. In this thesis, we present an embodied approach for visual analytics designed to teach students exploring alternative 2D projections of high dimensional data points using weighted multidimensional scaling. We proposed a novel concept, *Be the Data*, and its application to explore the possibilities of using human's embodied resources to learn from high dimensional data. In our 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 moves in a room with respect to others to interact with alternative projections and receive visual feedback. We conducted educational workshops with students inexperienced in relevant data analytical methods. Our findings indicate that the students were able to learn about high-dimensional data and data analysis process despite their low level of knowledge about the complex analytical methods. Similarly, we applied *Be the Data* to social meetings. We used the same techniques to analyze and display social-cluster related information to facilitate social interactions in real time.

This work was supported by NSF grant DUE-1141096.

Acknowledgments

One and a half years went in a blink of eyes. I would like to deliver my most sincere thanks to all those who have supported me on my way towards my Master's Degree in Computer Science and Applications.

First of all, Dr. Chris North and Dr. Leanna House, my Master thesis advisors, deserved great thanks for their understanding, support and guidance for me throughout my Master study. They have stimulated me with broad interests, inspired me with brilliant ideas, and guided me through all barriers in my way to the degree. Dr. North's encourages me to enter the field of information visualization, which is completely new to me. His generous support not ly helped me overcome the difficulties in changing the major from Education but also take advantage of my education background to conduct research in Computer Science. Dr. House always give me insightful feedback and patient instructions. My study would not have become reality without her accompany for several afternoons and nights in the Cube to test facilities. I am also very grateful to Dr. Polys for his insightful suggestions about my work. My dissertation would not have been done without the guidance of all my committee members: Dr. Chris North, Dr. Leanna House, and Dr. Nicholas Polys.

I have really enjoyed working with you in these years. I am really thankful for your invaluable guidance that shapes me into a good scholar.

Thanks are also sent to my colleagues and friends in Department of Computer Science. Jessica Self always helps me whenever I have questions about my research, study, and life. I will miss those days when we spent hours and hours together learning, chatting, and laughing. Maoyuan, Peng, Ji, Caleb, and Michelle are my lab mates and friends who always give me warm help and brilliant inspirations. My Master expedition would have been much more difficult without you. Thanks also go to all the faculty and students in Department of Computer Science I have asked for help.

Finally, my most sincere thanks belong to my parents, my husband, my son and my family. Thanks for your endless love, understanding, and support. I would not have today's achievement without you.

Thanks to all of you for what you have done for me!

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Big Data. Big Data. Big Data. In the news, online, and at work, we are constantly hearing the buzz phrase, "Big Data". With advances in technology, the amount of analyzable data is growing rapidly at low cost. Within these large datasets is information that we hope to derive scientific discoveries. However, as noted in the book Illuminating the Path [1], datasets are just tables of numbers without humans to discover, process, reflect, and communicate information in the data.

There is a clear need to promote education in knowledge discovery from big data. In practice, learning from data requires comprehensive critical thinking skills which (1) extend beyond the application of quantitative statistical or computational methods and (2) include qualitative forms of thought, such as formalizing potential biases, communicating personal judgment, exploring multiple solutions, assimilating new information with old, and assessing implications of discoveries. Unfortunately, current approaches in teaching data analytics focus primarily on its quantitative aspects to train students to master quantitative theory and methods. Students without strong math prerequisites may be excluded from the analytical classes. Even worse, the complexity of quantitative aspects scare students away from learning. Students normally have an unenthusiastic attitude towards learning data analytics if they do not have a strong mathematical background [2].

Immersive data exploration has the potential to motivate and reinforce quantitative and qualitative aspects of data analyses. To promote STEM outreach and attract students to learn data analytical skills, we designed and developed a novel combination of physical, virtual, and social worlds for immersive data exploration. Specifically, we propose a novel concept, Be the Data, which means an individual person embodies a unique virtual data point in a high-dimensional data set. As a proof of concept, we developed a system that immerses students as data points in a physical space. In our system, students enter a physical space to become individual data points, and the room becomes the low-dimensional projection. For example, if we consider a high-dimensional dataset about animals, each student becomes an animal data point. Their positions on the ground represent the two-dimensional projection of the high-dimensional data. That is to say, coordinates in the room are coordinates in a two-dimensional plane to which the high-dimensional data are projected.

In addition to educating people on data analysis, *Be the Data* has the potential to be meshed with our daily activities. The *Be the Data* system has been implemented into social meetings to facilitate social interactions. Each participant is a data point of a high-dimensional data set where participants' attributes are the dimensions. The system provides participants feedback about who are similar in what dimensions to help them find conversational topics and locate the people of interest.

1.1 Research Motivations

1.1.1 Immerse In the Data

In the presence of large datasets, research in immersive analytics is devoted to facilitating the comprehension of data by bringing data and the data analysis process into the physical world.

We are seeing immersive interfaces develop quickly to immerse analysts in the data for natural methods of data exploration and collaboration.

Interactive surfaces enable direct interaction with data which is easier and more intuitive than using a mouse on a desktop display [3]. These interfaces are also increasingly developed into large high-resolution displays (or high-pixel tiled display walls) [4, 5]. The combination of higher pixel count and multi-touch interaction allow embodied physical navigation that outperforms virtual navigation, especially when dealing with large datasets [6].

Emerging stereoscopic 3D display technologies, complemented with virtual reality techniques, immerse users in computer-generated scenes. For example, the design of CAVE2 [7] allows users to "walk through" and "fly-over" the hybrid reality scenes. Users in the AVIE interface are able to walk inside a 360-degree stereographic interface to manipulate digital archives in forms text, photos, images, and sound [8]. With tools embracing multi-modalities (e.g., audio, haptic, gestural), immersive analysis is not only about a visual experience, but becomes an integrated multi-sensory experience [9–11].

In addition to bringing users to the hybrid reality environments, attempts are being made to bring the virtual data into the physical world. Digital data are now made accessible in graspable and manipulative artifacts whose physical attributes (e.g., geometry, materials) encode data [12]. For instance, Professor Richard Burdett presented his Maps of City Population in wooden 3D models in which height property encodes population density [13]. It was an engaging way to represent mass statistical information that invited people to explore. As digital fabrication technologies made it possible and easy to create physical representations of the data (even large datasets), researchers increasingly investigate how to leverage a humans perception skills in exploring data in physical forms [14, 15].

By means of touching virtual data on an interactive surface, walking inside computer-generated

scenes of data, or exploring physical representations of data, existing immersive analytical approaches place users in their data. We call it *"Be In the Data"*.

1.1.2 Be the Data: A New Perspective

We propose a new facet of immersive data analytics that seeks to take immersion to the extreme. Unlike existing approaches that immerse users in the data, we immerse users to actually become data points. We call this new perspective "Be the Data".

"Be the Data" shares many similarities as *"Be In the Data"*. Users navigate and explore a physical-virtual hybrid space to analyze data. Users take advantages of collaborative work in a 3D space. However, Be the Data differs from Be In the Data in the perspective that the user is taking. Instead of looking into the data points, users are the data points. In our system, after students embody data points, they are able to take an egocentric role in conjecturing various relationships among the data. For example, for being a skunk, I may naturally separate myself from other animals because obviously I am very smelly. Students are able to discuss/negotiate with their neighbors to determine the positions of themselves based on their prior knowledge about animals and based on the context of specific problems. Data exploration naturally becomes a social process of users collaboratively reorganizing themselves in the room.

Research in virtual environments suggests that learning benefits from embodiment of an avatar. The activities performed by avatars inside virtual worlds render situated or authentic learning experiences as users would solve problems contextualized in real life situations. Similar to the idea of an "avatar", here users become a "datatar" as we focus on data. We seek to explore if and how people interact with data in embodied ways through the "datata" could render an engaging, collaborative, and effective experience, which could lead to deep insights about data and analytical

processes.

1.2 Contributions

This thesis work focuses on the innovative education and application of high-dimensional data analysis with interactive visualizations for non-experts. The key contributions are:

- **the novel concept *Be the Data***, which means that an individual person embodies a unique data point in a high-dimensional dataset, to empower interactive data exploration and statistical concepts interpretation.
- **the novel system as a proof of concept** for educational and social purposes. The system leverages users' embodied resources with computer-based visual feedback to invoke learning and conversations.
- **the identification of users' analytical strategies that employ this form of embodiment**. Users were found to have an effective and engaged experience interacting with the system to explore/analyze a high-dimensional data set to solve their problems.

1.3 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 describe the *Be the Data* system specifics. Chapter 3 presents an embodied approach designed to teach students about exploring 2D projections of high dimensional data points using the *Be the Data* system. Chapter 4 presents an application of the *Be the Data* system to promote social meetings. Chapter 5 concludes the study and proposes future work.

Chapter 2

Be the Data System

Be the Data is an extension of a desktop-based application called Andromeda [16] that explores high dimensional-data in a professional manner. The desktop-based platform has its educational limitations. It is difficult to imagine that a novice student would engage in such an advanced interface. Moving data points on a screen could turn into a tedious and meaningless task. Also, it is challenging to conceptualize an abstract mapping from the virtual data to the virtual visualizations. The key concepts and insights are veiled behind small screen portals and simplistic interaction mechanics suggested by mouse and keyboard. Therefore, we invented Be the Data for immersive analytics. Within the physical space, students have an intuitive and egocentric spatial perception to judge the physical distance: walking toward people that seem similar to me while staying away from people that seem different from me. The physical metaphor near is similar matches the conceptualization of the underlying mathematical model. The concrete experience provides a physical medium for students to reason about the abstract Euclidean distance. Moreover, the shared space brings multiple learners for collaboration and the sharing of ideas.

To implement *Be the Data*, we exploit a multi-media physical room, the Cube [17] in the Institute for Creativity, Arts, and Technology at Virginia Tech. Relying on advanced interactive technologies for physical-virtual cross-overs, our system is comprised of a collocated physical space (the

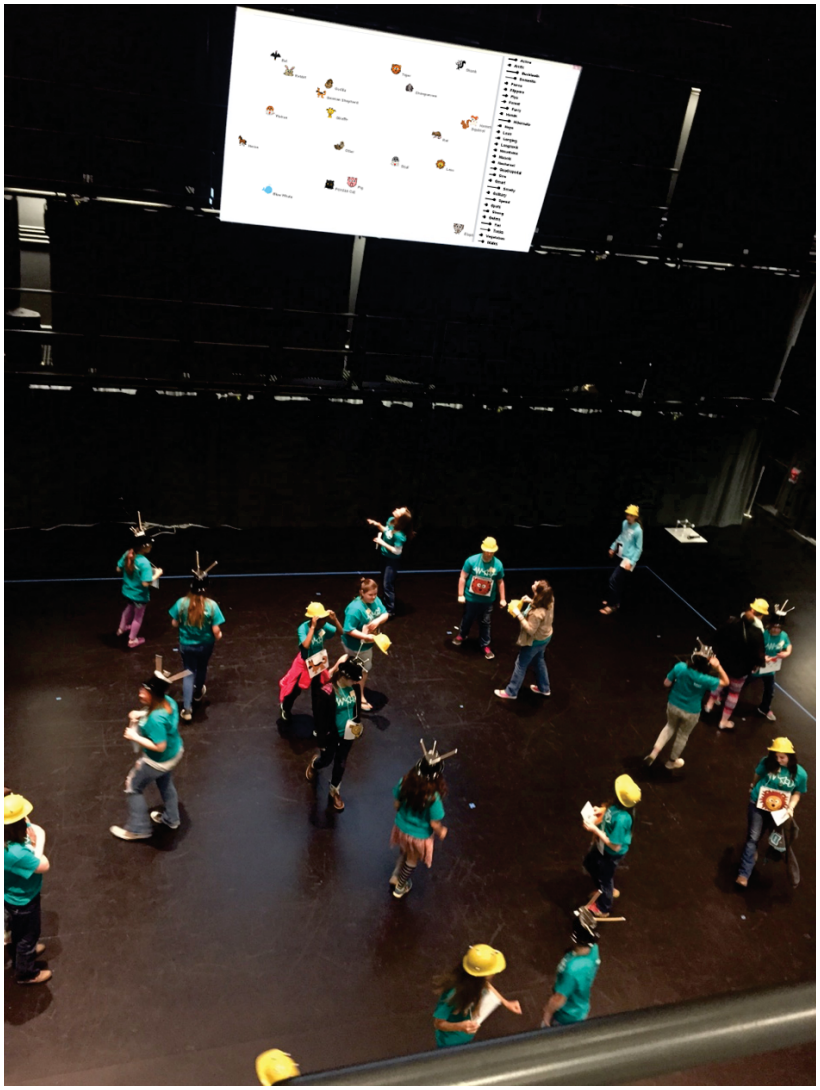


Figure 2.1: In *Be the Data*, students become individual data points. A birds-eye view of their locations in the room is displayed on the large display above them.

Name	Walks	Vegetation	Tail	Speed	Smelly
Persian Cat	65.69	6.25	66.8	26.98	7.86
Horse	55.58	51.05	70.42	81.68	33.07
Blue Whale	0	0	26.42	21.42	13.75
Skunk	64.86	44.38	83.33	30.21	100

Table 2.1: A portion of the animal dataset (20-25 animals X 30 variables). Example in this table shows 4 animals and 5 variables.

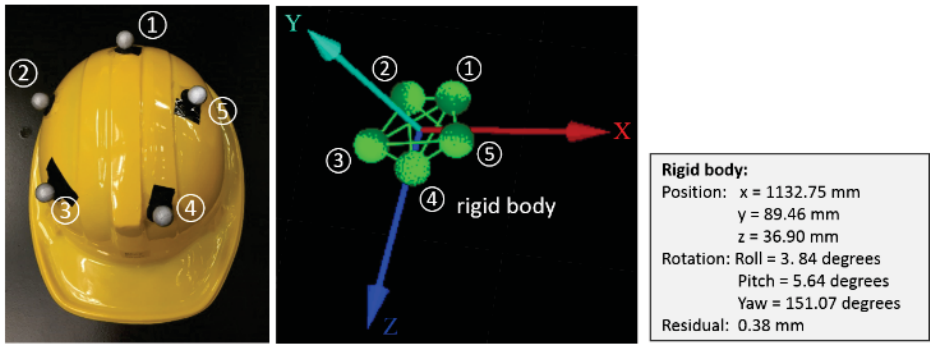
Cube), a motion tracking system, several trackable hats, a backend software layer, and an overhead large display (Figure 2.1).

2.1 Motion Tracking Facilities

Be the Data uses an OptiTrack motion tracking system, which includes 24 Oqus cameras, reflective markers, and the Qualisys Track Manager (QTM) software. QTM is used to collect and process motion capture data from the cameras. We retrieve data from the QTM server over a UDP/IP connection in real-time by following the Open Sound Control (OSC) protocol.

2.2 Trackble Hats

To simultaneously track multiple individuals and differentiate them from each other, we made our trackable hats (Figure 2.2a). Each hat is a rigid body that has its own particular and definite space. It is defined by a particular placement of 4-6 reflective markers (Figure 2.2b). Each rigid body in 3D space has six degrees of motion freedom (Figure 2.2c). We determine the 2D coordinates of individuals in the room by streaming the x and z values of the rigid body in real time. The



(a) (b) (c)

Figure 2.2: The hat used to track participants' positions in the Cube. (a) A trackable hat. Its unique structure is defined by the placement of reflective markers on it. (b) A rigid body presentation of the hat in 3D views. (c) The two-dimensional coordinates are from x and z values in the rigid body.

current implementation of the tracking system and the trackable hats allow for accurate tracking and differentiation of more than 50 objects.

2.3 Backend Software Layer

Be the Data is supported by the backend software layer called Andromeda [16], a desktop-based application for professional data analysis. By applying Visual to Parametric Interaction (V2PI) [18], Andromeda allows users to communicate their ideas about the high-dimensional data by manipulating data points in the visualization, which is a 2D projection of the high-dimensional data. For example, users can drag data points to change the pairwise Euclidian distances among them. Users convey the judgment that data points are similar by pulling them closer and data points are different by pushing them further apart. In turn, the system runs the inverted MDS algorithm to provide visual feedback: a set of weights that describe the visualization.

V2PI shields users from the technicalities of mathematical models so that users may focus on exploring data based on what they know, hypothesize, or learn from the data in an iterative way.

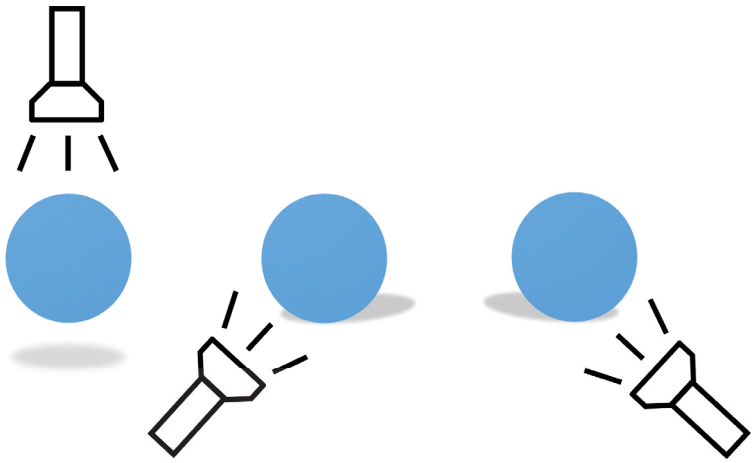


Figure 2.3: A simplified illustration of projection. A 3-dimensional ball is projected to the 2-dimensional plane, producing different shadows based on the position of the flashlight.

Be the Data integrates the Andromeda software to immerse users as movable data points in a physical immersive environment. With Be the Data, users employ their whole body as portable input that works from any location within the defined area in the Cube. The inputs to the system are users' positions in the Cube captured in real time via the trackable hats. The outputs are interactive visualizations as described in the next section.

2.4 Interactive Visualization

The interactive visualization for the *Be the Data* system includes two essential parts: (1) a WMDS plot and a dimension chart, organized left and right respectively on the large display (Figure 2.4), and (2) a dynamic clustering plot on the top of the WMDS plot (Figure 2.6). The visualization is displayed on a 14' diagonal Stumpfl projection screen with a 1920x1080 resolution.

2.4.1 WMDS Plot

WMDS plots a low-dimensional spatialization of the data in 2D Euclidean space to represent how the data spread in the high-dimensional space. The 2D layout is determined by weight parameters of p dimensions, which reflects the relative importance of each dimension in a visualization. The coordinates r of a WMDS plot for high-dimensional data d relies by minimizing a stress function:

$$r = \min_{r_1,\ldots r_n} \sum_{i=1}^{n} \sum_{j>1}^{n} |dist_L(r_i, r_j) - dist_H(\omega, d_i, d_j)|, \tag{2.1}$$

where n is the number of data points, $dist_L(r_i, r_j)$ is a distance between 2D points r_i and r_j, and $dist_H(\omega, d_i, d_j)$ is a distance measured between high-dimensional points d_i and d_j.

Effectively, the solution for r given ω defines a shadow of the high-dimensional data, much like a flashlight, as illustrated by Figure 2.3. As weights change, the flashlight moves and new shadows result.

To interpret the shadow (i.e., WMDS plot), the relative distances between data points reflect their similarity or difference: near suggests relatively similar while far suggests relatively different in the dimensions that are emphasized (i.e., variables that are weighted more). All weights are set equal and ordered alphabetically in the default layout (Figure 2.4a). As users change the layout by rearranging themselves in the room, the weights get updated to explain users choice of positions (Figure 2.4b). 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 Figure 2.4a and Figure 2.4b, the Tiger moves closer to the Pig, thus the Tiger is now considered more similar to the Pig than the remaining animals in the up-weighted dimensions, such as Flipper, Hibernate, and Size.

To update the weights, the system employs the inverse WMDS algorithm [18] to map layout

changes to new values for weights. That is, the inverse algorithm solves weight ω given adjusted low-dimensional coordinates $r*$,

$$\omega = \min_{\omega_1,\ldots\omega_p} \sum_{i=1}^{n} \sum_{j>1}^{n} |dist_L(r_i^*, r_j^*) - dist_H(\omega, d_i, d_j)|. \tag{2.2}$$

Because the algorithm considers the relative distance, not the absolute distance between data points, the size of the Cube does not affect the performance of the algorithm. The inverted algorithm runs fast enough to get results in real time.

In *Be the Data*, students adjust their low-dimensional coordinates (as determined by equation 2.2) by rearranging themselves in the Cube. In turn, they are provided with new weights for the dimensions (as solved from equation 2.1) that explain their choice of locations. We summarize this data exploration process in Figure 2.5. When students move several times, they are effectively exploring data from multiple perspectives that is defined by different 2D projections and the updated weights. This is a clear advantage of our system that students are shielded from the technicalities of mathematical models and may focus on exploring and learning from data based on their domain-specific questions.

2.4.2 Dynamic Clustering

Although the WMDS plot reveals up-weighted dimensions that characterize users choice of grouping, it does not show information about how groups distribute on these dimensions. Therefore, we implement dynamic clustering to visualize clusters of data on top of the WMDS plot (Figure 2.6a). That is to say, given the projected coordinates on the two-dimensional plane, the system automatically reveals clusters of data points based on their Euclidean distance in real time. We focus

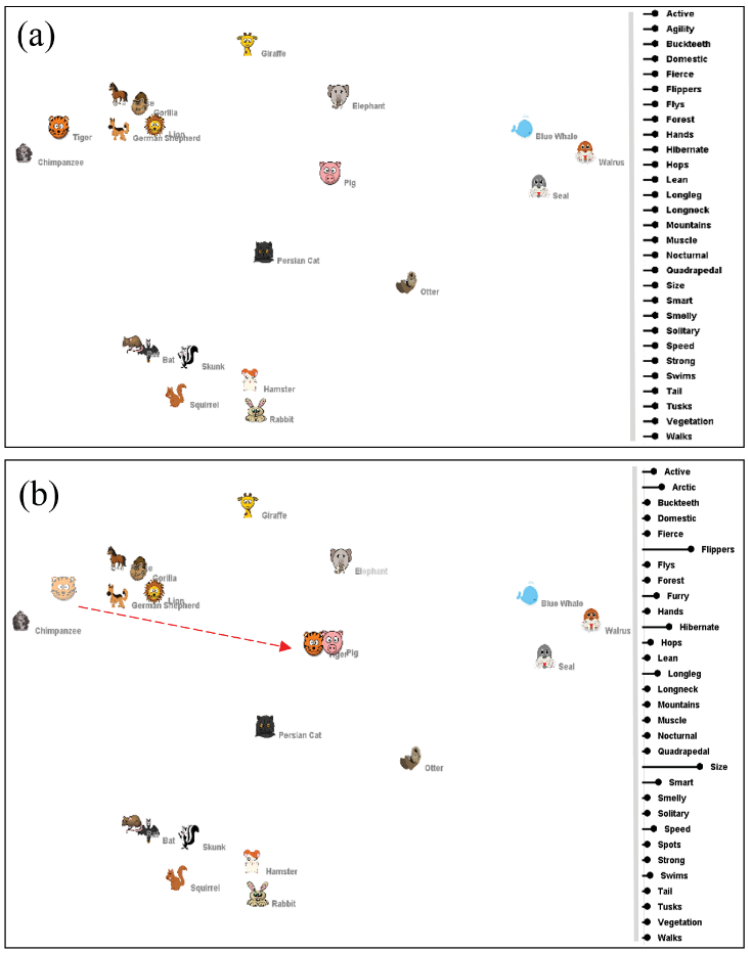


Figure 2.4: Interactive visualizations. (a) A clear image shown on the overhead large display to visualize students locations in the room. (b) When students move in the room, they are changing the two-dimensional coordinates of the WMDS plot and relative weights of dimensions.

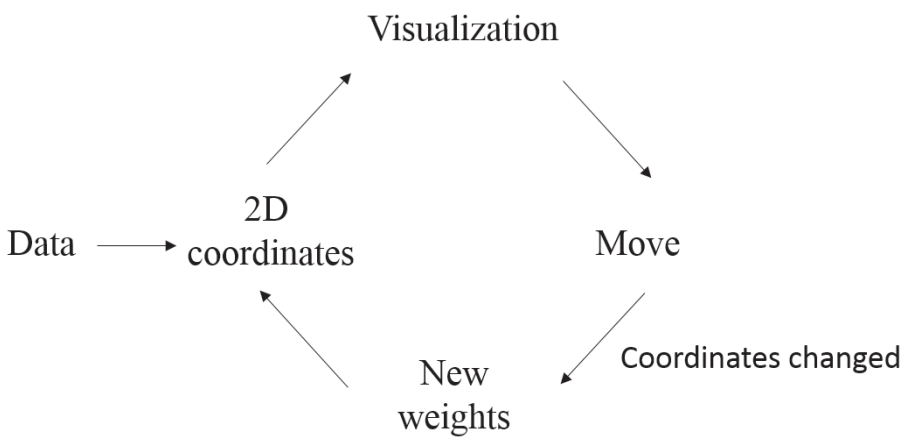


Figure 2.5: A simplified illustration of the *Be the Data* loop.

clustering on the 2D view space, not the high-dimensional space.

Dynamic clustering is calculated by an optimized k-means: the number of clusters (k) is determined at scene. We apply the heuristic elbow method [19] to automatically refine k to improve the quality of clustering. The elbow method plots an error measure (also called percentage of variance) against k. The error measure decreases as the number of clusters k increases; but starting with some k, the decrease suddenly flattens and the appropriate k is the one that hits this "elbow".

Centralized cluster values (i.e. the mean value for a given dimension of all data points in a cluster) are calculated. We show relative centralized values on the top highest weighted dimensions. For example in Figure 4a, cluster 2 ranks highest on the Swims dimension, suggesting that cluster 2 differentiates with other clusters because animals in this cluster tend to be good swimmers. With the dynamic clustering feature, the dimension chart is set to be sorted based on the weights. Therefore, users are able to identify cluster distributions on the most up-weighted dimensions that characterize the clustering.

Label switching (if clusters 1, 2, 3 change their color encoding from Figure 2.6a to Figure 2.6b affects users to track their cluster characteristics on the dimension chart. For the first time the

dynamic clustering is executed, colors are randomly assigned to the clusters. After that, when there is a update in the layout, clusters will appropriately restore the color encoding from the previous clustering. For example, from Figure 2.6a to Figure 2.6b, the German Shepard moves from cluster 1 to 3, the Skunk and Chimpanzee move away from their original clusters to form the cluster 4, and the Bat becomes the cluster 5. We see the current clusters 1, 2, 3 in Figure 2.6b preserve their original colors from Figure 2.6a. The German Shepard changes to blue as it merges into the cluster 3. Clusters 4 and 5 are assigned new colors.

We preserve colors from the previous clustering by comparing the centroids (centers) of current and previous clusters. Specifically, for each current cluster, we iterate its centroid over previous centroids, from which we find one located most closely to the current centroid. If more than two clusters share the same closest centroid, the cluster that appears closest to this previous centroid inherits its color. Figure 4c illustrates how colors are restored from Figure 2.6a to Figure 2.6b. In Figure 2.6c, blue triangles are the centroids of current clusters as mapped in Figure 4b while black triangles are centroids of previous clusters as mapped in Figure 2.6a. Both current cluster 3 (the blue triangle 3) and cluster 5 (blue triangle 5) have the previous centroid 3 (the black triangle 3) as their closet centroid. Because cluster 3 is closer than cluster 5 to the previous centroid 3, cluster 3 preserves the color from the previous cluster 3 while cluster 5 is assigned a new color.

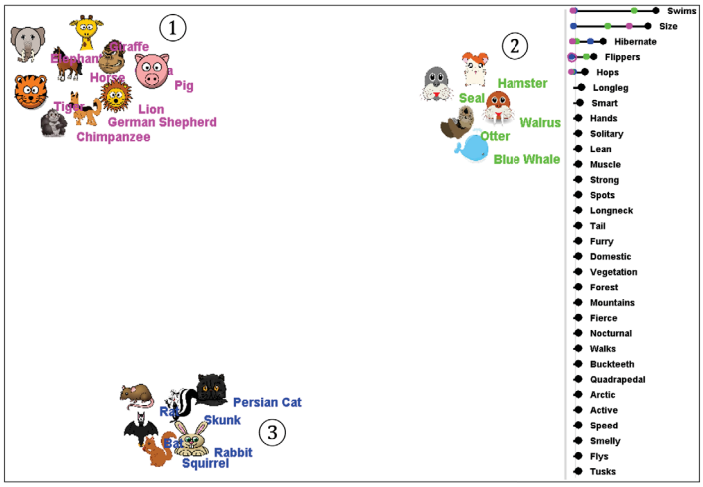
Dynamic clustering helps students reveal cluster distributions on important dimensions. It also provides an opportunity to verify themselves within and outside of a cluster. We strive for simplicity in our algorithms for linear algorithmic time complexity. Cluster detection is performed real time.

2.5 Data Exploration Using the System

We presented Be the Data in several STEM outreach workshops, as invited by various organizations, including the Center for Human-Computer Interaction, the Association for Women in Computing, the Center for the Enhancement of Engineering Diversity, and the Student Transition Engineering Program which is a summer orientation for incoming freshmen to the College of Engineering.

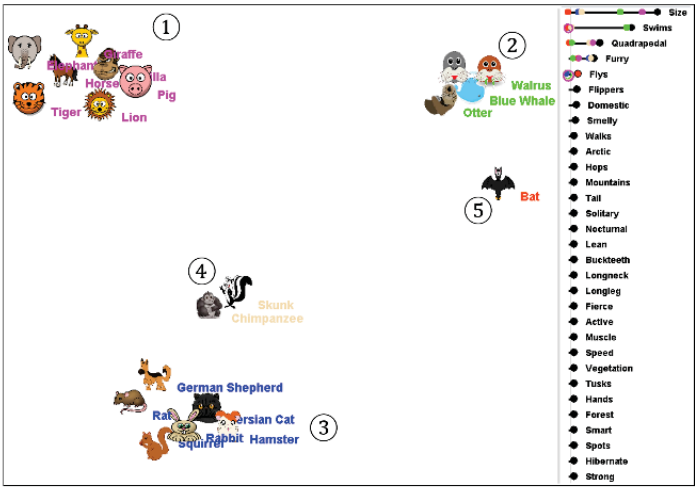
The goal of our workshop was to encourage and foster further interest in data-related disciplines. We reached over 100 students, ranging from 7th grade middle school, through pre-college, undergraduate, and graduate students. The majority of students participating in our workshops were new to high-dimensional data analysis. They had not learned the MDS algorithm before, with the exception of a few graduate students. We began with enabling students to explore high-dimensional data about animals. Each student embodied one animal in the Cube. Students worked collaboratively to explore the data with the system. A sub-group of students congregated in the space (clustering themselves) to discover virtual feedback about what made their data points similar to each other. Some students wandered away from others to identify what made her/him unique.

Through this bi-directional process of posing queries via proactive movement and understanding results through reactive movement, students understood numerous complex and latent relationships in the animal data. They collectively answered many questions about the data, such as “What make some animals good to eat?”, “What makes animals more attractive to humans?”, “What differentiates predators, prey, neither or both”, “How are vegetarians, carnivores, and omnivores different and similar?”.



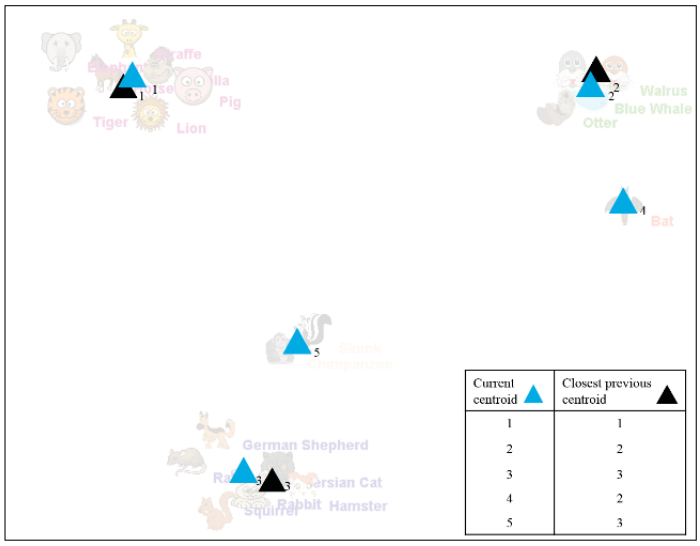
(a)

(a) Dynamic clustering of 2D points



(b)

(b) Cluster 1, 2, 3 preserve their colors while cluster 4, 5 are assigned new colors.



(c)

(c) Color preserved by comparing current cluster centroids to previous centroids.

Figure 2.6: Dynamic clustering illustration.

Chapter 3

Be the Data: Embodied Visual Analytics

3.1 Introduction

As today's data is becoming more and more complex, there is a clear need to advance research and education in knowledge discovery from large data. Educators are called upon to teach data analytical techniques to students. Current analytical methods rely on certain mathematical models to formalize understanding from data. For example, analysts normally use dimension reduction algorithms (e.g., multidimensional scaling, principal components analysis) to project high-dimensional data onto fewer (often two or three) dimensions to visualize relationships of data.

Yet, it is difficult to learn from high-dimensional data for students not experienced in data analytics. The students often do not have prior mathematical knowledge to understand the complexity of reducing dimensions from high to two orthogonal axes that have no explicit meanings [20, 21]. Even though a visualization is in a two dimensional space, it still uses all the original dimensions. The reduced dimensions may not correspond to actual dimensions of the data. This is a cognitively complex abstraction to grasp. The complexity of the subject may further prevent students from

learning it. Also, students who find statistical techniques difficult may confront non-cognitive factors such as an unenthusiastic attitude towards learning data analysis and try to avoid the subject [2, 22].

Learning from high-dimensional data requires comprehensive critical thinking skills that extend beyond the application of mathematical methods. As conceptualizing high-dimensional data is inherently subjective and uncertain [23], it requires formalizing alternative hypotheses, communicating personal judgement, exploring multiple solutions, and assessing implications of discoveries. Unfortunately, current approaches in teaching data analytics primarily focus on its quantitative aspects. Students do not experience comprehensive analytical processes and how complex quantitative summaries may advance knowledge. Even worse, data analytics are not approachable to students until they have mastered necessary quantitative theory and methods. Educators are called to have innovative and effective instructional methods to bring interest towards data analytics, make learning high-dimensional data analysis more approachable to the untrained population (low proficiency in mathematics or statistics), and provide hands-on experiences for students to engage in the analysis process [21, 22].

Therefore, we present a new approach to apply embodied interaction [24] to visual analytics for education and outreach. Specifically, we adopted an interactive system (Figure 2.1) [25] to teach abstract analytical concepts in understanding high-dimensional data. In this system, each student, in a physical space, represents a data point. The positions of students in this physical space represents a 2D projection of the high-dimensional data. Students can explore alternative projections by physically moving themselves, and hence the corresponding data points, in the space. They receive real-time visual feedback that explains their changes in positions. Students interact with each other in collaborative groups to explore interactions among data points. Therefore, they can pose hypotheses about the data and further explore and understand it.

The goal of this research is to explore how students exploit this novel approach to learn about data and analytical processes. We are inspired by embodied mathematical cognition perspectives informed by Lakoff [26]. He suggested that "*mathematics emerges through the interaction of the mind with the world*" and "*A large number of the most basic, as well as the most sophisticated, mathematical ideas are metaphorical in nature*" (p.364). Lakoff further emphasized that abstract conceptual knowledge is *"embodied"* and *"mapped within our sensory-motor system"* [27]. Human understanding of mathematical concepts is rooted in physical embodied interaction [26]. Since childhood, we are encouraged to play with physical objects such as building blocks to comprehend integers, use fingers to count, and employ solid beads to gain a symbolic understanding of "adding" and "deducting". Young students were able to solve fraction problems facilitated by manipulating physical pieces while they could not complete the same task by calculating on the paper [28]. The bodily experience is essential when students learn to reason and think abstractly, relying on more concrete metaphors, such as in-out, up-down, near-far [29]. We assert that the conceptual understanding of a complex analytical method is stymied by its traditional strict limitation to the virtual world. The comprehensive analysis processes are veiled behind small screen portals, simplistic interaction mechanics suggested by mouse and keyboard, and the single-user environment. Therefore, we merge interactive data visualization with a motion tracking system to evoke embodied learning in a co-located collaborative space.

This research contributes a new approach for data analytics by combining embodiment with interactive visualizations. It is an initiative for demonstrating the real uses of teaching students to learn high-dimensional data via embodied interaction. The key contributions are:

- the new form of embodiment applied in visual analytics to invoke embodied learning.
- the identification of students' analytical strategies that employ this form of embodiment.

- the qualitative/quantitative evidence of students improvement in understanding high-dimensional data.

3.2 Related Work

3.2.1 Immersive Interface

With the technological breakthrough in virtual reality and mixed reality, the immersive interface is developed as the combination of real and virtual that immerses users in real-time natural interactions, enhancing the sense of users actually being within it [30, 31]. By extending interaction beyond the traditional desktop and into the real environment, the immersive environment opens up tremendous opportunities to facilitate learning through embodied interaction: creating, manipulating, and sharing meaning through interacting in a physical and social environment [24].

3.2.2 Embodied Interaction

Human cognition is derived from interactions with our physical and social environments [32]. Computer scientists developed two research trends, tangible computing and social computing, to study the human-computer interaction in physical and social aspects, respectively. Tangible computing refers to the design that enables direct interaction with digital information via tangible objects in the physical environment [24]. It extends interaction beyond the traditional desktop and into the real environment. Social computing is concerned with incorporating the understanding of social context into the design of interactive systems [24]. Embodied interaction emerges two as a more recent research trend [24]. Paul Dourish [24] describes embodied interaction as “the creation, manipulation, and sharing of meaning through engaged interaction with artifacts”. He

further explained it as an attempt to integrate physical and social reality of our everyday world into computing. From his definition, it is clear that embodied interaction capitalizes upon both ideas of tangible computing and social computing.

Several examples of embodied interaction for improving user performance are known [6, 33, 34]. Ball and his colleagues [6] discussed advantages of physical navigation (moving eyes, head, body) over virtual navigation (zooming, panning, flying) in navigating a map on the large display. They found that with virtual navigation, users easily lost track of where they are in the map. In contrast, with physical navigation, users maintained spatial orientation without difficulties. This is because the perception of a step distance is an embodied resource that has been fully exploited by ourselves. Even the perception of step distance is beneath the awareness of an individual, it was naturally applied in maintaining spatial orientation when users moved back and forth from different parts of the map. It happened so quickly that users were not conscious of or notice it. However, users' perception of step distance is not applicable when they used zooming or panning. The results of study showed that physical navigation improved user performance and was more favorable than virtual navigation. In particular, if physical or virtual navigation could be used to complete the task, physical navigation was chosen all the time. As better performance with physical navigation has been confirmed by more studies [35–37], it is likely that a display allowing users' physical navigation and interaction, will be beneficial in large information spaces.

3.2.3 Physical 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 natural data exploration and collaboration [38]. Interactive surfaces enable direct interaction with

data which is more intuitive than using a mouse [3]. Physical navigation of data on a large high-resolution display with head tracking techniques outperforms virtual navigation [6]. Physical actions (e.g. movement) promote users' satisfactory experiences and they prefer physical navigation over virtual navigation [6]. Emerging stereoscopic 3D display technologies allow users to "walk through" or "fly over" the computer-generated scenes to manipulate digital information [7]. It not only gives users a visual experience, but an integrated multi-sensory experience. Attempts are also being made to physicalize virtual data. Digital data are now made accessible by manipulative artifacts whose physical attributes (e.g., size, shape, materials, motion) encode data and relationships of data are represented spatially [12].

Current studies in information visualization demonstrate many different physical embodied ways to interact with data. But all of them share the similarity that they place users in their data. We call this perspective "*Be In the Data*". Our work differs from existing approaches in taking the embodiment to even further. Instead of placing users in the data, users actually become data points. Therefore, we call this new facet of visual analytics "*Be the Data*".

"*Be the Data*" differs from "*Be In the Data*" in the perspective that the user takes during analysis. Rather than "looking into" the data points, users themselves become data points. It may give users a more egocentric perception to conjecture various relationships with other data points. It has the potential to render a concrete personalized engaging experience that could lead to deep insights.

3.2.4 Co-located Collaborative Visual Analytics

Co-located collaboration has been shown to be beneficial for complex visual analytics tasks [39–43]. In an exploratory study of supporting group collaboration around a shared tabletop display [42], researchers found that task success was highly correlated with time spent working collaboratively. Close collaborative teams more frequently shared their discoveries, reported more correct facts, required fewer hints, and gained a higher task score than loosely coupled teams.

Ownership and awareness of collaborators' actions is important to coordinate group efforts [44]. For example, in an collaborative analysis of numerous documents, individual work is marked with a unique user-dependent color. Therefore, users were able to distinguish documents that team members had read or found [42]. When collaborators' interests differ, users were able to open personal views on the shared territoriality and change their own copies of data [40, 44].

Leveraging spatial affordance of the co-located space, our work enables both ownership and awareness in collaborative visual analytics. Being a data point, each individual is responsible and has absolute control of the position of his/her own data point. Meanwhile, everyone is aware of others' positions as they are located in the same place and their positions are visualized on the shared large display. Students can only determine their own positions based on relative distances from others. They negotiate with each other and move freely in a coordinated manner. Exploring data naturally becomes cognitive and social interactions of students collaboratively organizing themselves to experiment on the alternative projections of the data.

3.2.5 Application of Embodied Interaction in Education

Embodied interaction has been increasingly applied in the education arena. Most of these studies focused on teaching and learning in abstract problem domains. Moso tangibles are a set of interactive physical artifacts for children to learn music abstract concepts: pitch, volume, and tempo [45]. With an ongoing tone, students could point a pitch artifact upward or downward to make the pitch higher or lower, squeeze a volume artifact wildly or quietly to make the volume louder softer, and shake a tempo artifact fast or slowly to make the tempo faster or slower. The study found that while not every child was able to verbally express their understanding of the targeted concepts, all of them could reproduce sound examples with required pitch, volume and tempo by interacting with the artifacts. It indicated that children can understand these abstract concepts in terms of their familiar movement related concepts (i.e., low/high, quietly/wildly, slow/fast). It also indicated that students were able to reason about these concepts using proper movement rather than words.

Embodied interaction is particularly applicable to mathematical education because the understanding of mathematical concepts is rooted in embodied interaction [26]. Since childhood, people are encouraged to play with building blocks to comprehend integers, use fingers to learn how to count, and slid beads on an abacus to gain a symbolic understanding of “addition” and “deduction” concepts. Martin and his colleagues [28] showed how children’s understanding of fraction concepts was facilitated by embodied interactions. In their study, children worked individually either with physical manipulative materials or pictorial materials to answer fraction problems, for example, show $\frac{1}{4}$ of 12. With physical materials, a child was given 12 pieces of small objects. With pictorial materials, a child was given a picture of 12 pieces, a paper and a pencil. Results showed that children tried more strategies, created more partitioned adaptions, solved more fraction problems, and gave more accurate interpretations with the manipulative materials than with the pictorial materials. Most importantly, the same child who could not complete a fraction problem by calculating

with the picture and pencil was able to solve the similar problem when s/he manipulated physical pieces. Rearranging physical objects into groups is a familiar action for children. In the physical condition, children constantly unitized and portioned objects into groups. Groups with the same size is intuitive for children and therefore they equally subdivided the pieces into a number of groups and that number was the denominator of the given fraction. By taking advantages of embodied resources (e.g., spatial organization), physical manipulation facilitated children to overcome their prior misapplied interpretation: choosing the whole number in the fraction number (e.g., 1 and/or 4) to answer the problem (e.g., $\frac{1}{4}$ of 12).

Howison and his colleagues [46] studied how to apply body movement to help young students understand proportional equivalence (e.g., $\frac{2}{3} = \frac{4}{6}$). In their study, students were asked to position two arms above a desk so that the distance between the arms and the desk could make and maintain a certain proportion. For example, to maintain a 1/2 ration, for every 1 units of distance on the left arm, it is 2 units of distance on the right arm (e.g., $\frac{1}{2} = \frac{2}{4} = \frac{3}{6}$). With a natural perception of relative distance between two arms and desk, all the participants succeeded in getting correct proportions. While arm-movement related interaction has not yet been exploited in teaching and learning, this study provides empirical evidence that it is able to evoke basic arithmetic operations to understand proportional equivalence.

Embodiment is important to cognitive activities, as our mind and body are deeply integrated. There exists an action-mind compatibility effect between people's physical and mental states [32, 47]. For example, reading the word "lick" stimulates brain parts that control the motor activity of the mouth while reading the word "pick" stimulates brain parts that manage the hand [48]. When embodiment and cognition is redundant, cognitive process is speeded up. For example, people' response time is shorter when they were asked to indicate liking by pulling the object towards them than pushing the object away from them [47]. Users unconsciously applied bodily experience (e.g., distance perception, gesturing, spatial relationship, social discourse) to facilitate their cognitive process [6,

28, 45, 46, 49]. The physical concepts, such as up-down, near-far, slow-fast were used as embodied metaphors in these studies. Embodiment provides learners with a concrete physical medium to reason about the targeted abstract concepts. After the concepts become grounded in embodied action, learners are able to apply the new interpretation to re-purpose many environments to solve similar problems in other contexts [28].

3.2.6 Teach High Dimensional Data using Visualization

Teaching high-dimensional data to non-statistician students is a challenge. A traditional problem is the lack of interest and mathematical background among these students [2, 22]. Interactive visualizations have been used for easy consumption as they allow students to learn and apply data analytical techniques in a visually interactive manner. Andromeda [50] is such a desktop application. It was developed to help learners understand dimension reduction techniques by encoding similarity with spatial proximity, and thus learners are not required to understand the underlying dimension reduction algorithm to conduct multivariate analysis. With Andromeda, learners communicate the judgment that data points are similar by pushing them closer and data points are different by pulling them further apart. In turn, Andromeda runs the inversed dimension reduction algorithm to calculate a set of variables that describe users' judgment and provides visual feedback. Andromeda shields learners from the complexity of mathematical models, so that they are able to explore various projections of high-dimensional data based on their conjectures of relationships of the data.

Our work is an extension of Andromeda [50] for visual analytics beyond the desktop [51]. The desktop-based application has several educational limitations for learners who have not experienced data analytics yet. It is difficult to imagine that a novice learner would engage in moving data points and tuning parameters on a screen. Analyzing data would easily turn into a tedious

task. It is challenging for learners to conceptualize an abstract mapping from the virtual data to the virtual visualizations through simplistic interaction mechanics suggested by a mouse and a keyboard [27].

Learning new mathematical concepts is metaphorically structured with physical interaction [52] and fully embedded in body actions [26]. However, we see few applications of visual analytics that use other technology than standard desktop/laptop computers. In our everyday life, we are heavily accustomed to processing information integrated from vision, hearing, touching, moving, and etc. It is desirable to exploit the same approach to interact with data [51]. Here, we took a more proactive approach to employ our own bodies as movable data points so that we are able to map data analysis to an integrated sensory/motor activity, not just to visual.

3.3 Be the Data

We developed the *Be the Data* system [25] to explore how students exploit embodied interaction to learn from high-dimensional data.

3.3.1 System Overview

Be the Data exploits a unique new physical space, the multi-media Cube which is comprised of advanced interactive technologies for physical-virtual cross-overs. Our system includes a large overhead display, a motion tracking system, and the backbone software adapted from Andromeda [16] for direct manipulation of virtual high-dimensional data models (Figure 2.1).

To use the system, students enter the Cube and embody virtual data points by wearing a trackable hat so that their positions will be detected. Students are able to manipulate the layout of data

points by walking around in the Cube. We have a large display above head where visualizations are displayed to show the representations of students' movements. For example, if we consider a high-dimensional dataset about animals (Table 2.1), each student represents an animal data point and his/her position in the Cube is visually reflected on the display (Figure 2.1).

The underlying algorithm of *Be the Data* relies on Weighted Multi-Dimensional Scaling (WMDS) [53] to map three or more dimensions to two dimensions. WMDS visually plots the data in 2D Euclidean space to represent the data spread in the high-dimensional space. *Be the Data* takes advantage of the inverse WMDS algorithm [18] so that we can map the layout changes to the value adjustment of weights. Since the algorithm considers relative distances among data points, the actual size of the Cube does not impact its performance. Students adjust two dimensional coordinates by rearranging themselves in the Cube. In turn, they are provided real time feedback (i.e., a new set of weights) that best describe their current layout. When students move several times to adjust the projection, they are effectively exploring the same dataset from multiple perspectives (Figure 3.2, Figure 3.3).

3.3.2 User Interface to Support Embodied Interaction

There are two important types of user interface in the *Be the Data* system: physical and virtual. The former includes the Cube, the tracking system, and the large display as mentioned earlier. The latter highlights interactive visualizations. Different from traditional input devices (e.g., the mouse and keyboard), students interact with the system by walking in the Cube. As discussed before, such physical movement in the Cube is captured in real time via the tracking system, and then transformed as coordinates, which work as the input for the inverse WMDS and dynamic clustering algorithms. The virtual layout and the calculated weights are displayed as interactive visualizations on the large display.

Interactive WMDS Visualization

The interactive WMDS visualization includes two essential parts: a WMDS plot and a dimension chart, organized left and right respectively on the large display (Figure 3.2). The WMDS plot reflects the current physical layout in the Cube (from a bird's eye view). The dimension chart lists the dimensions in alphabetical order and reveals their current relative weights.

We use the near-far metaphor to interpret WMDS plot; near data points suggests that they are relatively similar and far suggests that they are relatively different in the variables that are emphasized (i.e., variables that are weighted more). The default layout is set by considering all the dimensions equally weighted. As students walk around in the Cube, the plot updates simultaneously. The underlying WMDS algorithm resolves the parameters (i.e., variable weights) to meet user-generated layouts. For example, in the left image of Figure 3.2c, the *cat*, *hamster*, and *rabbit* move closer to each other and separate from the rest , indicating that students are considering these three animals more similar than other animals in some way. After calculation, the system shows in the weight chart that they are more similar because of the emphasized variables: *Buckteeth*, *Domestic*, *Hops*, and *Solitary*.

In line with common situations in real life, both parts of visualization do not show actual values (e.g., coordinates and weights values). People have the sense of relative distance even without knowing exact values [54]. We expect that students do not need to know their exact coordinates to generate the visualization. Instead, they simply need to make an estimation of the spatial relationships as relative to others.

Dynamic Clustering

Although the WMDS visualization reveals dimension weight changes that characterize students' choice of locations, it does not tell any information about the attributes of students' generated clusters. Therefore, we implemented dynamic clustering to visualize cluster features on top of the WMDS plot (Figure 3.2b).

Clusters are determined real time by an optimized method of *k-means* [25] based on the projected coordinates in the two-dimensional view space. Each cluster is assigned a unique color. Its centralized values (i.e., the average value of all data points in that cluster) are shown on the top up-weighted dimensions. For example, on the right image of Figure 3.2b, the magenta cluster ranks high on the Buckteeth dimension, the grey, magenta, and yellow clusters rank high on the Domestic dimension. Once the dynamic clustering feature is enabled, the dimension chart is sorted based on their weights. Therefore, learners are able to quickly identify features that differentiate the clusters.

3.4 Evaluation

We conducted educational workshops to explore how novice students (i.e., students without a mathematical background in WMDS algorithm) employ embodied interaction to learn and analyze high-dimensional data. Specifically, we seek to answer the following questions:

1. Did students learn key concepts about high-dimensional data?
2. How did students exploit *Be the Data* to learn about data and data analytics processes?

Workshop	Grade Level	Number of Participants
ICAT	2nd grade	more than 100
AWC	6th or 7th grade	62
CEED	10th or 11th grade	50
STEP	pre-college	33
DA	undergraduate	49

Table 3.1: Participants information.

3.4.1 Participants

We recruited more than 250 participants at various ages in several STEM outreach activities held at our institution (Table 3.1). We recruited more than 100 2nd grade participants at the ICAT (Intitute for Creativity, Arts, and Technology) Day, 62 6th or 7th grade participants at the Association for Women in Computing (AWC) workshop, 50 10th or 11th grade participants at Center for the Enhancement of Engineering Diversity (CEED) workshops, 33 participants entering the college in the Student Transition Engineering Program (STEP) workshop, and 49 undergraduate participants in a data analytic (DA) introduction workshop, and

Participants were new to the MDS algorithm. For AWC participants, neither the 2nd nor 6th nor 7th grade curriculum had covered WMDS related concepts before. For CEED, STEP and DA participants, we asked their familiarity with MDS. In 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 student checked "expert on it."

3.4.2 Procedure

In groups of 20-30, students were asked to analyze a high-dimensional dataset of about 20-30 animals with 31 dimensions (Table 2.1) using our system. Each dimension reflects the degree (on

a scale of 0-100) to which animals could be described by a characteristic (e.g, Skunks were rated 100 in the dimension Smelly, whereas Horses were rated 33).

The workshop started with a short introduction on high-dimensional data. The instructor explained what is high-dimensional data using the animal data shown in a table (Table 2.1, Appendix B), and identified dimensions as columns in the table. Then, the instructor explained how to use the system. Specifically, she explained that the visualization on the screen is a 2D projection of the given high-dimensional data and positions of animal icons on the visualization represents 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 near-far metaphor to interpret the WMDS visualization with up-weighted dimensions. The dynamic clustering feature was implemented later and was only available for the DA workshop.

After the introduction, students performed group based analytical tasks to learn from the data. Students were allowed to use the system to answer any animal data related analytical question that interested them. When students had difficulty coming up with a question, the instructor gave them a question to solve. We gave students the freedom to learn from data by solving their own questions. We expected open-ended tasks would inspire students to derive insights from a particular perspective that engages them.

Before the workshop started, students completed a pre-survey. After the workshop, they completed a post-survey. Due to workshop schedules, ICAT participants did not take pre-survey or post-survey.

3.4.3 Data Collection and Analysis

To answer the research questions, we collected qualitative and quantitative data from recorded video, pre-surveys, and post-surveys. The video recordings kept 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 learning data. Also, differences in pre and post survey answers are 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-surveys. 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 take a Bayesian approach (Appendix A). There has been some debate over the use of classical methods which rely on p-values to make inference [55]. P-values can easily be miscalculated and/or misunderstood when comparing them to a type I error, i.e., a significance level α. Thus, for this paper, we prefer to use Bayesian models to analyze the quantitative data. In many (if not all) cases, a classical approach would result in the same, final inferences as those that we report in Section 5. Thus, to support those unfamiliar with Bayes, we report p-values when applicable.

The quantitative data from the surveys are 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 and wrong answers per question and survey with Beta-Binomial Bayesian models [56]. From these models, we learn the posterior distribution for the probability of a correct answer and then 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 a variable" (for the STEP workshop). The *maximum a posteriori* (MAP) difference is 0.02 with a credible interval (-0.24, -.024) (row 4 in Table 3.2). This means that the most probable difference between the pre and post probabilities of being correct is merely 0.02. And, because the credible interval overlaps 0, we can infer that students' understanding did not change before and after 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.

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 Normal model with reference priors [56] to analyze responses per question and 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" is -0.75 with a credible interval (-0.85, -0.64). This means that most probable difference in mean is -0.75 and 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 and recorded videos, we had two researchers encode the data independently and compared their codes to make interpretations.

3.5 Results

3.5.1 Question 1: Did students learn key concepts about high-dimensional data

Students learned key concepts: variable, relative distance, dimension reduction, data exploration

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 of them 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 3.2. It is suggested that students gained knowledge about following concepts: variable, relative distance, dimension reduction, and data exploration. It is also suggested that students increased interests and confidence in learning high-dimensional data.

To assess whether students were able to identify variables in high-dimensional data, they were asked to identify a variable in a spreadsheet containing a high-dimensional dataset. In the AWC workshop, 0.42 proportion of the students answered correctly on the post-survey. In the STEP workshop, there was no improvement. In the CEED workshop, there was significant evidence of improvement. In the DA workshop, there was some evidence of improvement (credible interval (0.00, 0.35)). We hypothesize that these results are relatively low in comparison to the other results because students understood variables used in the plots, but some were confused about variables in a spreadsheet. It is reasonable because students conducted all their activities by interacting with the visualizations. They were not exposed to a spreadsheet to analyze high-dimensional data.

Workshop	Question	Pre Correctness	Post Correctness	Expected Difference	Credible Interval	p-value
Key concept: variable						
AWC	identify a variable	—	0.42	—	—	—
STEP	identify a variable	0.36	0.36	0.02	(-0.24, 0.24)	0.9241
CEED	identify a variable	0.45	0.79	0.33	(0.15, 0.51)	<0.001*
DA	identify a variable	0.61	0.79	0.17	(0.00, 0.35)	0.0541
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, -0.0.09)	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 3.2: A quantitative summary of students' understanding about the key concepts (i.e., variable, relative distance, dimension reduction, data exploration), interests, and confidence towards learning high-dimensional data before and after the workshop. DR stands for dimension reduction. Column 3 and 4 are observed proportion of correct answers for the pre and post surveys. Column 5 and 6 are the expected difference and the credible interval for the difference in proportions. Column 7 are the p-values from a two-tailed two sample t-test. The * in column 6 and 7 flags questions when there is important difference in pre and post.

To assess students' understanding of relative distance in the visualization, they were asked to interpret similarities of data points in a dimension reduction visualization based on their relative distances. In all the workshops, students were asked questions on a dimension reduction (DR) plot with all weights equal. In the AWC workshop, 0.92 proportion 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, there was no improvement due to the high correct proportion (0.96) in the pre-survey. The relatively high correct percentages in all workshops indicate that the "near is similar" metaphor is an intuitive idea that can be exploited for usable data analytics.

In the CEED and DA workshops, we asked three additional questions to probe their understanding of relative distance in more details. First, students were asked to interpret similarities of data points on a dimension reduction plot with weights not equal. There was no significant improvement due to the high correct proportion (0.96 in CEED, 0.94 in DA) in the pre-survey. Second, they were asked to explain if and why their answer changed between the equal weight plot to the not equal weight plot. In qualitatively analyzing 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). Eight students who did not answer the question or wrote *"I don't know why"* in the pre-survey, answered correctly in the post survey with

- *"the weights changed"* or
- *"the smelly and walks variables played a greater part"*.

Sixteen students who answered incorrectly in the pre-survey because they only restated the distance changes on the graph (e.g., *"the points appear closer on this plot"*, *"they are now more closely together"*) answered correctly in the post survey by considering weighted variables:

- *"it changed because we are comparing the [points] based on different weighted categories than before. Before, all categories were equal, and now they have weighted values"* or

- *"the weight of the variables changed, maybe because of a different research question that was asked*)".

Third, they were asked what needs to change if they want one particular data point (e.g., Seal) to be closer with a cluster (e.g., blue whale and otter). In qualitatively analyzing 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 or wrote *"not sure"* in the pre-survey answered correctly in the post survey:

- *"much higher weight on 'swims'"*
- *"strength and swims variable are weighted heavier"*.

Seven students who incorrectly it wrong 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 student 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"*).

To assess students' knowledge about dimension reduction plots in comparison to standard 2D and 3D scatterplots, they were asked which dimensions are used to plot data in a given 2D scatterplot, 3D scatterplot and dimension reduction 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, 0.57, 0.50, and 0.78 proportions of students, respectively, answered these questions correctly afterwards. Students in the STEP workshop showed a significant improvement in understanding the dimension reduction plot (credible interval (0.22, 0.67)). Before the instruction, the students understood the 2D scatterplot and 3D scatterplot, but not the dimension reduction plot. After the instruction, they understood the dimension reduction

plot approximately equally as well as the other scatter plots. In the DA and CEED workshops, we only asked the same question on a dimension reduction plot. There was strong evidence of improvement (credible interval (0.12, 0.42) in CEED, (0.09, 0.36) in DA).

To assess students' understanding of the exploratory nature of dimension reduction plots, 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, which plots are right and which are wrong. This question was not asked in the pre-survey because the question was meaningless before students had the opportunity to explore the data. In qualitatively analyzing their written responses, we found that 0.86, 0.91, and 0.90 proportions of students in the STEP, CEED, and DA workshops understood that many different plots could be created to investigate different questions. Students answered that,

- *"there are no wrong plots"* or
- *"depending on the research question asked, different variables are taken into consideration more heavily than others"* or
- *"there are different interpretations of the same data"*.

Two students further elaborated that visualization helps humans see the data in an easier way because we tend to focus on only a few variables when there are often many more in play.

Students increased interests and confidence in learning high-dimensional data

In the CEED and DA workshops, we studied students' attitudes in learning 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, expected difference, credible intervals, and p-values are shown in Table 3.2. For the statement that "Analyzing data is boring", the Bayesian approach suggests

that there is notable difference (credible interval (-2.96, -1.21) in CEED, (-1.48, -0.22) in DA) while the classical approach does not ($p = 0.0519$ in DA, slightly above 0.05). However, if we were to consider a significance level at 0.10, both methods would agree on the presence of a significant difference that students disagreed the statement more after taking the workshop. We will investigate this question further in future work.

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

Results from the question that "I know what is meant by the term, high-dimensional data" showed strong evidence that students perceived that they knew high-dimensional data better after taking 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, and even expanded the use of *Be the Data* to a real life case:

- *"we use multidimensional data all the time without realizing it"* or
- *"anybody is capable of using analyzing high-dimensional data and applies this more often than one would initially think"* or
- *"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 also mentioned other high-dimensional data (e.g., sports, countries) that could be analyzed like this.

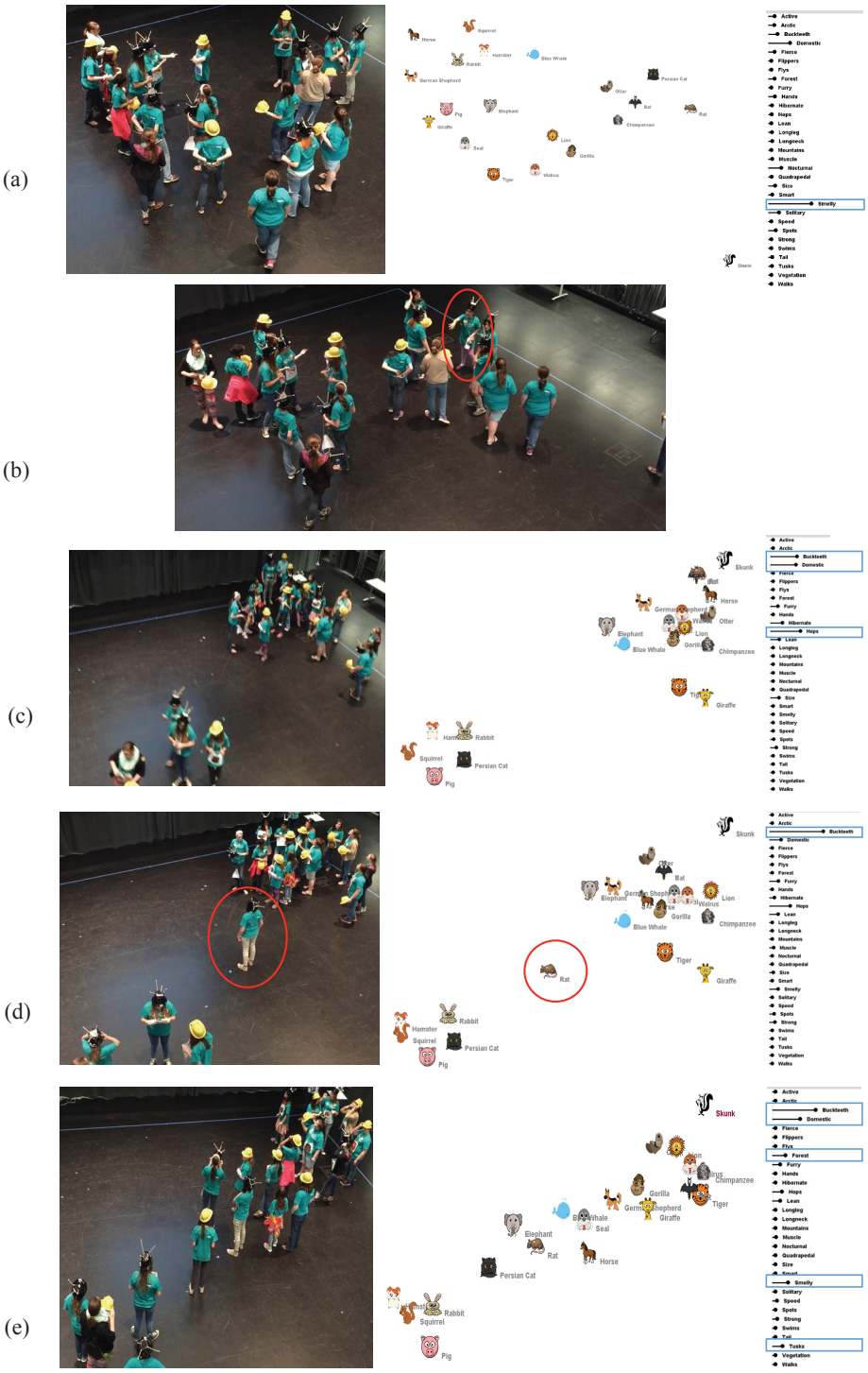
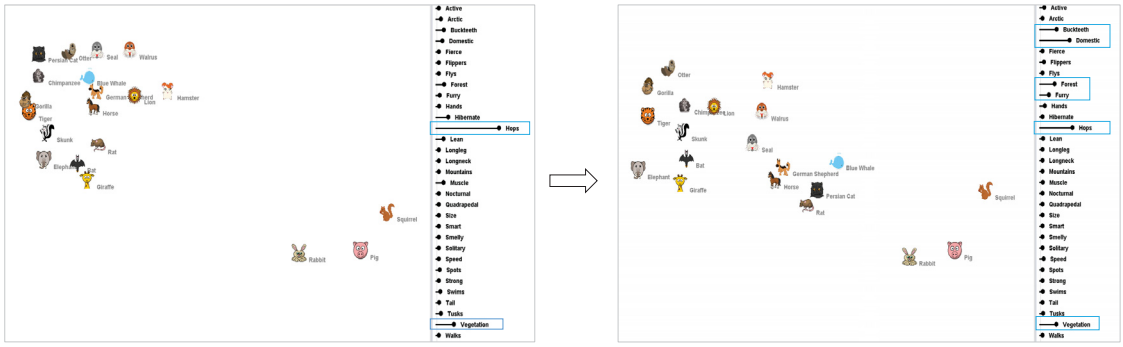
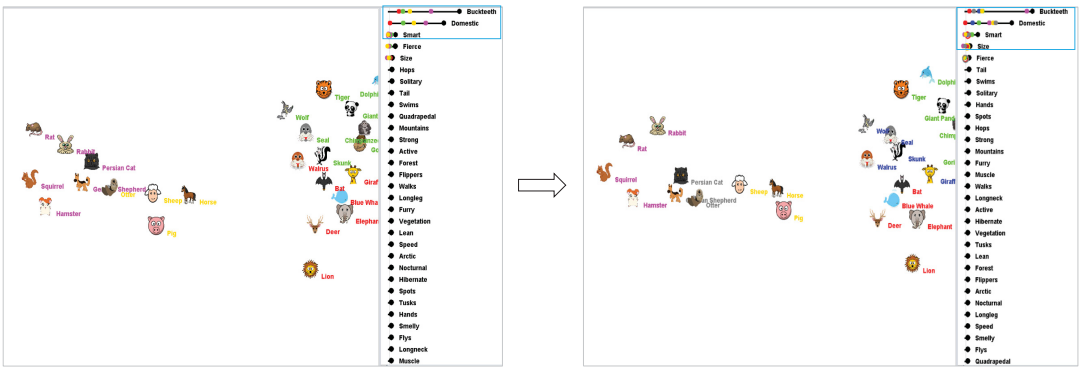


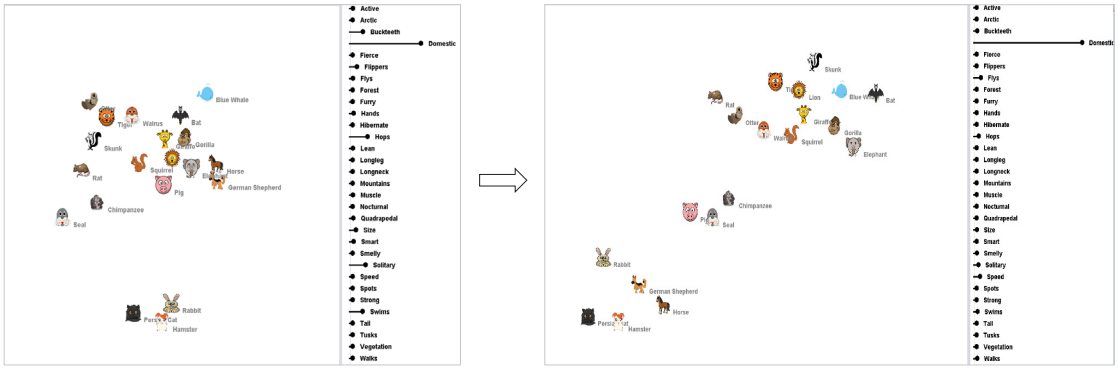
Figure 3.1: Students exploit embodiment to solve multivariate problems (a) Students explored the first hypothesis where skunk was an outlier. The student who embodied skunk was in the lower-right corner (not shown)). (b) Students discussed the alternative positioning in large group collaboration and one student took the leadership. (c) Students moved to binary groups to explore the second hypothesis. (d) Conflicts arose as a student raised different opinions. (e) Students worked in small group collaboration to adjust their positions. They made the collective decision for the third hypothesis.



(a) Students' analysis at two stages to answer "what make some animals good to eat"

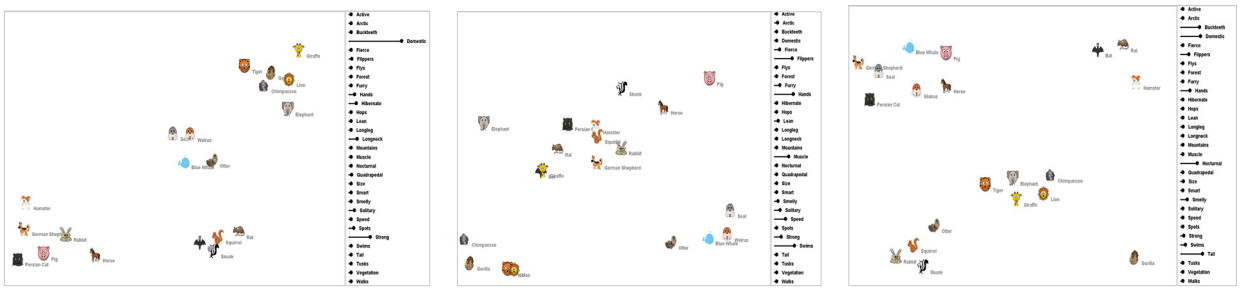


(b) Students' analysis at two stages to answer "what make animals a good pet"

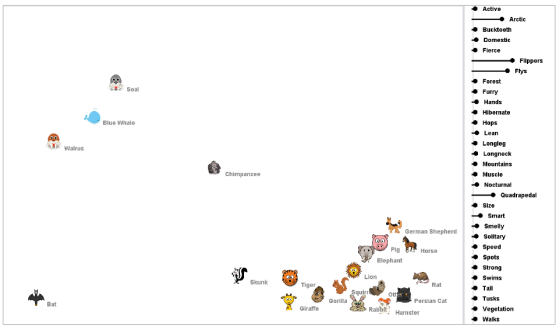


(c) Students' analysis at two stages to answer "what differ wild and domestic animals"

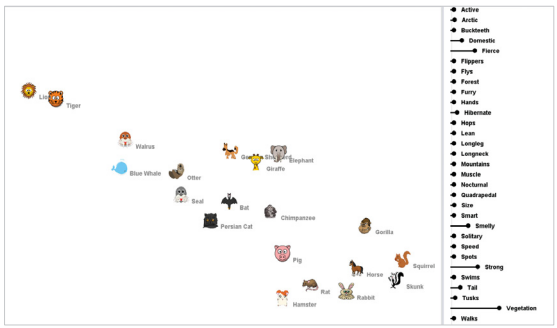
Figure 3.2: Students' analysis at two stages for a given question.



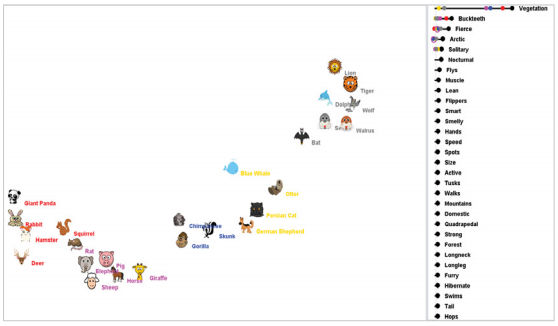
(a) Three groups' three answers to the question "what explains where animals live"



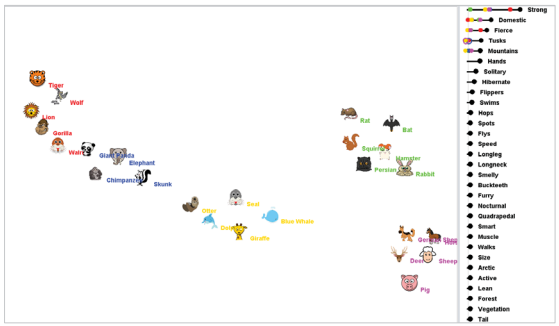
(b) One group answered "what differentiate four-leg animals with others"



(c) One group answered "what differentiate predators, prey, neither or both"



(d) One group answered "how vegetarians, carnivores, and omnivores differ"



(e) One group answered "what animal you want to encounter on an island"

Figure 3.3: Students generated various inferences and visualizations from the same dataset.

3.5.2 Question 2: How did students exploit *Be the Data* to learn

An example case to illustrate how students exploit embodiment

Students exploited embodied interaction to collaboratively solve multivariate problems, as evidenced by their behaviors during the course of their analytic process. We demonstrate these behaviors with one group who answered the question "What make some animals good to eat". Their

process is shown in (Figure 3.1).

In the initial exploratory phase of their process, one student who represented the "Skunk" immediately separated herself as being definitely not edible. This led to the identification of "Smelly" as an important variable in edibleness (Figure 3.1a), because the skunk is an outlier in that dimension with an extreme value of 100, while the other animals range from 1 to 51 in Smelly. Thus, the system up-weighted the single dimension "smelly". However, students were not satisfied with this result.

In the next phase, each student began discussing with their neighbors in the room whether their animal was edible, based on their external prior knowledge about their own animals. This represents students taking ownership of their data point. Then they gathered together in one group to discuss alternative positioning. Here, some students gradually took a more dominant role in directing the movement of others. For example, one student (as noted in Figure 3.1b) spoke up to direct the crowd and pointed with her hand, *"So less edible animals move here, more edible animals move there"*. Instead of focusing on one dimension, students' attention was directed to the embodied objects, the animals. They moved into two clusters: non-edible animals in the upper right corner and edible animals in the lower left corner (Figure 3.1c). This produced the up-weighting of dimensions Buckteeth, Domestic, and Hops. From Figure 3.1a to Figure 3.1c, dimensionality of their hypothesis was increased from one to three. However, some students still argued with this simple binary classification.

Next, a student (as noted in Figure 3.1d) who embodied the rat did not feel that she belonged to either of the groups. She explained that although rat was normally not good to eat, it might be a dish in certain food cultures. She then stepped out from her original cluster to identify what made her unique: Buckteeth. This idea, conflicting with the previous hypothesis, caused other students to also refine their positions in the room. The non-edible group negotiated to spread themselves out

more, and they discussed in small groups with their neighbors to refine their positions (Figure 3.1e). Also the girl representing the Persian Cat in the edible group decided on her own that she should not be as edible as the rabbit or the pig. This produced a layout representing more of a spectrum of edibility, rather than a binary edibleness. The system then revealed that Buckteeth, Domestic, Forest, Smelly, and Tusks were significantly up-weighted as compared to others in order to produce this layout.

In this scenario, the visualization evolved as students continuously interpreted and changed their judgments of data to test different hypotheses. The gradual transition in layouts from one outlier (Figure 3.1a), to a binary group (Figure 3.1c), to a spread spectrum (Figure 3.1e) clearly demonstrated progression of the students' thinking to experiment alternative solutions and consider more associated dimensions.

We believe that students exploited embodied interaction in conducting their analysis. In addition the example case shown above, we observed that students were engaged in embodied interaction (as summarized in Table 3.3) to conduct data analysis. Students took the ownership of their data point through embodiment. It raised students' personal responsibility for the animal they embody, which might force them to think in details and demonstrate individual initiative. Yet, all of students managed to arrive at an agreement in a collaborative effort. The embodied interaction elicited discussions for students to collaboratively organize themselves, from which some students took the leadership to direct the movement of others. Other students came to an agreement and implemented the agreement to find their positions. During the process of re-organizing, students might move and interact with different groups of people as apposed to sit and interact with a same group people sit close in a classroom. It was evident that students took advantage of the physical space to organize data points. Without knowing the exact coordinates, students could still rely on the sense of spatial relationship in a physical space to interpret similarity and difference [54]. It was intuitive to see their movement on the visualization and became aware of others' positions. While moving,

students looked at each other and employed natural social interactions (pointing, voice, gaze, facial expressions, etc.) to coordinate their movements as relative to others. Once they found a spot, they looked up to the screen to evaluate their positions in the group, gain feedback from the system about dimensions, and determine the next step. The pointing behaviors, accompanied by instructions directing physical locations appeared to indicate that students exploited their spatial capabilities in physical positions to aid their understanding.

Students generated various interpretations of the data

Students were able to interact with data to solve many different multivariate problems from the same data (Figure 3.2, Figure D3.3). They produced more than 20 visualizations from which we observed four typical structures, including outliers (e.g., Figure 3.3b), binary groups supported by a boolean logic (e.g., Figure 3.2a_left), multiple clusters (e.g., Figure 3.3a_left), and linear spectra (e.g., Figure 3.3d). One visualization may include one or more structures to demonstrate the relationships of data points (e.g., Figure 3.2c_right has multiple clusters spread on a linear spectra).

With the aid of instantaneous feedback from the system and collaborators, students continuously reflected on their assumptions and changed their judgments of data. In addition to the example illustrated in Figure 3.1, we observed that such adjustments occurred across groups and questions. In Figure 3.2, each row represents students' analysis for a given question at two stages in their analytical process, an early stage on the left and a later stage on the right. In these examples, students split their aggregations into a more spread structure or smaller subgroups. Figure 3.2c asked "what make wild animals and domestic animals different" and "Domestic" happened to be one of the dimensions. After the adjustment, students answered this question more correctly because "Domestic" was up-weighted more while other dimensions were down-weighted more.

Embodied Interaction
Students took ownership of the data point they embodied. • Students became a data point to think about the problem. *"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."* • Students had the authority to determine their own position. *"I don't think I [gorilla] am a good pet."* *"I (rat) may not belong to the non-edible group [and wandered away from the original cluster]."* *"I am pretty sure I [chimpanzee] am a omnivore."* • Students thought in details about their own animal to refine positions with their neighbors. • Students demonstrated individual initiatives (e.g., the rat in Figure 3.1).
Some students took leadership to organize themselves. *"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 need to move that way."*
Students collaborated in large and small group discussions • Students moved around to discuss with different groups of people • Large group collaboration spontaneously occurred to reorganize themselves in the space. • Small group collaboration spontaneously occurred to refine their positions with close neighbors.
Students made use of the physical space. • Students exploited their spatial capabilities. *"They gotta be in that corner."* *"Cat come over here."* • Students employed natural social communications.

Table 3.3: A summary of students' embodied interaction during the activity.

The progression of students' analysis to interpret the same question suggested that they were able to form alternative hypotheses and derived different variables associated with the same topic.

We found in three cases that students progressively considered more dimensions of the data during the course of their exploratory analysis process (Figure 3.1, Figure 3.2a, Figure 3.2b). In Figure 3.1, the gradual transition of the visualizations from one outlier (Figure 3.1a) to binary groups (Figure 3.1c) to a linear spectrum (Figure 3.1e) demonstrated that dimensions being considered (shown as the blue highlighted dimensions) increased from one to three to five. In Figure 3.2a and Figure 3.2b, in the early stages students tended to focus two dimensions in the data. In the later stages they had expanded their focus to four or five dimensions of the data. The increase in considered dimensions represents more comprehensive interpretation of the data by the students, thus more complex insights gained. We only observed increased dimensions in the questions that seemed more subjective than the other questions. For example, the interpretation of whether an animal is good to eat or a good pet is subjective while the interpretation of whether an animal is domestic is objective.

Moreover, different groups derived different insights from the data. For example, three groups answered the same question "what explains where animals live" in different ways (Figure 3.3a); Four groups asked and answered different questions to interpret the same data from various perspectives (Figure 3.3b-e). Students in separate groups possessed varying sources of expertise and experiences that guided data exploration distinctly from others. It coincided with the exploratory nature of high-dimensional data that different interpretations can be right in many ways.

3.5.3 Usability issues

Be the Data system was easy to use. It only took about 3-5 minutes for students to learn to use it. Students soon found ways to engage each other through the system (e.g, attempting to move away from the group or swing back and forth to see how it affects). Unfortunately, the size of the hats did not fit all the participants. In some circumstances, the hat would fall off. A few users had to hold the hat with their free hands, and when tired of holding them, would just take them off.

Be the Data system provided a pleasant and engaged user experience. Students liked 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", "interesting" [or "fun", "engaged", "cool"] were the main factors participants liked, each being mentioned 53, 68, 69 times respectively. Students described *Be the Data* as

- *"extremely interesting"*
- *"way more fun than class"*
- *"beats going to class and sitting through a lecture"*
- *"not having to sit at a desk"*
- *"being more interactive in class"*
- *"engaged in the activity (as opposed to a classic lecture)"*.

Therefore, students felt that

- *"data can be fun, not always tedious and boring"*.

We observed that students (especially younger students) excelled when they were provided opportunities to move. In addition, students mentioned 26 times that they liked "collaboration" and "teamwork". One student appreciated the opportunity for knowing more my classmates that she has never spoken to.

We observed that when students found desired results collaboratively, they whistled, gave "high-fives" or made other exclamations. We also observed active collaborative participation where students determined their own location, made adjustment with neighbors, helped others to locate a position, and contributed to the group solutions in a coordinated effort. A middle school teacher observed his students' activities and commented that

- *"I have never seen my students being so engaged"*.

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 is the] most unique data organization tool I've seen"*
- *"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 doing multi-dimensional analysis and the difficulties of describing and quantifying data details. One student disliked the workshop.

3.6 Discussion

It was our goal to explore this new form of embodiment to learn about data and analytical processes. 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 offers an intuitive medium for students to reason about abstract data. Students moved around and they got immediate visual feedback of dimension changes. They built intuitions about relationships between their movement and the dimension changes. Feedbacks coming from more than one sense force the brain to engage in a different way. They were able to experiment and test hypotheses to increase their understanding of the data. Gigerenzer describes such learning as "Gut Feelings" [57]. 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 the concepts and were able to draw multivariate insights that were not stemmed from the algorithm formula. It might lead to a better understanding later on when they are confronted with the mathematical underpinnings.

Be the Data allows cognition and computation to work together, interacting through embodied interaction and visualization. The embodied near-far metaphor is familiar and matches the conceptualization of the underlying mathematical model. Children as young as 6th or 7th grade identified with such mappings. By walking around and having social communication, students injected their semantic expert knowledge into the visualization and interpreted the calculations. The real-time feedback parallels students' mental processing of the relationships in data and allows them to continuously reconcile changes in visualizations to explore many aspects of the data. *Be the Data* exploits students' spatial awareness and capabilities in the real world [58, 59] that may benefit a convincing setting for interpreting the spatial organization of data. It is an attempt to unify physical world experience and computational experience in a natural way with the new form of embodiment.

Be the Data encourages ownership and awareness which are important for collaborative work [41]. The co-located space and interactive visualization allows students to see all the data points and coordinate their movements. It also makes individual's input salient so that the consequences of

everyone's decision is visible and counts for the group result. The embodiment gives individual responsibility for the data point he embodies. It may decrease students' inclinations to obey authority, encourage individual initiatives, and promote conversations among collaborators.

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 an hour. But with *Be the Data*, students were engaged in data exploration throughout the one-hour workshop. The complexity of the data model often scares students away from learning it [22], but students may enjoy learning data that do not require much mathematics [2, 60]. *Be the Data* shields users from the technicalities of mathematical models so that they may focus on exploring the data by directly manipulating the visualization based on their domain-specific questions and interpretations. The interactive visualization explicitly documents students' thinking processes in analyzing data. Results from our user study suggests that *Be the Data* made the complex analytical method approachable to novices, made the data analysis tasks appealing, sparked their interests, and encouraged their further experiences towards the subject. Therefore, we expect that *Be the Data* could be applied to reach and interest a broad population of learners who are not necessarily knowledgeable about multivariate analysis algorithms.

There are many ways that *Be the Data* could be improved and extended. From a perspective of application functionality, we may give each student a hand-held device for parametric interactions (e.g., adjust weight parameters by dragging the weight bar). Therefore, students are able to tune parameters or modify choices of distance metrics to reveal new projections. We may also project data points on the floor and let students chase their data points as the mathematical model updates. From an embodied perspective, we could combine other kinesthetic sensors and explore how other bodily actions (e.g., pointing, waving, jumping) evoke embodied data analytics. From an education perspective, more *Be the Data* studies could be conducted to teach other analytical methods (e.g., classification). Versions of *Be the Data* can be designed for students varying from artists to

engineers who have a range of analytical experience.

Our study has several limitations. It did not have a control group to compare. More work is needed to understand how learners would perform differently when given a desktop application. It is unknown if embodied physical interaction improved the collaborative understanding of information over purely virtual interactions. However, our research is valuable in the identification of learners' analytical/collaborative strategies (qualitative) that employ this form of embodiment, accompanied by the evidence of learning (quantitative). A comparison of different applications/educational methods is beyond the scope of current exploration.

3.7 Conclusion

Be the Data is transformational from both a research and educational standpoint. We proposed a novel concept: embodied individuals as unique data points of a high-dimensional dataset. We implemented this concept via embodied interaction and visualization. We demonstrated its effectiveness in an educational usage scenario. *Be the Data* is able to empower students, who have relatively low analytical sophistication in the underlying data model, to learn and draw various inferences from the data. We believe that *Be the Data* achieves embodiment, enjoyment, and engagement that may benefit learning.

Chapter 4

Be the Data: Characterizing Social Meetings with Visual Analytics

4.1 Introduction

Social meetings are common to see in our daily life. They provide important venues for people to get connected and exchange information. Social meetings are in large part about finding, initiating, and sustaining conversations [61]. However, depending on the pre-existing relationships of attendees, opportunities for such interactions may not be evenly distributed among attendees. Challenges exist for attendees as to find people with similar interests, to identify participants whose expertise they seek, or to start a conversation with little embarrassment [62]. Being unaware of information about other attendees acts as a barrier for such social interactions. People may miss opportunities for communication and may not benefit from the interactions as much as they could.

We present a new approach to apply embodied interactive visualization to informal social meetings. Specifically, we employed an interactive system called *Be the Data* [25] to track social meeting attendees and provide real-time feedback about social clusters (Figure 4.1). In this system, each

attendee, in a physical space, is a data point of a high-dimensional dataset that contains attendees' information (Table 4.1). Attendees' quantitative responses to each question are the dimensions of the data points. The positions of attendees in this physical space represent the 2D projection of their high-dimensional data. During the social meeting, attendees are free to walk to any other attendees to social. They could alter the projections by physically moving themselves, and hence the corresponding data points, in the space. They receive real-time visual feedback that explains their changes in positions when they socialize with different groups of people. Therefore, they could have more opportunities to know about other attendees and locate people of interest.

During the social meeting, the system tracks participants' positions in the Cube and provides two kinds of feedback in real time visualizations (Figure 4.2). On the visualization, the left panel shows the changing visualizations of data points on a shared display which matches participants' layout in the meeting room. The right panel shows the changing weights of the data dimensions and relative cluster values. During the real studies, participants' headshot images along with their name tags were shown on the visualization for easy identification. However, in this paper, we used dots and id number for anonymous purposes.

The visualizations are designed to enable participants to know about others, to provide talking points with little embarrassment, and further to create and display topics for all participants to view. By reading clusters and dimension weights, participants may have a better understanding about what make they cluster together or separately. Therefore, they may be able to locate people with similar interests to socialize. They may also feel less embarrassed to initiate a conversation by talking about commonalities within the cluster.

The target for this application is informal social meetings, such as coffee breaks in a conference, where there is a mix of strangers and people who know each other. The primary focus of this

Figure 4.1: With the system, participants become individual data points of a high-dimensional dataset about themselves. A birds-eye view of their locations in the room is displayed on the large display above them. The visualization shows participants' headshot images, name tags, and a dimension weight chart.

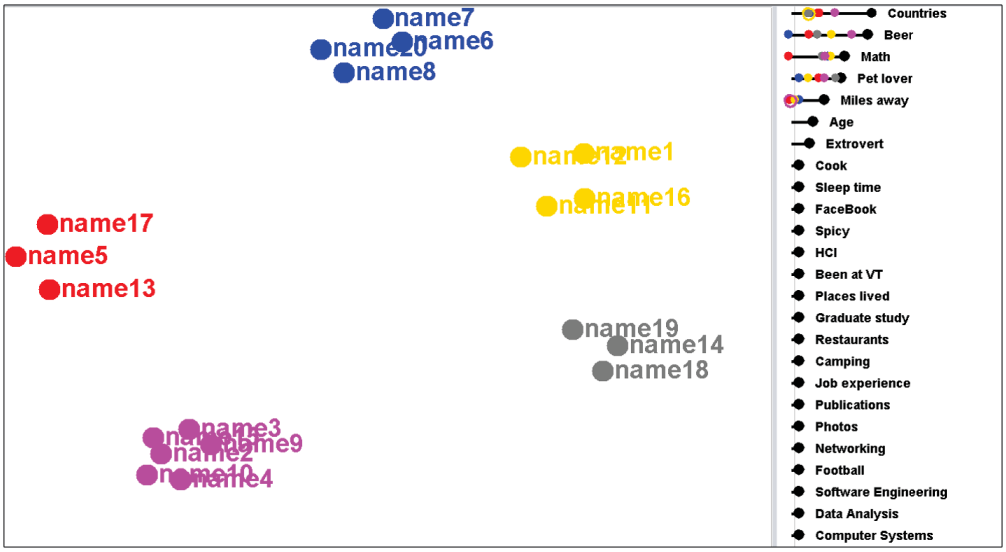


Figure 4.2: The real time visualization presented for participants.

exploratory study is a qualitative evaluation of the *Be the Data* applications deployed during informal social meetings. We aimed to find out how socialization could be mediated or empowered by interactive visual analytics. Specifically, we are interested in how people socialize with each other using the *Be the Data* system and how does the system impact socialization. The key contributions are:

- The demonstrative use of interactive visual analytics meshed with common social interactions.
- The identification of participants' strategies that employ the interactive visual analytics in socializing.
- The qualitative evidence of system's usefulness in facilitating social interactions.

Name	Age	Beer	Countries	Facebook	HCI	Math	Networking
ID1	26	50	1	100	0	100	50
ID2	21	75	8	646	65	80	15
ID3	26	100	2	0	50	100	0
ID4	19	1	4	80	50	100	50
ID5	24	1	5	318	0	80	0
ID6	27	2	5	300	1	80	0
ID7	27	75	10	600	100	80	0
ID8	31	75	9	739	100	72	0
ID9	33	60	9	400	100	50	0
ID10	32	50	3	750	100	100	20
ID11	27	25	10	800	20	50	0
ID12	25	60	3	800	90	70	0
ID13	24	65	7	400	80	80	0
ID14	24	60	3	1000	70	100	0
ID15	19	50	5	0	50	50	50
ID16	33	1	6	380	0	90	0
ID17	25	50	3	260	60	80	50
ID18	32	1	3	0	10	80	0

Table 4.1: A portion of the high-dimensional dataset that describes participants in one social meeting. The dataset is generated from 18 participants' quantitative responses to a list of questions in the pre-survey. Each participant is a data point and each question is a dimension. The example in this table shows 18 participants (all of the participants in one meeting) and 7 (out of 26) dimensions. Questions for the above dimensions were asked as below:
Age Your age?
Beer How much do you like beer? 1=I don't drink it at all; 100=I have it every day; 50=don't care.
Countries How many different countries have you visited in your lifetime?
Facebook How many FaceBook Friends and/or Google+ friends do you have? 0=i don't use Facebook/Google+
HCI On a scale of 1-100, my study (research) is related with human computer interaction (HCI). 0=not relevant; 100=I'm a HCI researcher; 50=ok or don't care.
Math Do you like math? 1=hate; 100=love; 50=ok or don't care.
Networking On a scale of 1-100, my study (research) is related with networking. 0=not relevant; 100=I'm a HCI researcher; 50=ok or don't care.

4.2 Related Work

4.2.1 Augment physical social space

Researchers haven been attempted to make use of physical space to promote social interactions among people. Proactive displays are a means to augment public physical social space. They are normally large public displays coupled with motion sensors. The displays are able to detect and respond to people standing nearby. For example, Ticket2Talk [61] is a proactive display designed to bring up interesting topics during coffee breaks in academic conferences. An attendee is able to intentionally move closer to the display, so that the display will show his interest to bring up discussions. Another example of proactive display is AutoSpeakerID [61] used during the question and answer session following a paper/panel presentation. The display will show the name, affiliation, and photo (if provided) of the person who approach the display in oder to ask the speaker questions. It provides a quick and brief background of the questioner to facilitate future follow-ups of speaker or audience with the questioner.

Interactive technologies have been extended beyond the proactive displays to maximize the use of larger physical space. Wearable and hand-held devices haven been applied to augment physical social spaces during academic conferences. These devices take advantage of sensor technology (e.g., radio frequency) to provide location-based services. They are often made small enough to be carried on with conference attendees like a conference badge or a cellular device, such as IntelliBadge [63], CharmBadge (www.charmed.com), Meme Tags [64], and SpotMe(www.spotme.ch). Some of these devices facilitate the one-on-one or person-finding activities. For example, the Meme Tags (a badge with small LED displays) allows people to share memes in person-to-person transactions [64]. Some of these devices focus on the aggregated data of all attendees to create dynamic visualizations for pubic views. For example, the IntelliBadge system tracks conference

attendees and provides real-time snapshot of the conference events attendance [63].

The *Be the data* application differs from the tools described above both in terms of the motion tracking technology and more importantly in terms of end-user information analysis techniques. First, the *Be the data* system uses an OptiTrack motion tracking system to collect and process motion capture data from 24 motion cameras. Second, in addition to simple aggregated statistics to summarize attendees, the *Be the Data* system coupled the Weighted-Multidimensional Scaling (WMDS) techniques with a prior knowledge about the attendees to interpret dynamic social clusters in real time visualizations. For example, as Figure 4.2 shows, attendees clustered in red seem to share the similarity of disliking math as compared to attendees in other clusters. As a probe of context-aware technology for the space that "comes into being through interaction" [65], the *Be the Data* system intends to augment the physical social space with such interactive technologies, so that attendees are able to observe and react to the dynamics of social clusters.

4.2.2 Social Computing

Social computing is concerned with incorporating the understanding of social context into the design of interactive systems [24]. Driven by the needs for computerizing aspects of social behaviors to promote communication and interaction among groups of people [66], social computing seeks to integrate technologies with humans' interaction improvised naturally in real time and real space [24]. It has been increasingly influenced by social and psychological theories as an analytical perspective to understand interaction and the use of interactive tools [24]. Social computing denotes a collaborative status where users work collectively to construct understanding [67]. One of the most influential application of social computing is the development of the field of Computer-Supported Cooperative Work (CSCW) [24]. Our application is an attempt of social computing within regular social meetings to trigger and respond to users' social reactions.

Embodied interaction is extended from the work of social computing [24]. Paul Dourish [24] describes embodied interaction as "the creation, manipulation, and sharing of meaning through engaged interaction with artifacts." He further explained it as an attempt to integrate physical and social reality of our everyday world into computing. People have developed sophisticated perceptions and skills acting in the world. However, they are rarely embraced for visual analytics to facilitate social interaction in an intentional and natural way. Therefore in this study, we present an application of the embodied system to mesh natural embodied social interactions with visual analytics. With the *Be the Data* system, we highlight how people naturally interpret the relative distance in the physical space (i.e., the "near is similar, far is dissimilar" metaphor) to comprehend clustering information. While people socialize with others, they embody their virtual data points to manipulate the underlying mathematical model. The real world social experiences is unified with the computational experience in a seamless way.

4.3 System Overview

The system exploits a multi-media Cube which includes a large overhead display, a motion tracking system, and the backbone software adapted from Andromeda [50] for direct manipulation of virtual high-dimensional data models (Figure 4.1).

To use the system, participants enter the Cube and embody their virtual data points by wearing a trackable hat so that their positions will be detected. Food tables and stand tables are randomly placed in the Cube. Participants are able to manipulate the layout of data points by walking around in the Cube. There is a large display above head where visualizations are displayed to show the representations of participants' social clusters. For example, if we consider a high-dimensional dataset about themselves (Table 4.1), each participant represents their data point and his/her position in the

Cube is visually reflected on the display (Figure 4.1). The underlying algorithm of visualizations relies on Weighted Multi-Dimensional Scaling (WMDS) [53]. WMDS visually plots the data in 2D Euclidean space to represent the data spread in the high-dimensional space. The system takes advantage of the inverse WMDS algorithm [18] so that we can map the layout changes to the value adjustment of weights. Participants adjust two dimensional coordinates by rearranging themselves in the Cube. In turn, they are provided with real time feedback (i.e., a new set of weights) that best describe their current layout.

In addition to the WMDS plot that reveals up-weighted dimension that characterize participants' choice of group, the system further provides dynamic clustering to visualize cluster information. That is to say, given participants' positions (the projected coordinates on the two-dimensional plane) during their social activities, the system automatically reveals clusters of data points. Colors are randomly assigned by the system to differentiate clusters. Centralized cluster values (i.e. the mean value for a given dimension of all the data points in the cluster) are calculated and visualized on some of the dimensions that are highly weighted. For example, in Figure 4.2, the blue cluster ranks lowest on the Beer dimension, suggesting that the blue cluster differentiates with other clusters because people in this cluster do not not Beer compared to other clusters. The dimension chart is sorted based on their weights (high to low) for easy identification of cluster distributions that characterize the clustering.

4.4 Evaluation

We conducted two informal social meetings to explore how participants deployed interactive visual analytics to socialize with others. Specifically, we seek to answer the following questions:

- Did participants learn about others through the use of interaction with the system?

- What different strategies participants took to socialize?

4.4.1 Participants

We held informal social meetings at our institutions and recruited 18 participants attended our study. Participants are a mix of strangers and people who had already known some of the other participants. They were from various departments, including Computer Science, Statistics, Industrial Engineering, and some other disciplines.

4.4.2 Procedure

The study includes three parts. First, participants were asked to answer a list of questions about themselves upon their arrival. Those questions were phrased to require a quantitative answer (e.g., Do you like to cook? 1 = I'd rather starve; 100 = I will be on the next Chef Wars TV show; 50 = don't care). The numeric responses were turned into a high-dimensional data set (Table 4.1) and used to generate interactive visualizations.

Second, participants will socialize with others. The social meeting starts with a short introduction about how to use the system. The research team introduces that the interactive visualizations on the screen were generated by a high-dimensional dataset created from their responses to the pre-survey. Each of them represent their data point in the data set and their responses to different questions are values in different dimensions of their data point. The research team also explained the near-far metaphor and weight changes in dimensions to interpret the visualizations. After the introduction, participants started social activities. There was no specific required task for participants to complete during the study. Participants moved in the Cube to talk with others on a completely voluntary basis.

Third, participants answered a post-survey probing their experiences with the visualized social meeting.

4.4.3 Data Collection and Analysis

To answer the research questions, we collected qualitative data from a post-survey and a recorded video of the meeting. The post-survey include short answer questions probing participants' experiences and reflections about using the system to socialize. The video documented participants' movement and social clusterings during the meeting.

To analyze qualitative data from surveys' open-ended questions, we had two authors encoded the data independently and compared their codes to draw interpretations. To analyze qualitative data from recorded videos, we had two authors watched the replay and collaboratively identified important social strategies participants used.

4.5 Results

After the studies, 15 out of 18 attendees returned post-survey. The results suggested that participants learned new information about other from the system, tried different strategies to socialize, and felt the system could be useful to facilitate social interactions.

4.5.1 Question 1: Did participants learn about others from the system?

Participants were asked if they learned something new about others from the system. In 15 answers, 13 attendees felt that the information presented by the system helped them learn something

new about another attendee. They learned attendees by using the visualization information. Participants learned others' names by observing the moving data points on the screen. They were aware of what grouped or separated them by checking the clusters and relative cluster values on weight bars made. They mentioned that they observed colored clusters and compared the corresponding colorful dots on the weight bars to see which clusters were relatively high or low in certain dimensions. Participants also learned who shared common interests from the cluster information. For example, they observed the "math" dimension was highly weighted at a point and located attendees who like math. One attendee found a group high in camping interested him. Moreover, participants got suggestions for conversational topics from the visualization. They asked others questions related to the data set to verify information suggested by the system, which led to further discussions.

4.5.2 Question 2: What different strategies participants took to socialize?

In participants' survey responses, 8 out of 15 attendees said that they talked with someone they would normally not talk and one attendee particularly mentioned that he/she did something different to socialize. One participant explained that he/she was trying to find what he/she had in common with other people by moving closer. One participant said that he/she introduced him more new people because he/she was curious about how the system would react when he/she approached to a new person.

With the aid of the *Be the Data* system, participants used the information present on the visualization to locate who to socialize with. For example, they observed the highly weighted dimensions, noticed the group that was high in the dimension of his/her interested, approached the group and socialized with people in that group. This strategy seemed natural to take place when people were provided with information about a group.

We also noticed that the system provided extra conversational topics for participants to start with. Normally people would talk about weather, sports, or local news to start their conversations. We did not expect they would start a conversation like "Do you also like spicy food?" because of the spicy dimension was up-weighted when one participant walked towards to a cluster of people.

Some of the participants took advantage of the visualization a step further. They attempted to alter who they were socializing to see how they impacted the visualization. One participant explained that he/she was trying to find what he/she had in common with other people. The other participants said that he/she introduced him more new people because he/she was curious about how the system would react when he/she approached to a new person. We also observed that some participants played games to get a certain dimension to the top of the list. For example, one participant noticed the "Math" variable might be an interesting factor and asked participants who also liked math to get closer to herself. One attendee indicated that there was difficulty in getting the dimension they wanted as not every attendee would coordinate to play the game. It was surprising to us that participants experimented to move around to create certain groups instead of just seeing what were there on the visualizations.

4.5.3 Usability Issues

The system was easy to understand and use. After the researcher's 1-minute short introduction, participants learned to engage each other through the interactive visualizations. Participants showed interests in the system when they found their icons could move on the screen paralleled with their movement in the room. Some of participants deliberately moved to understand how did the system.

Participants evaluated their overall impression about whether the system helped or hindered their socialization in their post-surveys. From 15 answers, 7 attendees believed that it helped, 6 thought

it neither helped or hindered, and 2 felt that it hindered the social activities. Participants who answered "helped" thought the system helped them to find the people/group to talk with and fostered discussion with extra conversational topics. Within a cluster, the clustering information highlighted their similarities, and therefore participants felt easier to initiate a conversation. Being in the same color might lead to a feeling of being related with others. Two participants mentioned that the colored clusters motivate them to join and talk with different groups and meet people they would not normally. Participants who answered "hindered" had different opinions. One thought the visualizations were distracting. He spent long time to look at the screen to check the system's updates, which disrupted their normal routine of socializing. He complained that he paid more attention to the system than the people. The other participant was not satisfied with the physical conditions (light, ceiling) of the Cube for social activities. For participants who had a mixed feeling (neither helper or hindered) about the system, one explained that they did not use the visualization to social. Some though the system could be more helpful when attendees know fewer people because it provided a starting point of question to talk about. However, these conversations might not last long because people were busy to check system updates. The system might also hinder the existing relationship among attendees.

The dimension chart needs to be resized according to the display size and distance. Four participants complained that dimension names and the distribution of clusters along those weight bars were not clear enough.

The trackable hats ran small and were uncomfortable for some participants. But even they chose to carry the hat in their hands, it did not affect the tracking or the visualization. The head gear fitting problem could be fixed by employing elastic band to fit a larger range of heads securely.

During the activity, participants might gathered around the food table to get food (not for socializing purpose), which had an effect on the visualization results. The random number and distribution

of the stand tables might have an effect on how people were gathered. One participant suggested to add music to create a social atmosphere.

4.6 Discussion

It was our goal to explore how socialization could be mediated or empowered by interactive visual analytics. We focused on an exploratory qualitative analysis of how participants used the *Be the Data* system to socialize. *Be the Data* was able to be implemented into regular social meetings with minimal required actions from attendees. Our results suggested participants used the information from real time visualizations to learn other attendees, to start conversations. While they used the visualizations, they tended to alter their socializing routines to meet new people.

The visualized social party might be more useful when most attendees don't know each other, so that they could potentially maximize the use of visualization information to set up new connections. Attendees may ignore the visualizations when they have strong existing relationships. In addition, the prior information should be collected relevant to the topics and purposes of the meeting to facilitate meaningful conversations. The interactive visualization might on the other hand hinder social activities. With the visualization, users' attentions are directed towards the screen. The information was not connecting on a personal level, which might also affect the interactions. More guided questions could be added to prompt social activities. For example, ask attendees to move to different areas of the room according to their prior information and explain the visualizations with the movement, so that attendees would understand how the system works, how variables differentiate, and how accurate they are.

4.7 Future Work

Current application for visualized social meetings could be improved in many ways to guide and facilitate users' social activities. Reading from the graph (on the left side) and check cluster values (on the right side) is not efficient enough. To reduce human effort, it would be better to add theme labels near clusters of dots and highlight cluster themes within the graph. Rich semantic information within the graph is desired to enable various real time feedback. For example, the system could add the feature for highlighting the area when lots of people share some commonalities. The entire area could be turned into mapped regions to fuse the similarity of the data points with respect to their attributes. The system could also highlight people on the screen who move actively from group to group, which may entice others to move and meet others.

Moreover, the system could provide more personal-level information, in addition to current cluster-level information. For example, probably with the aid of individual hand-held device, the system could show who are the people similar to each participant in the dimensions of his/her interest. It will help participants to locate interesting individuals. It also would be interesting to animate similarity (e.g. flash) when the participant is approaching someone, indicating a potential conversation topic. The integration of individual hand-held devices could potentially allow private view on for each individual participant, on which they could click the data points or drag the weight bars for parametric interaction. We expect personalized queries and feedback to provide more social information for participants.

In addition to get responded to users' movement, the system could be designed to direct users' movement more aggressively. Additional prompting (e.g., ice breaker games) could be implemented to break up clusters of people. The system may automatically send alert to ask participants to move to a specific area if the conversation is finished.

This study has several limitations. It did not have a control group to compare the social activities with static prior information (e.g., in excel). More work is needed to understand how participants use the prior information differently without the aid of the *Be the Data* system. However, this research is valuable in identifying attendees' usage scenarios and strategies that used this form of embodied visualization to interact with others. A comparison of different methods is beyond the scope of current exploration.

4.8 Conclusion

We implemented the *Be the Data* concept into visualized social meetings with real time feedback. We demonstrated its usages in informal social meetings at our institution. Our results suggested that *Be the Data* was able to facilitate social interactions.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This thesis focuses on the interactive data exploration via embodied visual analytical techniques. Specfically, we proposed a novel concept, called *Be the Data*, and its application, to explore the possibilities of using embodied interaction to learn from high-dimensional data. In *Be the Data*, each participant embodies a data point and the position of students in a physical space represents a 2D projection of the high-dimensional data. Participants physically move in a room with respect to each other to interact with alternative data projections and receive visual feedback.

The primary goal of *Be the Data* is to promote interactive data exploration for STEM education and outreach. With the rise of big data, it is becoming increasingly important to educate students about data analytics. In particular, students without a strong mathematical background usually have an unenthusiastic attitude towards high-dimensional data and find it challenging to understand relevant complex analytical methods, such as dimension reduction. Attracting students to learn data analytical skills is an important national need. We can imagine that on a desktop computer, moving data points around would be tedious. But with Be the Data, data points are students who

move themselves, either by their own volition or based on instructions from a collaborator. It engages students in an otherwise potentially boring data exploration. Moreover, for students who have no analytical experience, the concrete engaging experience of being a data point makes an abstract data analytical concept such as WMDS approachable to them, instead of scaring them away.

The second goal of *Be the Data* is to explore its innovative application to promote social interactions. We use *Be the Data* to characterize social gatherings in informal social meeting with a mix of people who did and did not know each other. Our findings indicate that participants can use the information from real time visualizations to learn about other attendees, alter their socializing routines, and encourage conversational topics.

5.2 Future Work

There are many ways *Be the Data* can be improved and extended. For example, we learned studies that showed changes in dimension bars successfully told students that data are clustered based on up-weighted dimensions, but failed to provide enough information about how clusters distribute on the dimensions.

Be the Data, as intended, enables social interaction to enable knowledge discovery from data. However, it does not open opportunities for self reflection, self processing, or personalized inquiry. That is to say, possibly a balance of group and individual learning is better than one or the other. This explains *Be the Data* to include hand-held devices that might result in improved learning. With the device, students might tune weights or modify choices of distance metrics, see the entire dataset, look up facts that come to mind while exploring data, and etc. The system may project data on the floor, and let students chase their data points as the mathematical model updates.

Appendix A

Difference in Probability for being Correct

To model questions with right or wrong answers, we use a Beta-Binomial Bayesian model.

$$
\begin{aligned}
y_{jis} &= \begin{cases} 1 & \text{if student } j \text{ for question } i \text{ on survey } s \text{ is correct} \\ 0 & \text{otherwise} \end{cases} \\
z_{is} &= \sum_{j=1}^{n_s} y_{jis} \\
n_s &= \text{number of responses for survey } s \\
s &\in \{pre, post\}
\end{aligned}
$$

$$
\begin{aligned}
z_{js}|\rho_{is} &\sim \text{Binomial}(\rho_{is}, n_s) \\
\rho_{is} &\sim \text{Beta}(0.5, 0.5),
\end{aligned}
$$

The idea of being a data point can be applied to other data analytical problems, such as factorial design, classification, and clustering where participants have their features and move into different groups. There are more open-ended questions stemming from Be the Data. For example, does physical interaction improve the collaborative understanding of information over purely virtual interactions? What and how can other bodily actions (e.g., pointing, waving, jumping) be applied to interact with data?

To break the limitation of a walking space, *Be the Data* techniques has the potential to track any movable object in a larger space. For example, with the wide application of car GPS, individual vehicle could be easily turned into a data point with many dimensions (e.g., various vehicle features). Instead of running intentionally designed driving scenarios, investigators could investigate important traffic issues under a naturalistic driving context [68, 69]. Visual data analytical techniques could be applied to explore and visualize the interactive effects of safety-threatening factors, such as hand-held cell phone [70], sleep habits [71], hormone responses [72], age [73], and other distraction factors.

where $\Pr[y_{jis} = 1] = \rho_{is}$. The prior distribution for ρ_{is} is a standard objective prior, known as the reference or Jeffreys prior. We assess the difference between $\rho_{i,pre}$ and $\rho_{i,post}$ by estimating the distribution of Δ_i where,

$$\Delta_i = \rho_{i,pre} - \rho_{i,post}.$$

A.1 Difference in Mean Attitude

To model questions that reflect attitude, we use a Normal Bayesian model.

$$\begin{aligned} y_{jis} &= \text{attitude score from student } j \text{ for question } i \text{ on survey } s \\ s &\in \{pre, post\} \end{aligned}$$

$$\begin{aligned} y_{jis}|\mu,\phi &\sim \text{Normal}(\mu_{is}, \phi_{is}^{-1} \\ f(\mu_{js}) &\propto 1 \\ f(\phi_{js}) &\propto \phi^{-1}, \end{aligned}$$

where μ_{is} and ϕ_{is} are the mean and precision of attitude scores, respectively. Again, we assess the difference between $\mu_{i,pre}$ and $\mu_{i,post}$ by estimating the distribution of Γ_i where,

$$\Gamma_i = \mu_{i,pre} - \mu_{i,post}.$$

Appendix B

The Animal Dataset

The high-dimensional data about animals was retrieved from [74].

Name	Walks	Vegetation	Tusks	Tail	Swims	Strong	Spots	Speed	Solitary	Smelly	Smart	Size	Quadra-pedal	Nocturnal	Muscle	Mountains	Longneck	Longleg	Lean	Hops	Hibernate	Hands	Furry	Forest	Flys	Flippers	Fierce	Domestic	Buckteeth	Arctic	Active
Persian Cat	65.69	6.25	0	66.8	6.25	12.58	6.25	26.98	36.6	7.86	38.62	6.25	66.15	12.5	6.25	11.25	8.75	8.89	14.54	6.25	1.56	0	90.19	8.98	0	0	13.98	72.88	0	0	13.89
Horse	55.58	51.05	0	70.42	0	63.13	15.8	81.68	16.78	33.07	37.28	71.15	70.35	1.11	51.14	16.22	44.94	70.57	47.96	0	1.11	0	40.58	10.76	0	0	15.51	59.33	27.6	0	55.95
German Shepherd	76.51	7.5	0	72.3	0	62.33	11.59	57.02	35.11	23.41	57.53	54.88	82.93	7.5	50.69	11.25	0	23.31	38.05	0	0	0	66.05	17.89	0	0	57.44	68.55	0	6.25	58.15
Blue Whale	0	0	0	26.42	71.82	55.26	23.75	21.42	25.89	13.75	39.06	86.46	0	6.25	25.37	0	0	0	11.88	0	0	0	0	0	0	64.87	7.5	5	0	32.75	7.88
Skunk	64.86	44.38	0	80.33	8.33	3.12	1.25	30.21	47.86	100	12.5	0	70.83	35.62	3.75	5.62	2.78	1.39	16.88	0	33.88	8.33	80	47.86	0	0	17.28	8.89	18.12	0	22.5
Tiger	73.92	4.63	0	66.65	5	84.04	6.08	76.21	30.21	39.36	28.52	36.59	78.36	77.12	82.34	14.88	8.33	19.2	48.87	0	14.24	0	66.23	21.32	0	0	63.81	5.64	21.11	1.25	59.45
Elephant	66.19	55.85	70.47	51.97	0	67.45	1.25	3.75	6.25	49.67	22.81	85.48	70.61	7.5	25.56	0	1.25	39.14	1.25	0	0	0	1.14	7.58	3.75	0	20.63	6.88	11.25	0	6.48
Gorilla	63.43	53.31	0	10	2.5	73.42	0	37.06	7.5	42.6	56.19	70.48	37.17	8.75	65.79	17.14	3.75	33.57	12.68	7.5	0	615	79.21	33.19	0	0	46.23	18.93	11.98	1.25	49.22
Seal	4.06	7.81	11.25	41.14	81.51	23.18	23.12	29.32	6.25	6.25	44.49	30.91	11.25	2.5	10.45	0	6.25	0	10.94	8.75	0	2.5	20.45	0	0	78.12	13.75	47.03	7.5	74.38	28.98
Chimpanzee	59.72	67.15	0	54.26	3.75	41.19	0	56.94	8.75	49.83	84.36	28.29	44.53	9.81	41.83	27.35	0	65.34	45.75	10.31		0	77.77	81.98	63.58	0	23.11	36.62	11.84	1.25	77.8
Hamster	62.24	60.33	0	26.34	0	3.89	8.5	39.33	35	26.23	18.37	0	72.49	0	0	10	0	0	5.56	23.61	26.88	7.5	86.82	2.5	0	0	8.75	71.44	62.08	0	47.28
Squirrel	35.21	64.72	0	84.69	3.75	14.17	2.5	54.93	22.58	1.95	18.49	2.54	42.67	10	15	4	3.89	1.39	19.59	23.02	51.26	17.79	79.78	66.53	13.02	0	7.92	14.76	49.44	3.75	62.01
Rabbit	19.48	75.67	0	61.11	1.39	4.79	9.26	63.26	18.89	12.71	11.47	2.27	43.13	16.61	13.96	31.25	2.78	4.51	5.06	67.59	19.68	0	60.89	28.32	0	0	4.17	60.52	60.52	12.38	47.86
Bat	2.5	36.46	0	15.11	0	7.5	1.25	80.96	19.97	30.33	32.3	1.25	6.48	98.86	35.73	0	5.42	10.96	45.69	2.5	45.3	5.28	36.75	58.01	96.9	0	32.57	5.56	8.12	0	36.24
Giraffe	64.07	69.61	0	44.8	0	31.43	77.08	31.86	9.57	22.09	16.95	78.23	77.12	0	21.52	0	96.71	84.01	68.82	2.02	0	0	17.17	11.11	4.44		1.11	3.33	18.32	0	24.22
Rat	70.75	11	0	73.69	0	18.19	1.25	57.14	40.87	48.31	24.99	1.88	73.62	67.81	20.24	7.5	0	0	28.76	8.12	0	3.34	48.79	36.71	0	0	59.44	16.69	56.08	0	56.88
Otter	5	7.78	0	43.61	85	11.39	0	40.97	30.56	12.64	27.22	11.25	35.56	12.5	11.39	0	1.25	0	30.14	0	10.14	10.28	46.94	25.83	0	47.36	6.25	7.78	28.75	35.56	41.11
Pig	52.77	48.85	5	40.56	2.5	27.44	21.23	16.24	4.19	51.17	23.04	48.17	68	2.5	14.4	6.25	1.25	1.25	0	0	0	0	17.05	2.5	0	0	25.48	49.95	8.44	3.75	15.53
Lion	54.06	8.75	0	62.22	0.25	77.19	0	65.47	11.88	30	25.7	71.88	65.55	9.77	50.94	21.25	0	10.62	22.81	6.25	0	0	67.03	22.5	3.75	0	70.47	3.75	0	0	33.44
Walrus	6.25	9.62	90.74	20.76	76.85	49.94	13.12	24.76	11.19	28.12	23.96	76.46	11.88	2.5	16.55	0	4.38	0	9.11	5	6.88	1.25	18.99	0	0	70.92	20	18.75	28.12	81.45	8.78
Deer	63.33	67.66	0	43.85	0	23.88	48.08	73.73	6.51	6.53	40.05	39.88	68.71	2.63	32.58	25.16	34.17	48.98	42.52	10.24	0	0	52.33	69.22	0	0	70.3	10.27	9.03	1.39	50
Leopard	62.77	1.25	0	48.22	5.28	55.18	74.97	85.26	33.84	11.02	37.16	54.64	78	25.81	62.22	26.03	6.81	40.39	73.48	0	14.76	0	43.27	18.05	0	0	70.3	6.25	0	2.5	71.82
Giant Panda	71.96	66.01	0	9.03	0	46.11	24.17	5	55.38	20.23	37.3	75.01	55.2	9.63	17.69	18.12	2.5	2.5	1.25	1.25	18.66	14.31	84.15	48.03	0	0	6.25	29.58	38.12	5	5.5
Weasel	51.87	30.55	0	30.55	14.93	8.89	8.75	61.5	26.26	22.39	36.56	1.25	57.65	28.32	22.11	9.82	11.25	6.25	54.05	7.5	3.57	0	45.41	31.66	0	0	59.37	3.75	6.94	0	72.09
Wolf	32.95	0	0	62.59	0	61.63	0	63.13	41.37	4.17	61.48	40.21	80.84	63.39	51.14	23.47	8.14	1.56	41.55	0	4.17	0	67.2	34.25	0	0	75.75	5	0	36.11	77.68
Horse	55.58	51.05	0	70.42	0	69.13	15.8	81.68	16.78	33.07	37.28	715	70.35	1.11	51.14	16.22	44.94	70.57	47.96	0	1.11	0	40.58	10.76	0	0	15.51	59.33	27.6	0	55.95
Cow	58.21	61.33	4.95	62.77	0.46	47.02	31.35	10.89	5.04	34.8	13.97	68.31	62.42	7.14	18.63	9.17	6.35	13.72	5.13	0	0	0	32.05	10.21	0	0	3.34	72.89	14.58	0	20.93
Fox	62.75	5.28	0	66.65	3.72	26.72	0	67.51	48.78	11.66	79.26	5.23	66.89	216	43.24	13.12	0	7.59	50.61	0	21.91	6.25	66.35	61.82	0	0	44.9	4.38	6.25	19.28	54
Sheep	73.69	63.91	0	11.65	0	10	0	4.66	0	23.75	10.62	16.53	80.14	8.41	49.8	49.8	0	10.8	2.5	1.25	0	0	61.16	1.25	1.25	0	0	30.68	15	0	3.75
Dolphin	0	7.38	0	41.85	60.36	41.88	0.8	59.09	3.96	7.57	60.38	47.82	4.8	0.62	26.29	0	5.76	28.2	1.29	0	0	0	0.62	0	0.82	86.44	10.73	37.98	2.55	20.34	53.15
Grizzly Bear	64.85	16.88	0	9.38	2.5	78.48	0	46.97	58.64	11.25	24	86.69	36.77	24.32	48.89	43.85	0	9.01	0	0	86.14	11.65	82.37	77.4	0	0	6.18	11.39	15	6.65	29.95
Raccoon	57.43	26.91	0	69.78	0.62	10.62	22.87	38.21	35.95	19.97	48.69	7.5	65.87	13.11	1.14	25.86	0	0	1.14	0	29.83	12.27	67.37	63.95	0	0	34.46	5	15.59	0	70.54
Rhinoceros	52.54	40.18	24.59	17.05	0	55.38	1.39	10.21	25.17	22.52	10.59	82	58.54	12.04	0	10	6.94	9.9	0	0	0	0	1.14	0	0	0	42.5	0	0	0	2.64
Zebra	77.61	74.3	0	61.43	0	23.85	0	70.05	1.25	9.55	22.63	49.82	89	0	38.76		21.68	39.02	38.74	4.17	0	0	36.4	8.75	0	0	0	3.94	14.38	0	42.51
Spider Monkey	37.85	52.61	0	72.26	0	24.17	0	58.12	9.38	16.98	55	20	32.4	5	24.06	3.75	0	55.69	44.93	0	0	67.64	64.93	33.26	0	0	17.71	12.78	0	0	74.9

Figure B.1: The high dimensional data about animals

Bibliography

[1] J. J. Thomas, *Illuminating the path:[the research and development agenda for visual analytics]*. IEEE Computer Society, 2005.

[2] 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, pp. 1–19, 2011.

[3] 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, 2013.

[4] 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*, ACM, 2010, pp. 55–64.

[5] C. Andrews and C. North, "Analyst's workspace: An embodied sensemaking environment for large, high-resolution displays", in *Visual analytics science and technology (vast), 2012 ieee conference on*, IEEE, 2012, pp. 123–131.

[6] R. Ball, C. North, and D. A. Bowman, "Move to improve: Promoting physical navigation to increase user performance with large displays", in *Proceedings of the sigchi conference on human factors in computing systems*, ACM, 2007, pp. 191–200.

[7] A. Febretti, A. Nishimoto, T. Thigpen, J. Talandis, L. Long, J. Pirtle, T. Peterka, A. Verlo, M. Brown, D. Plepys, *et al.*, "Cave2: A hybrid reality environment for immersive simulation and information analysis", in *Is&t/spie electronic imaging*, International Society for Optics and Photonics, 2013, pp. 864 903–864 903.

[8] J. Bennett, *T_visionarium: A user's guide*. UNSW Press, 2008.

[9] J. Thompson, J. Kuchera-Morin, M. Novak, D. Overholt, L. Putnam, G. Wakefield, and W. Smith, "The allobrain: An interactive, stereographic, 3d audio, immersive virtual world", *International journal of human-computer studies*, vol. 67, no. 11, pp. 934–946, 2009.

[10] Y. Tankaka, H. Yamauchi, and K. Amemiya, "Wearable haptic display for immersive virtual environment", in *Proceedings of the jfps international symposium on fluid power*, vol. 2002, 2002, pp. 309–314.

[11] R. Kehl and L. Van Gool, "Real-time pointing gesture recognition for an immersive environment", in *Automatic face and gesture recognition, 2004. proceedings. sixth ieee international conference on*, IEEE, 2004, pp. 577–582.

[12] 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*, ACM, 2015, pp. 3227–3236.

[13] *Global cities at Tate Modern creative review*, `https://www.creativereview.co.uk/cr-blog/2007/july/global-cities-at-tate-modern/`, Accessed: 2016-06-10.

[14] F. Taher, J. Hardy, A. Karnik, C. Weichel, Y. Jansen, K. Hornbæk, and J. Alexander, "Exploring interactions with physically dynamic bar charts", in *Proceedings of the 33rd annual acm conference on human factors in computing systems*, ACM, 2015, pp. 3237–3246.

[25] X. Chen, J. Z. Self, L. House, and C. North, "Be the data: A new approach for immersive analytics", in *Virtual reality workshop on immersive analytics*, 2016, to appear.

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

[27] 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.

[28] T. Martin and D. L. Schwartz, "Physically distributed learning: Adapting and reinterpreting physical environments in the development of fraction concepts", *Cognitive science*, vol. 29, no. 4, pp. 587–625, 2005.

[29] D. Pecher and R. A. Zwaan, *Grounding cognition: The role of perception and action in memory, language, and thinking*. Cambridge University Press, 2005.

[30] R. T. Azuma, "A survey of augmented reality", *Presence: Teleoperators and virtual environments*, vol. 6, no. 4, pp. 355–385, 1997.

[31] J. N. Bailenson, N. Yee, J. Blascovich, A. C. Beall, N. Lundblad, and M. Jin, "The use of immersive virtual reality in the learning sciences: Digital transformations of teachers, students, and social context", *The journal of the learning sciences*, vol. 17, no. 1, pp. 102–141, 2008.

[32] A. M. Glenberg, "Embodiment as a unifying perspective for psychology", *Wiley interdisciplinary reviews: Cognitive science*, vol. 1, no. 4, pp. 586–596, 2010.

[33] J. Cassell, T. Bickmore, H. Vilhjálmsson, and H. Yan, "More than just a pretty face: Affordances of embodiment", in *Proceedings of the 5th international conference on intelligent user interfaces*, ACM, 2000, pp. 52–59.

[15] S. Stusak, A. Tabard, F. Sauka, R. A. Khot, and A. Butz, "Activity sculptures: Exploring the impact of physical visualizations on running activity", *Visualization and computer graphics, ieee transactions on*, vol. 20, no. 12, pp. 2201–2210, 2014.

[16] J. Z. Self, X. Hu, L. House, S. Leman, and C. North, "Designing for interactive dimension reduction visual analytics tools to explore high-dimensional data", 2015. [Online]. Available: `http://people.cs.vt.edu/~jazeitz/jzself_vast2015_tech_report.pdf`.

[17] *the Cube icat*, `https://www.icat.vt.edu/content/cube-0`, Accessed: 2016-07-15.

[18] S. C. Leman, L. House, D. Maiti, A. Endert, C. North, *et al.*, "Visual to parametric interaction (v2pi)", *Plos one*, vol. 8, no. 3, e50474, 2013.

[19] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2001, vol. 1.

[20] M. Verleysen *et al.*, "Learning high-dimensional data", *Nato science series sub series iii computer and systems sciences*, vol. 186, pp. 141–162, 2003.

[21] 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.

[22] N. S. Ashaari, H. M. Judi, H. Mohamed, M. T. Wook, *et al.*, "Student's attitude towards statistics course", *Procedia-social and behavioral sciences*, vol. 18, pp. 287–294, 2011.

[23] J. Choo, H. Lee, Z. Liu, J. Stasko, and H. Park, "An interactive visual testbed system for dimension reduction and clustering of large-scale high-dimensional data", in *Is&t/spie electronic imaging*, International Society for Optics and Photonics, 2013, pp. 865 402–865 402.

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

[34] W.-J. Lee, C.-W. Huang, C.-J. Wu, S.-T. Huang, and G.-D. Chen, "The effects of using embodied interactions to improve learning performance", in *Advanced learning technologies (icalt), 2012 ieee 12th international conference on*, IEEE, 2012, pp. 557–559.

[35] C. Andrews and C. North, "The impact of physical navigation on spatial organization for sensemaking", *Visualization and computer graphics, ieee transactions on*, vol. 19, no. 12, pp. 2207–2216, 2013.

[36] R. Ball and C. North, "The effects of peripheral vision and physical navigation on large scale visualization", in *Proceedings of graphics interface 2008*, Canadian Information Processing Society, 2008, pp. 9–16.

[37] M. Jakobsen and K. Hornbæk, "Proximity and physical navigation in collaborative work with a multi-touch wall-display", in *Chi'12 extended abstracts on human factors in computing systems*, ACM, 2012, pp. 2519–2524.

[38] B. Lee, P. Isenberg, N. H. Riche, and S. Carpendale, "Beyond mouse and keyboard: Expanding design considerations for information visualization interactions", *Visualization and computer graphics, ieee transactions on*, vol. 18, no. 12, pp. 2689–2698, 2012.

[39] 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.

[40] 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*, ACM, 2005, pp. 119–122.

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

[42] P. Isenberg, D. Fisher, S. A. Paul, M. R. Morris, K. Inkpen, and M. Czerwinski, "Co-located collaborative visual analytics around a tabletop display", *Visualization and computer graphics, ieee transactions on*, vol. 18, no. 5, pp. 689–702, 2012.

[43] M. Tobiasz, P. Isenberg, and S. Carpendale, "Lark: Coordinating co-located collaboration with information visualization", *Visualization and computer graphics, ieee transactions on*, vol. 15, no. 6, pp. 1065–1072, 2009.

[44] P. Isenberg and S. Carpendale, "Interactive tree comparison for co-located collaborative information visualization", *Visualization and computer graphics, ieee transactions on*, vol. 13, no. 6, pp. 1232–1239, 2007.

[45] S. Bakker, E. van den Hoven, and A. N. Antle, "Moso tangibles: Evaluating embodied learning", in *Proceedings of the fifth international conference on tangible, embedded, and embodied interaction*, ACM, 2011, pp. 85–92.

[46] 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*, ACM, 2011, pp. 1989–1998.

[47] L. W. Barsalou, P. M. Niedenthal, A. K. Barbey, and J. A. Ruppert, "Social embodiment", *Psychology of learning and motivation*, vol. 43, pp. 43–92, 2003.

[48] F. Pulvermüller and L. Fadiga, "Active perception: Sensorimotor circuits as a cortical basis for language", *Nature reviews neuroscience*, vol. 11, no. 5, pp. 351–360, 2010.

[49] P. Isenberg, D. Fisher, M. R. Morris, K. Inkpen, and M. Czerwinski, "An exploratory study of co-located collaborative visual analytics around a tabletop display", in *Visual analytics science and technology (vast), 2010 ieee symposium on*, IEEE, 2010, pp. 179–186.

[50] J. Z. Self, N. Self, L. House, S. Leman, and C. North, "Improving students' cognitive dimensionality through education with object-level interaction", 2014. [Online]. Available: `http://people.cs.vt.edu/~jazeitz/vast_self_tech_report.pdf`.

[51] 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.

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

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

[54] T. Munzner, *Visualization analysis and design*. CRC Press, 2014.

[55] R. L. Wasserstein and N. A. Lazar, "The asa's statement on p-values: Context, process, and purpose", *The american statistician*, no. just-accepted, pp. 00–00, 2016.

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

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

[58] 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*, ACM, 2006, pp. 140–149.

[59] 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*, ACM, 2008, pp. 201–210.

[60] 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, pp. 1–29, 2010.

[61] J. F. McCarthy, D. W. McDonald, S. Soroczak, D. H. Nguyen, and A. M. Rashid, "Augmenting the social space of an academic conference", in *Proceedings of the 2004 acm conference on computer supported cooperative work*, ACM, 2004, pp. 39–48.

[62] X. Chen, Y. Fang, and B. Lockee, "Integrative review of social presence in distance education: Issues and challenges", *Educational research and reviews*, vol. 10, no. 13, pp. 1796–1806, 2015.

[63] D. Cox, V. Kindratenko, and D. Pointer, "Intellibadgetm: Towards providing location-aware value-added services at academic conferences", in *Ubicomp 2003: Ubiquitous computing*, Springer, 2003, pp. 264–280.

[64] R. Borovoy, F. Martin, S. Vemuri, M. Resnick, B. Silverman, and C. Hancock, "Meme tags and community mirrors: Moving from conferences to collaboration", in *Proceedings of the 1998 acm conference on computer supported cooperative work*, ACM, 1998, pp. 159–168.

[65] D. Svanaes, "Context-aware technology: A phenomenological perspective", *Human–computer interaction*, vol. 16, no. 2-4, pp. 379–400, 2001.

[66] F.-Y. Wang, K. M. Carley, D. Zeng, and W. Mao, "Social computing: From social informatics to social intelligence", *Intelligent systems, ieee*, vol. 22, no. 2, pp. 79–83, 2007.

[67] M. Parameswaran and A. B. Whinston, "Social computing: An overview", *Communications of the association for information systems*, vol. 19, no. 1, p. 37, 2007.

[68] F. Guo, Y. Fang, and J. F. Antin, "Older driver fitness-to-drive evaluation using naturalistic driving data", *Journal of safety research*, vol. 54, 49–e29, 2015.

[69] F. Guo and Y. Fang, "Individual driver risk assessment using naturalistic driving data", *Accident analysis & prevention*, vol. 61, pp. 3–9, 2013.

[70] G. M. Fitch, S. A. Soccolich, F. Guo, J. McClafferty, Y. Fang, R. L. Olson, M. A. Perez, R. J. Hanowski, J. M. Hankey, and T. A. Dingus, "The impact of hand-held and hands-free cell phone use on driving performance and safety-critical event risk", Tech. Rep., 2013.

[71] G. X. Chen, Y. Fang, F. Guo, and R. J. Hanowski, "The influence of daily sleep patterns of commercial truck drivers on driving performance", *Accident analysis & prevention*, vol. 91, pp. 55–63, 2016.

[72] M. C. Ouimet, T. G. Brown, F. Guo, S. G. Klauer, B. G. Simons-Morton, Y. Fang, S. E. Lee, C. Gianoulakis, and T. A. Dingus, "Higher crash and near-crash rates in teenaged drivers with lower cortisol response: An 18-month longitudinal, naturalistic study", *Jama pediatrics*, vol. 168, no. 6, pp. 517–522, 2014.

[73] F. Guo, Y. Fang, and J. F. Antin, "Evaluation of older driver fitness-to-drive metrics and driving risk using naturalistic driving study data", *Nstsce; 15-um-036*, 2015.

[74] *the Animal Dataset data*, `http://attributes.kyb.tuebingen.mpg.de/`, Accessed: 2016-07-15.