Beyond3D: Visualizing High-Dimensional Data Sets

David Durst

2015

Advised by Mark Braverman, Sebastian Maass, and Alex Wu

Submitted to: Princeton University

Department of Computer Science

This thesis represents my own work in accordance with University Regulations.

Date of Submission: April 29, 2015

Abstract

I present Beyond3D, a system that explains high-dimensional data sets to non-technical users with domain-specific knowledge through intuitive, interactive visualizations. Beyond3D surpasses current systems which fail to handle large numbers of dimensions, like matrices of scatterplots and histograms, or require technical training, such as PCA. A combination of user testing and a case study demonstrates Beyond3D's effectiveness. In the user study, non-technical test subjects with significant financial knowledge used the tool to answer questions about a financial data set. The case study demonstrates how a non-technical investor could use Beyond3D to decrease losses during market crashes.

Dedication

Thank you to my advisors: Mark Braverman, Sebastian Maass, and Alex Wu. Thank you to Brian Kernighan and Kai Li for reading drafts. Thank you to the game that started it all.[1]

[1]The image is from the video game Ridge Racer. [35]

Contents

List of Figures

Chapter 1

Introduction

Modern data analysis narrowly focuses on "big data": data sets which have many points and rapidly accumulate new ones. [40] Software developers and designers demonstrate this focus by creating tools for analyzing trends in arbitrarily large, low-dimensional data sets. [6] [22, pp.1,5,8-12][1] These tools explain data sets with many points, each of which is insignificant and has few values. However, big data analysis techniques do not work for data sets which fail to satisfy certain assumptions. High-dimensional data sets[2] require other methods of analysis because the following assumptions fail to hold: (1) a small combination of dimensions can explain each trend and (2) sufficient data exists to describe the relationships between dimensions. [53, slide 4]

The first assumption prevents big data techniques from succinctly explaining high-dimensional data sets. Since there are so many dimensions in a high-dimensional data set, each trend may result from a complex combination of them. Big data techniques take the overly simplistic view that a few dimensions can explain each trend. This assumption holds for data sets with only 5 to 10 dimensions, for any trend's explanation requires at most 10 dimensions. The low dimensionality enables the use of standard visualization tools such as matrices of scatterplots and heatmaps. [22,

[1] While Grinstein et al. claim to present high-dimensional visualizations, their evaluation methods make unreasonable assumptions such as that computer displays have "unlimited resolution". [22, p.10] I focus on high-dimensional visualizations that function without unrealistic assumptions.

[2] A high-dimensional data set is one in which each data point has many different values associated with it. Each value is a different dimension. Dimensions are also known as categories or variables.

pp.2-3] These tools explain trends by creating an entry in the matrix for each pair of dimensions. The rate at which the number of pairs of dimensions increases is the square of that for the number of dimensions. Therefore, using these visualization techniques for high-dimensional data sets causes the matrices of graphs to become pages and pages of unmanageable colored boxes. A successful tool for analyzing high-dimensional data sets must identify all the dimensions that explain each trend in a comprehensible manner without making the first, simplifying assumptions.

The second assumption limits the ability of big data techniques to even process high-dimensional data sets. According to the assumption, data sets with more dimensions must have more points. The techniques need the extra points to explain the additional relationships between dimensions. A linear increase in the number of dimensions requires an exponential increase in the amount of data point for big data techniques to draw conclusions with the same accuracy. [53, slide 4] However, the exponential increase may not be feasible as, for example, a big data technique requires 100^{10} data points to analyze a 10 dimensional data set with the same accuracy as a 1 dimensional data set with 100 data points. [53] Any analysis technique will struggle to handle that much data. Even if I assume that every data point takes one byte of space, 100^{10} bytes is more than 5 exabytes, the amount of information ever spoken by humans. [31] As shown by the example, big data techniques may struggle to process sufficient amounts of information to draw conclusions about high-dimensional data sets. Successful techniques for analyzing high-dimensional data sets must function with reasonable amounts of data.

The currently available tools for analyzing high-dimensional data sets overcome the issues associated with other techniques. Numerical techniques, such as Principal Component Analysis (PCA), can handle small data sets and identify trends that involve complicated combinations of dimensions. Previous research has shown that appropriately tuning the PCA algorithm allows it to handle data sets with limited numbers of data points relative to the number of dimensions. [27, pp.117,123] Additionally, PCA can explain trends that involve complicated combinations of dimensions through pairs of vectors and scalars. PCA produces a list of <vector, scalar> pairs. Each vector is a combination of dimensions in the original data set that explains some trend. The

associated scalar value is the strength of that trend relative to others. [15] These features of PCA make it a solution that technically trained individuals can use to analyze high-dimensional data sets.

The flaw of numerical techniques is that they require users to understand mathematical concepts that may be beyond their abilities. Explanations of PCA are too complicated for non-technical users. [52] Articles that describe PCA-based algorithms to general audiences will briefly skim over the technique before demonstrating its application. Kritzman et al. use PCA to develop a novel method to measure risk in financial markets. However, they do not sufficiently describe PCA for the reader to be able to reproduce their algorithm. They merely explain it enough so that the casual reader can follow the rest of the paper. [33, pp.4-7] I see an opportunity to create a tool that can analyze high-dimensional data sets, like PCA, but also presents its results in a manner that is understandable to non-technical users.

Many non-technical users need an intuitive tool to analyze high-dimensional data sets. Users ranging from biologists and financial analysts look for patterns in such data sets. Biologists have microarrays of genetic data where the number of dimensions is an order of magnitude larger than the number of data points. [4] Financial analysts can view snapshots of the markets as individual data points. A data set in which each point contains the price of every stock in the S&P 500 at the time the market closes would have about 500 dimensions. [44] Analysts who understand these data sets can make more profitable and less risky investments. Both of these non-technical users need a data analysis tool.

This paper presents a tool for those users: Beyond3D. Beyond3D uses a combination of interactive visualizations to intuitively explain high-dimensional data sets to non-technical users while also enabling them to apply their own domain-specific knowledge. The system consists of visualizations that allow users to: (1) understand general relationships between all dimensions, (2) focus in on subsets of dimensions, and (3) investigate the values of individual data points across all dimensions. The visualizations provide general and specific views that together enable the user to understand a data set. I will further explain Beyond3D and demonstrate its effectiveness through the follow parts of the paper:

1. **Background Information:** This chapter enables the reader to understand the rest of the paper by defining terminology and providing a general overview of the field of data visualization.
2. **Beyond3D System:** This chapter describes how to use Beyond3D, why it is a successful data visualization tool, and the technology behind the tool. This information enables the reader to reproduce my work and understand how it improves upon previous research.
3. **User Study:** This chapter demonstrates Beyond3D's effectiveness through the results of a study that tested non-technical users' abilities to analyze data sets using the tool. It also explains how the tool evolved to address flaws highlighted by user feedback.
4. **Case Study:** This chapter provides an example of how non-technical investors can use the tool to analyze financial data sets and minimize losses during market crashes. Like the user study, it demonstrates the effectiveness of Beyond3D.

Chapter 2

Background Information

In this chapter, I provide the context necessary to understand Beyond3D: the terminology of data visualization,[1] the effect of data storage methods on visualization techniques, and the major problems in data visualization. The reader will be able to understand my work and its relationship with existing research after reading this section.

2.1. Terminology

1. **Data Set:** A collection of values that the user wants to analyze. [49]
2. **Database:** A collection of data sets. Databases store and organize data sets in a method that is "convenient [for] access." [1]
3. **Dimension:** In tabular data, a dimension is a column. A dimension is also known as a variable, as in an explanatory or response variable.
4. **Data Point:** In tabular data, a data point is a row.
5. **Response Dimension:** The dimension whose value the user is trying to predict. This paper assumes that there is one response dimension. [46]
6. **Explanatory Dimension:** A dimension whose value may predict that of the response dimension. This paper assumes that all dimensions except for the response one are explanatory. [46]
7. **Axis:** A line that displays the values of the data points in one dimension.

[1] Some of these definitions are based on the readme section of an early version of Beyond3D. [18]

8. **Glyph:** "A [graphical] symbol ... that conveys information". [3]
9. **Histogram:** A visualization that represents the distribution of data points in one dimension using two axes and rectangles. Data points are binned together based on their values in the dimension. Each rectangle represents a bin. The x-axis denotes the range of values that are binned together into a rectangle. The y-axis denotes the number of data points in a bin. [2]
10. **Bar Chart:** A visualization that represents the distribution of data points in one dimension relative to another using two axes and rectangles. Data points are binned together based on their values in one of the dimensions. Each rectangle represents a bin. The x-axis denotes the range of values that are binned together into a rectangle. The y-axis denotes some information about the values of each bin's data points in the other dimension. In this paper, I will use it to represent the average value of the bin's data points in the second dimension. [34]
11. **Scatterplot:** A visualization that represents the distributions of data points in two dimensions using two axes and glyphs such as dots. Glyphs represent individual data points. The location of a glyph represents a data point's values in the dimensions associated with the axes. The shape and color of a glyph may represent values of a data point in other dimensions. [22, p.2]
12. **Heatmap:** A visualization that is similar to a scatterplot except that, instead of drawing each data point, it divides the visualization into colored cells. The visualization bins together the data points in each cell. The color of a cell is a function of the data points in its bin. [22, p.2]

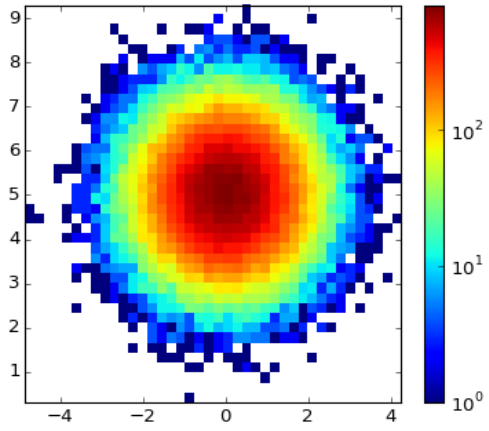


Figure 2.1: An example of a heatmap [11]

13. **Matrix**[2]**:** A method for combining visualizations to explain a multidimensional data set. Typically, the matrix is a grid of visualizations that each handle two dimensions, such as scatterplots or heatmaps. This grid contains a visualization for each pairwise combination of dimensions. Sections of the grid where a dimension is paired with itself have a one dimensional visualization, such as a histogram. [22, p.2]

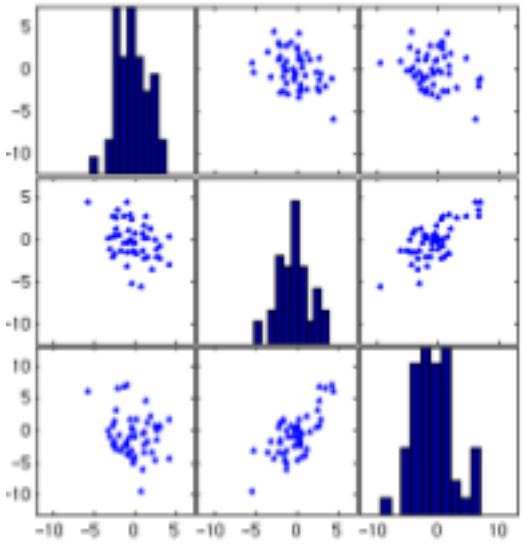


Figure 2.2: An example of a matrix of scatterplots and histograms [22, p.2]

14. **Radar Chart [20]**[3]**:** A visualization that represents the distributions of data points in an arbitrary number of dimensions[4] using N axes organized like spokes in a tire. Each axis's minimum value is in the center of the visualization. Values increase as they get farther away from the center. N-sided polygons represent individual data points. The location where a polygon intersects an axis represents the value of the data point in that dimension. The color of a polygon may represent the value of the data point in a dimension that lacks an axis. [22, p.5]

[2]Matrices are also known as lattices in R. [47]
[3]Radar charts are also known as polar charts. [22, p.5]
[4]Let N be the number of dimensions.

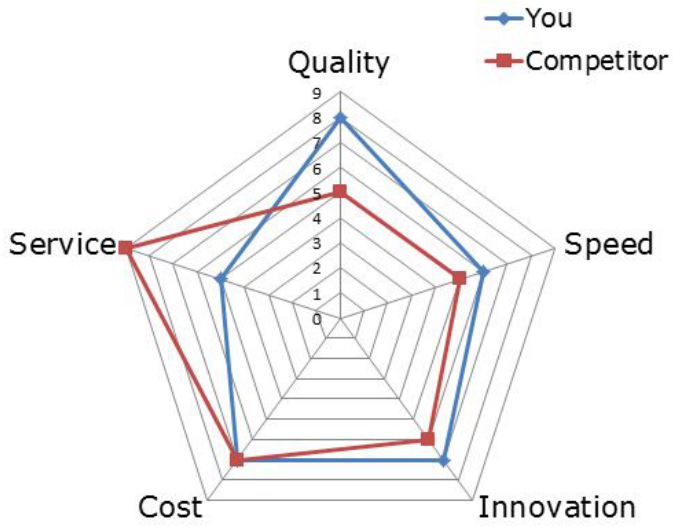


Figure 2.3: An example of a radar chart [14]

15. **Parallel Coordinates:** A visualization that represents the distribution of data points in an arbitrary number of dimensions[5] using N parallel axes. Each axis's minimum value is at the bottom of the visualization. The higher sections of an axis are for larger values. Lines represent individual data points. The location where a line intersects an axis represents the value of the data point in that dimension. The color of a line may represent the value of the data point in a dimension that lacks an axis. [22, p.2]

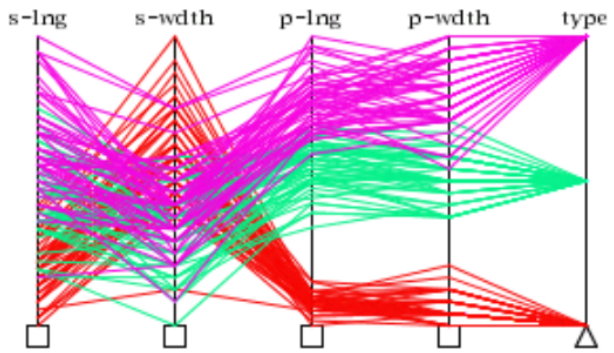


Figure 2.4: An example of a parallel coordinates visualization [22, p.4]

[5]Let N be the number of dimensions.

2.2. Databases and Visualization Techniques

Columns are Dimensions

Rows are Data Points

	A	B	C	D	E	F
1	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5	Dimension 6
2	1	10	9.1513198847	10	1	1
3	2	9	1.7047422659	1	1	2
4	3	8	4.4622738706	9	1	3
5	4	7	6.066768826	8	1	4
6	5	6	9.3368000095	2	1	5
7	6	5	7.1661181864	3	5	6
8	7	4	3.6269606254	7	5	7
9	8	3	5.6987510412	6	5	8
10	9	2	5.2078995993	4	5	9
11	10	1	5.9597376036	5	5	10

Figure 2.5: An example of a relational database

The type of database used to store a data set limits the visualization techniques that can successfully explain it. Visualization techniques explain data sets to users by converting the numbers into images. Users have a hard time drawing conclusions from tables of numbers, such as the one in figure 2.5. Visualization techniques explain the numbers through images that the user intuitively understands. However, the techniques must make some assumptions about the structure of the data set to create the images. The radar chart in figure 2.3 uses the assumption that each data point has exactly one value in every dimension to create the n-sided polygon. Since databases store data sets, they determine the structure. There are two main types of databases: relational and object-oriented. [45] Relational databases such the one in figure 2.5 store data sets in tables. Each row of the table is a data point and each column is a dimension. Object-oriented databases store each data point in a separate object. The values in each dimension are either attached to one object or stored as a relation between multiple objects. [5] Visualizations target one type of database so that they can make assumptions about the structure of the data set.

I developed Beyond3D to visualize relational databases because they store multidimensional data in a more accessible format. Beyond3D can easily identify the value associated with each data point and dimension because relational databases explicitly list those as rows and columns.

Beyond3D would have a more difficult time parsing the values associated with each dimension in object-oriented databases. Beyond3D would need a component that traces networks of objects to handle dimensions whose values are stored as relations between objects. Targeting relational databases enables me to focus on representing high-dimensional data and to limit the scope of my research to exclude the extraneous problem of parsing networks.

A reader who is interested in visualizations of data sets stored in object-oriented databases should look at other papers. There is already a significant amount of research into visualizing these data sets as networks of edges and nodes. These methods usually display each object as a node. The color or shape of a node represents values associated with the individual object. The visualizations represent a value stored as a relation between objects as an edge. The color or length of the edge may denote some information about the value. [43, pp.733-5] While these visualizations are helpful when examining certain data sets, they are beyond the scope of this paper.

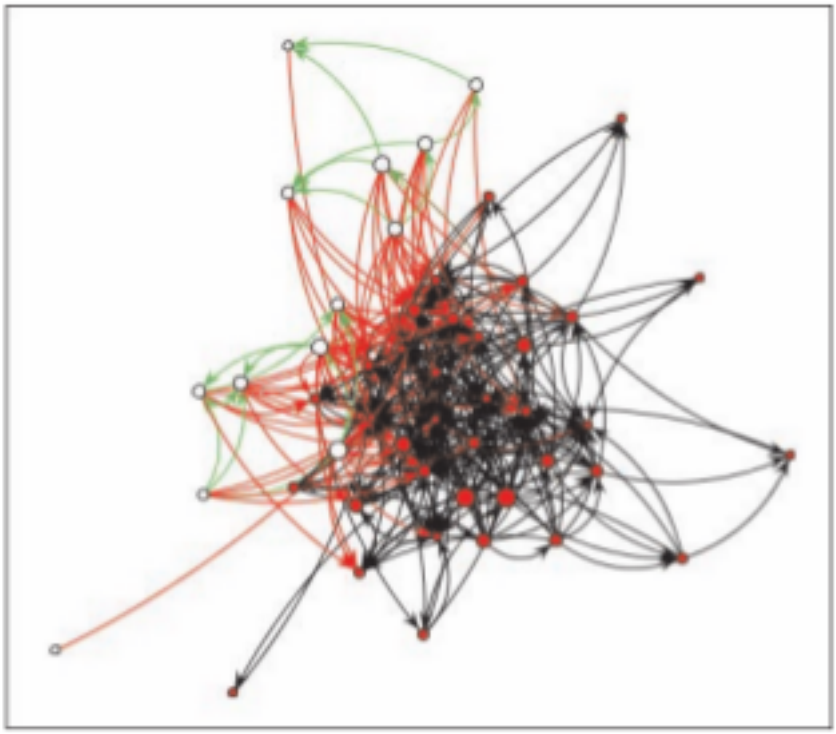


Figure 2.6: An example of a network visualization in which each node is a legal case and each edge represents one case citing another [43, p.736]

2.3. Data Visualization Problems

The field of data visualization consists of two problems: Data Viewing, creating individual visualizations that at least partially explain a data set; and Data and View Manipulation, designing interactions for and combinations of visualizations that completely explain a data set. [12, pp.79-80], [25, pp.1-2] The example programs provided in previous papers solve these problems for low-dimensional data sets. Their techniques, such as matrices of scatterplots seen in figure 2.2, will not scale to data sets with many dimensions. Nevertheless, the techniques and the problems they attempt to solve provide the basis for Beyond3D.

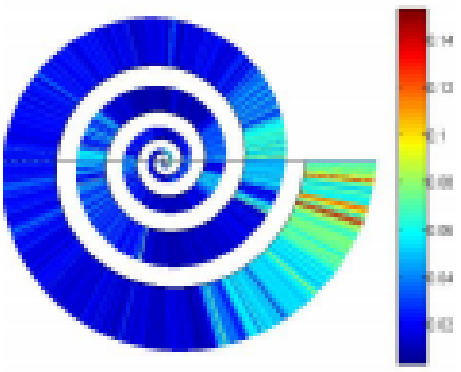
2.4. Data Viewing

Data Viewing is the problem of creating individual visualizations that explain part or all of a data set. Traditional research into Data Viewing takes an overly theoretical approach that does not lead to functional, high-dimensional visualizations. Grinstein et al. demonstrate the flaws of this approach with the intrinsic dimension metric: "the largest k, $k \leq n$, for which a set of k unit vectors in [a] n-dimensional space can be uniquely identified ... in the visualization." [22, p. 7] The metric claims to allow for precise evaluation of visualizations, but its unrealistic assumptions lead to inaccurate conclusions. In defining the metric, Grinstein et al. assume that the screens displaying the visualizations are "arbitrarily large ... with infinite resolution." [22, p. 8] These are not reasonable as screens have defined sizes and resolutions. As a result of the assumptions, visualizations such as parallel coordinates have arbitrarily large intrinsic dimension values. These visualizations do not perform as well as predicted on real data sets. [22, pp. 10, 12] Thus, the metric exemplifies how a too theoretical approach fails to create high-dimensional visualizations.

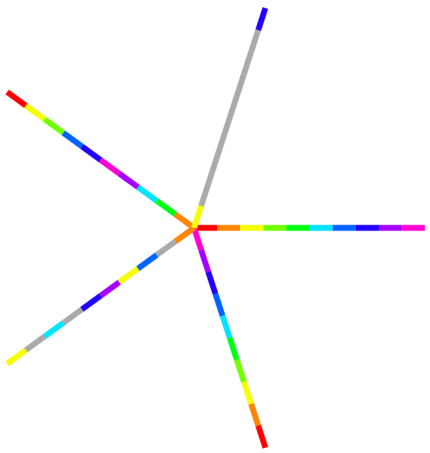
Despite their theoretical origins, some of the visualizations from Grinstein et al. play a practical role in Beyond3D. A relative of the paper's Radviz visualization, radar charts, can handle large numbers of dimensions by tightly packing axes together. Parallel coordinates can succinctly display up to about 10 dimensions simultaneously. The visualization can complement more general views of the entire data set by highlighting a few dimensions. Combining these visualizations with others

will hide their flaws. Radar charts and parallel coordinates work best when displaying a limited number of points. Both produce indecipherable images when attempting to show all the points in a data set. While they are not a complete solution, the techniques in Grinstein et al. are a starting point for my high-dimensional visualizations.

Recent developments in Data Viewing fill the holes in traditional research. In addition to radar charts and parallel coordinates, I need a visualization that displays an entire data set by itself without obscuring major trends. Therefore, I want something that can handle many dimensions but displays distributions rather than individual data points. The polar parallel coordinates technique that I will demonstrate later in the paper is based on a similar visualization in the Shenghui et al. paper. Figure 2.7a from their paper shows the distribution of a single axis over time. [41, p.179] Each line in the shell is the axis at a moment in time. The farther out in the shell, the greater the value of the time parameter. I can modify this to show the distributions of many dimensions simultaneously by making all axes have the same origin and length. Figure 2.7b demonstrates how I use the work of Shenghui et al. Each axis is a different dimension and the colors represent the distribution of points along that dimension. Polar parallel coordinates provide a general overview that polar charts and radar charts complement to explain an entire high-dimensional data set.



(a) The polar graph from Shenghu et al., [41, p.182]



(b) My version of the polar parallel coordinates visualization

Figure 2.7: Polar Visualizations

2.5. Data and View Manipulation

Data and View Manipulation is the problem of combining multiple visualizations in an interactive manner so that users understand a data set. This problem has two parts. The first issue is allowing users to identify the parts of a data set that they find interesting. The second is combining multiple visualizations so that the user can examine each part separately and see how they fit together. Unlike Data Viewing, previous papers' solutions to these issues work for high-dimensional data sets. Therefore, I can directly apply them to Beyond3D.

The degree of interest (DOI) function solves the problem of enabling users to select subsets of data sets. The user defines the DOI function to specify which portions of the data set are the focus of his investigation, which provide context, and which are irrelevant. [16, pp.239-40] The program must provide a method for the user to define such a function. Technical users may feel comfortable creating DOI functions by writing queries in a text box. However, non-technical users require a more an intuitive implementation. One such implementation is drawing arrows on parallel coordinates visualizations. [23, pp.3-4] The user drawing these arrows can select the data points that are greater than a certain value in some dimension. Another DOI implementation allows users to draw angles between axes to filter based on multiple dimensions. A user drawing these can select data points that make an angle between axes that is greater than a certain number of degrees. [23, pp.3-4] Figure 2.8 demonstrates this idea. Thus, an intuitive interface allows non-technical users to easily specify the desired subsets of data.

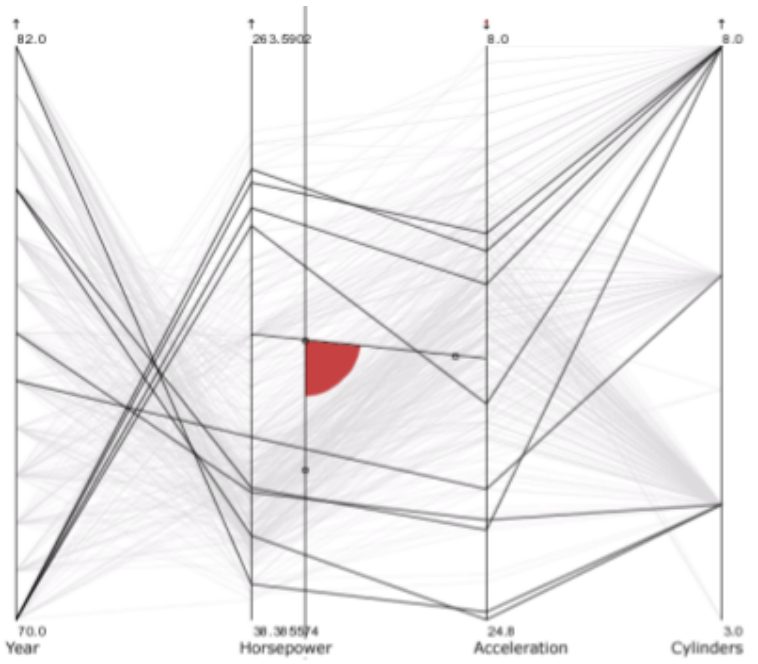


Figure 2.8: This image is an example of a DOI function in a parallel coordinates visualization. The user creates the red polygon to indicate the angles of interest. [23, p.3]

The Gestalt Laws specify how to combine the visualizations that explain each subset of data. The eight rules consist of seemingly obvious requirements such as that all visualizations use a consistent color scheme, logically organize data, and clearly separate foreground and background. [51] However, these rules helped me correct some design mistakes. I previously used different color schemes for different visualizations. Also, I combined visualizations in a confusing order. Explicitly stating and considering these rules helped me to create a tool that is intuitive for non-technical users.

Chapter 3

Beyond3D System

In this chapter, I explain Beyond3D's user interface, how it solves the problems of Data Viewing and Data and View Manipulation, and its relationship with previous research. I also describe Beyond3D's technology stack. The reader will understand how and why Beyond3D works after reading this section.

The reader should note that I am not trying to replace all previous data analysis tools with Beyond3D. I am targeting non-technical users with domain-specific knowledge about the visualized data set. Beyond3D should not be evaluated in the same manner as tools targeting advanced users, such as PCA. Additionally, Beyond3D assumes that the user always wants to explain one dimension, the response dimension, using the rest of the dimensions in the data set, the explanatory dimensions. This assumption is reasonable because my tool's target users have domain-specific knowledge. They likely already know the dimension of interest.

The assumption of one response dimension simplifies the problem. It forces the data set to take a specific structure. This structure enables me to follow the Gestalt Law of Similarity: "[u]se similar characteristics (color, size, shape, etc.) to establish relationships and to encourage groupings of objects." [51] Since I know that exactly one dimension is a response dimension, I can reserve color to explain its value. Position in each visualization represents the values of all other dimensions.

The process of translating a response dimension value to a color takes several steps. First,

Beyond3D divides the range between the minimum and maximum of all the response dimension values in a data set into 10 bins of equal width. Next, the value is placed into one of the bins. Finally, the bin is converted into a color using the mapping in figure 3.1.

Figure 3.1: The bin-to-color mapping[1]

3.1. User Interface Overview

The user interface has four components which together explain an entire data set. The first three components are the Polar Parallel Coordinates, Parallel Coordinates, and Radar Chart visualizations. The user controls these by interacting with them and the fourth component, the sidebar. This combination enables users to find trends in a data set and to apply their domain-specific knowledge.[2]

3.2. Polar Parallel Coordinates

This section explains the functionality of the polar parallel coordinates component and how the user should interpret the visualization.

[1] I recognize that this mapping may have some issues including placing similar colors in adjacent bins and not being color-blind friendly. Future iterations of Beyond3D may change the mapping to address these problems, but the current one functions reasonably well in practice.

[2] Figure 3.8 is a screenshot of Beyond3D that demonstrates how these components fit together.

3.2.1. Functionality

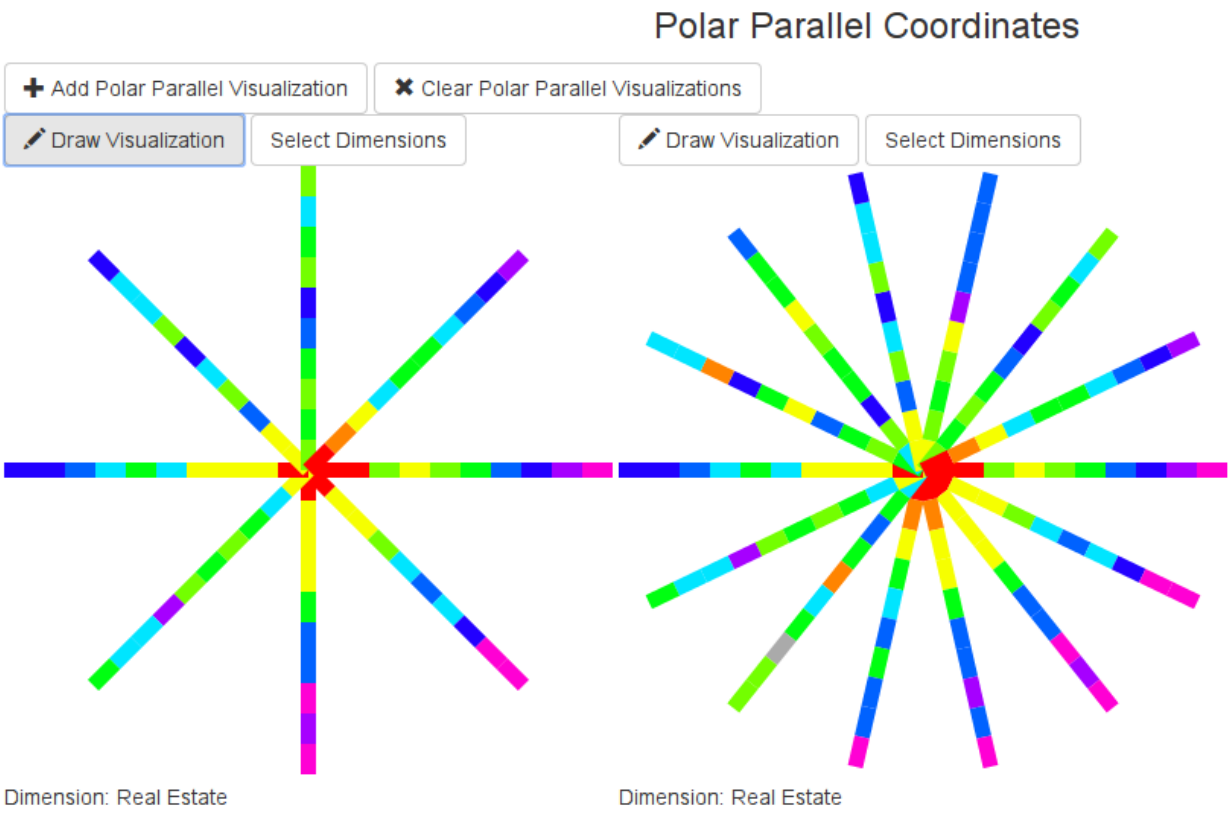


Figure 3.2: The polar parallel coordinates component

The polar parallel coordinates visualization displays the distribution of an entire data set in many explanatory dimensions relative to the response dimension. Each axis represents an explanatory dimension and consists of 10 line segments. For every explanatory dimension, data points are divided into equal-width bins with each one corresponding to a line segment. The segments closest to the center have lesser values and those farther away have greater values. Each segment's color is based on the response dimension values of the points in its bin. The average of those values is converted to a color using the process described above.[3]

This component allows the user to manipulate the number of visualizations and the data in each one. Figure 3.2 demonstrates the ability to simultaneously render multiple visualizations. Each

[3] A line segment is gray if its bin contains no data points.

visualization in that figure displays a different subset of the dimensions in the data set. The user can make these changes using the following buttons:

1. **Add Polar Parallel Visualization:** This button adds space for another polar parallel visualization. Then, it resizes all the existing visualizations so that they form a grid on the screen. Please note that the newly added space is blank until the user clicks the Draw Visualization button.
2. **Clear Polar Parallel Visualizations:** This button deletes all polar parallel visualizations.
3. **Draw Visualization:** This button draws the polar parallel visualization with the dimensions selected in the Select Dimensions dropdown. Each space for a polar parallel visualization receives one of these buttons.
4. **Select Dimensions:** This button allows the user to select which explanatory dimensions should be shown in the below polar parallel coordinates visualization. Each space for a polar parallel visualization receives one of these buttons.

The polar parallel coordinates visualization supports the following interactions:

1. **Identify a Dimension:** The user does this by moving the cursor mouse over an axis. The name of the corresponding explanatory dimension will appear in the text "Dimension:".
2. **Zoom:** The user does this by scrolling. The visualization will zoom in towards the area pointed to by your cursor.
3. **Pan:** The user does this by holding shift, pressing down the left mouse button, and dragging the cursor around. The visualization will pan in the direction of the drag.
4. **Focus On Axis:** The user does this by clicking on an axis. This will add the corresponding explanatory dimension to the "Currently Focused Dimensions" list in the sidebar and to the Parallel Coordinates visualization. Clicking the axis again will unfocus the dimension.

3.2.2. How To Interpret

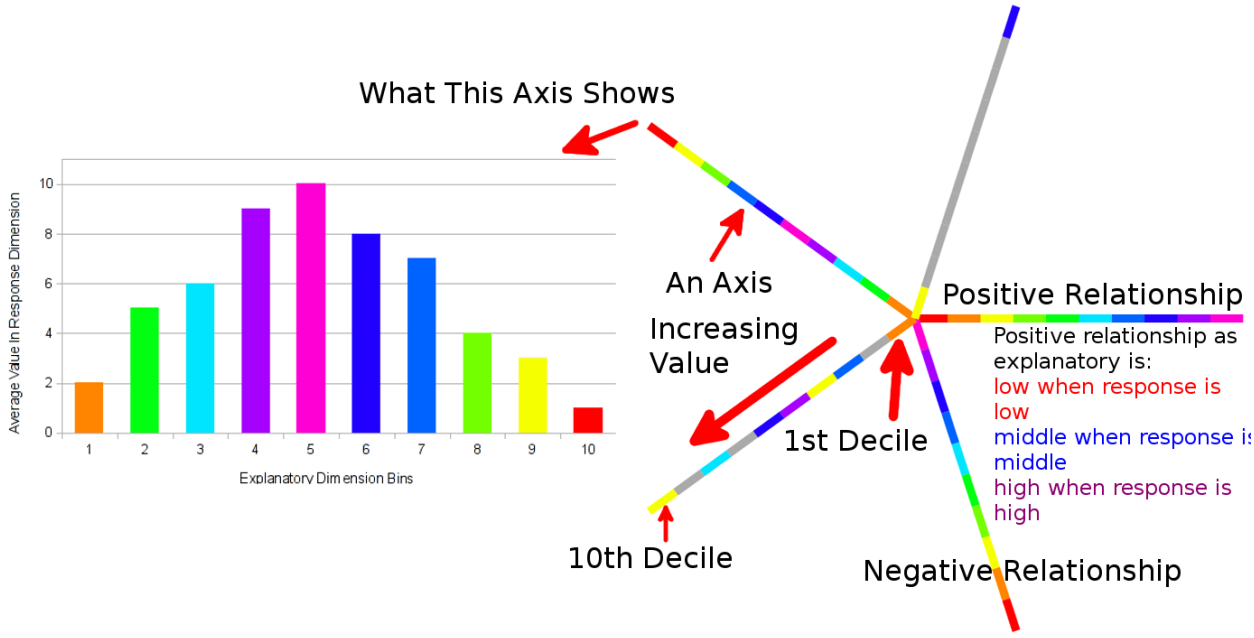


Figure 3.3: Interpreting a polar parallel coordinates visualization

The goal of the polar parallel coordinates visualization is to explain the relationship between each explanatory dimension and the response dimension. As the left part of figure 3.3 shows, each axis in the visualization is a bar chart with colors replacing column height. These axes take less space than regular bar charts. Packing them together in a circle succinctly provides a general overview of an entire data set.

Users analyzing the individual axes can determine which explanatory dimensions have positive and negative relationships with the response dimension. A positive relationship is one in which the explanatory dimension's value increases with the response dimension. This appears in the visualization through the color moving up the bin-to-color mapping in figure 3.1 as the bins get farther from the center. A negative relationship, where the explanatory dimension decreases as the response dimension increases, appears as the color moving up the mapping as the bins get closer to the center. Combinations of these patterns can also appear in a single axis. The top, left axis has a positive relationship in the first half of the axis and a negative one in the second half. Users who

recognize these color patterns can quickly develop a general understanding of a data set.

3.3. Parallel Coordinates

This section explains the functionality of the parallel coordinates component and how the user should interpret the visualization.

3.3.1. Functionality

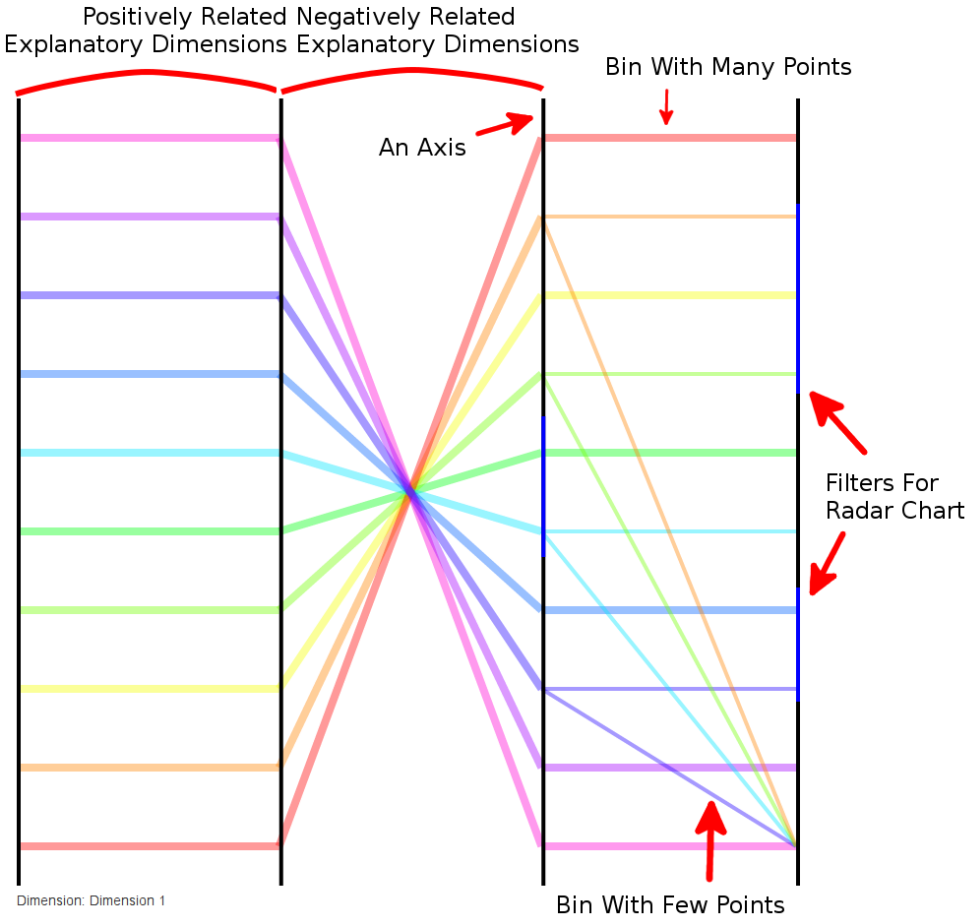


Figure 3.4: The parallel coordinates visualization

The parallel coordinates visualization displays the distribution of an entire data set in a few explanatory dimensions relative to each other and the response dimension. Each vertical, black line is an axis representing an explanatory dimension. Each axis has 10 locations where it can intersect the

colored lines. The colored lines connecting each pair of axes represent the relationship between the corresponding dimensions. The colored lines are two-dimensional bins of data points. To create these bins, the parallel coordinates visualization first divides each explanatory dimension into 10 equal-width, one-dimensional bins, one for each intersection location in the corresponding axis. The intersections at the bottom of the visualization indicate smaller values and those at the top are for larger values. Next, it creates a two-dimensional bin for each pairwise combination of one-dimensional bins in dimensions with adjacent axes. The visualization then bins data points twice. Every data point is one-dimensionally binned for each dimension and then two-dimensionally binned for each pair of dimensions with adjacent axes. Finally, the visualizations draws the lines between each pair of axes. For each pair of axes, the 15 bins with the most data points are drawn.[4] The thickest lines represent the bins with the most data points.[5] The colors of the lines are determined in the same way as the segments of polar parallel coordinates visualization: the response dimension values of the points in a bin are averaged, and that average is converted to a color using the process described at the beginning of this chapter.[6]

The parallel coordinates visualization supports the following interactions:

1. **Identify a Dimension:** The user does this by moving the cursor over an axis. The name of the corresponding explanatory dimension will appear in the text "Dimension:".
2. **Rearrange Axes:** The user does this by moving the cursor over an axis, holding shift, pressing down the left mouse button, and dragging the cursor left or right. When the axis gets at least half way to another axis, the user should release the left mouse button. This will cause the axes to switch places.
3. **Create a Filter for the Radar Chart Visualization:** The user does this by moving the cursor over an axis, pressing down on the left mouse button, and dragging up or down. All data points

[4] If fewer than 15 bins contain data points, then the visualization only draws lines for those that are not empty.

[5] Almost all the bins in figure 3.4 have an equal number of points. Therefore, there is little variation in line widths. Line widths vary more in other data sets.

[6] I developed this component based on techniques that placed histograms in parallel coordinates visualizations. [10, p.2] [21, p.1-2] I choose to draw 15 lines between each pair of axes because that was the most I could place in the image without making it too cluttered. Future systems may choose other methods for determining the number of lines, but the constant value of 15 is sufficient for my current research.

whose values in the axis's corresponding dimension are within the range covered by the blue highlight will be included in the Radar Chart visualization. One axis can have multiple filters. Multiple axes can have filters simultaneously. The user should double click on an axis to delete all filters applied to it.

3.3.2. How To Interpret

The goal of the parallel coordinates visualization is to explain the important relationships involving the response dimension and small combinations of explanatory dimensions. The parallel coordinates visualization complements the polar parallel one with a more narrow focus. This focus enables the visualization to examine more complicated relationships. The colored lines visualize the relationships between both the explanatory and response dimensions and the explanatory dimensions themselves. The visualizations of these relationships enable the user to understand multidimensional trends.

Users analyzing the parallel coordinates visualization can find and evaluate the significance of positive and negative relationships involving pairs of explanatory dimensions and the response dimension. A positive relationship between the two dimensions appears in the visualization through straight lines. These show that the data points are placed in the same one-dimensional bins for both dimensions. A negative relationship appears as crossed lines. These shows that that the data points are placed in opposite one-dimensional bins. The colors of the lines reveal how the two explanatory dimensions relate to the response dimension. A pink line that travels diagonally from the top of one axis to the bottom of another one shows that high values in the response dimension correspond with high values in one explanatory dimension and low values in the other one. The width of the colored lines indicate how many data points are in each relationship. The thickest lines represent the most significant relationships while thinner ones indicate potential noise. The many features associated with the colored lines enable users to both identify and determine the importance of relationships between dimensions.

3.4. Radar Chart

This section explains the functionality of the radar chart component and how the user should interpret the visualization.

3.4.1. Functionality

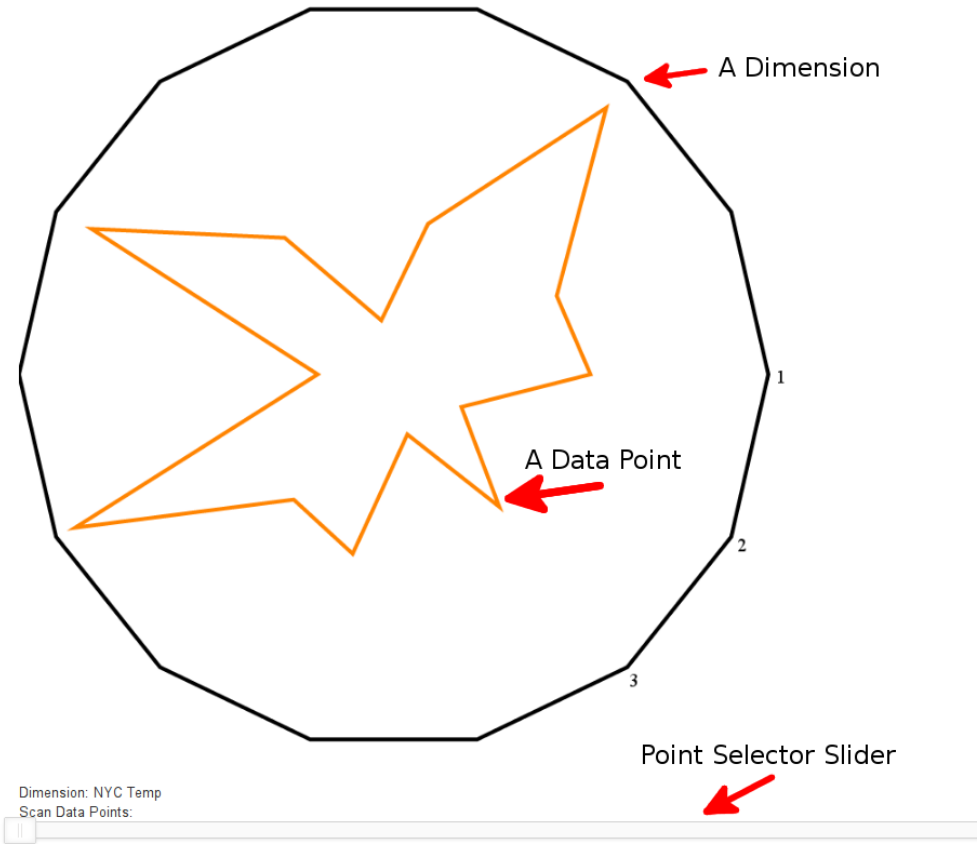


Figure 3.5: The radar chart visualization

The radar chart visualization displays the values of a single data point in all explanatory dimensions and the response dimension. Each corner of the black polygon represents an explanatory dimension. The numbers next to the corners correspond to the "Currently Focused Dimensions" list in the sidebar. These labels help users find their dimensions of interest in the black polygon. The colored polygon inside the black one represents the data point. The distance of each of the colored polygon's corners from the center indicates the point's value in an explanatory dimension. Corners farther

from the center indicate greater values. The data point's value in the response dimension determines its polygon's color. The visualization converts the point's response dimension value into a color using the process explained at the beginning of this chapter. This combination of a color and corners displays all the information associated with a single data point.

The user can change the currently visualized data point using the slider at the bottom of the screen. The filter applied in the parallel coordinates visualization creates a subset of data points for the radar chart. Dragging the slider scrolls through the points in the subset. This feature enables users to quickly visualize multiple data points.

The parallel coordinates visualization supports the following interactions:

1. **Identify a Dimension:** The user does this by moving the cursor over an axis. The name of the corresponding explanatory dimension will appear in the text "Dimension:".
2. **Zoom:** The user does this by scrolling. The visualization will zoom in towards the area pointed to by the cursor.
3. **Pan:** The user does this by pressing down the left mouse button and dragging the cursor around. The visualization will pan in the direction of the drag.

3.4.2. How To Interpret

The goal of the radar chart is to highlight individual data points. This component is the last step in the process of analyzing a data set, complementing the more general polar parallel coordinates and parallel coordinates visualizations. Since the other visualizations both examine the binned distribution of the points in a data set, the radar chart focuses on individual points. This detailed view enables users to study the data points that make up the larger trends.

Users analyzing the radar chart can find patterns in subsets of a data set using the slider. Positive and negative relationships between explanatory dimensions and the response dimension appear through the changing, colored polygon. For example, a positive relationship between an explanatory dimension and the response dimension appears as a corner moving outward as the color goes up the bin-to-color mapping. A negative relationship between two explanatory dimensions appears as one

corner moving inwards while the other goes out. Thus, this component complements the others by providing a more nuanced view of a data set.

3.5. Sidebar

This section explains the three features of the sidebar: the bin-to-color mapping, the dimension range selector, and the Currently Focused Dimensions list.

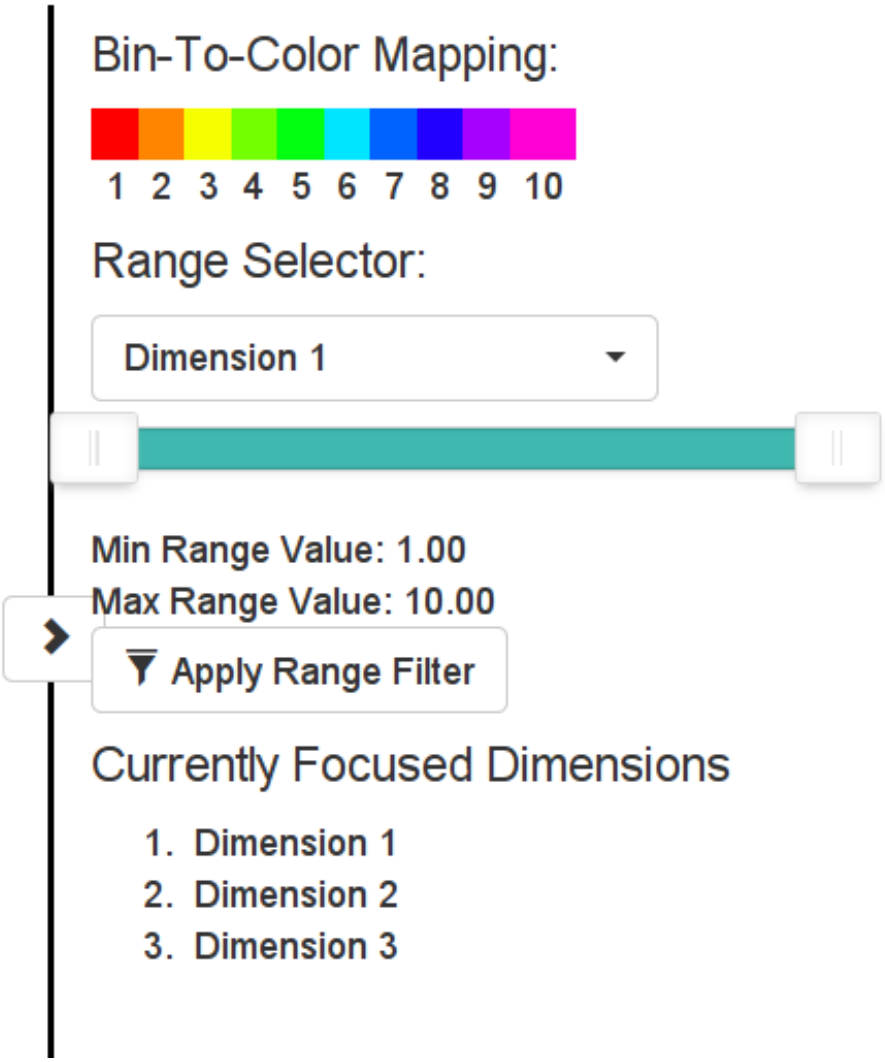


Figure 3.6: The sidebar

3.5.1. Bin-To-Color Mapping

This feature places the bin-to-color mapping in a convenient place for the user. The user can remind themselves of the mapping without navigating away from their current visualizations.

3.5.2. Dimension Range Selector

This feature enables the user to filter the data displayed in all three visualizations. To create a filter, the user should select a dimension to filter on, drag the sliders to the desired minimum and maximum values, and then press the Apply Range Filter button. All three visualizations will refresh to display only the data that fits within the filter. Example use cases include filtering by time. If a user is looking at stock performance, he can filter the data to examine performance by quarter. He can see how performance changes between quarters by shifting the minimum and maximum values of the time filter.

3.5.3. Currently Focused Dimensions

This feature lists the dimensions that the user has focused on by clicking their corresponding axes in the polar parallel coordinates visualization. The numbers match those displayed in the radar chart so the user can easily find the dimensions in that visualization.

3.6. How Beyond3D Addresses Data Viewing and Data and View Manipulation

The components of Beyond3D form a data funnel to solve the problems of the data visualization. Each of the three visualizations is a different layer in the funnel. The sidebar provides additional information and enables users to manipulate the funnel. Together, the four components explain a data set by presenting interactive visualizations with perspectives ranging from general views of distributions to narrow depictions of individual points.

The three visualizations occupy separate parts of the funnel and, by presenting different perspectives, each explain part of a data set and address the Data Viewing problem. The polar parallel coordinates visualization occupies the top of the funnel. It provides a general view of the distribution of a data set in all dimensions. This shows users which trends they should explore further. The visualization does not address the relationships between explanatory dimensions that cause the trends. The next step in the funnel, the parallel coordinates visualization, handles that. This

component explains how a subset of dimensions fit together. The colored lines identify and evaluate the relative strengths of the relationships between the explanatory dimensions. The colored lines also show how combinations of explanatory dimensions affect the response dimension. The user can look at the details in these relationships through the smallest part of the funnel, the radar chart. The radar chart individually shows each data point as a colored polygon. The user can look at these polygons to understand how individual data points cause the previously identified trends. Having seen all perspectives of a data set as it travels through the funnel, the user now understands the general trends and the dimensions and data points that cause them.

The organization and interactivity of the visualizations solves the Data and View Manipulation problem by providing users with a comprehensive understanding of a data set. As explained above, the data funnel explains a data set from many perspectives. By organizing the visualizations from general to specific,[7] Beyond3D displays the components in such a way that the user can combine their perspectives into one complete depiction of a data set. The interactive features enable users to fill any holes in their understanding. For example, suppose that a user is examining a financial data set where the daily returns of his portfolio are the response dimension and the daily returns of the S&P 500 are an explanatory dimension. The user wants to understand the relationship between his portfolio's returns and the largest daily returns of the S&P 500 during the month of May. The user would apply a filter in the sidebar to focus only on data points with values in May for their time dimension. Then, he would zoom and pan to focus on the top half of the S&P 500 daily returns axis in the polar parallel visualization. The combination of the visualizations and the sidebar provides the comprehensive, interactive experience necessary to address the Data and View Manipulation problem.

Beyond3D solves the two problems of data visualization through its interactive combination of the polar parallel coordinates, parallel coordinates, and radar chart visualizations. In the following chapters, I will demonstrate the ability of this system to explain data sets to users in practice.

[7] See figure 3.8 for an example of this organization.

3.7. Beyond3D and Previous Attempts To Explain High-Dimensional Data Sets

Beyond3D's novel contribution is its focus on explaining all aspects of high-dimensional data sets to non-technical users.[8] As I discuss in the introduction, previous research focused on visual and numerical methods for explaining these data sets. Research in the field of data visualization led to tools which can address parts of high-dimensional data sets. However, the research failed to develop methods for combining these pieces into a comprehensive explanation. The papers' brute force techniques, which explore all combinations of dimensions, do not scale to high-dimensional data sets. While numerical techniques can scale, research in the area has not focused on producing user-friendly results. The following data set demonstrates the unique ability of Beyond3D to explain high-dimensional data to non-technical users.[9]

Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5	Dimension 6
1	10	9.1513198847	10	1	1
2	9	1.7047422659	1	1	2
3	8	4.4622738706	9	1	3
4	7	6.066768826	8	1	4
5	6	9.3368000095	2	1	5
6	5	7.1661181864	3	5	6
7	4	3.6269606254	7	5	7
8	3	5.6987510412	6	5	8
9	2	5.2078995993	4	5	9
10	1	5.9597376036	5	5	10

Table 3.1: The example data set

[8] Beyond3D's target users both are non-technical and possess domain-specific knowledge about the currently visualized data set. However, Beyond3D demonstrates a novel method for explaining high-dimensional data sets to non-technical users regardless of their knowledge of the data set. Therefore, the user's knowledge of the data set is not relevant in the current section.

[9] Please note that, in this data set, dimension 6 is the response dimension and the rest are explanatory dimensions.

Previous techniques fail to provide a simple, comprehensive explanation of the data set even though it is neither extraordinarily large nor high-dimensional. The matrix technique exemplifies the issues with previous data visualization research. I apply this technique to the data set in figure 3.7. The 36 graphs are too confusing to provide a comprehensive explanation. Users can't identify which pairs of dimensions are of interest and which are noise. Additionally, it is difficult to understand how each dimension relates to the response dimension. Since the matrix is a static image and lacks a zoom feature, users can't focus on the rightmost column comparing each explanatory dimension to the response dimension. This visualization is not useful because it provides too much information.

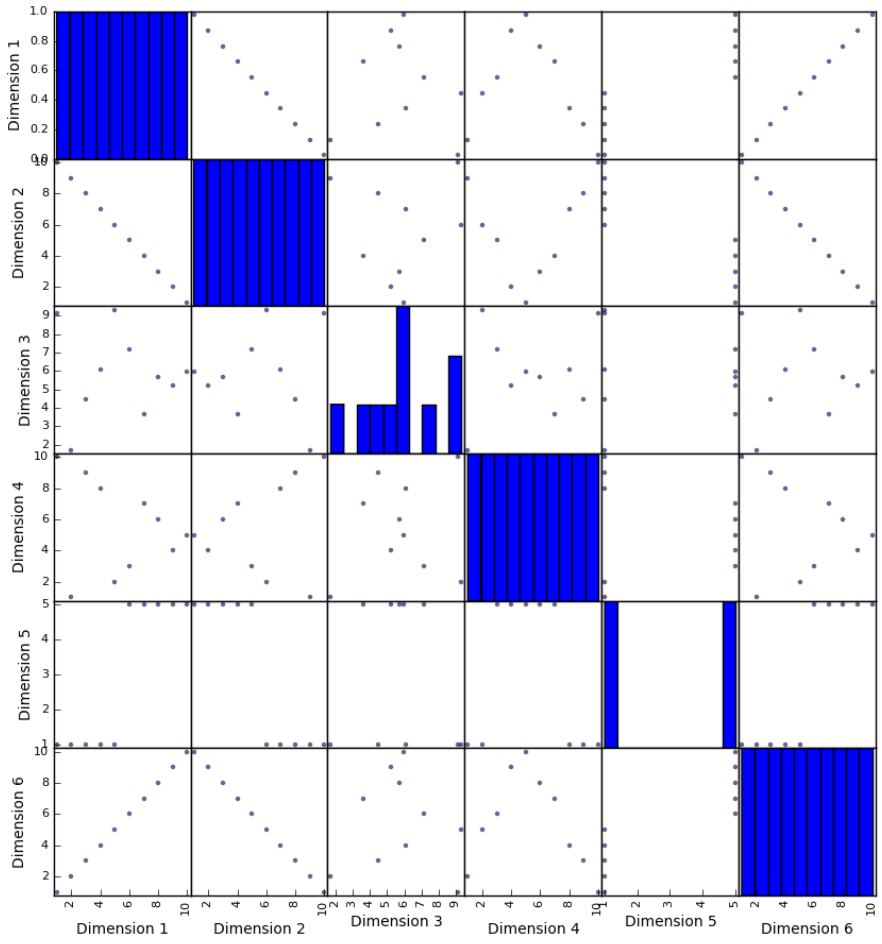


Figure 3.7: A visualization of the data set using a matrix of scatterplots and histograms

PCA demonstrates the inability of numerical techniques to explain high-dimensional data sets to non-technical users. Tables 3.2 and 3.3 contain the outputs of scikit-learn's PCA function when it is applied to the example data set.[10] They are completely incomprehensible to non-technical users. Each row of the first table is an eigenvector, a vector describing how a combination of dimensions explains a trend in the data set. Each row in the second table is an eigenvalue, a scalar describing the importance of the associated eigenvector. While PCA documentation can provide basic explanations of terms such as eigenvectors and eigenvalues, these are not sufficient for a user to understand the output. Users need to understand the linear algebra behind these terms and how PCA applies that math. Non-technical users lack the math skills necessary to make sense of these tables of eigenvectors and eigenvalues.

Eigenvector	Dim. 1	Dim. 2	Dim. 3	Dim. 4	Dim. 5	Dim. 6
0	0.53365	-0.53365	-0.04256	-0.18003	0.33379	0.53365
1	0.09601	-0.09601	0.30061	0.93552	0.08239	0.09601
2	-0.01416	0.01416	-0.94765	0.29931	0.10853	-0.01416
3	-0.19781	0.19781	0.09894	-0.05304	0.93275	-0.19781
4	-0.80772	-0.50725	0.00000	0.00000	0.00000	0.30048
5	-0.11938	0.63982	0.00000	0.00000	0.00000	0.75920

Table 3.2: The eigenvectors [11]

[10] I have modified the outputs slightly. I made the headers of the tables more meaningful. Scikit-learn just puts the index of the column in each header.

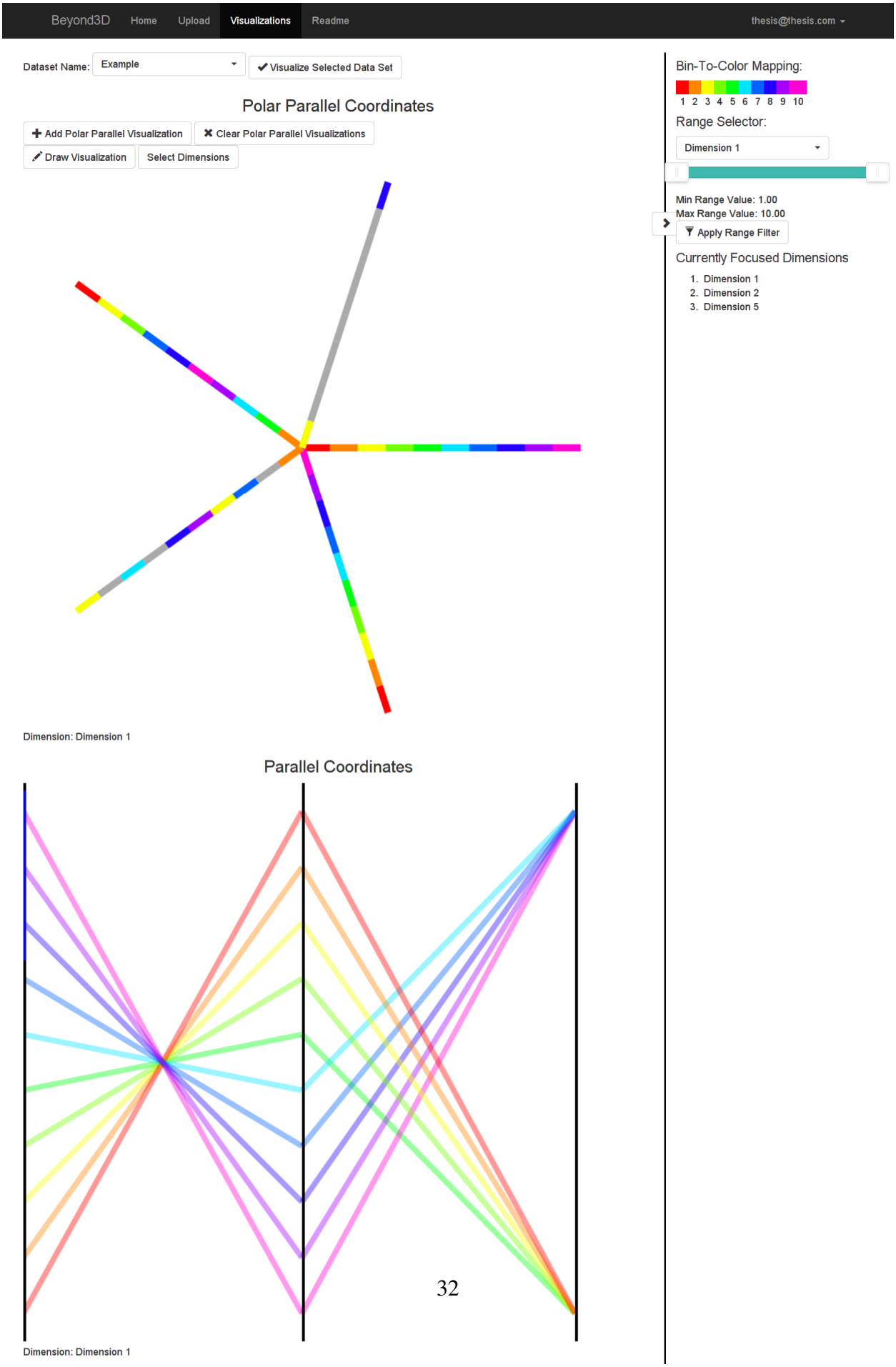
[11] Dim. is short for dimension.

Eigenvector	Eigenvalue
0	0.68201
1	0.18817
2	0.11048
3	0.01934
4	0.00000
5	0.00000

Table 3.3: The eigenvalues

Beyond3D successfully explains this data set using simple visualizations that depict the data set from multiple perspectives without overloading the user. Unlike PCA, the polar parallel coordinates visualization explains the relationships between the explanatory dimensions and the response dimension in a way that is comprehensible to non-technical users. Even users without strong mathematical backgrounds can recognize the color patterns and understand how they depict relationships between dimensions. Beyond3D surpasses the matrix technique by only focusing on the most important relationships between explanatory dimensions. First, the polar parallel coordinates visualization helps users separate the important explanatory dimensions, those with meaningful color patterns, from the noisy ones, those with random patterns. In figure 3.8, the middle, right axis has a meaningful pattern and the bottom, left one has a noisy one. Then, the user applies the parallel coordinates visualization to study the relationships between this limited subset of explanatory dimensions. Finally, the user zooms in on a few data points causing these trends using the radar chart. This interactive process helps Beyond3D avoid making the matrix's mistake of overloading the user with combinations of dimensions. Although the interactive features do not work in the below picture, they stitch together the actual components of Beyond3D into a comprehensive tool for explaining high-dimensional data sets to non-technical users.[12]

[12]Please note that the image stretches across two pages. The bottom part of the first page overlaps with the top part of the second page to show how the images fit together.


Beyond3D
Home
Upload
Visualizations
Readme
thesis@thesis.com
Dataset Name:
Example
Visualize Selected Data Set
Polar Parallel Coordinates
Add Polar Parallel Visualization
Clear Polar Parallel Visualizations
Draw Visualization
Select Dimensions
Dimension: Dimension 1
Parallel Coordinates
Dimension: Dimension 1
Bin-To-Color Mapping:
1 2 3 4 5 6 7 8 9 10
Range Selector:
Dimension 1
Min Range Value: 1.00
Max Range Value: 10.00
Apply Range Filter
Currently Focused Dimensions
1. Dimension 1
2. Dimension 2
3. Dimension 5

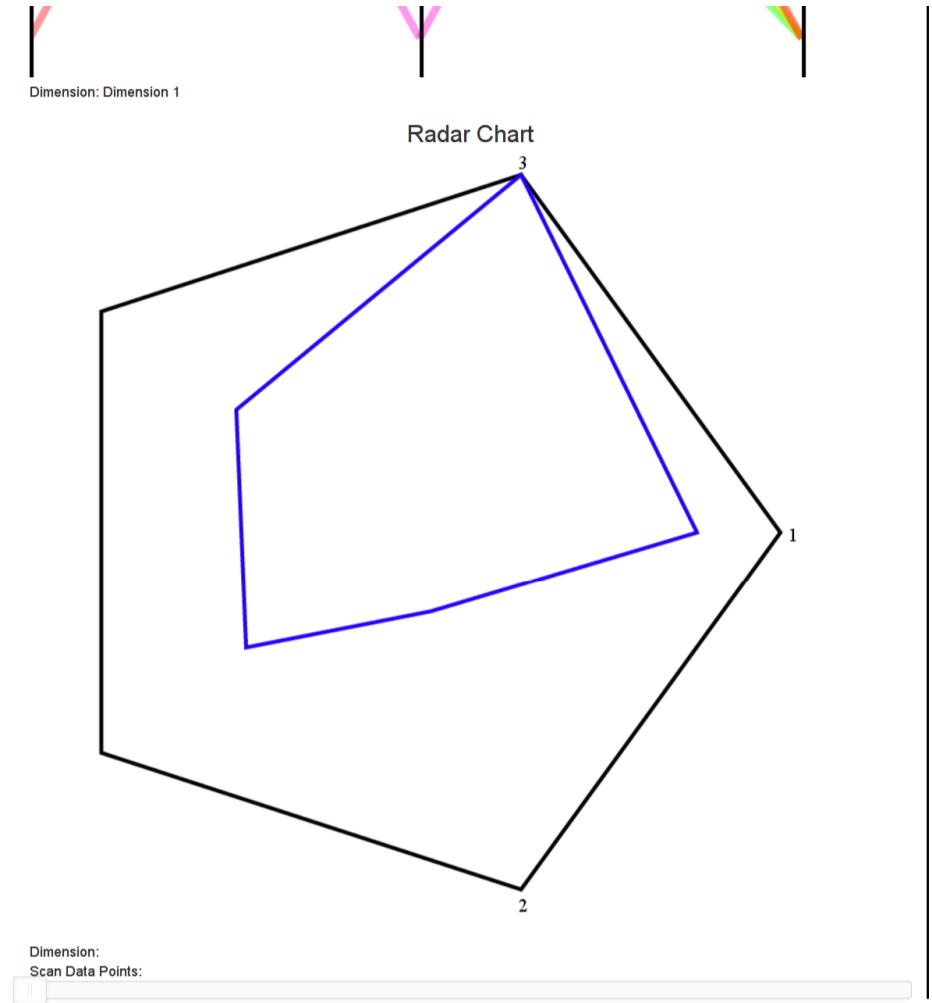


Figure 3.8: A complete example of Beyond3D

3.8. Technology

The technology stack behind Beyond3D consists of three components: a server that stores data sets and delivers code to the user, a webpage to visualize data sets, and another page to process uploaded data sets.[13] I use Meteor for my server software because it provides a full-stack solution that decreases my workload and simplifies my code. It does this by abstracting away issues such as maintaining a database. [37] Meteor natively integrates with MongoDB so that I don't have to spend time setting up a database to store the data sets and user accounts. [38] Its libraries also handle issues such as account security by allowing users to login through external services like

[13] Beyond3D is hosted at "beyond3d.herokuapp.com". An example account exists with the email address "thesis@thesis.com" and password "thesis".

GitHub. [38] In addition to helpful abstractions, Meteor further simplifies my codebase by only using the JavaScript language. I can write code once and run it on either the client or the server. I do not need to add any more languages to my codebase beyond CSS, HTML, and JavaScript. Fewer languages means fewer interfaces between languages and less time debugging those connections. As the JavaScript focus exemplifies, the Meteor server provides the development environment I need to rapidly build a research project.

I used the time I saved by choosing Meteor to build the visualization page using KineticJS. KineticJS is a library that enables drawing complex, interactive shapes in the HTML canvas element. [39] I chose KineticJS over other libraries because its API provides the best tools for drawing novel, high-dimensional visualizations. Other libraries are too opinionated and focus too much on visualizations which do not effectively explain high-dimensional data sets stored in relational databases. [8] Developers who had previously compared KineticJS to other JavaScript drawing libraries came to similar conclusions. [13] They also recommended KineticJS because it had much better documentation than other libraries. Unexpected, well-documented features such as KineticJS's ability to separately draw different layers in an image allow me to more efficiently render the visualizations. The KineticJS APIs and documentation make it easy to utilize such optimizations without writing them myself. Features such as these make KineticJS the optimal graphics library for data visualization.

The Papa Parse library sped the development of Beyond3D's data upload page. This page accepts a CSV file containing a data set from the user's computer, converts it into a JavaScript object using the Papa Parse library, [26] and saves the object in the MongoDB database. The CSV file must have the following format for my system to process it: each row is a data point, each column is a dimension, and the first row is reserved for the names of each dimension. The visualization page reads the data set objects from the database. This is a simple part of my system that heavily relies on the Papa Parse library. Like the other, non-visualization components of my system, I designed it as quickly as possible so that I could spend most of my development time on the visualizations.

Chapter 4

User Study

In this chapter, I demonstrate that Beyond3D is reasonably successful at explaining high-dimensional data sets to non-technical users with domain-specific knowledge. I do this by analyzing the results of a study in which test subjects answered questions based on Beyond3D's visualization of a data set. After reading about this study, the reader will understand both Beyond3D's ability to function in practice and the improvements that I made to the system in response to user feedback.

The reader should note that I used a different color scheme in the version of Beyond3D given to test subjects. In this, older version, I converted response dimension values to colors without binning. Figure 4.1 demonstrates the old color range. I changed Beyond3D's color scheme from the one in figure 4.1 to the one in figure 3.1 in response to feedback from Professor Ben Shneiderman, an expert in the field of data visualization. [42] [1]

Figure 4.1: The old value-to-color mapping

[1] He suggested that I convert the gradient into five or six discrete colors. However, I wanted a finer binning process. I thought that users would struggle to differentiate adjacent bins if I used 10 of them and assigned each one a color from the blue-to-red gradient. Therefore, I chose the current scheme with very different colors.

4.1. Methodology

The user study had three parts: gathering users, building an effective questionnaire, and distributing it to users. My finance advisor, Sebastian Maass, and I collaborated to recruit users. I identified Master in Finance candidates as ideal test subjects due to their non-technical backgrounds and domain-specific knowledge in the field of finance. Since Mr. Maass is a candidate, he recruited 12 of his friends to take the questionnaire. The seven who completed it formed my user base.

I designed the questionnaire to gather subjective and objective feedback that I could use to evaluate Beyond3D. The test subjects provided subjective feedback using text fields at the end of the questionnaire. Through these fields, they made general comments about Beyond3D such as suggesting which visualizations required the most improvements. The questionnaire gathered objective feedback using four questions which each had unambiguously correct and incorrect answers. There were two questions for the polar parallel coordinates visualization, one for the parallel coordinates visualization, and one for the radar chart visualization. I designed these questions around a data set consisting of the daily returns of 14 asset classes, the daily ratings for ABC, and the daily high temperature in NYC. The response dimension was the daily returns of the US Lg Cap Growth asset class. Since the test subjects had domain-specific knowledge about the financial data set, Beyond3D's performance in the questionnaire was indicative of its effectiveness for other target users.

The process for creating the data set had two steps. First, I randomly generated values for the daily returns of the US Lg Cap Growth asset class, the daily ratings, and the daily high temperature. I intended the ratings and temperatures to be random noise. Then, I generated values for the daily returns of the other 13 asset classes using random number generators that were correlated with the US Lg Cap Growth returns to the degree described by Morningstar. I created these correlated random number generators in LibreOffice Calc using a technique described by Deevy Bishop. [9] The technique only functioned correctly after I removed the NORMSINV function call from the formula given at the bottom of the blog. I verified that my process produced a data set which

matched Morningstar's correlations to a reasonable degree using Calc's CORREL function.

I distributed the user study by email. I sent the test subjects an email with a link to Beyond3D, a link to a readme explaining the site,[2] the email address and password for an account on Beyond3D to which I had already uploaded the sample data set, and a form containing the questions. While only 7 of the 12 users responded, I consider this enough data to draw conclusions from because Ben Shneiderman recommended that I only needed a small study to test Beyond3D. [42]

4.2. Polar Parallel Coordinates Questions

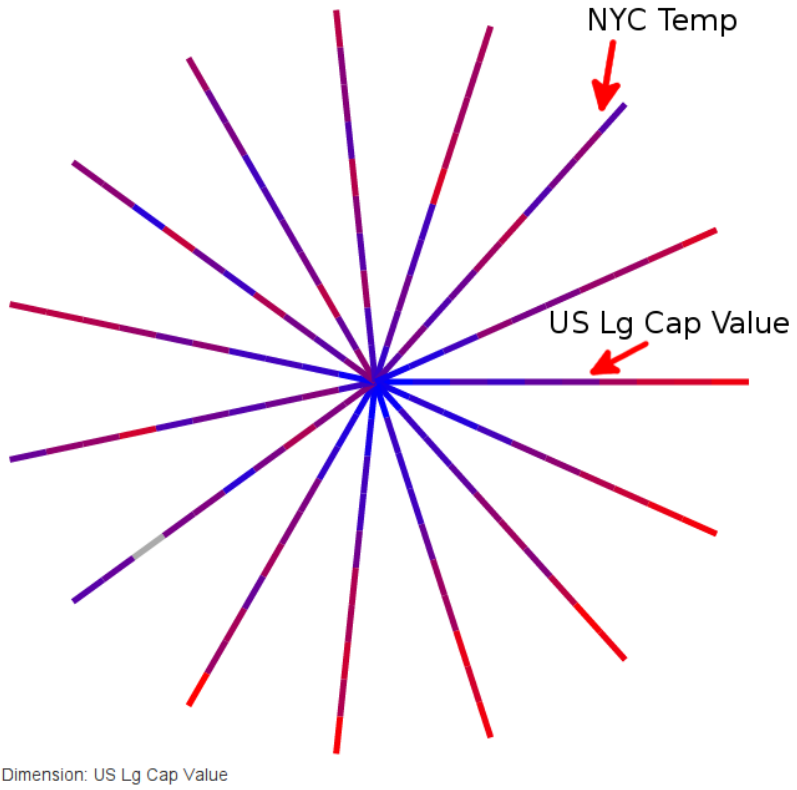


Figure 4.2: The polar parallel coordinates visualization used to answer the questionnaire

Users answered the first two questions using the visualization shown in figure 4.2.[3] The questions required users to label the following statements as true or false. They were also given the option to

[2]Go to http://ndimvis.herokuapp.com/readme to see the readme distributed with the questionnaire. All but one of the test subjects had no previous experience with or training for Beyond3D other than the readme.

[3]The labels and arrows in figure 4.2 were not present for the users in the study. They were able to identify the dimensions by moving their cursors over the axes. Since that functionality does not exist in a paper, I added the arrows and labels for the reader's benefit.

say that they weren't sure of the answer instead of guessing:

1. NYC Temp is strongly related to the response dimension.[4]

   Hint: use the Polar Parallel Coordinates visualization.

2. The US Lg Cap Value asset class is strongly related to the response dimension.

   Hint: use the Polar Parallel Coordinates visualization.

Users' answers demonstrate that they understood the polar parallel coordinates visualization. The first statement is false and the second one is true. Six out of the seven users correctly answered both questions. The other user answered the first one correctly and was unsure of the answer to the second question. Based on these answers, I decided not to significantly alter the visualization.

4.3. Parallel Coordinates Question

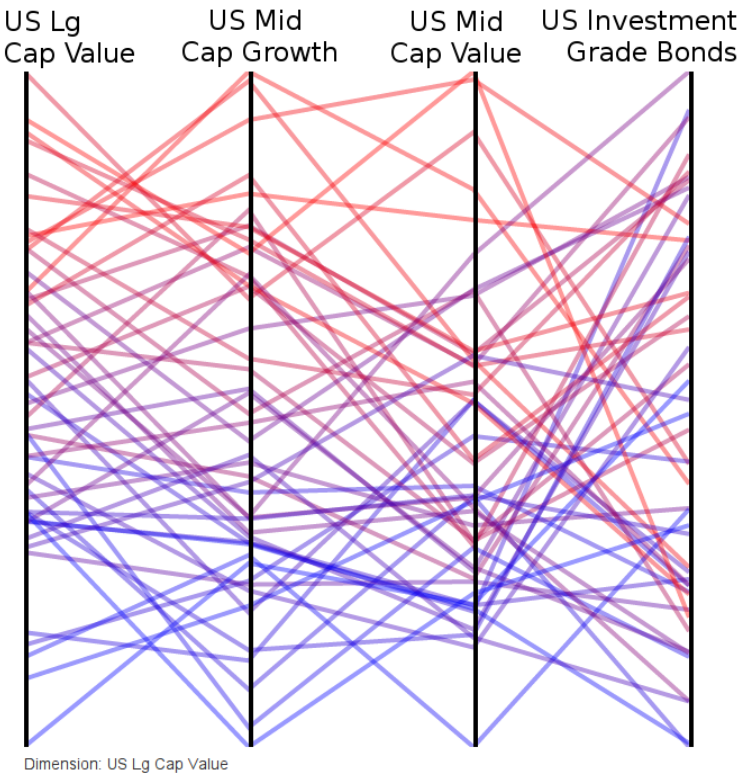


Figure 4.3: The parallel coordinates visualization used to answer the questionnaire

[4]There were a few, minor errors in the questionnaire. I corrected them for this paper.

Users answered the third question using the visualization shown in figure 4.3.[5] The question was: Of the dimensions US Lg Cap Value, US Mid Cap Growth, US Mid Cap Value, and US Investment Grade Bonds, which pairs are strongly related? The question had the hint: create a Parallel Coordinates visualization with the four categories in this question. The users answered by selecting any combination of the following choices:

1. US Lg Cap Value and US Mid Cap Growth
2. US Mid Cap Growth and US Mid Cap Value
3. US Mid Cap Value and US Investment Grade Bonds
4. US Lg Cap Value and US Investment Grade Bonds

Users' answers demonstrate that they completely understood the polar parallel coordinates' interactive features but only partially understood the parallel coordinates visualization. The correlations for each of the four choices were 0.65, 0.58, 0.17, and 0.30. Therefore, I consider the first two to be the correct choices. Every user selected at least one of the correct choices and only one selected an incorrect choice. However, only two users selected both correct choices.[6] The fact that users consistently selected at least one correct choice means that almost all of them could successfully render the parallel coordinates visualization. Causing the visualization to render required clicking on the axes in the polar parallel coordinates visualization. Therefore, the users must have understood the interactive features of the polar parallel coordinates visualization. However, they did not totally understand the parallel coordinates visualization because only two users got the question completely correct.

I redesigned the parallel coordinates visualization to make it easier for users to understand. The version in the user study drew a line for every data point. This made the visualization too cluttered for users to find trends. The current version bins data points into at most 15 lines. The smaller number of lines enables users to more easily detect trends.

[5] As above, I added the labels and arrows in figure 4.3 for the reader's benefit.

[6] One of the users who gave only one, correct answer stated in the subjective feedback section that they could not get the parallel coordinates visualization to work. I discount his issue because none of the other users reported having trouble rendering the parallel coordinates visualization.

4.4. Radar Chart Question

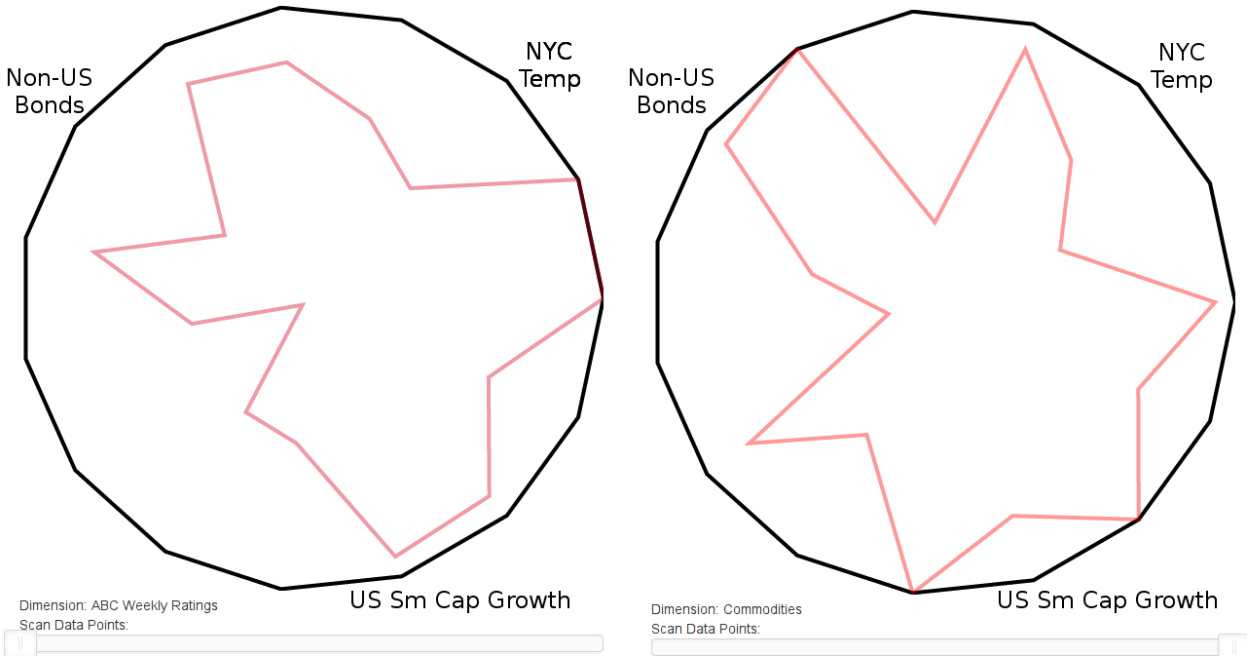


Figure 4.4: The radar chart visualizations that should've been used to answer the questionnaire

Users should've answered the fourth question using the visualization shown in figure 4.4.[7] The question was: For the two data points with the largest values in the US Lg Cap Value dimension, which one of the following dimensions do both of the data points also have large values? The question had the hint: create a Polar Chart[8] visualization by making a filter on the US Lg Cap Value dimension in the Parallel Coordinates visualization from question 3. The users answered by selecting any one of the following choices:

1. NYC Temp
2. US Sm Cap Growth
3. Non-US Bonds

Users' answers demonstrate that they did not understand the parallel coordinates' interactive features. Only one user chose the correct answer, US Sm Cap Growth, without also reporting that the radar chart malfunctioned. Every other user stated that they had issues rendering the visualization. I

[7] As above, I added the labels in figure 4.4 for the reader's benefit.
[8] I called the radar chart the polar chart in this version of Beyond3D.

know that the feature technically worked on users' browsers. For example, I showed one user how to create the radar chart on his computer after he reported that the feature malfunctioned as he took the questionnaire. Thus, the issue was a UI problem rather than a technical one.

I addressed the UI problem by making it easier to draw filters in the parallel coordinates visualization. Users no longer need to be as precise when clicking on the axes. They can click close to an axis to begin drawing a filter. Also, future readmes will better explain the feature so that users understand where to click to draw the filters. While future users will be able to view the radar chart, I can't draw any conclusions about how to improve it because only one user successfully reached that part of the questionnaire.

Chapter 5

Case Study

In this chapter, I carry out a case study to demonstrate the effectiveness of Beyond3D at explaining high-dimensional data sets to users who lack technical training but have domain-specific knowledge and completely understand the system. Unlike the user study's test subjects who had no previous experience with Beyond3D, I play the role of a non-technical investor who thoroughly understands it. The investor applies the system to the problem of decreasing losses during stock market crashes. The case study assumes that the investor lacks knowledge of advanced data analysis techniques but has some financial knowledge. Additionally, the investor only needs Beyond3D, widely available data, and the holding period return and financial turbulence metrics. Most investors can understand these metrics as they require only a basic knowledge of statistics. After reading this chapter, the reader will understand the power of Beyond3D when utilized by a target user who has experience with the tool.

Investors can gain a lot from tools that modify their investment portfolios to minimize losses during crashes. The S&P 500 index lost 22% of its value over six trading days in October 2008. [48] Investors must be creative to minimize losses during such crashes as the S&P 500 contains so many influential companies that the rest of the US equity market typically reflect its movements. [19] If the index performs poorly, then many portfolios will lose money regardless of their specializations within the US equity market. Tools for decreasing losses can protect these portfolios.

Previous research has focused on minimizing losses for technical investors while ignoring those with weaker mathematical skills. The methods put forth by researchers such as Kritzman and Li use numerical techniques for analyzing high-dimensional data sets stored in relational databases. As I explained earlier in my paper, these techniques are too complex for non-technical users. Kritzman and Li require multiple pages of dense mathematics to explain their combinations of financial turbulence with mean-variance and full-scale portfolio optimization techniques. [32, pp. 37, 39-40] Previously developed visualization techniques also fail non-technical investors. Like the techniques I explored in previous chapters, financial visualizations do not scale to the appropriate number of dimensions. [50] Non-technical investors need a method for analyzing these data sets and minimizing losses.

Beyond3D fulfills the role of helping non-technical investors understand and utilize metrics such as financial turbulence. Beyond3D's combination of interactive visualizations can explain high-dimensional data sets to users and enable them to derive actionable investment strategies. However, the reader must recognize that the combination of Beyond3D and financial turbulence is not strictly better than the optimization techniques of Kritzman and Li. I am not offering a method for predicting the future nor one for optimal loss minimization. Additionally, the Beyond3D approach will not be as rigorous as traditional, numerical techniques. It requires users to interpret visualizations. Each user may have a unique interpretation. Despite these flaws, Beyond3D enables non-technical users to understand high-dimensional data sets of metrics such as financial turbulence. They can use this understanding to make better investment decisions that will lead to reasonable decreases in losses.

I demonstrate the success of the Beyond3D and financial turbulence technique by showing how its portfolio alterations outperform the S&P 500 during the worst market crashes. Since the technique successfully avoids some losses under these conditions, it should also work during less significant downturns. I focus on the stock market crashes of 1987 and 2008 as they are the two largest in the US over roughly the last 30 years. [7] The alterations' successful performances during major crashes that are separated by significant periods of time show that the Beyond3D visualization

technique can be used to aid the creation of crash-resistant portfolios in the future.

5.1. Financial Metrics

The case study's investor uses Beyond3D to visualize two metrics: daily holding period return and financial turbulence. Holding period return (HPR) is a percentage that expresses how much money an investor gains or loses for owning a stock over a period of time. These changes include both price movements and income such as dividends. [28] Thus, HPR takes into account issues such as the drop in a stock's price after a company issues a dividend. A simpler metric such as percent change in stock price would state that investors lost money and ignore the dividend. Daily HPR enables Beyond3D to accurately display the total return an investor would have received for holding each stock every day.

Financial turbulence helps investors understand how long HPR will remain negative following a market crash. Conceptually, financial turbulence is "a condition in which asset prices, given their historical patterns of behavior, behave in an uncharacteristic fashion, including extreme price moves, decoupling of correlated assets, and convergence of uncorrelated assets." [32, p.30][1] [2] Figure 5.1 shows how financial turbulence increases during market crashes.

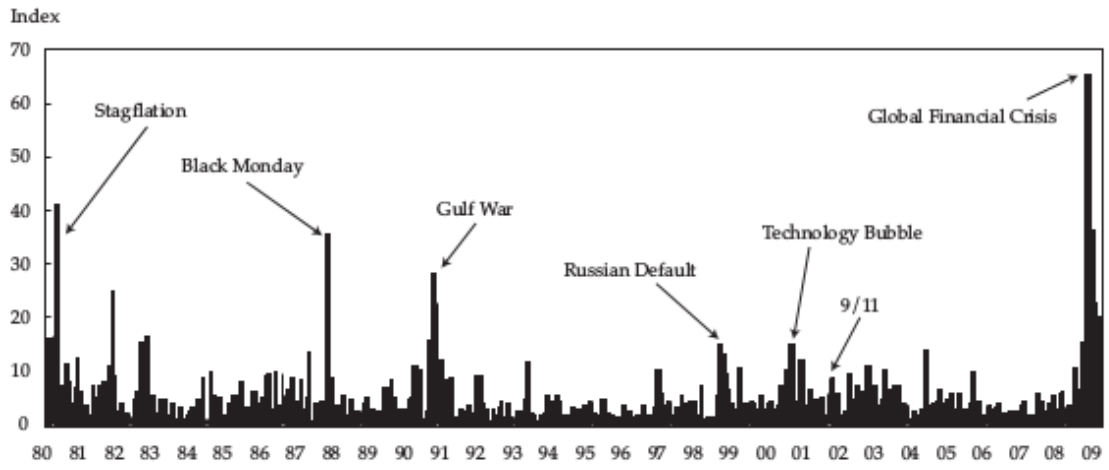


Figure 5.1: Financial turbulence over time [32, p.34] [3]

[1] Stocks are the only type of asset that I use in this paper. While the Kritzman and Li paper defines the financial turbulence metric for all asset prices, I only apply it to stock prices.

[2] This is the section of the case study that requires some mathematical knowledge. I assume that most investors have a basic knowledge of statistics that includes z-scores.

The mathematical interpretation, the turbulence index, is a type of distance function that measures how far asset prices are from the average day. [32, pp.31,34] A day's turbulence index is the "multivariate z-score" [30] of its stock returns. Larger turbulence index values indicate that a day's stock returns are unusual in terms of terms of magnitude and correlation. The formula for turbulence index is[4] [5]:

$$d_t = (y_t - \mu)\Sigma^{-1}(y_t - \mu)' \tag{5.1}$$

$$d_t = \text{turbulence for a particular time period t (scalar)} \tag{5.2}$$

$$y_t = \text{vector of asset returns for period t (1 } \times \text{ n vector)} \tag{5.3}$$

$$\mu = \text{sample average vector of historical returns (1 } \times \text{ n vector)} \tag{5.4}$$

$$\Sigma = \text{sample covariance matrix of historical returns (n } \times \text{ n matrix)} \tag{5.5}$$

Financial turbulence's persistence and correlation with negative returns make it a tool for avoiding continued losses. As Kritzman and Li note, stocks generally perform poorly during financially turbulent times regardless of which area of the market increases the turbulence index. [32, p.34] Therefore, investors can treat financial turbulence as a proxy for poor portfolio performance. Investors who could predict financial turbulence would be able to avoid the losses by moving to safer stocks before their portfolios lost money. Unfortunately, investors cannot predict when turbulence will increase. However, highly turbulent days are likely to be followed by other turbulent days. [32, pp.34-5] An investor who tracks the turbulence index will lose money on the first highly turbulent day, but can act to improve returns in the future. During the six day crash of October 2008, the investor would have recognized the high value of financial turbulence on the first day, adjusted

[3] The chart clearly shows that the largest values of financial turbulence coincide with the periods labeled Stagflation, Black Monday, Gulf War, and Global Financial Crisis. Other periods of financial distress have financial turbulence values which are difficult to differentiate from noise, such as 9/11. Nevertheless, this chart demonstrates that many financial crises and spikes in financial turbulence occur at the same time.

[4] The formula and definitions are quoted from [32, p. 31].

[5] I only use daily financial turbulence. Therefore, t is always 1 day.

his portfolio, and prevented losses on the following five days.[6] [48] Beyond3D's visualizations provide non-technical investors a way to act on the financial turbulence metric by visually explaining its relationships with stocks' HPRs.

I use financial turbulence instead of other metrics of financial risk because it offers a more complete perspective on the state of the market. Other metrics, such as VIX, measure risk without a historical perspective. Each day's VIX value is a measure of market volatility derived from options on the S&P 500. Each value is not related to previous ones. [29] Financial turbulence provides a perspective that is calibrated with historical values. It is a z-score, so it is explicitly calculated using previous stock returns. Kritzman and Li demonstrate that this more complete perspective translates into better results. They show that investment strategies based on financial turbulence have better returns than those using other metrics. [32, p.38] Thus, I pair Beyond3D with the best metric for minimizing losses, financial turbulence.

5.2. Application of Beyond3D and Financial Turbulence

5.2.1. 2008 Crash

An investor analyzing financial turbulence on the morning of September 19th, 2008 would have decided to immediately alter his portfolio to handle a market crash. Financial turbulence on the 15th was higher than on any day between 9/18/06 and 9/12/08.[7] The value on the 17th was within 0.21% of the maximum value between 9/18/06 and 9/12/08. Finally, financial turbulence on the 18th was about 33 times the mean value from 9/18/06 to 9/12/08 and roughly 90% greater than the maximum value over that time period. The investor would have recognized the consistently high values ending with an unmistakable outlier on 9/18/08 and decided that he needed to alter his portfolio to decrease losses during a market crash.

The investor would have taken three actions with Beyond3D to find stocks to protect him from

[6] Financial turbulence may be able to predict when to reenter the market. Krtizman and Li are unclear on whether low financial turbulence is as persistent as high financial turbulence. [32, p.34-5] Regardless, the issue is beyond of the scope of this chapter because I am focusing on minimizing losses during crashes, not reentering the market to take advantage of rallies.

[7] 9/12/08 is the last trading day before the 9/15/08.

the 2008 crash. The first step is comparing the financial turbulence of each day over the past two years[8] with the mean daily HPR of every stock in the S&P 500 over the following 10 days.[9] [10] [11] By comparing these metrics, the investor could have found stocks which performed well following days with high financial turbulence. Figure 5.2 shows how Beyond3D displays such a comparison in the polar parallel coordinates visualization. Each axis is a stock. Each day is binned for an axis according to the mean of the stock's daily HPRs over the following 10 days. Each color represents the average financial turbulence of the days in a bin. I have filtered out the data points with turbulence values in the bottom 5 bins as the investor only would have been interested in how stocks perform following highly turbulent days. This removes the colors from bottom half of the bin-to-color mapping in figure 3.1. The investor would have looked for axes with gray near the center, meaning that the stocks never performed poorly following turbulent days, and blue and purple near the edges, meaning that the stocks performed well following turbulent days. If he found areas of the visualization that were interesting, he would need to see them in more detail to identify individual stocks.

[8] While I would like to include more data, Beyond3D cannot handle it. Beyond3D is a research system whose technology stack sends lots of data over the web. The system cannot store and process larger data sets.

[9] The investor would not have considered the most recent days of the last two years. These days would have had less than 10 days of HPR data available and so I removed them to ensure that every HPR calculation had the same amount of data.

[10] The investor could have looked at 10-day HPRs rather than the means of the daily HPRs. However, one does not need to hold a stock for 10 days after purchasing it. Therefore, I think that the investor would have found an average of a stock's performances to be more useful than an exact measure of its return over a period of time. [17]

[11] The number of days in the HPR calculation is arbitrary.

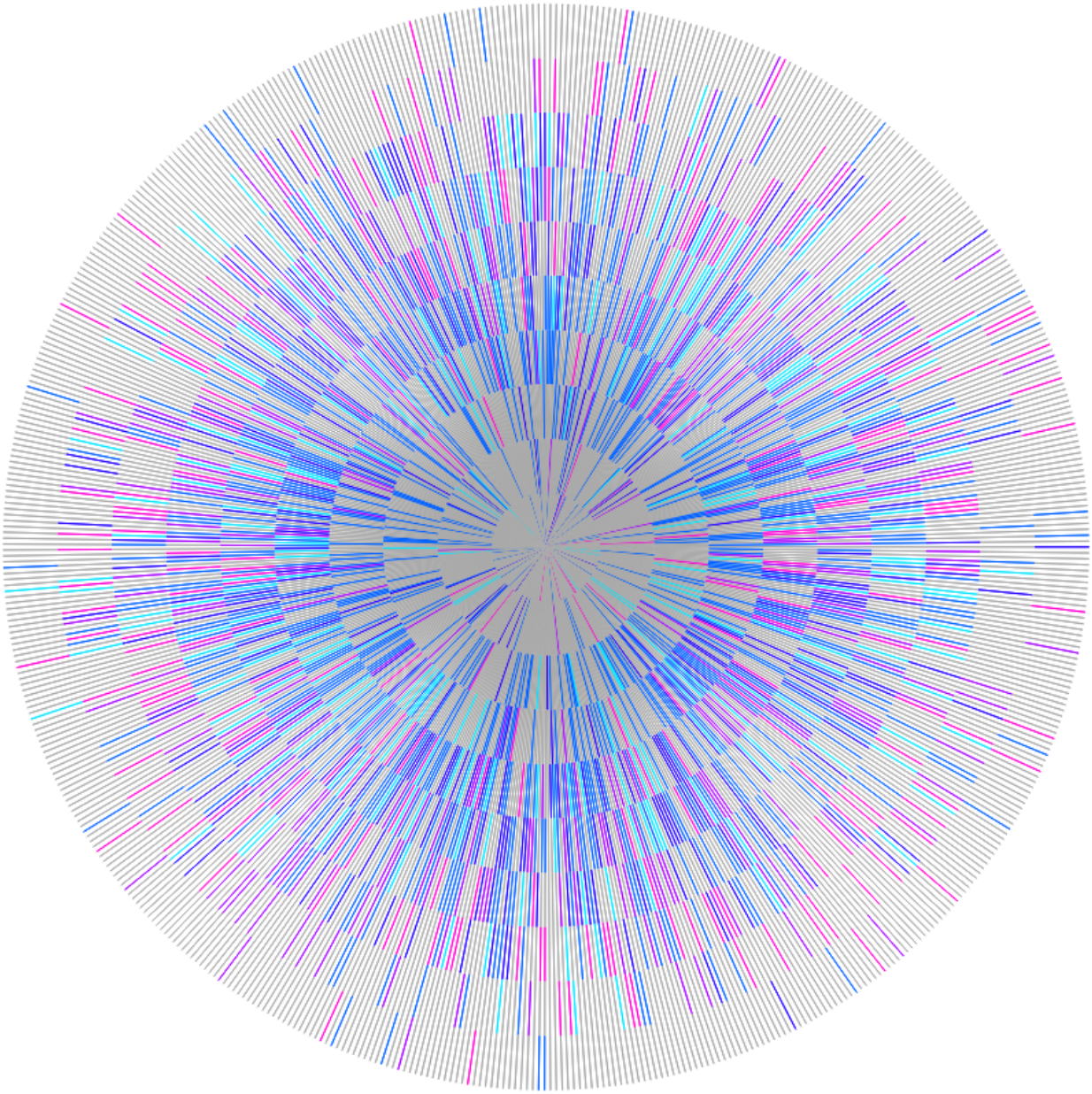


Figure 5.2: Beyond3D's visualization of HPR values and financial turbulence

The second step is zooming into the axes of interest as shown in figure 5.3. The figure clearly shows that Intel performed well following turbulent days. When I followed this step, I found six companies of interest: MasterCard Inc., Intel Corp., International Business Machines Corp., Medco Health Solutions Inc., Xcel Energy Inc., and Compuware Corp.[12] The investor would have found a

[12]I used Google Finance and Search to correct these capitalizations. My data set only used capital letters for companies' names.

similar set of companies as he was looking at the same visualization.

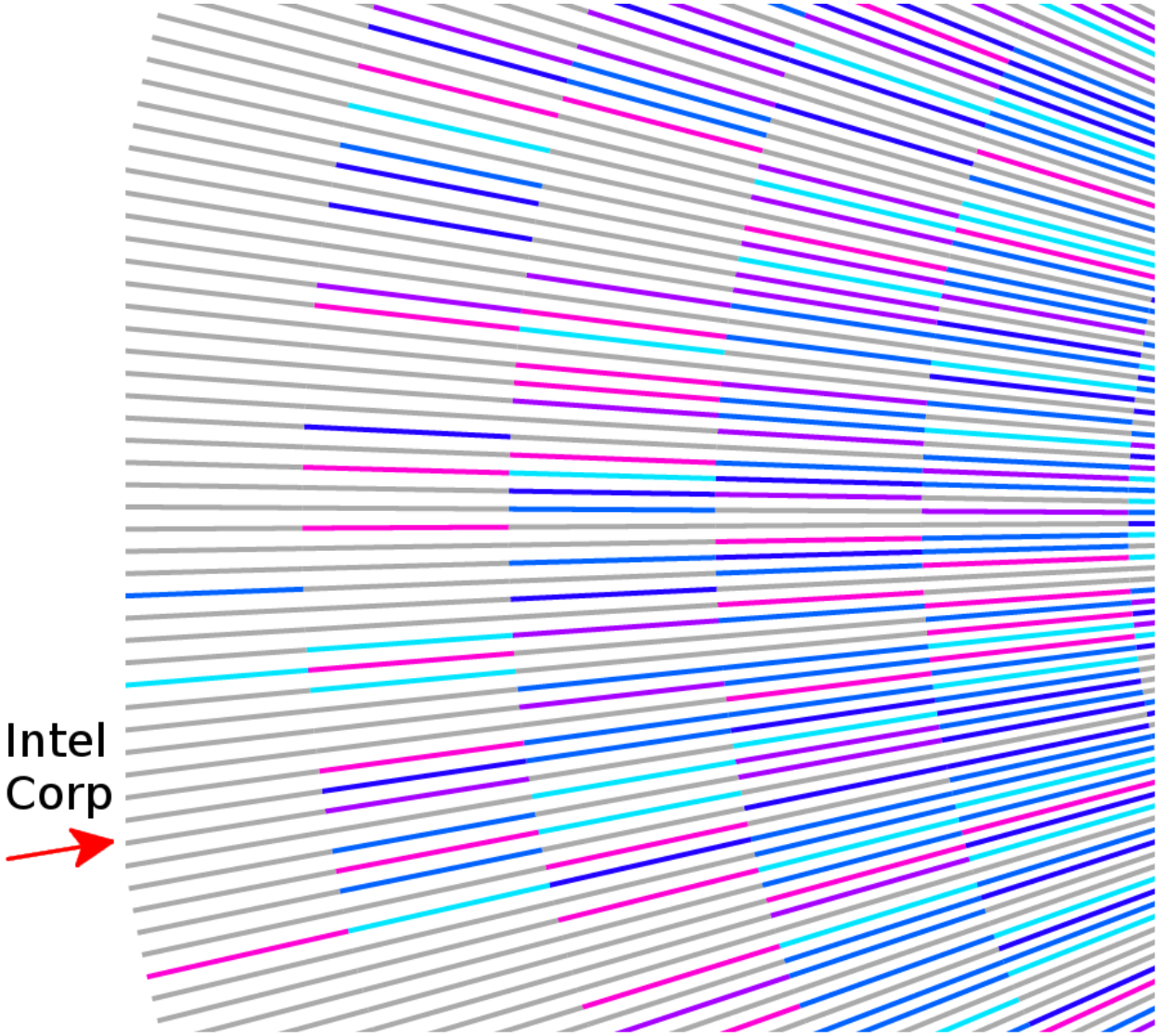


Figure 5.3: A more focused view of the data from figure 5.2

The last step of the analysis is comparing the companies from step two using the parallel coordinates visualization. Figure 5.4 shows such a comparison.[13] This is the final step of the analysis and, among my six companies, no one seems strictly worse than the rest. The investor would have reached a similar conclusion and altered his portfolio to include these six companies.

[13] I added the labels in figure 5.4 for the reader's benefit.

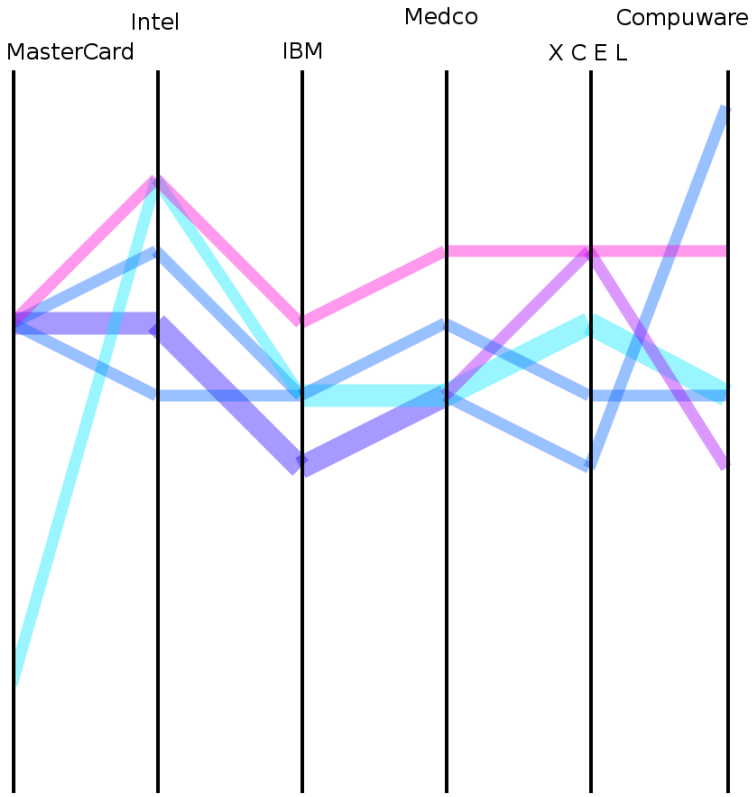


Figure 5.4: Beyond3D's visualization of the HPR values of the 6 companies

The investor's portfolio alterations outperformed the S&P 500 over the 15 trading days starting on 9/19/08. The alterations outperformed the S&P 500 from 9/19/08 to 10/9/08 by 6.2 points, a -21.9% return compared to a -28.1% one. As seen in table 5.1, every stock outperformed the S&P 500 over that period of time. The histogram in figure 5.5 further emphasizes the success. This is a histogram of the returns of every stock in the S&P from 9/19/08 to 10/9/08. The green line is the average of the returns. The red rectangle shows the area of the histogram that contains the stocks selected by the investor. Every stock outperformed a significant portion of the market. This distribution emphasizes that the investor's choices would have decreased losses during the 2008 crash.

Individual or Collection of Stocks	Return
S&P 500	-0.28093
Average of Portfolio Alterations	-0.21855
MasterCard Inc.	-0.25266
Intel Corp.	-0.19003
International Business Machines Corp.	-0.22689
Medco Health Solutions Inc.	-0.22669
Xcel Energy Inc.	-0.18122
Compuware Corp.	-0.23381

Table 5.1: The returns of the portfolio alterations compared to those of the S&P 500 over the 15 trading days starting on 9/19/08

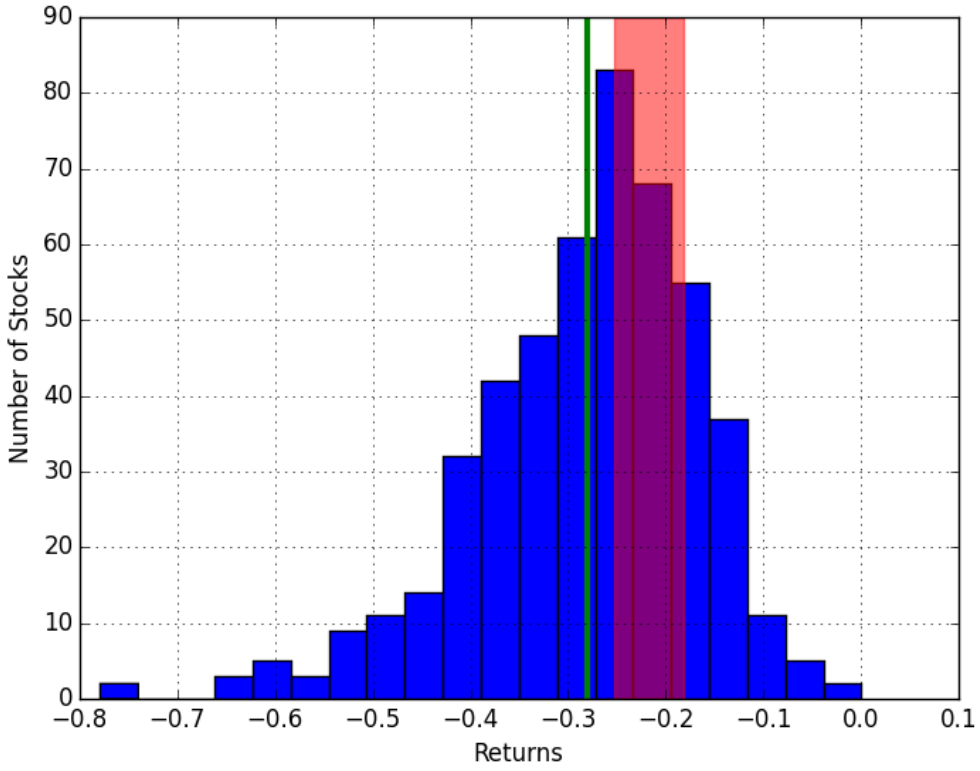


Figure 5.5: The returns of the stocks selected by the investor compared to those of every stock in the S&P 500 over the 15 trading days starting on 9/19/08

5.2.2. 1987 Crash

An investor analyzing financial turbulence on the morning of October 19th, 1987 also would have decided to alter his portfolio. Financial turbulence on the 14th and 15th were about 9 and 7 times higher than the mean value from 10/16/85 to 10/13/87. Financial turbulence on the 16th was about 35 times the mean value from 10/16/85 to 10/13/87 and roughly 27% greater than the maximum value over that period of time.[14] As with the 2008 crash, the investor would have seen the clear trend of high financial turbulence and recognized that he needed to protect his portfolio against losses.

To protect his portfolio against the 1987 crash, the investor could have taken the same three actions with Beyond3D as I describe in the 2008 crash section. Figure 5.6 and 5.7 show the visualizations of the data. For these visualizations, I allowed in some less turbulent days. There were too few data points with turbulence values in the upper bins to draw conclusions without the more calm days. This is the reason why some of the colors from figure 3.1 for the lower bins appear in the visualization. Nevertheless, I was still able to identify six companies that the investor could have added to his portfolio: Cross & Trecker Corp., The Coca-Cola Co., Dun & Bradstreet Corp., Echlin Inc., Southern Co., Amerada Hess Corp.

[14] 10/16/87 is the last trading day before the 10/19/87.

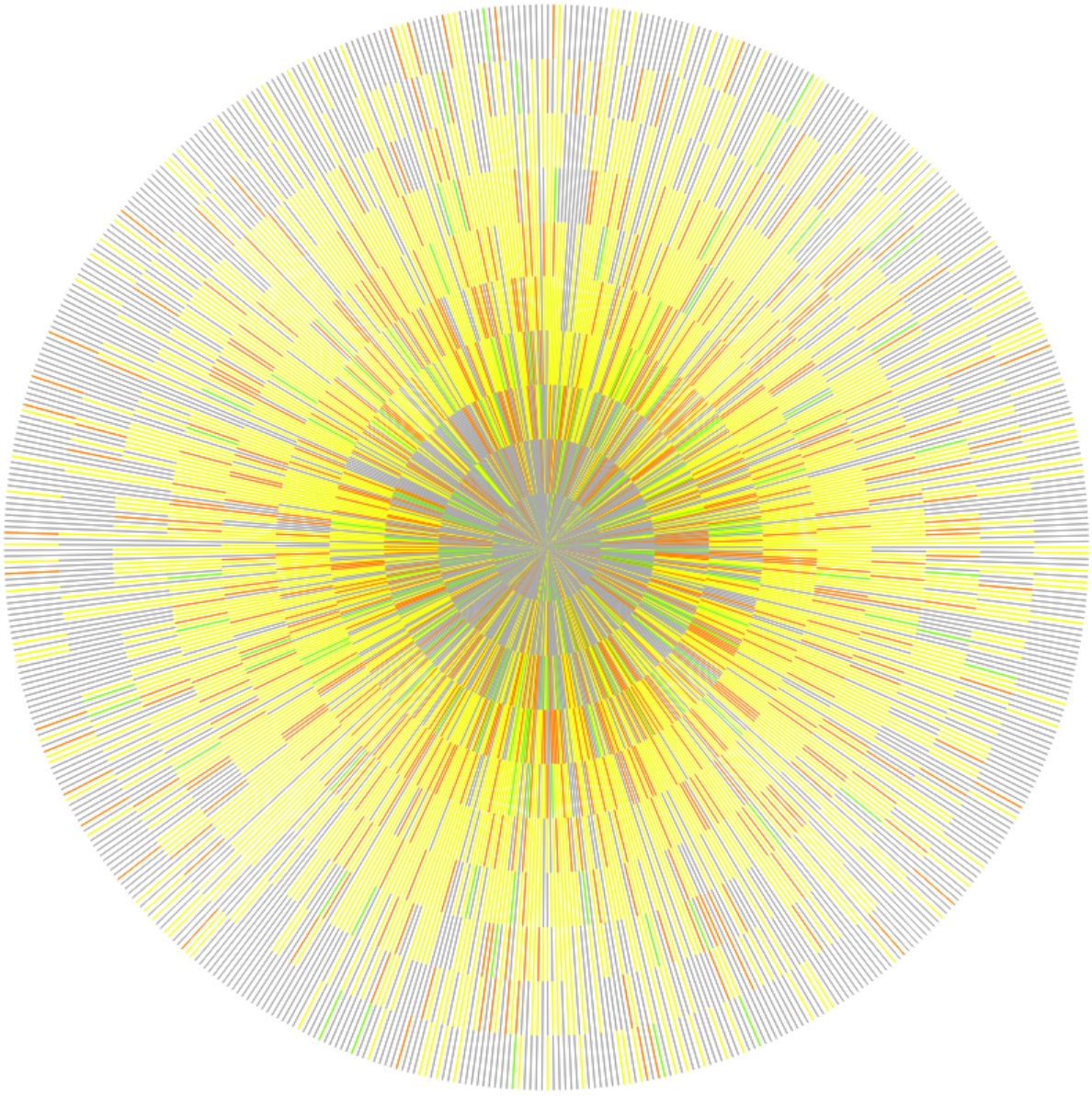


Figure 5.6: Beyond3D's visualization of HPR values and financial turbulence

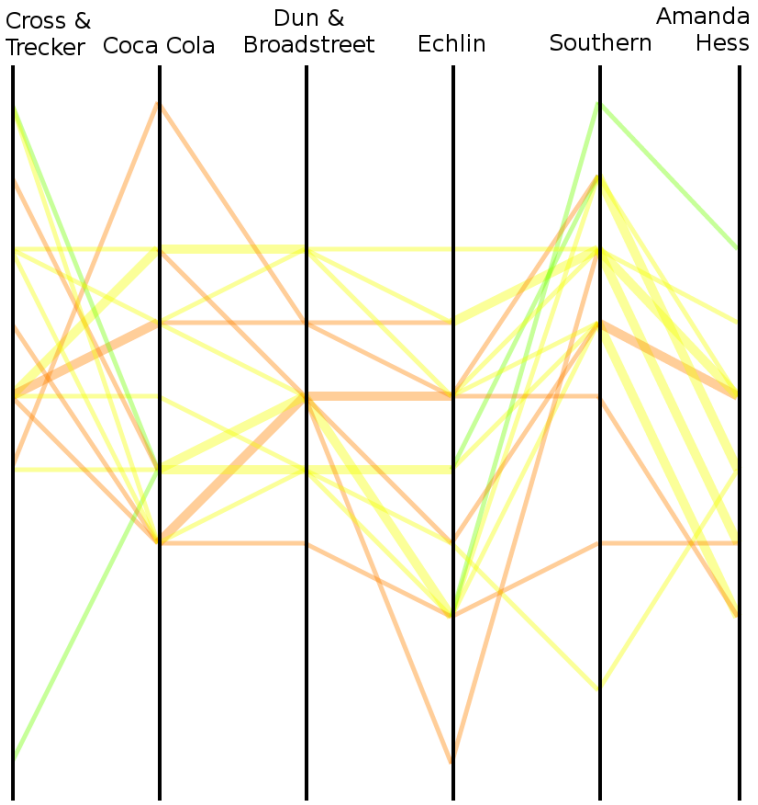


Figure 5.7: Beyond3D's visualization of the HPR values of the 6 companies

The investor's portfolio alterations moderately outperformed the S&P 500 over the 15 trading days starting on 10/19/87. The 1987 alterations outperformed the S&P 500 from 10/19/87 to 11/6/87 by 4.4 points, a -10.0% return compared to a -14.4% one. As seen in table 5.2, four out of the six stocks outperformed the S&P 500 over that period of time. Only one of the six seriously underperformed the S&P 500. The histogram in figure 5.8 confirms this reasonable success. This is a histogram of the returns of every stock in the S&P from 10/19/87 to 11/6/87. The green line and red rectangle have the same meanings as in figure 5.5. I consider the investor's alterations a success because only one laggard, Cross & Trecker Crop., dragged down performance. Most of the stocks outperformed the S&P 500, and the best choices were significantly farther above the mean than the worst ones were below it. Thus, the combination of Beyond3D and financial turbulence would've helped the investor.

Individual or Collection of Stocks	Return
S&P 500	-0.14433
Portfolio Alterations	-0.09995
Cross & Trecker Corp.	-0.20313
The Coca-Cola Co.	-0.01852
Dun & Bradstreet Corp.	-0.13152
Echlin Inc.	-0.13281
Southern Co.	0.03668
Amerada Hess Corp.	-0.15040

Table 5.2: The returns of the portfolio alterations compared to those of the S&P 500 over the 15 trading days starting on 10/19/87

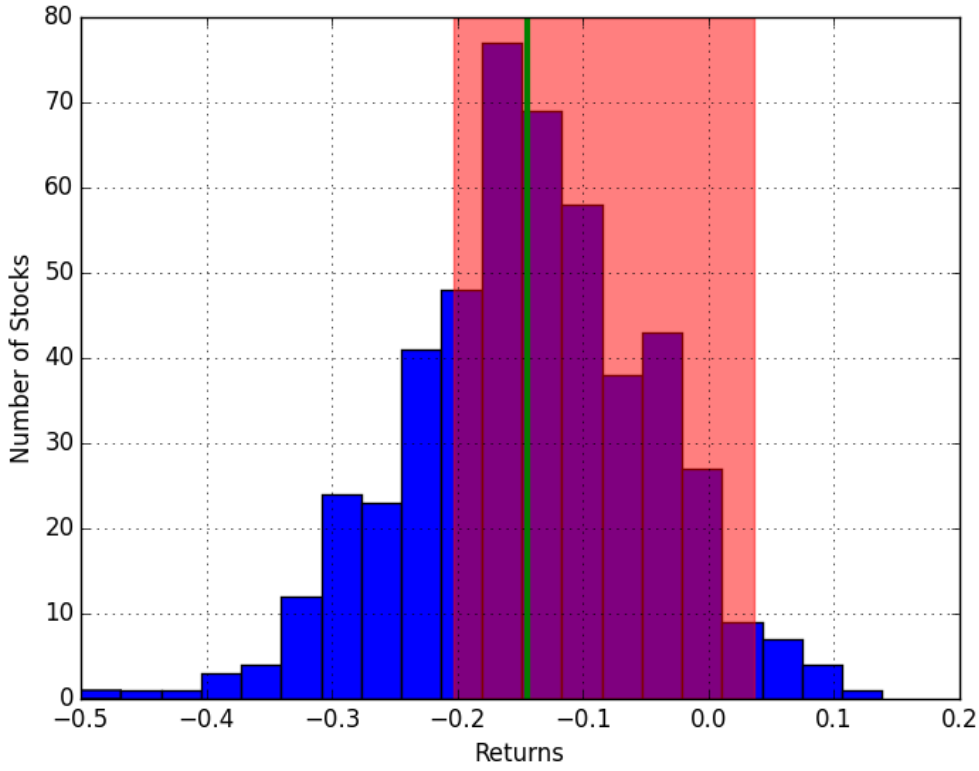


Figure 5.8: The returns of the stocks selected by the investor compared to those of every stock in the S&P 500 over the 15 trading days starting on 10/19/87

5.3. Effectiveness Of Beyond3D and Financial Turbulence

The Beyond3D and financial turbulence technique would have succeeded during the crashes of 2008 and 1987. Both of the technique's portfolio alterations outperformed the S&P 500 over the 15 trading days following their implementations.[15] These improved returns would have helped investors through the tough financial periods of 1987 and 2008.

Investors should feel comfortable using the technique in the future. The 1987 and 2008 crashes demonstrate the flexibility of the Beyond3D and financial turbulence technique. The crashes were large, [7] separated by long periods of time, and occurred for different reasons. [36] Future crashes shouldn't be too large for the Beyond3D and financial turbulence technique to develop portfolio alternations that decrease losses. Also, the unique situations that cause future crashes shouldn't frustrate the technique. As Kritzman and Li note, "returns to risk are substantially lower during turbulent periods than during nonturbulent periods, irrespective of the source of turbulence." [32, p.34] Figure 5.1 confirms this quote. Financial turbulence rose when markets performed poorly due to diverse factors such as the Gulf War and the Global Financial Crisis. Thus, Beyond3D and financial turbulence can help investors decrease losses during crashes, regardless of the size or cause,

[15] I calculated the returns of the S&P 500 by evenly weighting all the stocks in the index. While the actual S&P 500 uses a different system for weighting stocks [24], I used this scheme for consistency with how I calculated returns for my portfolio alterations. I evenly weighted all six stocks for both alterations.

Chapter 6

Conclusion

My thesis makes two contributions: a user-oriented process for researching high-dimensional visualizations and the resulting Beyond3D tool. I have demonstrated the value of Beyond3D for users without strong technical backgrounds. The above two studies exemplify how such users can apply the tool to high-dimensional data sets to derive actionable information. Both the Master in Finance candidates and the hypothetical investor used the tool to understand a high-dimensional data set. Then, they made mostly correct decisions based on this knowledge: the candidates performed well on three of the four questions in their questionnaire; the investor's portfolio alterations outperformed the S&P 500. Their successes mean that Beyond3D fulfilled its initial goals.

The development process that built Beyond3D consists of creating a prototype and then repeatedly testing it with users and making improvements using their feedback. The development of the parallel coordinates visualization exemplifies this process. The prototyping stage began with reading previous papers on the visualization and sketching out a few examples of my own implementation. I then built a functional prototype. My advisor tested the initial version and stated that it only made sense after I verbally explained it. I realized that this feedback applied to every visualization and responded with a comprehensive, illustrated readme.[1] The next iteration of the testing-development loop began with the user study in chapter 4. I redesigned the entire parallel coordinates visualization

[1] Chapter 3 is a descendant of this document.

based on their feedback. The new version hid some data points and binned others together rather than displaying every one individually so that users could identify important trends. The updated interactive features required less precise user input and had a clearer written explanation. The process of listening and reacting to user feedback culminated in the case study. Pretending to be a non-technical user, I applied the parallel coordinates visualization as part of a successful analysis of a high-dimensional data set. The positive outcome demonstrates that the iterative, user-oriented process leads to effective visualization techniques.

I will use this process to make Beyond3D more accessible for both users and other developers. My development of Beyond3D has taught me that users care most about whether a tool is simple and provides actionable information. Personal experience dictates that developers care about how well a tool integrates with other technologies. Converting Beyond3D into an easy-to-use and widely applicable platform will take three steps. First, I will redesign the technology stack to make it handle larger and more varied data sets. Currently, the tool only accepts data sets in which there is one data point per line and one dimension per column. This change will make Beyond3D simpler as users will not need to reformat their data before the tool visualizes it. Second, I will enable users to create and export hypotheses through Beyond3D. I had to build an external system for making and testing my portfolio alterations in the case study. A component that does this in Beyond3D will aid users' abilities to derive actionable information from the visualizations. They will be able to formalize and test their understandings of the data in the same workflow as the visualizations themselves. Finally, I will make Beyond3D more modular. I want to support developers that only need the polar parallel coordinates or parallel coordinates visualization. Making Beyond3D more modular will enable developers to integrate the tool into a wider range of projects. These three improvements to Beyond3D will enable non-technical users and developers to more effectively apply the system to their own work.

A more extreme improvement is expanding the scope of Beyond3D so that technically trained users can combine it with machine learning tools. Beyond3D will provide an overview of the data set. Technically trained users can use this to understand which dimensions are most interesting and

to get a general sense of their relationships. These users will then be able to more effectively apply their own tools, such as PCA. Their understanding of the data set from Beyond3D's visualizations will enable them to intuitively interpret the output of the mathematical techniques. Human intuition combined with mathematical computations will perform better than either one in isolation.

Beyond3D succeeds because it combines human intuition and mathematical computations. It does not perform the most complicated computations, but it does the ones that help human users understand their data. Continuing the user-oriented development process explained in this paper will improve Beyond3D's ability to aid human intuition. The work in this thesis is a step in the process of building a complete tool for addressing the high-dimensional data problem.

Appendix A

Websites I Used In Developing Beyond3D and Writing This Thesis

1. http://docs.meteor.com/#/basic/
2. http://pandas.pydata.org/pandas-docs/stable/
3. http://stackoverflow.com/questions/2998784/how-to-output-integers-with-leading-zeros-in-javascript
4. http://stackoverflow.com/questions/57803/how-to-convert-decimal-to-hex-in-javascript
5. http://stackoverflow.com/a/13663169
6. http://www.storminthecastle.com/2013/07/24/how-you-can-draw-regular-polygons-with-the-html5-canvas-api/
7. http://stackoverflow.com/questions/17004094/how-to-make-rounded-corners-on-polygon-with-kineticjs
8. http://robertdickert.com/blog/2013/11/14/why-is-my-meteor-app-not-updating-reactively/
9. http://stackoverflow.com/a/23387846
10. http://stackoverflow.com/
11. http://matplotlib.org/api/pyplot_api.html
12. http://pandas.pydata.org/pandas-docs/stable/

Bibliography

[1] "Dictionary.com unabridged," Apr 2015. [Online]. Available: http://dictionary.reference.com/browse/database

[2] "Dictionary.com unabridged," Apr 2015. [Online]. Available: http://dictionary.reference.com/browse/histogram

[3] "glyph," *Merriam-Webster.com*, 2015. [Online]. Available: http://www.merriam-webster.com/dictionary/glyph

[4] G. I. Allen, "Examples of High-Dimensional Data," Jan. 2011. Available: http://www.stat.rice.edu/~gallen/examples_high_dim_data.pdf

[5] M. Allen, "Object oriented databases." Available: http://www.comptechdoc.org/independent/database/basicdb/dataobject.html

[6] Apache Spark, "Apache spark - lightning-fast cluster computing," 2015. Available: https://spark.apache.org/

[7] V. Bajaj and M. M. Grynbaum, "For Stocks, Worst Single-Day Drop in Two Decades," *The New York Times*, Sep. 2008. Available: http://www.nytimes.com/2008/09/30/business/30markets.html

[8] N. Belmonte, "Javascript infovis toolkit," 2015. Available: http://philogb.github.io/jit/

[9] D. Bishop, "BishopBlog: The joys of inventing data," Oct. 2011. Available: http://deevybee.blogspot.com/2011/10/joys-of-inventing-data.html

[10] J. Blaas, C. P. Botha, and F. H. Post, "Extensions of parallel coordinates for interactive exploration of large multi-timepoint data sets," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 14, no. 6, pp. 1436–1451, 2008.

[11] O. Botvinnik, "Implementation of typographic and design principles in matplotlib and ipython notebook," Apr. 2013. Available: http://nbviewer.ipython.org/gist/olgabot/5357268

[12] A. Buja, D. Cook, and D. F. Swayne, "Interactive high-dimensional data visualization," *Journal of Computational and Graphical Statistics*, vol. 5, no. 1, pp. 78–99, 1996.

[13] J. Burton, "Current state of javascript canvas libraries?" 2013. Available: http://stackoverflow.com/a/14527100

[14] J. Dager, "A Quick Tool for Value Analysis." Available: http://business901.com/blog1/a-quick-tool-for-value-analysis/

[15] G. Dallas, "Principal Component Analysis 4 Dummies: Eigenvectors, Eigenvalues and Dimension Reduction." Available: https://georgemdallas.wordpress.com/2013/10/30/principal-component-analysis-4-dummies-eigenvectors-eigenvalues-and-dimension-reduction/

[16] H. Doleisch, M. Gasser, and H. Hauser, "Interactive feature specification for focus+ context visualization of complex simulation data," in *Proceedings of the symposium on Data visualisation 2003*. Eurographics Association, 2003, pp. 239–248.

[17] M. Dumenko, "Financial Returns: Holding Period vs. Arithmetic vs. Geometric," Nov. 2013. Available: http://www.financialanalystwarrior.com/financial-returns/

[18] D. Durst, "N-Dimensional Visualization." Available: http://ndimvis.herokuapp.com/

[19] R. Ferri, "S&P 500: A Great 2nd Place Index Fund." Available: http://www.forbes.com/sites/rickferri/2014/08/28/sp-500-a-great-2nd-place-index-fund/

[20] FusionCharts, "Radar (Spider) Chart." Available: http://www.fusioncharts.com/chart-primers/radar-chart/

[21] Z. Geng, J. Walker, and R. S. Laramee, "Markov Chain Driven Multi-Dimensional Visual Pattern Analysis with Parallel Coordinates," in *Vision, Modeling and Visualization*, M. Goesele *et al.*, Eds. The Eurographics Association, 2012, pp. 191–198.

[22] G. Grinstein, M. Trutschl, and U. Cvek, "High-dimensional visualizations," in *Proceedings of Workshop on Visual Data Mining, ACM Conference on Knowledge Discovery and Data Mining*, 2001, pp. 1–14.

[23] H. Hauser, F. Ledermann, and H. Doleisch, "Angular brushing of extended parallel coordinates," in *Proceedings of the IEEE Symposium on Information Visualization (InfoVis' 02)*. IEEE Computer Society, 2002, p. 127.

[24] K. Hawkins, "S&P 500 ETFs: Market Weight Vs. Equal Weight." Available: http://www.investopedia.com/articles/exchangetradedfunds/08/market-equal-weight.asp

[25] J. Heer and B. Shneiderman, "Interactive dynamics for visual analysis," *Queue*, vol. 10, no. 2, p. 30, 2012.

[26] M. Holt, "Papa parse - powerful csv parser for javascript," 2015. Available: http://papaparse.com/
[27] D. Hoyle and M. Rattray, "Pca learning for sparse high-dimensional data," *EPL (Europhysics Letters)*, vol. 62, no. 1, p. 117, 2003.
[28] Investopedia, LLC., "Holding Period Return/Yield Definition." Available: http://www.investopedia.com/terms/h/holdingperiodreturn-yield.asp
[29] Investopedia, LLC., "VIX (CBOE Volatility Index) Definition." Available: http://www.investopedia.com/terms/v/vix.asp
[30] W. Kinlaw and D. Turkington, "Correlation surprise," *Journal of Asset Management*, vol. 14, no. 6, pp. 385–399, Dec. 2013. Available: http://www.palgrave-journals.com/jam/journal/v14/n6/full/jam201327a.html
[31] V. Klinkenborg, "Trying to Measure the Amount of Information That Humans Create," *The New York Times*, Nov. 2003. Available: http://www.nytimes.com/2003/11/12/opinion/12WED4.html
[32] M. Kritzman and Y. Li, "Skulls, financial turbulence, and risk management," *Financial Analysts Journal*, vol. 66, no. 5, pp. 30–41, 2010.
[33] M. Kritzman *et al.*, "Principal components as a measure of systemic risk," 2010.
[34] D. M. Lane, "Bar charts," *Online Statistics Education: A Multimedia Course of Study*. Available: http://onlinestatbook.com/2/graphing_distributions/bar_chart.html
[35] S. Lethbridge, "Ridge Racer." Available: http://obsoletegamer.com/ridge-racer/
[36] R. McKeon and J. Netter, "What caused the 1987 stock market crash and lessons for the 2008 crash," *Review of Accounting & Finance*, vol. 8, no. 2, pp. 123–137, 2009. Available: http://search.proquest.com/docview/215640085?accountid=13314
[37] Meteor Development Group, "Meteor." Available: https://www.meteor.com/
[38] Meteor Development Group, "Meteor accounts." Available: https://www.meteor.com/accounts
[39] E. Rowell, "Part 3: HTML5 Canvas KineticJS Tutorials Introduction," 2013. Available: http://www.zeali.net/mirrors/html5canvastutorials/kineticjs/html5-canvas-events-tutorials-introduction-with-kineticjs/index.html
[40] SAS Institute Inc., "What Is Big Data?" Available: http://www.sas.com/en_us/insights/big-data/what-is-big-data.html
[41] C. Shenghui *et al.*, "The polar parallel coordinates method for time-series data visualization," in *Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on*. IEEE, 2012, pp. 179–182.
[42] B. Shneiderman, "RE: Data Visualization Research," private communication, Feb. 2015.
[43] B. Shneiderman and A. Aris, "Network visualization by semantic substrates," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 12, no. 5, pp. 733–740, 2006.
[44] S&P Dow Jones Indices LLC, "S&P 500® - S&P Dow Jones Indices." Available: http://us.spindices.com/indices/equity/sp-500
[45] F. Stajano, "A Gentle Introduction to Relational and Object Oriented Databases," Olivetti & Oracle Research Laboratory, ORL Technical Report TR-98-2, 1998. Available: http://www.cl.cam.ac.uk/~fms27/db/tr-98-2.pdf
[46] State University of New York at Oswego, "Variable types." Available: http://www.oswego.edu/~srp/stats/variable_types.htm
[47] steve, "Conditioning and Grouping with Lattice Graphics," Feb 2014. Available: http://www.r-bloggers.com/conditioning-and-grouping-with-lattice-graphics/
[48] B. Steverman, "Stock Market Crash: Understanding the Panic," *Bloomberg Business*. Available: http://www.bloomberg.com/bw/stories/2008-10-10/stock-market-crash-understanding-the-panicbusinessweek-business-news-stock-market-and-financial-advice
[49] Study.com, "Data Set in Math: Definition, Examples & Quiz." Available: http://study.com/academy/lesson/data-set-in-math-definition-examples-quiz.html
[50] B. Sylvester, "The Visualization of Financial Data: A review of information visualization tools in the financial data domain," Rutgers University, 2008. Available: http://comminfo.rutgers.edu/~aspoerri/Teaching/InfoVisOnline/Resources/projects/FinancialData_Sylvester.ppt
[51] S. C. Udhaya Padmanabhan, "How to make data visualization better with gestalt laws," July 2013. Available: http://sixrevisions.com/usability/data-visualization-gestalt-laws/
[52] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1, pp. 37–52, 1987.
[53] H. H. Zhang, "Lecture 10: Curse of Dimensionality," 2014. Available: http://math.arizona.edu/~hzhang/math574m/2014Lect10_curse.pdf