MASTER THESIS

EVALUATING AYASDI’S
TOPOLOGICAL DATA ANALYSIS
FOR BIG DATA

Author: Hee Eun Kim
Advisor 1: Prof. Dr. rer. nat. Stephan Trahasch
Advisor 2: Prof. Dott. Ing. Roberto V. Zicari

31 AUGUST 2015

OFFENBURG UNIVERSITY OF APPLIED SCIENCES
GOETHE UNIVERSITY FRANKFURT
Abstract

Data is emerging exponentially and becoming bigger and more complex than ever before. Stored data sets conserve too much of noise since complexity reduction is not a priority while a large-scale of data is stored. As a result, extracting meaningful knowledge is more critical than ever before in an emerging field of knowledge discovery in databases (KDD). Whereas there are multiple ready-made methods existing to reduce the noise and discover insights, this thesis will introduce new technique, previously not yet seen. This technique is called topological data analysis and it provides a hidden insight in intuitive and fast manner by converting the complex data sets into a topological network. The goals of this thesis are to understand a functionality of TDA and evaluate its performance by comparing it to conventional algorithms.
Declaration of Authorship

I declare in lieu of an oath that the Master thesis submitted has been produced by me without illegal help from other persons. I state that all passages which have been taken out of publications of all means or unpublished material either whole or in part, in words or ideas, have been marked as quotations in the relevant passage. I also confirm that the quotes included show the extent of the original quotes and are marked as such. I know that a false declaration will have legal consequences.

Acknowledgment

Thank you to:

Alexander Zylla
Matthew Eric Bassett
Roberto V. Zicari
Sascha Niro
Stephan Trahasch
Todor Ivanov

Thank you to everyone who contributed to the overwhelming supports
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>II</td>
</tr>
<tr>
<td>DECLARATION OF AUTHORSHIP</td>
<td>III</td>
</tr>
<tr>
<td>ACKNOWLEDGMENT</td>
<td>IV</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>V</td>
</tr>
<tr>
<td>LIST OF FIGURES AND ILLUSTRATIONS</td>
<td>IX</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>X</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>XI</td>
</tr>
<tr>
<td>1. CHAPTER 1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 MOTIVATION</td>
<td>1</td>
</tr>
<tr>
<td>1.2 AYASDI</td>
<td>1</td>
</tr>
<tr>
<td>1.3 AYASDI IN DATA MINING PROCESS</td>
<td>2</td>
</tr>
<tr>
<td>1.3.1 Business Understanding</td>
<td>3</td>
</tr>
<tr>
<td>1.3.2 Data Understanding</td>
<td>3</td>
</tr>
<tr>
<td>1.3.3 Data Preparation</td>
<td>3</td>
</tr>
<tr>
<td>1.3.4 Modeling</td>
<td>3</td>
</tr>
<tr>
<td>1.3.5 Evaluation</td>
<td>4</td>
</tr>
<tr>
<td>1.3.6 Deployment</td>
<td>4</td>
</tr>
<tr>
<td>1.4 THESIS STRUCTURE</td>
<td>5</td>
</tr>
<tr>
<td>2. CHAPTER 2 TOPOLOGICAL DATA ANALYSIS (TDA)</td>
<td>6</td>
</tr>
<tr>
<td>2.1 TOPOLOGY</td>
<td>6</td>
</tr>
<tr>
<td>2.1.1 Three Fundamental Invariances</td>
<td>8</td>
</tr>
<tr>
<td>2.2 ANALYSIS PIPELINE OF TDA</td>
<td>9</td>
</tr>
<tr>
<td>2.2.1 Finite Data set (a)</td>
<td>10</td>
</tr>
<tr>
<td>2.2.2 Point cloud (b)</td>
<td>10</td>
</tr>
<tr>
<td>2.2.2.1 METRICS FOR CONTINUOUS VALUES</td>
<td>12</td>
</tr>
<tr>
<td>2.2.2.2 EXERCISE OF CONTINUOUS VALUES</td>
<td>13</td>
</tr>
<tr>
<td>2.2.2.3 METRICS FOR CATEGORICAL VALUES</td>
<td>14</td>
</tr>
<tr>
<td>2.2.2.4 EXERCISE OF CATEGORICAL VALUES</td>
<td>14</td>
</tr>
<tr>
<td>2.2.3 Simplicial Complex (c)</td>
<td>16</td>
</tr>
<tr>
<td>2.2.4 Topological Invariant (d)</td>
<td>16</td>
</tr>
<tr>
<td>3. CHAPTER 3 TOPOLOGICAL DATA ANALYSIS OF AYASDI</td>
<td>21</td>
</tr>
</tbody>
</table>
3.1 CONSTRUCT TOPOLOGICAL NETWORK .............................................................................. 21
  3.1.1 Mapper ......................................................................................................................... 21
    3.1.1.1 Filter and metrics ....................................................................................................... 21
    3.1.1.2 Resolution .................................................................................................................. 22
    3.1.1.3 Clustering .................................................................................................................. 22
    3.1.1.4 Connection ............................................................................................................... 22
  3.1.2 Elements of Output Network ...................................................................................... 23
  3.1.3 Metric Selection ......................................................................................................... 23
  3.1.4 Filter Selection ........................................................................................................... 24

3.2 INSIGHT DISCOVERY ...................................................................................................... 26
  3.2.1 Color Spectrum .......................................................................................................... 26
  3.2.2 Subgroup Comparison ............................................................................................... 26
    3.2.2.1 Rank methods ........................................................................................................... 27
    3.2.2.2 Insight by Ranked Variables .................................................................................... 28

4. CHAPTER 4 USE CASES ................................................................................................... 29
  4.1 CASE 1: PREDICTING ENERGY CONSUMPTION .......................................................... 29
    4.1.1 Business Understanding ............................................................................................ 29
    4.1.2 Data Understanding .................................................................................................. 29
    4.1.3 Data Preparation ....................................................................................................... 30
      4.1.3.1 Format Data ........................................................................................................... 30
    4.1.4 Applying Ayasdi ......................................................................................................... 31
      4.1.4.1 Color Spectrum ..................................................................................................... 31
      4.1.4.2 Discovered Insights ............................................................................................. 31
  4.2 CASE 2: CUSTOMER CHURN ANALYSIS ...................................................................... 32
    4.2.1 Business Understanding ............................................................................................ 32
    4.2.2 Data Understanding .................................................................................................. 32
      4.2.2.1 Describe Data ........................................................................................................ 32
    4.2.3 Data Preparation ....................................................................................................... 33
      4.2.3.1 Format Data ........................................................................................................... 33
    4.2.4 Applying Ayasdi ......................................................................................................... 33
      4.2.4.1 Color Spectrum ..................................................................................................... 33
      4.2.4.2 Subgroup Comparison .......................................................................................... 34
      4.2.4.3 Discovered Insights ............................................................................................. 34
  4.3 CASE 3: EARTHQUAKE PREDICTION ........................................................................... 36
    4.3.1 Business Understanding ............................................................................................ 36
    4.3.2 Data Understanding .................................................................................................. 36
      4.3.2.1 Collect Raw Data .................................................................................................... 36
      4.3.2.2 Select Data ............................................................................................................. 37
      4.3.2.3 Describe Data ........................................................................................................ 37
      4.3.2.4 Verify Data Quality .............................................................................................. 41
    4.3.3 Data Preparation ........................................................................................................ 41
A.1. USED SYNTAX ........................................................................................................... 64
A.2. CONSTRUCT DATA ...................................................................................................... 65
  A.2.1. Earthquake ........................................................................................................... 65
  A.2.2. Active Fire ........................................................................................................... 66
  A.2.3. Volcano ............................................................................................................... 69
  A.2.4. Heat Flow ............................................................................................................. 69
A.3. INTEGRATE DATA ........................................................................................................ 70
  A.3.1. Earthquake, Solar wind and Magnetic ............................................................... 70
  A.3.2. Earthquake, Volcano and Heat flow ................................................................. 71

B. APPENDIX B MAPE IN R ............................................................................................. 72
List of Figures and Illustrations

Figure 1: Ayasdi in Data Mining Process (“Ayasdi Core,” 2015) .................................................................................. 2
Figure 2: CRISP-DM (“Making Better Business Decisions with Analytics and Business Rules,” 2014) ........................................... 2
Figure 3: Supervised Learning (left) and Unsupervised Learning (right) (“Machine Learning - Stanford University,” 2015) .................................................................................................................. 3
Figure 4: Topological representation of Königsberg bridge (Weisstein) ............................................................................... 6
Figure 5: The curvature of the sphere is not apparent to ant ....................................................................................................... 7
Figure 6: Complexes for two different scale of ε (Paul et al., 2013) .......................................................................................... 8
Figure 7: Pipeline of TDA (Zomorodian and Zomorodian, 2012) .............................................................................................. 9
Figure 8: Homeomorphisms Objects (Rickert, 2014) ................................................................................................................. 16
Figure 9: The first three Betti numbers .................................................................................................................................. 17
Figure 10: Complexes for various ε values ............................................................................................................................... 18
Figure 11: The barcode plot of the data shown in Figure 10 .................................................................................................... 18
Figure 12: Barcode plot of the Iris flowers data. $H_0$ is on the top, $H_1$ is on the bottom. ...................................................... 19
Figure 13: Difference pipeline between TDA and Ayasdi’s TDA ............................................................................................. 22
Figure 14: Elements of topological network .......................................................................................................................... 23
Figure 15: A metaphor of filter function to summarize original input data set ........................................................................ 24
Figure 16: Color spectrum corresponding to the chosen attribute of Titanic data set .................................................................. 27
Figure 17: Subgroup comparison .................................................................................................................................................. 27
Figure 18: Color spectrum corresponding to the chosen attribute of Case 1 ............................................................................ 31
Figure 19: Color spectrum corresponding to the chosen attribute of Case 2 ............................................................................ 33
Figure 20: Determined Two subgroups ................................................................................................................................... 34
Figure 21: Simple data transmission in Hue ................................................................................................................................ 37
Figure 22: Color spectrum of integrated data (Earthquake, Solar wind and Magnetic) ................................................................. 45
Figure 23: Color spectrum of integrated data (Earthquake, Volcano and Heat flow) .................................................................. 46
Figure 24: Venn diagram of feature engineering techniques .................................................................................................... 49
Figure 25: Graphical representation of comparison methodology .................................................................................................. 51
Figure 26: Evaluation process of Case 1 in RapidMiner ................................................................................................................ 52
Figure 27: Evaluation process of Case 2 in RapidMiner ................................................................................................................ 52
A. Appendix A Pig Scripts .................................................................................................................................................. 64
B. Appendix B MAPE in R................................................................................................................................................ 72
List of Tables

TABLE 1: THREE FUNDAMENTAL PROPERTIES OF TOPOLOGY ................................................................................. 9
TABLE 2: METRICS FOR CONVERSION TO THE POINT CLOUD .................................................................................. 10
TABLE 3: SAMPLE DATA SET OF CONTINUOUS VALUES ................................................................................................. 13
TABLE 4: SAMPLE DATA SET OF CATEGORICAL VALUES ............................................................................................... 14
TABLE 5: TRANSFORMED TO THE NUMERICAL DATA MATRIX FROM TABLE 4 ....................................................... 15
TABLE 6: DISCRIMINATE KOREAN CONSONANT LETTERS .......................................................................................... 17
TABLE 7: A SAMPLE OF IRIS DATA SET ..................................................................................................................... 19
TABLE 8: DECISION CRITERIA FOR METRIC SELECTION .......................................................................................... 24
TABLE 9: DECISION CRITERIA OF FILTER FUNCTION ............................................................................................ 25
TABLE 10: VARIABLE RANKING BY KS-STATISTIC .................................................................................................. 28
TABLE 11: SAMPLE DATA SET OF ENERGY CONSUMPTION ...................................................................................... 30
TABLE 12: TRANSFORMED DATA SET FROM TABLE 11 .............................................................................................. 30
TABLE 13: SAMPLE OF TRAINING DATA SET .......................................................................................................... 32
TABLE 14: RANKED NUMERICAL VALUES BY KS-STATISTIC .................................................................................. 35
TABLE 15: RANKED CATEGORICAL VALUES BY P-VALUE THAT BIGGER THAN 0.05 ................................................... 35
TABLE 16: SAMPLE OF EARTHQUAKE DATA (1904-2015) .................................................................................. 38
TABLE 17: SAMPLE OF SOLAR WIND PROTON HISTORY (1998-2015) ................................................................. 38
TABLE 18: SAMPLE OF INTERPLANETARY MAGNETIC (1997-2015) ..................................................................... 39
TABLE 20: SAMPLE OF VOLCANO .......................................................................................................................... 40
TABLE 21: SAMPLE OF HEAT FLOW ......................................................................................................................... 40
TABLE 22: CRITERIA OF DATA QUALITY ................................................................................................................. 41
TABLE 23: DIFFERENT DATA SETS USE DIFFERENT TIME FORMATS .................................................................. 41
TABLE 24: CONSTRUCTED DATA SETS FROM TABLE 23 .......................................................................................... 43
TABLE 25: INTEGRATED DATA OF EARTHQUAKE, SOLAR WIND AND MAGNETIC .................................................... 43
TABLE 26: INTEGRATED DATA OF EARTHQUAKE, VOLCANO AND HEAT FLOW ..................................................... 44
TABLE 27: FORMAL DEFINITION OF FEATURE ENGINEERING (MOTODA AND LIU, 2002) ............................... 48
TABLE 28: ADVANTAGES (+) AND DISADVANTAGES (-) OF FEATURE SELECTION TECHNIQUES (SAEYS ET AL., 2007) ......................................................................................................................... 49
TABLE 29: EXPERIMENT VARIABLE OF USE CASE 1 .............................................................................................. 53
TABLE 30: PERFORMANCE COMPARISON OF USE CASE 1 .................................................................................... 54
TABLE 31: EXPERIMENT VARIABLE OF USE CASE 2 ............................................................................................ 55
TABLE 32: PERFORMANCE COMPARISON OF CUSTOMER DATA ........................................................................... 56
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSV</td>
<td>Comma Separated Values</td>
</tr>
<tr>
<td>CRISP-DM</td>
<td>Cross Industry Standard Process for Data Mining</td>
</tr>
<tr>
<td>HDFS</td>
<td>Hadoop Distributed File System</td>
</tr>
<tr>
<td>KDD</td>
<td>Knowledge Discovery in Databases</td>
</tr>
<tr>
<td>KS</td>
<td>Kolmogorov-Smirnov</td>
</tr>
<tr>
<td>MAPE</td>
<td>Mean absolute percentage error</td>
</tr>
<tr>
<td>MRMR</td>
<td>Minimum Redundancy and Maximum Relevance</td>
</tr>
<tr>
<td>PCA</td>
<td>Principle Component Analysis</td>
</tr>
<tr>
<td>RF</td>
<td>Random forest</td>
</tr>
<tr>
<td>RDBMS</td>
<td>Relational DataBase Management System</td>
</tr>
<tr>
<td>SVD</td>
<td>Singular Value Decomposition</td>
</tr>
<tr>
<td>TSV</td>
<td>Tab Separated Values</td>
</tr>
<tr>
<td>TDA</td>
<td>Topological Data Analysis</td>
</tr>
</tbody>
</table>
Chapter 1
Introduction

1.1 Motivation

The aim of the thesis is to understand, apply and evaluate new data analysis technique called topological data analysis (TDA) of Ayasdi. This method was invented by Gurjeet Singh, Facundo Mémoli and Gunnar Carlsson at Stanford University. (Singh et al., 2007) TDA of Ayasdi was able to capture great attention of data scientists and it is regarded as a proper tool to analyze complex data set in an intuitive and fast manner. This thesis pursues to deliver an unbiased opinion of TDA of Ayasdi featuring three applied cases and a comparison of conventional data analysis techniques.

1.2 Ayasdi

Ayasdi is a spinoff of a Stanford research project which started in the mathematics department. It pioneered to use a unique combination of topological data analysis and multiple data mining algorithms in order to extract unknown insights. Nevertheless, it remains a topological data analysis since it preserves three properties of Topology, which are introduced in subsection 2.1.1. Ayasdi offers either automate or manual analysis without any prior assumptions within a few steps. The automated analysis can make an algorithm selection that can formulate a decent portfolio of topological networks. Due to this approach with a fully interactive interface, the entry barrier for the non-expert is relatively lower than for other analysis methodologies.
1.3 Ayasdi in Data Mining Process

Figure 1 shows a position of Ayasdi in a data mining process presented on the official webpage of Ayasdi (“Ayasdi Core,” 2015). Ayasdi is an additional or optional step in the complete process. The knowledge of a data mining process can enhance due to the usage of Ayasdi, since the subsequent phases benefit from the previous ones, as the lessons learned during the process can trigger new processes. For instance, the input data set prepared during the previous phases determines the quality of insights determined by Ayasdi. Likewise, gained insights can improve the modeling process.

![Figure 1: Ayasdi in data mining process (“Ayasdi Core,” 2015).](image1)

The process of data mining exists in variations. CRISP-DM stand for Cross Industry Standard Process for Data Mining (Chapman et al., 2000; Wirth and Hipp, 2000) and is the most established process used by experts across the industry (KDnuggets, 2002). DRISP-DM as illustrated in Figure 2 defines six phases to tackle problems. The sequence of the phases is not a linear process and moving back and forth. The arrows in the diagram indicate the dependencies between phases. (“Cross Industry Standard Process for Data Mining,” 2015)

![Figure 2: CRISP-DM (“Making better business decisions with analytics and business rules,” 2014).](image2)
1.3.1 Business Understanding

This initial phase is a preliminary plan designed to achieve the objectives. It defines a goal of data mining as well as business objectives, resources assessment and a project plan with a detailed perspective view.

1.3.2 Data Understanding

The next phase collects the initial data. Then the collected data is analyzed by describing a schema and verifying data quality. Initial insights are gained in this phase. The quality of the data is determined and it is checked whether there are any missing or incorrect values. If one owns a given training data set, this phase can be skipped.

1.3.3 Data Preparation

Most data scientists are investing about 60 to 80% of their time and effort into data preparation (Jermyn et al., 1999). In general, collected data consist of different formats of data values, because each organization has different database, database schema, data management policy and more. Data preparation aims to integrate all data sets by repetitively processing sub-tasks such as data selection, data cleaning and transformation.

1.3.4 Modeling

In order to create a model, various algorithms are applied in an iterative manner with adjusting values of parameters. As shown in Figure 3, two kinds of learning algorithms exist to create a model: supervised learning and unsupervised learning. In the former, a numerical attribute is predicted while in the latter a category (a class) is predicted. The type of label determines the analysis type.

![Figure 3: Supervised learning (left) and unsupervised learning (right) ("Machine Learning - Stanford University," 2015)](image-url)
• Supervised learning: data set is labeled. The aim is to infer a function from labeled training data (Mohri et al., 2012). The left plot in Figure 3 shows categorical labels that are either circle or plus. The created model can classify a newly arriving label.

• Unsupervised learning: data set does not have a labeled value. The task is to try to find a hidden structure in unlabeled data without prior assumptions. The generated model is able to summarize data.

Note that Feature engineering (Section 0) is a part of the modelling process, not a data preparation process. More concretely, one should apply the feature engineering techniques on a training data set, but not on an entire data set. Otherwise, the result tends to be biased and overfitting, since the entire data set consists of a training and a test data set. Particularly, dimensional reduction is an essential step in a high dimensional data set. Two potential problems of the high dimensional data set are as follows: First, it hinders the building of a decent prediction model. Second, it demands high computational costs and more time. Therefore, dimensional reduction can be used to reduce the irrelevant variables and to capture insights that can hint at underlying variables.

1.3.5 Evaluation

This phase thoroughly executes the created model and evaluates the performance result if it achieves the data mining objectives. The criteria of evaluation is dependent on the chosen model type. For instance, confusion matrix with classification models, mean error rate or mean percentage error with regression models.

1.3.6 Deployment

In many cases, the end users are not the data analyst. Therefore, the gained knowledge needs to be represented in a useful way to the targeted audience. Depending on the goal of data mining defined in initial phase, the output of deployment can be as simple as generating a report or as complex as implementing a data mining procedure. Nevertheless, the simple report should include the simulated results when model is actually applied.
1.4 Thesis Structure

While this chapter introduced basic ideas about TDA, chapter 2 focuses on the mathematical basis of standard TDA in order to understand where the TDA of Ayasdi stems from. Chapter 3 examines TDA of Ayasdi with few experiments. Chapter 4 proposes three concrete use cases. Each use case consists of data set of different complexity level. In chapter 5 compares TDA to other feature selection methods to evaluate the quality of insights that captured from the previous chapter 4. Finally, conclusion and future research extensions are discussed in chapter 6.
Chapter 2
Topological Data Analysis (TDA)

This chapter introduces the basis of Topological Data Analysis (TDA) in order to establish groundwork. This step will allow one to know where Ayasdi’s TDA stems from and ultimately help to understand Ayasdi’s TDA better. In case you are interested in more detailed explanation of Topological Data Analysis, please refer to the following papers (Carlsson, 2009; Zomorodian and Zomorodian, 2012). If you are already familiar with the Definition 1, you can proceed to Chapter 3. Topological Data Analysis of Ayasdi.

Definition 1. (Topological Data Analysis). Topological Data Analysis is the use of persistent homology to study a parameterized range of simplicial complexes determined from a point cloud.

2.1 Topology

Topology is a subfield of geometry and it was introduced in 1736 by mathematician Leonhard Euler with the investigation of certain questions in geometry about the Seven Bridges of Königsberg as shown in Figure 4. Euler kept only key information of connection with nodes and edges, but discarded irrelevant information such as house, size of land. After this, Topology has grown in popularity in mathematics.

Figure 4: Topological representation of Königsberg bridge (Weisstein)
Figure 5: The curvature of the sphere is not apparent to ant

Topology studies the space. While Geometry studies features of that shape, a shape sits within a space. Therefore, the assumption of Topology is more general than Geometry because it is blind to ideas about the shape and does not consider a manifold structure, smoothness or curvature. A sphere is an example of a the surface of a shape that defines a space. In geometry, one can look at the curvature of the sphere in unknown space. However, in Topology, one looks at the sphere as if it were an ant walking across it. Figure 5 illustrates that the Ant perceives the curvature of the sphere as a flat plane, likewise the curvature of the earth to human.

The main object of topology are properties of space that are preserved by a continuous map \( f : X \to Y \) from a Topological Space \( X \) to a topological space \( Y \). Topology only mentions Topological Spaces because that is where Continuous Functions live. These properties are called topological invariants, some of which are most easily identified through Homology Theory. Continuity allows one to understand the need for \( \epsilon \)-balls. Corresponding to the scale of the radius, the homological features will appear and disappear. For example, too small radius leads no complex structure, but too big radius joins all points together.

Figure 6 is a visual aid to understand the concept above. Let \( X \) as a sample of data in \( \mathbb{R}^2 \) with \( \epsilon \)-radius balls around each point. Two points on the top are joined if two corresponding \( \epsilon \)-balls intersect, and triangle on the bottom formed by three lines is colored if three corresponding \( \epsilon \)-balls all intersect.
In terms of $\epsilon$-balls, one can describe a continuous function as follows: A function $f : X \rightarrow Y$ is continuous if for each $\epsilon$-ball $B \subset Y$, the set $f^{-1}(B)$ is an $\epsilon$-ball in $X$. A continuous map preserves key properties of a space. With the notion of a continuous map, topology studies the properties of a space that are invariant under a continuous map. The topological invariants are discussed at following subsection 2.2.4.

### 2.1.1 Three Fundamental Invariances

Table 1 represents three fundamental invariants of Topology. **Coordinate invariance** considers only property of shape rather than coordination and rotation. Therefore, three ellipses in Table 1 are topologically same. **Deformation invariance** does not change the identity of object by stretching or squashing unless it is not teared or glued. For instance, a perfect round circle is identical to a squashed one like an ellipse. Finally, **compressed representations** focus on connectivity and continuity. This information allows the topology to recognize relevant data patterns. For instance, a circle has infinite amounts of vertices whereas the shape of hexagon is still similar to a circle, but consists of six nodes and six edges.

Studies of topological invariants fall into two categories: analytic (point-set) studies and algebraic studies. Algebraic studies again subdivide into studies of Homotopy and Homology. While the former is irrelevant to this thesis, the Homology to Topological Invariants are discussed here.
### Table 1: Three fundamental properties of topology

<table>
<thead>
<tr>
<th>Coordinate invariance</th>
<th><img src="image1" alt="Coordinate invariance" /></th>
</tr>
</thead>
<tbody>
<tr>
<td>Deformation invariance</td>
<td><img src="image2" alt="Deformation invariance" /></td>
</tr>
<tr>
<td>Compressed representations</td>
<td><img src="image3" alt="Compressed representations" /></td>
</tr>
</tbody>
</table>

#### 2.2 Analysis Pipeline of TDA

Figure 7 shows a procedure of TDA that is subdivided into three-step pipeline. The first step is creation of **Point cloud (b)** from the input **Finite data set (a)** by calculating the similarity value by a distance metric. The next step approximates and constructs the given input (b) into a **Simplicial complex (c)**. The final step computes topological **Invariant (d)** shown as a hole in this case. It is a summary of the original Data set S. In other words, TDA considers a Finite data set (a) as a noisy Point cloud (b) taken from an unknown dimensional space X. TDA assumes that both space X and Point cloud (b) are topological spaces. Then TDA recovers the topology of X by constructing or gluing the Simplicial complex (c) corresponding to the most robust interval of persistent homology (Zomorodian and Zomorodian, 2012).

![Figure 7: Pipeline of TDA (Zomorodian and Zomorodian, 2012)](image4)
2.2.1 Finite Data set (a)

The format of Data set is a matrix such as a spreadsheet with data samples in the rows and multiple attributes in the columns. The data points, data sample, or tuple refer to the rows in the matrix. Similarly, dimensions, fields, variables, features and attributes indicate the columns in the Data set. The type of attributes can be one of the following: numeric, categorical, or string. For instance, numerical attributes have a continuous scale of values such as weight, price, and profits, whereas categorical attributes have a discrete values such as gender, age group, product type and title.

The type of attributes is critical since it is not capable to measure heterogeneous attribute types at the same time. In order to avoid this limitation, change the categorical type to a numerical type (e.g. bad: -1, ok: 0, good: 1), so that it can be measured together with numerical type.

2.2.2 Point cloud (b)

Point cloud Y is a representation of a special feature of a manifold that is a specialized topological space. Informally, every row in the Finite data set is extracted into a single data point in the point cloud. Please note that plotted point cloud is synthesized. In reality, the point cloud is placed in high-dimensional space with a single vector value. One can convert the finite data set into the Point cloud by one of following metrics as shown in Table 2.

<table>
<thead>
<tr>
<th>Type of Data</th>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>Euclidean Distance</td>
</tr>
<tr>
<td></td>
<td>Manhattan Distance</td>
</tr>
<tr>
<td></td>
<td>Bray Curtis Distance</td>
</tr>
<tr>
<td></td>
<td>Canberra Distance</td>
</tr>
<tr>
<td></td>
<td>Cosine Distance</td>
</tr>
<tr>
<td></td>
<td>Correlation Distance</td>
</tr>
<tr>
<td>Boolean</td>
<td>Binary Hamming Distance</td>
</tr>
<tr>
<td></td>
<td>Binary Jaccard Dissimilarity</td>
</tr>
<tr>
<td></td>
<td>Russell Rao Dissimilarity</td>
</tr>
<tr>
<td></td>
<td>Rogers Tanimoto Dissimilarity</td>
</tr>
<tr>
<td></td>
<td>Matching Dissimilarity</td>
</tr>
<tr>
<td></td>
<td>Dice Dissimilarity</td>
</tr>
<tr>
<td>String</td>
<td>Hamming Distance</td>
</tr>
<tr>
<td></td>
<td>Categorical Cosine Distance</td>
</tr>
<tr>
<td></td>
<td>Smith Waterman Similarity</td>
</tr>
<tr>
<td></td>
<td>Damerau Levenshtein Distance</td>
</tr>
<tr>
<td></td>
<td>Jaccard Dissimilarity</td>
</tr>
<tr>
<td></td>
<td>Edit Distance</td>
</tr>
<tr>
<td>Image</td>
<td>Image Distance</td>
</tr>
<tr>
<td>Color</td>
<td>Color Distance</td>
</tr>
</tbody>
</table>

Table 2: Metrics for conversion to the point cloud
Each metric calculates similarity in its own manner between data points. Following subsections introduce most widely used metrics corresponding exercises. The common syntax used in all metric equations is defined as follows:

- $X, Y =$ data points
- $X_i, Y_i =$ attributes of each data point to be compared
- $N =$ number of attributes (i.e., dimensionality of $X$)
- $M =$ number of points
2.2.2.1 METRICS FOR CONTINUOUS VALUES

- **Euclidean Distance Similarity**

  
  \[
  \text{Euclidean Distance, } L_2(X, Y) = \sqrt{\sum_{i=1}^{N} (X_i - Y_i)^2}
  \]

  Euclidean Distance Similarity generalizes the usual distance formula studied in 2-dimensional and 3-dimensional spaces. It measures geometrically the most valid similarity in Euclidean space. The similarity values range between zero and infinity. If the output distance has a lower bound of zero, this indicates a perfect match. This metric is particularly good to apply to data set that is not directly comparable. In other word, each column has a disparate quantities or wide variance.

- **Variance Normalized Euclidean**

  
  \[
  \text{Variance Normalized Euclidean, } VNE(X, Y) = \sqrt{\sum_{i=1}^{N} (X_i - Y_i)^2} \over \sum_{i=1}^{N} V_i
  \]

  Variance Normalized Euclidean is a derivational Euclidean Distance Similarity. It is able to give better performance when data set contains heterogeneous scale variables. For instance, height, weight, and shoes size are heterogeneous scale variables. This metric finds rescales the variables by subtracting the coordinate mean and dividing by the corresponding standard deviation.

- **Cosine Similarity**

  
  \[
  \text{Cosine Similarity, } \cos(\theta) = \sqrt{\sum_{i=1}^{N} X_i \times Y_i} \over \sqrt{\sum_{i=1}^{N} X_i^2} \times \sqrt{\sum_{i=1}^{N} Y_i^2}
  \]

  Unlike the Euclidean Distance Similarity, Cosine Similarity does is sensitive for recognizing whether two data samples are collinear rather than the absolute magnitude of differences. For instance, Cosine Similarity can be applied to the purchasing activity of a group of people in order to detect behavioral trends.
Table 3: Sample data set of continuous values

Table 3 is a sample spreadsheet, which is composed of three data points with four numerical attributes in order to measure similarity of continuous values. Let us compare the similarity of each point in an iterative manner using particular metrics: Euclidean Distance Similarity and Cosine Similarity.

- **Euclidean Distance Similarity (derivation)**

The result is a distance. It approaches to zero when data samples share the most similarity. Accordingly, sample A and B are the most similar in Table 3, since the distance between them is the lowest. On the other hands, sample A and C are the least similar in Table 3.

\[
L_2(A, B) = \sqrt{(3 - 4)^2 + (7 - 8)^2 + (10 - 11)^2 + (12 - 13)^2} = \sqrt{3 + 7 + 10 + 12} = \sqrt{334} \approx 18.275
\]

\[
L_2(B, C) = \sqrt{(4 - 13)^2 + (8 - 11)^2 + (11 - 8)^2 + (13 - 4)^2} = \sqrt{13 + 7 + 9 + 9} = \sqrt{37} = 6.083
\]

\[
L_2(A, C) = \sqrt{(3 - 13)^2 + (7 - 11)^2 + (10 - 8)^2 + (12 - 4)^2} = \sqrt{10 + 4 + 6 + 8} = \sqrt{28} = 5.303
\]

- **Cosine Similarity (derivation)**

The outcome is bounded in \([0, 1]\) since cosine similarity is used in positive space. Note that the result is a judgment of orientation, not a magnitude. For instance, a derived result is approaching to 1 when two data samples have an similar orientation (cosine of 0°). Thus, sample A and B are the most similar, while sample A and C are the least similar in Table 3.

\[
\cos(\angle AOB) = 0.99
\]

\[
\cos(\angle BOC) = 0.76
\]

\[
\cos(\angle AOC) = 0.73
\]
2.2.2.3 METRICS FOR CATEGORICAL VALUES

- **Cosine Similarity for Categorical values**

The formula of cosine similarity metric for categorical values is same as one for continuous values. However, a prior step of data preparation is necessary. Depending on the frequency appearances of each column, categorical value converts into a numerical value.

- **Jaccard Dissimilarity**

\[
\text{Jaccard}(X, Y) = 1 - \frac{\text{Set}(X) \cap \text{Set}(Y)}{\text{Set}(X) \cup \text{Set}(Y)}
\]

Jaccard metric measures the dissimilarity of asymmetric information on non-binary variables. While two selected rows are regarded as union set, each row is regarded as a subset of objects. Since this metric treats the columns as a delimiter, presence or absence of value is not a matter. Jaccard metric is also widely used to determine the similarity of words in a string.

- **Hamming Distance**

The Hamming distance \(H\) is defined for two strings of equal length, besides Damerau Levenshtein Distance is used for the strings of unequal length. Hamming distance determines how many symbols or bits are different at which the corresponding symbols are positioned. For instance, the Hamming distance between "Roberto" and "Alberto" is 2 and the distance between 2173896 and 2233796 is 3.

2.2.2.4 EXERCISE OF CATEGORICAL VALUES

<table>
<thead>
<tr>
<th></th>
<th>attribute 1</th>
<th>attribute 2</th>
<th>attribute 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample A</td>
<td>orange</td>
<td>blue</td>
<td>tea</td>
</tr>
<tr>
<td>Sample B</td>
<td>orange</td>
<td>pink</td>
<td></td>
</tr>
<tr>
<td>Sample C</td>
<td>orange</td>
<td>blue</td>
<td></td>
</tr>
<tr>
<td>Sample D</td>
<td>banana</td>
<td>pink</td>
<td></td>
</tr>
<tr>
<td>Sample E</td>
<td>banana</td>
<td>blue</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Sample data set of categorical values

Table 4 consists of five data points with three categorical attributes. Likewise, we will compare the similarity of each point in an iterative manner with particular metrics: Cosine Similarity, Jaccard Dissimilarity and Hamming distance.
• **Cosine Similarity for Categorical values**

<table>
<thead>
<tr>
<th></th>
<th>attribute 1:</th>
<th>attribute 1:</th>
<th>attribute 2:</th>
<th>attribute 2:</th>
<th>attribute 3:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>orange</td>
<td>banana</td>
<td>blue</td>
<td>pink</td>
<td>tea</td>
</tr>
<tr>
<td>Sample A</td>
<td>log(3/5)</td>
<td>0</td>
<td>log(3/5)</td>
<td>0</td>
<td>log(1/5)</td>
</tr>
<tr>
<td>Sample B</td>
<td>log(3/5)</td>
<td>0</td>
<td>0</td>
<td>log(2/5)</td>
<td>0</td>
</tr>
<tr>
<td>Sample C</td>
<td>log(3/5)</td>
<td>0</td>
<td>log(3/5)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sample D</td>
<td>0</td>
<td>log(2/5)</td>
<td>0</td>
<td>log(2/5)</td>
<td>0</td>
</tr>
<tr>
<td>Sample E</td>
<td>0</td>
<td>log(2/5)</td>
<td>log(3/5)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table 5: Transformed to the numerical data matrix from Table 4**

First, the categorical data set should be prepared as a numerical data set as shown in Table 5. The orientation between D and E is more similar than one between B and C since the points D and E shares the more rare value of "banana" than “orange” at the attribute 1.

\[
\cos(\angle BOC) = \frac{(\log(3/5)\times\log(3/5))+(0\times\log(3/5))+(\log(2/5)\times0)}{\sqrt{\log(3/5)^2+\log(2/5)^2}\times\sqrt{\log(3/5)^2+\log(3/5)^2}} = \frac{0.05}{\sqrt{0.21\times0.1}} = 0.05
\]

\[
\cos(\angle DOE) = \frac{(\log(2/5)\times\log(2/5))+(0\times\log(3/5))+(\log(2/5)\times0)}{\sqrt{\log(2/5)^2+\log(2/5)^2}\times\sqrt{\log(2/5)^2+\log(3/5)^2}} = \frac{0.16}{\sqrt{0.32\times0.21}} = 0.16
\]

• **Jaccard Dissimilarity (derivation)**

The result ranges from zero to one. It approaches to zero when data samples are the most similar. On the other hands, the result is one when two data samples do not have any intersection. By the derived result, Sample A and C share the most the similar properties.

\[
J(A,B) = 1-(1/4) = 3/4 \quad , \quad 1 \text{ intersection } \{\text{orange}\} \text{ out of the 4 union values } \{\text{orange, blue, pink, tea}\}
\]

\[
J(A,C) = 1-(2/3) = 1/3 \quad , \quad 2 \text{ intersections } \{\text{orange, blue}\} \text{ out of the 3 union values } \{\text{orange, blue, tea}\}
\]

\[
J(A,D) = 1-0 = 1 \quad , \quad \text{no intersection}
\]

\[
J(A,E) = 1-(1/4) = 3/4 \quad , \quad 1 \text{ intersection } \{\text{blue}\} \text{ out of the 4 union values } \{\text{orange, banana, blue, tea}\}
\]

• **Hamming Distance (derivation)**

Sample A cannot be compared due to the length inequality. Sample B and C are the most similar since the derived distance is the shortest compared to others.

\[
H(B,C) = 4 \quad , \quad \text{the distance between orange, pink and orange, blue is 4}
\]

\[
H(B,D) = 6 \quad , \quad \text{the distance between orange, pink and banana, pink is 6}
\]

\[
H(B,E) = 10 \quad , \quad \text{the distance orange, pink and banana, blue is 10}
\]
2.2.3 Simplicial Complex (c)

Simplicial Complex is a form of wireframe approximation of the manifold. Intuitively, it is similar to a hypergraph or wired frame, where it is represented as a relationship between (n+1) nodes with n-dimensional simplex. Persistent Homology, a recent breakthrough idea, extends Homology theory to work across a range of parameterized Simplicial Complexes, like the one arising from a point cloud, instead of just a single, isolated complex. It looks for topological invariants across various scales of a topological manifold. Persistent Homology is a chief scientific idea in TDA and it makes TDA a practical tool for data analysis.

Unfortunately, this chapter does not explain the detailed process of the idea behind Persistent Homology. Ayasdi’s TDA focuses on the creation of point cloud (subsection 2.2.2) and clustering, but optionally uses the Persistent Homology if one would need to build a higher dimensional complex using more functions on the data (Singh et al., 2007). Nevertheless, one may reference following article (Ghrist, 2008) as a background material of Persistent Homology. The author describes the following questions: How does one turn a point cloud into a Simplicial Complex? How does one parameterize this process? How can the tools of Algebraic Topological, in particular, Simplicial Homology be extended to analyze a parameterized range of complexes?

2.2.4 Topological Invariant (d)

Once you have Simplicial Complex, the tools of Algebraic Topology is used to construct Homology Groups that are algebraic analogues of certain properties of the manifold. Algebraic topology is a branch of topology that identifies topological spaces by algebraic invariants. For example, Homology Groups can describe the algebraic analogues of the holes in a manifold in various dimensions. Therefore, the number of holes can characterize the topology of the manifold. For instance, Figure 8 classifies random 2-dimensional objects into different or equivalent space according to the number of holes. Objects remain equivalent by stretching or shrinking.

![Figure 8: Homeomorphisms Objects (Rickert, 2014)](image)
**Betti numbers** is a compact method to present this Homology Groups by investigating the properties of the topological spaces. It distinguishes topological spaces according to the connectivity of $n$-dimensional simplicial complexes. The $n^{th}$ Betti number represents the rank of the $n^{th}$ homology group, denoted as $H_n$. Informally, the $n^{th}$ Betti number refers to the number of $n$-dimensional holes on a topological surface. Figure 9 shows that the first three Betti numbers have the following definitions for 0-dimensional, 1-dimensional, and 2-dimensional simplicial complexes: $\beta_0$ is the number of connected components $\beta_1$ is the number of holes (one-dimensional) and $\beta_2$ is the number of two-dimensional "voids".

![Betti Numbers Diagram]

**Figure 9: The first three Betti numbers**

The case of Korean letters in Table 6 shows the Homology Groups. One can use Homology as a tool for the discrimination task. In this example, the 1<sup>st</sup> homology group ($H_1$) is able to distinguish a single letter of “ㅃ” since it is the only letter that has first Betti number as two.

<table>
<thead>
<tr>
<th>$\beta_1$</th>
<th>Letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>{ㄱ,ㄲ,ㄴ,ㄷ,ㄸ,ㄹ,ㅁ,ㅂ,ㅅ,ㅆ,ㅈ,ㅉ,ㅊ,ㅋ,ㅌ,ㅍ,ㅎ}</td>
</tr>
<tr>
<td>1</td>
<td>{ㅁ,ㅂ,ㅇ,ㅍ,ㅎ}</td>
</tr>
<tr>
<td>2</td>
<td>{ㅃ}</td>
</tr>
</tbody>
</table>

**Table 6: Discriminate Korean consonant letters**

A **barcode** plot is a qualitative visualization of a homology group that looks like a collection of horizontal line segments. The x-axis represents the parameter $\epsilon$ and the y-axis is an arbitrary ordering of homology generators. The length of line is a key element to exam a barcode plot and shows that the $\epsilon$-window for which a feature is a part of the space. Short lines are noise because it cannot last long enough over the windowing. Longer lines that span a good portion of the x-axis can represent features of data.
Figure 10: complexes for various $\epsilon$ values

Figure 11: The barcode plot of the data shown in Figure 10

Figure 10 and Figure 11 are taken by the following thesis (Paul et al., 2013). Let $X$ be a sample of points in $\mathbb{R}^2$. Figure 11 represents the barcode plot of Figure 10. While the red lines indicate the $\epsilon$-windows of the 0-dimensional features, the green line is for the 1-dimensional features. The blue intersecting vertical lines are the complex at a particular $\epsilon$.

Table 7 is an Iris flower data set ("Iris flower data set," 2015) consists of 50 samples of three different species of Iris flowers: Iris setosa, Iris versicolor and Iris virginica. It is plotted in R with package of phom (persistent homology) excluding species.
<table>
<thead>
<tr>
<th>Sepal length</th>
<th>Sepal width</th>
<th>Petal length</th>
<th>Petal width</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>0.2</td>
<td><em>I. setosa</em></td>
</tr>
<tr>
<td>4.9</td>
<td>3</td>
<td>1.4</td>
<td>0.2</td>
<td><em>I. setosa</em></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>6.4</td>
<td>3.2</td>
<td>4.5</td>
<td>1.5</td>
<td><em>I. versicolor</em></td>
</tr>
<tr>
<td>6.9</td>
<td>3.1</td>
<td>4.9</td>
<td>1.5</td>
<td><em>I. versicolor</em></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>6.3</td>
<td>3.3</td>
<td>6</td>
<td>2.5</td>
<td><em>I. virginica</em></td>
</tr>
<tr>
<td>5.8</td>
<td>2.7</td>
<td>5.1</td>
<td>1.9</td>
<td><em>I. virginica</em></td>
</tr>
</tbody>
</table>

**Table 7: A sample of Iris data set**

Figure 12 is a barcode plot of 0-dimensional and 1-dimensional of the iris flower data set. One of intersecting vertical lines at (i), one can assume that there are three complexes without hole. Three barcode lines in 0-dimensional indicate three complexes and no particular line in 1-dimensional indicates no hole at (i). Similarly, one can assume that there are two complexes with one hole at (ii).

![Barcode plot of the Iris flowers data. $H_0$ is on the top, $H_1$ is on the bottom.](image)

*Figure 12: Barcode plot of the Iris flowers data. $H_0$ is on the top, $H_1$ is on the bottom.*
Following R code is used in order to plot the barcode of 0-dimensional. For the barcode of 1-dimensional, one can alter the value of max_dim from 0 to 1.

```r
install.packages("phom") # phom installation
library(phom) # load a library
data = as.matrix(iris[, -5]) # load the iris data excluding the species

metric_space = "euclidean"
max_dim = 0 # define a dimension of persistent homology
max_f = 1 # define a maximum scale of epsilon
mode = "vr" # type of filtration complex, either "vr" or "lw"
# "vr": Vietoris-Rips
# "lw": lazy-witness, two additional parameters should be defined

intervals = pHom(data, max_dim, max_f, mode, metric_space)
plotBarcodeDiagram(intervals, max_dim, max_f, title="Barcode plot of Iris data")
```
Chapter 3
Topological Data Analysis of Ayasdi

3.1 Construct Topological Network

3.1.1 Mapper

The core process of Ayasdi, which is the creation of topological networks, is called Mapper, proposed by Gurjeet Singh who is one of Ayasdi founders, back then was a PhD student in Mathematics department at Stanford University. Mapper compresses high dimensional data set and constructs topological networks that are discrete and combinatorial objects. While Ayasdi’s TDA keeps the three essential features of Topology, the internal procedure is different from conventional TDA. As shown in Figure 13, Ayasdi’s TDA is more intuitive than conventional TDA. At the same time, it does not demand a deeper understanding of mathematics.

3.1.1.1 Filter and metrics

The first step is converting raw data set into a point cloud data, using filter and metric. Ayasdi’s TDA is less sensitive to the metric selection compared to conventional TDA since the projection is driven by the selected filter. Filter can be used as a standalone or it uses the output of metric as its input.
3.1.1.2 Resolution

The second step is defining a resolution for the range of partitioning and overlapping. Corresponding to the defined resolution scale, point cloud data preserves the kernel information to create nodes and edges for the remaining steps.

3.1.1.3 Clustering

The third step is clustering. Mapper uses any clustering algorithm from data mining discipline in order to create node elements. In other words, node is a clustered representation of group of data samples that is defined in the previous step.

3.1.1.4 Connection

Finally, Mapper constructs the output of topological network by referencing redundant data points. By using these overwrapped data points, Mapper is able to know how to connect each node. If node contains no shared data sample, it remains as singleton.
3.1.2 Elements of Output Network

The final output of a topological network consists of multiple nodes and edges. A node can contain multiple samples and samples can appear in multiple nodes. The edge contains the redundant data samples. An edge as shown in Figure 14 is a shared data sample that appears in node A and node B. Figure 14 shows nodes and edges in each step of the pipeline. Node A contains five data samples and node B contains three data samples.

3.1.3 Metric Selection

The construction of topological networks is automated in Ayasdi. Multiple outputs are constructed by applying all possible combination of metrics and filters. Therefore, the output provides sufficient information. Alongside, manual decision might be considered to decrease information redundancy. For example, let us assume that we have some domain knowledge of a gene expression profiling. In this domain, relative expression values among the genes or probes to conserve meaningful information more than absolute gene expression levels. Thus, instead of applying all metrics, one should select a metric that can measure a distance of numerical data points. The relevant metrics are Cosine, Angle, Correlation or Euclidean. In contrast, the metrics that measures a frequency of value in context or measures time series could not properly summarize the input data. Table 8 proposes critical considerations for metric selection.

<table>
<thead>
<tr>
<th>If you want to measure...</th>
<th>Cosine</th>
<th>Angle</th>
<th>Correlation</th>
<th>Euclidean</th>
<th>Jaccard</th>
<th>Else</th>
</tr>
</thead>
<tbody>
<tr>
<td>...a similarity of continuous values?</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>...a frequency of value in each column?</td>
<td>☑</td>
<td>☐</td>
<td>☐</td>
<td>☑</td>
<td>☑</td>
<td>☐</td>
</tr>
<tr>
<td>...a frequency of value in the corpus?</td>
<td>☑</td>
<td>☑</td>
<td>☐</td>
<td>☑</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

Table 8: Metric Selection Considerations

*Figure 14: Elements of topological network*
… time-series?  ☑  ☑  ☑  ☐  ☐  ☐  ☐
… a similarity of categorical value?  ☑  ☐  ☐  ☐  ☐  ☐  ☐  Hamming, Categorical
… a collection of things?  ☐  ☐  ☐  ☐  ☐  ☑  ☐  ☐

Table 8: Decision criteria for metric selection

3.1.4 Filter Selection

Filter functions summarize relevant information from the noisy original data. Each filter function summarizes the input data in a different manner. Accordingly, each summary is able to preserve different information. Moreover, multiple filters can be associated to build higher dimensional complex. M numbers of filters are able to find a covering of RM (M dimensional hypercube). They are considered as a Cartesian product.

For instance, Figure 15 represents a filter function as a metaphor (Note that the input data at the left top is a synthesized geometrical data). Filter function $f$ summarizes the input data from a vertical perspective whereas filter $g$ summarizes the input data from a horizontal perspective.

Similarly, a movie can be another informal metaphor. A movie can be described by multiple summarized information: story, characters, genre, language, director and more. Although one individual summary does not help a lot to capture the full picture, multiple summaries are able to provide a decent overview. Some filter functions (e.g. density estimator and centrality measurement) are universally informative across the board, while other filter functions produce summaries for a particular purpose. Table 9 can be taken into account when one has to select which filter function to use.

![Figure 15: A metaphor of filter function to summarize original input data set](image-url)
<table>
<thead>
<tr>
<th>Goal</th>
<th>Suitable Type</th>
<th>Filter</th>
<th>Description</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>to understand a mathematical context surrounding the data points in the space</td>
<td>geometric filter</td>
<td>Gaussian density</td>
<td>$\sum_y e^{-\frac{d^2(x, y)}{2h}}$</td>
<td>considers each row as Euclidean data to estimate the density</td>
</tr>
<tr>
<td>to magnify difference between groups</td>
<td>projection filter</td>
<td>L1 centrality:</td>
<td>$\sum_y d(x, y)$</td>
<td>measures row centrality relative to others</td>
</tr>
<tr>
<td></td>
<td></td>
<td>L-infinity centrality</td>
<td>$\max_y d(x, y)$</td>
<td>associates a maximum distance from X to others</td>
</tr>
<tr>
<td>to identify anomalous behavior</td>
<td>summary statistics</td>
<td>PCA COORD 1&amp;2</td>
<td>generate PCA coordinates with the highest and the second-highest variance component</td>
<td>- assumes data is using the Euclidean metric - use both together for XY coordinates of 2-d embedding</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Neighborhood 1&amp;2</td>
<td>generate a two-dimensional embedding of the k-nearest neighbor graph by connecting point to its nearest neighbors</td>
<td>- uses the selected metric - emphasize the metric structure of data - do not equalize filter function</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>computes an average across a row</td>
<td>- suitable to measure similar phenomena in different coordinates or columns have been normalized to make comparable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Variance</td>
<td>computes a variance of each column in a row</td>
<td>- suitable in similar circumstances as the Mean - provide more insights with Mean, Median or Max filters</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Approximate kurtosis</td>
<td>approximates a kurtosis (peak) of data whether flat or sharp</td>
<td>- suitable to treat the values of your data as points in a distribution to detect anomalous distribution</td>
</tr>
</tbody>
</table>

**Table 9: Decision criteria of filter function**
3.2 Insight Discovery

As opposed to the automated network creation from the previous section 3.1, the elements introduced in this section depend on manual investigation. In order to enhance the data understanding, there are two methods are available: Color spectrum and Subgroup comparison.

3.2.1 Color Spectrum

The varying color spectrum is a primary step used for the insight discovery. The range of color spectrum is able to deliver graphical intuitions of value distribution and correlation among the variables, which reach beyond statistical methods. Highly correlated variables are represented as similar color patterns over the topological network. Furthermore, we are able to recognize which shape contains more information than other shapes. The insights can be used for further exploration in the Chapter 3.2.2 Subgroup Comparison.

Color ranges over red to blue and it has different meanings, depending on the type of attributes. For the continuous values, color represents an average of value. A red node contains data samples that have higher average values. In contrast, a blue node contains lower average values. In contrast, for the categorical values, color represents a value concentration.

For example, Figure 16 shows the variation of color spectrums according to the chosen attribute over the output of topological network. The given topological network are extracted from the data set that contains the information of Titanic passengers consists of 891 samples with eight numerical attributes including a label. The label discriminates either the passenger that survived or died. Now, refer to the the color spectrum of the attribute “Sex”. The red node contains only male passengers (high concentration), whereas green node hardly contains any male passengers (low concentration) while the red node contains no male passengers at all.

3.2.2 Subgroup Comparison

The comparison of particular shapes of networks provides a ranking of information related to variables. Please note that a subgroup should contain a minimum of 30 samples in order to avoid a biased comparison. One can compare either a subgroup to another subgroup, or a subgroup to remaining shapes. The results of the comparison are always two lists of variable ranking: Numerical variable ranking and Categorical variable ranking. By reference the top ranked variables, we are able to recognize the underlying features that drives subgrouping.
3.2.2.1 Rank methods

Ayasdi provides two methods to rank a variable correlation: The first one is Kolmogorov-Smirnov (KS) statistic for the Continuous type values. The second one is referred to as p-
Value for the Categorical type. KS statistic is a test of the likelihood that two groups have the same distribution of values for a column. The result bounds between [0, 1] and cut off value is 0.5. If the value is less than 0.5, the variable is not significant and can most likely be considered “noise”. Additionally, the p-Value describes the probability that this categorical value would be as common in the group as observed by chance alone. The result is bounded between [0, 1] and the cut off value is 0.95. If a variable is not ranked in the top range from 0 to 0.05, it should be ignored (“Kolmogorov–Smirnov test,” 2015).

### 3.2.2.2 Insight by Ranked Variables

Going back to the Titanic example, the result of the KS-statistic show, that the variable “Sex” is the most strongly related to passengers death. We could generally assume that men conceded the places in lifeboats to women. Furthermore, it is feasible to deduct the subtle reasons of the death of each group. The passengers in group A died because of two reasons: they were man and the cabin class type was low. The passengers in the group B died because they were man. Finally, the passengers in the group C died because they were staying at third class even though most of them were women.

<table>
<thead>
<tr>
<th>ID</th>
<th>Class</th>
<th>Sex</th>
<th>Age</th>
<th>#SibSp</th>
<th>#ParCh</th>
<th>Fare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survivor vs Death</td>
<td>0.06</td>
<td>0.33</td>
<td>0.53</td>
<td>0.09</td>
<td>0.11</td>
<td>0.13</td>
</tr>
<tr>
<td>Survivor vs Death A</td>
<td>0.10</td>
<td><strong>0.65</strong></td>
<td>0.68</td>
<td>0.16</td>
<td>0.39</td>
<td>0.32</td>
</tr>
<tr>
<td>Survivor vs Death B</td>
<td>0.07</td>
<td>0.20</td>
<td><strong>0.67</strong></td>
<td>0.21</td>
<td>0.04</td>
<td>0.23</td>
</tr>
<tr>
<td>Survivor vs Death C</td>
<td>0.12</td>
<td><strong>0.53</strong></td>
<td>0.06</td>
<td>0.19</td>
<td>0.29</td>
<td>0.36</td>
</tr>
</tbody>
</table>

Table 10: Variable ranking by KS-Statistic
Chapter 4
Use Cases

This chapter introduces three applied cases. The data set of each one has a different level of complexity: there is a low dimensional data set, a high dimensional data set in terms of the number of attributes and a data set which is particularly large in size. Especially the third case is challenging since it uses Ayasdi as well as multiple tools running in a Hadoop ecosystem.

4.1 Case 1: Predicting Energy Consumption

4.1.1 Business Understanding

Can daily energy consumption be predictable based on weather and calendar information? The goal is to detect the critical variables that are highly correlated to the daily energy consumption. The accurate prediction ensures that an energy plant produces the right amount of energy in order to minimize the costs (“Forecasting Challenge 2015,” 2015).

4.1.2 Data Understanding

A training data set is given by RWE nPower, an energy utility/provider plant in U.K. Therefore, this use case does not include the initial phase of data collection. The given data consists of 1,096 rows with 9 variables including one label of volume. The label is a target variable that data scientists want to classify or predict. A sample of the data set is shown in Table 11. The entire data set is available online (“Training data of energy consumption,” 2015).
Table 11: Sample data set of energy consumption

4.1.3 Data Preparation

4.1.3.1 Format Data

The process of formatting data modifies the existing raw data values without changing its meaning. This is a particularly critical phase since TDA is not able to measure a correlation of different types. For instance, it is obvious that there is no similarity between numerical and categorical variables. Therefore, the values of day type are converted into numerical values as shown in Table 12. The value of “WE” is mapped to the 0, “SA” to the 1, and “SU” to the 2. Accordingly, the data set below contains only numerical attributes.

Table 12: transformed data set from Table 11
4.1.4 Applying Ayasdi

A normalized angle is selected as a metric since the data set consists of numerical attributes with different scales. Neighborhood 1&2 are selected in order to project a point cloud formulated by selected metrics into a 2-dimensionsal space.

4.1.4.1 Color Spectrum

The color spectrum of volume can be found on the top left in Figure 18. Red nodes contain the data samples of high-energy consumption, while blue node contains data of lower energy consumption. Some attributes such as temp and wind have arbitrary color patterns. Other attributes such as day_type and school_holiday show a similar color pattern as the label variable.

![Color Spectrum](image)

**Figure 18: Color spectrum corresponding to the chosen attribute of case 1**

4.1.4.2 Discovered Insights

The attributes day_type and school_holiday are highly correlated to the label attribute due to the similar pattern of color spectrum. Similarly, we are able to interpret that energy consumption is high during both the weekdays and non-school holidays. The other attributes are considered as noise due to the different color patterns.
4.2 Case 2: Customer Churn Analysis

4.2.1 Business Understanding

Can we predict potential clients who will terminate a given service? Predictive knowledge on customers is a primary element in order to be proactive rather than reactive. The goal is to detect the critical variables that have a high correlation with a target variable, known as label, out of a large number of input variables. The target variable in this case is churn, which represents the tendency of customers to switch to other providers. (Orange telecom, 2009)

4.2.2 Data Understanding

4.2.2.1 Describe Data

Data collection is not necessary due to a training data set which is given by the French telecom company, Orange. It is a large marketing database that consists of 50,000 rows with 230 attributes including heterogeneous noisy data (numerical and categorical values), and unbalanced class distributions. The label is churn, which is a binary categorical variable. The sample data set is shown in Table 13 and the entire data set is available online (“orange_small_train.data.zip,” 2009).

<table>
<thead>
<tr>
<th>no</th>
<th>churn (label)</th>
<th>Var1</th>
<th>Var2</th>
<th>Var3</th>
<th>Var4</th>
<th>Var5</th>
<th>Var6</th>
<th>Var7</th>
<th>...</th>
<th>Var228</th>
<th>Var229</th>
<th>Var230</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1</td>
<td>1526</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F2FyR07IdsN7I</td>
<td>xb3V</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>525</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F2FyR07IdsN7I</td>
<td>fKCe</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
<td>5236</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ib5G6X1eUxUn6</td>
<td>am7c</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F2FyR07IdsN7I</td>
<td>FSa2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F2FyR07IdsN7I</td>
<td>mj86</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Zy3gnGM</td>
<td>am7c</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>49999</td>
<td>-1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RAYp</td>
<td>F2FyR07IdsN7I</td>
<td></td>
</tr>
<tr>
<td>50000</td>
<td>1</td>
<td>1694</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RAYp</td>
<td>F2FyR07IdsN7I</td>
<td></td>
</tr>
</tbody>
</table>

Table 13: Sample of training data set
4.2.3  Data Preparation

4.2.3.1  Format Data

As mentioned in the previous subsection 4.1.3.1, this is essential step since TDA is not able to measure a similarity of different value types. However, it is unable to modify all categorical values to the numerical values due to the complexity. It requires time and hardware resources, which are beyond the one available for this thesis (CPU: Inter Xeon E5420 @ 2.5 GHz, RAM: 8 GB).

4.2.4  Applying Ayasdi

Normalized Correlation is selected as a metric since the data set consists of numerical attributes which have different scales. Neighborhood 1&2 are selected to project into a 2-dimensional space. Additionally, label attribute is selected as a data filter function in order to project a discretized shape. Data filter is an optional projection function that is similar to the supervised learning analysis.

4.2.4.1  Color Spectrum

Figure 19 illustrates different color spectrum patterns of chosen variables. The color is illuminated according to the preserved values. For example, variable 1 on the left does not illuminate colors over the network as opposed to variable 65. This means that most values of variable 1 contain a null value. Variable 26 highlights color over the particular nodes in a similar way than variable 1 does. This means that both variables have null values for similar samples. Additionally, the value distribution of variable 26 is narrower than variable 1, since it does not show a multiple color spectrum as variable 1 does.

![Figure 19: Color spectrum corresponding to the chosen attribute of case 2](image)
4.2.4.2 Subgroup Comparison

Discovering insights by color patterns is infeasible in this case since the data set contains too many attributes that are of a heterogeneous type. In such cases, a method of subgroup comparison should be applied. Depending on the label variable of churn, two subgroups are defined as shown in Figure 15. A subgroup of red is a churn group that conserves a value of “1”. Similarly, another subgroup of blue is a loyal group that contains a value of “-1”.

Figure 20: Determined Two subgroups

4.2.4.3 Discovered Insights

There is no insight that we can discover by a ranking of numerical variables, since all scores of the KS-statistic are less than cut off value (0.5) as shown in Table 13. On the other hand, a ranking of categorical variables delivers information of associated variables. Table 14 represents a list of categorical variables that are smaller than the cut off value (0.05). After one removed the duplicated values in the attribute “Column Name”, only 49 variables are remained. These 49 variables preserve a strong signal that drives the subgrouping.

<table>
<thead>
<tr>
<th>Column Name</th>
<th>KS Statistic</th>
<th>KS p-value</th>
<th>T-test p-value</th>
<th>Group 1 Mean - Group 2 Mean</th>
<th>KS Sign</th>
<th>Fraction Null in Group 1</th>
<th>Fraction Null in Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Var64</td>
<td>0.406667</td>
<td>0.044349</td>
<td>0.210897</td>
<td>11683.02</td>
<td>+</td>
<td>0.996675</td>
<td>0.995142</td>
</tr>
<tr>
<td>Var53</td>
<td>0.234969</td>
<td>0.028154</td>
<td>0.084799</td>
<td>-303042</td>
<td>-</td>
<td>0.98864</td>
<td>0.985792</td>
</tr>
<tr>
<td>Var168</td>
<td>0.220772</td>
<td>0.044349</td>
<td>0.455796</td>
<td>12.4286</td>
<td>+</td>
<td>0.98864</td>
<td>0.985792</td>
</tr>
<tr>
<td>Var92</td>
<td>0.198718</td>
<td>0.644019</td>
<td>0.908284</td>
<td>-6461.07</td>
<td>-</td>
<td>0.995844</td>
<td>0.996632</td>
</tr>
<tr>
<td>Var190</td>
<td>0.185897</td>
<td>0.496191</td>
<td>0.82272</td>
<td>1661.541</td>
<td>+</td>
<td>0.994181</td>
<td>0.993263</td>
</tr>
</tbody>
</table>
### Table 14: Ranked numerical values by KS-statistic

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Value</th>
<th>Fraction in Group 1</th>
<th>Fraction in Both Group 1 and Group 2</th>
<th>Count in Group 1</th>
<th>Count in Both Group 1 and Group 2</th>
<th>Hypergeometric p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>appetency</td>
<td>-1</td>
<td></td>
<td>0.982232</td>
<td>3609</td>
<td>49035</td>
<td>1.00E-12</td>
</tr>
<tr>
<td>upselling</td>
<td>-1</td>
<td></td>
<td>0.926245</td>
<td>3609</td>
<td>46240</td>
<td>1.00E-12</td>
</tr>
<tr>
<td>churn</td>
<td>1</td>
<td></td>
<td>0.072293</td>
<td>3609</td>
<td>3609</td>
<td>1.00E-12</td>
</tr>
<tr>
<td>Var202</td>
<td>PXLV</td>
<td>8.31E-04</td>
<td>6.01E-05</td>
<td>3</td>
<td>3</td>
<td>3.78E-04</td>
</tr>
<tr>
<td>Var199</td>
<td>Gai9IEF2Fr</td>
<td>0.012192</td>
<td>0.007171</td>
<td>44</td>
<td>358</td>
<td>4.19E-04</td>
</tr>
<tr>
<td>Var198</td>
<td>Z4hPoJV</td>
<td>0.00194</td>
<td>4.21E-04</td>
<td>7</td>
<td>21</td>
<td>4.82E-04</td>
</tr>
<tr>
<td>Var222</td>
<td>xjJRusu</td>
<td>0.00194</td>
<td>4.21E-04</td>
<td>7</td>
<td>21</td>
<td>4.82E-04</td>
</tr>
<tr>
<td>Var220</td>
<td>25c0nTz</td>
<td>0.00194</td>
<td>4.21E-04</td>
<td>7</td>
<td>21</td>
<td>4.82E-04</td>
</tr>
<tr>
<td>Var216</td>
<td>TDe_9Yi</td>
<td>0.006096</td>
<td>0.002844</td>
<td>22</td>
<td>142</td>
<td>5.66E-04</td>
</tr>
<tr>
<td>Var99</td>
<td>136</td>
<td>0.001108</td>
<td>1.40E-04</td>
<td>4</td>
<td>7</td>
<td>7.99E-04</td>
</tr>
<tr>
<td>Var198</td>
<td>IVL4sPt</td>
<td>0.001385</td>
<td>5.61E-04</td>
<td>5</td>
<td>28</td>
<td>0.047965</td>
</tr>
<tr>
<td>Var198</td>
<td>M0ICMvZ</td>
<td>0.001385</td>
<td>5.61E-04</td>
<td>5</td>
<td>28</td>
<td>0.047965</td>
</tr>
<tr>
<td>Var198</td>
<td>LJS7NLP</td>
<td>0.001385</td>
<td>5.61E-04</td>
<td>5</td>
<td>28</td>
<td>0.047965</td>
</tr>
<tr>
<td>Var222</td>
<td>alq7an</td>
<td>0.001385</td>
<td>5.61E-04</td>
<td>5</td>
<td>28</td>
<td>0.047965</td>
</tr>
<tr>
<td>Var222</td>
<td>y1zz20V</td>
<td>0.001385</td>
<td>5.61E-04</td>
<td>5</td>
<td>28</td>
<td>0.047965</td>
</tr>
<tr>
<td>Var222</td>
<td>Chlj72z</td>
<td>0.001385</td>
<td>5.61E-04</td>
<td>5</td>
<td>28</td>
<td>0.047965</td>
</tr>
<tr>
<td>Var220</td>
<td>k5Om0jM</td>
<td>0.001385</td>
<td>5.61E-04</td>
<td>5</td>
<td>28</td>
<td>0.047965</td>
</tr>
<tr>
<td>Var220</td>
<td>Af96s0w</td>
<td>0.001385</td>
<td>5.61E-04</td>
<td>5</td>
<td>28</td>
<td>0.047965</td>
</tr>
<tr>
<td>Var220</td>
<td>rDm3DH0</td>
<td>0.001385</td>
<td>5.61E-04</td>
<td>5</td>
<td>28</td>
<td>0.047965</td>
</tr>
<tr>
<td>Var197</td>
<td>yMvB</td>
<td>0.009421</td>
<td>0.007011</td>
<td>34</td>
<td>350</td>
<td>0.049324</td>
</tr>
</tbody>
</table>

### Table 15: Ranked categorical values by p-value that bigger than 0.05
4.3 Case 3: Earthquake Prediction

4.3.1 Business Understanding

Can we detect the precursors that can predict an earthquake? While major earthquakes have killed tremendous amounts of people, no scientists was yet able to predicted a specific time, location or magnitude of an earthquake. Most studies are based on the statistical estimation of earthquake history. Therefore, it is restricted to the prediction of an accurate hazard. The focus of prediction is general for a long-term hazard prediction to improve the safety of infrastructure. The mathematical algorithm or domain knowledge is not able to forecast precise earthquakes. Nevertheless, no one yet has experimented to understand earthquakes through analyzing a big data set in terms of volume, veracity and variety. The intention of this chapter is to explore this topic and provide valuable insights, which could contribute to future research.

The problem domain of this use case includes additional tedious issues. The first and most important one is that there is no given training data set for the problem domain. This is an initial issue that data scientists face across the discipline. Furthermore, the problem domain lies within a social sector, which is fundamentally challenging and complex. Data scientist must collect a quality of data sets that are open to the public. However, in most of the cases, the data remains in silos. As a consequence, data scientists cannot tackle the problem, although they have the knowledge and will to do so.

4.3.2 Data Understanding

4.3.2.1 Collect Raw Data

The initial task is to collect the raw data. Fortunately, the problem domain ranges from the social to the science sector. Useful data sets in social sector are seldom available. If they are, they are still rarely available online for the public. For instance, some scientists insist that animals with a keen sense can feel P waves before S waves arrive. Often press reports that some animals escaped days before a major earthquake. Besides no one has yet recorded this data, it is not possible to do so worldwide in real time. Therefore, this use case targets only scientific data sets collected by science institutions such as National Aeronautics and Space Administration (NASA), University, or related government agencies.
Once data sets are collected, all data sets should be transferred into a Hadoop Distributed File System (HDFS). Apache Hadoop is an open-source software framework that can control distributed and scalable computer clusters in order to process a massive amount of data. Two core functions of Apache Hadoop are a storage part (HDFS) and a processing part (MapReduce). There are multiple open-source expansion packs running on top of the Hadoop ecosystem, such as Hue, Pig, Spark, Hive et cetera. In order to store the collected data into the HDFS, one can use Hue. Hue has a user-friendly web interface that supports the most components of Hadoop. The Hue allows one to browse and manipulate files and directories in HDFS. Figure 21 shows a simple data transmission to the HDFS though Hue within three clicks rather than complex UNIX command lines.

**4.3.2.2 Select Data**

In general, raw data conserves too much noise to be used for analysis. Removing the irrelevant values enhances the quality of analysis result, saves computational resources and consequently time. The included variables are highlighted with a yellow background.

**4.3.2.3 Describe Data**

---

**Figure 21: Simple data transmission in Hue**

The image shows a screenshot of the Hue File Browser interface, demonstrating the process of uploading data. The screenshot includes the steps for data transmission, highlighting the simplicity of using Hue compared to complex UNIX command lines.
• Earthquake

The data set of earthquake history has been archived since 1904. It is provided by the United States Geological Survey (USGS) to reduce earthquake hazards on a global scale. The data set consists of 870,013 rows with 14 variables without a label. A sample of the data set is shown in Table 16. The entire data set, including a glossary, is available online (U.S. Geological Survey, 2015).

<table>
<thead>
<tr>
<th>time</th>
<th>latitude</th>
<th>longitude</th>
<th>depth</th>
<th>mag</th>
<th>magType</th>
<th>nst</th>
<th>gap</th>
<th>dmin</th>
<th>rms</th>
<th>net</th>
<th>id</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015-01-01T00:05:41.000Z</td>
<td>62.8808</td>
<td>-149.155</td>
<td>20.3</td>
<td>0.7</td>
<td>ml</td>
<td>12</td>
<td>108</td>
<td>0.52</td>
<td>ak</td>
<td>ak11477119</td>
<td>earthquake</td>
<td></td>
</tr>
<tr>
<td>2015-01-01T00:07:08.764Z</td>
<td>41.8476</td>
<td>-119.655</td>
<td>6.548</td>
<td>1.01</td>
<td>ml</td>
<td>4</td>
<td>224.88</td>
<td>0.15</td>
<td>0.2041</td>
<td>nn nn00474720</td>
<td>solar wind proton</td>
<td></td>
</tr>
</tbody>
</table>

Table 16: Sample of earthquake data (1904-2015)

• Solar Wind Proton

The data set contains densities, speeds, and thermal speeds since 1998 up to the present day. ACE Science Center (ASC) ensures that the data set are properly archived and publicly available. The sensor used is a Solar Wind Ion Composition Spectrometer (SWICS). It is optimized for measuring the chemical and isotopic composition of solar and interstellar matters. The data set consists of 757,149 rows with 6 variables. A sample of the data set is shown in Table 17. The entire data set including a glossary is available online (ACE Science Center, 2015a).

<table>
<thead>
<tr>
<th>year</th>
<th>day</th>
<th>hr</th>
<th>min</th>
<th>nH</th>
<th>vH</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>28</td>
<td>21</td>
<td>53</td>
<td>-1.00E+04</td>
<td>-1.00E+04</td>
</tr>
<tr>
<td>1998</td>
<td>28</td>
<td>22</td>
<td>5</td>
<td>-1.00E+04</td>
<td>-1.00E+04</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>

Table 17: Sample of solar wind proton history (1998-2015)
- Interplanetary Magnetic Field

The data set of interplanetary magnetics has been archived since 1997 up to the present. Data provider is ACE Science Center (ASC) same as the one provides Solar Wind Proton. It consists of 6,427 rows with 3 variables including time and magnetic field value. The value is measured by twin vector fluxgate magnetometers. They are mounted on a spacecraft at opposite sides. Data error is less than 0.1 nano-Tesla. A sample of the data set is shown in Table 18. The entire data set including a glossary is available online (ACE Science Center, 2015b).

<table>
<thead>
<tr>
<th>hyear</th>
<th>Day</th>
<th>Bmag</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>226</td>
<td>-999.9</td>
</tr>
<tr>
<td>1997</td>
<td>227</td>
<td>-999.9</td>
</tr>
<tr>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
</tr>
</tbody>
</table>

Table 18: Sample of interplanetary magnetic (1997-2015)

- Active Fire

The data set contains geographic locations, dates, and some additional information for each fire pixel. These information have been detected by Terra and Aqua MODIS sensors since 2001 up to present. The file size of active fire history is 3.51 GB. A sample of the data set is shown in Table 19. The entire data set including a glossary is available online (NASA, 2015).

<table>
<thead>
<tr>
<th>latitude</th>
<th>longitude</th>
<th>brightness</th>
<th>scan</th>
<th>track</th>
<th>acq_date</th>
<th>acq_time</th>
<th>satellite</th>
<th>confidence</th>
<th>version</th>
<th>bright_t31</th>
<th>frp</th>
</tr>
</thead>
<tbody>
<tr>
<td>-16.162</td>
<td>128.607</td>
<td>323</td>
<td>3.2</td>
<td>1.7</td>
<td>1/1/2001</td>
<td>123</td>
<td>T</td>
<td>56</td>
<td>5.1</td>
<td>303.2</td>
<td>52.8</td>
</tr>
<tr>
<td>-17.465</td>
<td>127.889</td>
<td>326.3</td>
<td>3.4</td>
<td>1.7</td>
<td>1/1/2001</td>
<td>123</td>
<td>T</td>
<td>72</td>
<td>5.1</td>
<td>305.8</td>
<td>61.2</td>
</tr>
<tr>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
</tr>
</tbody>
</table>

Table 19: Sample of active fire (2001-2015)
• Volcano

A data set is available in the Significant Volcanic Eruptions Database provided by National Geophysical Data Center (NOAA, doi: 10.7289/V5JW8BSH). It contains global eruptions that include volcano related information. The data set consist of 1,556 rows with 12 variables. A sample of the data set is shown in Table 20. The entire data set including a glossary is available online.

<table>
<thead>
<tr>
<th>Year</th>
<th>Mo</th>
<th>Dy</th>
<th>Tsu</th>
<th>EQ</th>
<th>Name</th>
<th>Location</th>
<th>Country</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Elevation</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1904</td>
<td>2</td>
<td>25</td>
<td></td>
<td></td>
<td>Karthala</td>
<td>IndianO-W</td>
<td>Comoros</td>
<td>-11.75</td>
<td>43.38</td>
<td>2361</td>
<td>Shield volcano</td>
</tr>
<tr>
<td>1905</td>
<td>3</td>
<td>10</td>
<td></td>
<td></td>
<td>Vesuvius</td>
<td>Italy</td>
<td>Italy</td>
<td>40.821</td>
<td>14.426</td>
<td>1281</td>
<td>Complex volcano</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 20: Sample of volcano

• Heat Flow

The heat flow data is maintained by the International Heat Flow Commission (IHFC) of the International Association of Seismology and Physics of the Earth's Interior (IASPEI). The most recent global compilation consists of 35,523 continental points and 23,013 marine points. A sample of the data set is shown in Table 21. The entire data set, including a glossary, is available online (International Heat Flow Commission, 2015).

<table>
<thead>
<tr>
<th>Data No</th>
<th>Site Name</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Elevation</th>
<th>minD</th>
<th>maxD</th>
<th>No. Temps</th>
<th>Gradient</th>
<th>No. Cond.</th>
<th>Conductivity</th>
<th>Year of Pub.</th>
<th>Reference</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>SMU-SP3</td>
<td>44.3278</td>
<td>-112.213</td>
<td>1795</td>
<td>10</td>
<td>99</td>
<td>55</td>
<td>5</td>
<td>2.06</td>
<td>1983</td>
<td>Brott_etal1983</td>
<td>Brott_etal1983</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 21: Sample of heat flow
4.3.2.4 Verify Data Quality

Verifying the quality of data is a key to solve the fundamental problems. Examine the quality of the data while addressing following characteristics: Availability, compatibility, public accessibility, reliability, relevance and quantity. The detailed explanation of each term is introduced in the following Table:

<table>
<thead>
<tr>
<th>Availability</th>
<th>Data is available to the public without charging, ideally without signing up</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compatibility</td>
<td>Data format is identified without an installation of specific software</td>
</tr>
<tr>
<td>Convenience</td>
<td>Data is available as an archived (compressed) format. Providing data as HTML documents force one to use a web crawler</td>
</tr>
<tr>
<td>Reliability</td>
<td>Data must be accurate. Misspellings or missing values may cause a biased result</td>
</tr>
<tr>
<td>Relevance</td>
<td>Data preserves a common set of values to be joined together</td>
</tr>
<tr>
<td>Quantity</td>
<td>Data contains a decent amount of volume. Small size of data might cause a biased analysis</td>
</tr>
</tbody>
</table>

Table 22: Criteria of data quality

4.3.3 Data Preparation

Data preparation is recommended to be carried out outside of Ayasdi in this applied case, since Ayasdi allows only restricted join operations. Table 23 presents a different time format of five relations shown in Table 16, Table 17, Table 18, Table 19 and Table 20. Ayasdi cannot integrate them since each relation has a different time format.

<table>
<thead>
<tr>
<th>Name of a relation</th>
<th>time zone offset</th>
<th>month/day/year</th>
<th>year</th>
<th>month</th>
<th>day</th>
<th>day of year</th>
<th>hour</th>
<th>minutes</th>
<th>minutes of day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earthquake</td>
<td>2015-01-01T00:05:41.000Z</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solar wind</td>
<td></td>
<td>1998</td>
<td>28</td>
<td>21</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Magnetic</td>
<td></td>
<td>1997</td>
<td>226</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active fire</td>
<td>1/1/2001</td>
<td>1994</td>
<td>2</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Volcano</td>
<td></td>
<td>1994</td>
<td>2</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 23: Different data sets use different time formats
4.3.3.1 Pig Introduction

This thesis uses Pig for the data preparation, while one may use other tools such as MapReduce or Cascading. Pig is a high-level software for creating MapReduce programs. The language of Pig is called Pig Latin, similar to the Structured Query Language (SQL) for Relational Database Management System (RDBMS). It supports regular expression and abstracts the object oriented programming language into the simple SQL alike queries. Therefore, one is familiar to the conventional database that can process the MapReduce without learning how to program the MapReduce code. The version used in this thesis is 0.12.0 installed in Cloudera distribution version of 5.2.0.

Pig commands are running on the Grunt shell. A prompt of “grun$” indicates that user is in the Grunt shell. In terms of performance and efficiency, one can execute a single Pig script from shell script instead of running each command in the Grunt shell. Pig scripts can execute multiple Pig command lines sequentially. Consider using an additional option of shell script, called nohup command when one plans to run the Pig script. The “nohup” command prevents a remote connection abortion and allows a continuous operation in the background in any case. This is particularly useful if job takes a long processing time.

4.3.3.2 Construct Data

Data construction includes operations such as attributes derivation, new records creation, or value transformation from the existing attributes. The goal of this applied case is to construct a uniform time format out of the original relations shown in in Table 23. The new values with yellow background are derived from the original values with red strikethrough. This process makes sure that the reconstructed relations in Table 24 contain two common attributes: year and day of year. These common values will be used as a keyword of join operation in data integration (Subsection 4.3.3.3).

Pig Latin supports a set of built-in functions such as eval, load/store, math, string, bag and tuple functions (Apache Pig, 2013). Unlike the user defined functions (UDFs), one does not have to register the built-in functions before using. The built-in functions used in this case are DaysBetween, ToDate, CONCAT and SUBSTRING. The entire Pig scripts are available in Appendix A.2.
<table>
<thead>
<tr>
<th>Name of a relation</th>
<th>time zone offset</th>
<th>month/day/year</th>
<th>year</th>
<th>Month</th>
<th>day</th>
<th>day of year</th>
<th>hour</th>
<th>minutes</th>
<th>minutes of day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earthquake</td>
<td>2015-01-01T00:05:41.000Z</td>
<td>2015</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solar wind</td>
<td></td>
<td>1998</td>
<td>28</td>
<td>24</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Magnetic</td>
<td></td>
<td>1997</td>
<td>226</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active fire</td>
<td>1/1/2004</td>
<td>2001</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Volcano</td>
<td></td>
<td>1994</td>
<td>2</td>
<td>25</td>
<td>56</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 24: Constructed data sets from Table 23

4.3.3.3 Integrate Data

As a result of the previous phase, all data sets exist individually in HDFS. The aim is to integrate each one into a single and large tabular file. Pig Latin supports the relational operators that are available in SQL. The function can join two or more relations based on common field values. Inner joins ignore null keys. The used Pig scripts for this applied case are available in the Appendix.

Earthquake, Solar wind and Magnetic

The first data set is integrated based on time information. JOIN operator is used to integrate two or more relations based on common field values. The size of output is 5.4 GB and consists of 73,009,872 rows with 9 variables. However, Ayasdi could not process it and produced an “out of memory” error. The subsequently reduced data set (1.4GB, 18,928,485 rows with 9 variables) is also not able to be analyzed. The second reduced data set (230.8MB, 2,953,600 rows with 9 variables) is capable to be processed. The entire Pig script used is available in Appendix A.3.1.

<table>
<thead>
<tr>
<th>Common field value</th>
<th>by Earthquake data set</th>
<th>by Solar wind data set</th>
<th>by Magnetic data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>year</td>
<td>doy</td>
<td>latitude</td>
<td>longitude</td>
</tr>
<tr>
<td>------</td>
<td>-----</td>
<td>----------</td>
<td>-----------</td>
</tr>
<tr>
<td>2015</td>
<td>1</td>
<td>61.9826</td>
<td>-151.475</td>
</tr>
<tr>
<td>2015</td>
<td>1</td>
<td>65.2671</td>
<td>-148.974</td>
</tr>
<tr>
<td>2015</td>
<td>77</td>
<td>64.0963</td>
<td>-149.601</td>
</tr>
<tr>
<td>2015</td>
<td>77</td>
<td>-32.8455</td>
<td>179.8031</td>
</tr>
</tbody>
</table>

Table 25: Integrated data of Earthquake, Solar wind and Magnetic
Earthquake, Volcano and Heat flow

The second data set is merged based on locational information. UNION operator merges the contents of two or more relations. It consists of 898,578 rows with 2 variables. The Pig script used is available in Appendix A.3.2.

<table>
<thead>
<tr>
<th>Constructed field</th>
<th>Common field value</th>
</tr>
</thead>
<tbody>
<tr>
<td>flag</td>
<td>latitude</td>
</tr>
<tr>
<td>Earthquake</td>
<td>34.6</td>
</tr>
<tr>
<td>Earthquake</td>
<td>32.5</td>
</tr>
<tr>
<td></td>
<td>:</td>
</tr>
<tr>
<td>HeatFlow</td>
<td>20.8</td>
</tr>
<tr>
<td>HeatFlow</td>
<td>18.3</td>
</tr>
<tr>
<td></td>
<td>:</td>
</tr>
<tr>
<td>Volcano</td>
<td>40.827</td>
</tr>
<tr>
<td>Volcano</td>
<td>40.821</td>
</tr>
<tr>
<td>Volcano</td>
<td>40.73</td>
</tr>
</tbody>
</table>

Table 26: Integrated data of Earthquake, Volcano and Heat flow

4.3.3.4 Store Data

Unfortunately, Ayasdi is not a tool that is running on the Hadoop Ecosystem and does not yet have a driver to access to the HDFS directly. It only accepts a text file format as an input such as Comma Separated Values (CSV) or Tab Separated Values (TSV). Pig Latin supports some built-in functions that can archive an output result with a specified delimiter in HDFS. This use case stores the output as a CSV format using CSVExcelStorage function. The default delimiter of this function is comma (',').

4.3.4 Apply Ayasdi

Once data is prepared as a text format, Ayasdi can create the corresponding topological networks as shown in Figure 22 and Figure 23. Each node consists of multiple samples, while same samples can contain multiple nodes. A bigger size of the node indicates more samples, while a smaller node contains fewer samples. Norm Correlation is selected as a metric since the data set consists of numerical attributes with a different scale. Neighborhood 1&2 are selected as a projection. Data filter functions are not defined since there is no label attribute.
4.3.4.1 Color Spectrum

Figure 22 and Figure 23 illustrate different color spectrum patterns of chosen variables. The color is highlighted correspondingly to its average values. Red represents a high average value, whereas blue represents a low value.

Figure 22: Color spectrum of integrated data (Earthquake, Solar wind and Magnetic)
4.3.4.2 Discovered Insights

The color patterns in Figure 22 are arbitrary. Besides, one cannot say that earthquakes are highly correlated to a certain attributes, one can describe the data set as follows:

- Magnetic or Proton density is sometimes active, but sometimes not
- No pattern of proton speed means that it is a noise signal
- Great earthquakes do not recently occur more often

The color highlighted in Figure 23 shows interesting patterns. The volcano records and the strong earthquake records (magnitude is greater than 7) have a similar color pattern. One may assume that a great earthquake has a high correlation with the volcano area.

4.4 Lesson Learned

Understanding a data set is simple and intuitive because TDA can analyze the data set without any prior assumptions. Thus, a user, who is not an expert of statistics, is also capable to
analyze data sets in a few steps with shape and color. Furthermore, the insight discovery of TDA is similar to the dimensional reduction techniques. Both methods aim at discovering qualitative insights out of the irrelevant noisy variables. The quality assessment will be evaluated in chapter 5 by comparing prediction performance.
Chapter 5
Comparison

Chapter 5 compares the quality of findings derived by TDA and other feature engineering techniques (Section 0). Other techniques used in this thesis are PCA, MRMR and RF. Each technique creates an unique set of reduced variables. The comparison methodology (Section 0) follows the modeling and validation process in a data mining introduced earlier (Section 1.3). The criteria of performance include the robustness of prediction and reduced feature rate, but excludes time efficiency.

5.1 Feature Engineering

A formal definition of each feature engineering technique is shown in Table 27. For a review, see the following literature (Liu and Motoda, 1998). Feature selection is the process of choosing a subset of the original set of variables. Meanwhile, feature extraction extracts a set of new features from the original set of variables through functional mapping. Finally, feature construction augments additional features to the original set of variables.

<table>
<thead>
<tr>
<th></th>
<th>original variables set</th>
<th>new feature set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature Selection</td>
<td>$A_1, A_2, ..., A_n$</td>
<td>$A_1, A_2, ..., A_m$ ($m &lt; n$)</td>
</tr>
<tr>
<td>Feature Extraction</td>
<td></td>
<td>$B_1, B_2, ..., B_k$ ($k &lt; n$), $B_i = f(A_1, A_2, ..., A_n)$</td>
</tr>
<tr>
<td>Feature Construction</td>
<td></td>
<td>$A_1, A_2, ..., A_n, A_{n+1}, ..., A_{n+m}$</td>
</tr>
</tbody>
</table>

Table 27: Formal definition of feature engineering (Motoda and Liu, 2002)
Figure 24: Venn diagram of feature engineering techniques

Figure 24 depicts a Venn diagram that represents an example of feature engineering and related subsets. All subset techniques can be mutually inclusive. For example, feature construction can be used as a pre-processing step for any dimensional reduction technique.

5.1.1 Feature Selection

Feature selection is one of the dimensional reduction techniques. It does not modify the original variables. In other words, it chooses a subset of original variables without any reformation. Thus, feature selection can be applied to a certain domain where modification variables are not recommended. There are a few well-known applications: sequence analysis, microarray analysis, mass spectrum analysis, single nucleotide polymorphism analysis and text mining.

<table>
<thead>
<tr>
<th>Type</th>
<th>Fast</th>
<th>Scalable</th>
<th>Feature dependency</th>
<th>Model dependency</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Univariate)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filter</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>Kolmogorov-Smirnov (KS) test t-test</td>
</tr>
<tr>
<td>Multivariate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filter</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>Minimum Redundancy and Maximum Relevance (MRMR)</td>
</tr>
<tr>
<td>Wrapper</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>Forward selection Backward elimination</td>
</tr>
<tr>
<td>Embedded</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Random forest (RF) Weight vector of SVM</td>
</tr>
</tbody>
</table>

Table 28: Advantages (+) and disadvantages (-) of feature selection techniques (Saeys et al., 2007).

Depending on the selection behavior, feature selection is categorized into three methods: filter methods, wrapper methods and embedded methods. Table 28 summarizes the advantages and
disadvantages of each technique. The detail description of each technique is introduced as follows (Guyon and Elisseeff, 2003):

5.1.1 Filter methods

It ranks each feature by calculating a relevance score. Afterwards, it selects only high-scoring features. It is an easy and fast method, but it ignores feature dependencies. In order to challenge these disadvantages, a number of multivariate filter methods are proposed. Multivariate feature selections are the modified methods that aim at the feature dependencies.

5.1.2 Wrapper methods

It is dependent on a model hypothesis as opposed to the filter methods. Multiple subsets of features are generated and evaluated by training and testing. Thus, wrapper methods need a large amount of computational power.

5.1.3 Embedded methods

It is sensitive to the model hypothesis. However, they are less computationally expensive than the wrapper methods due to the integration with subset features and models.

5.1.2 Feature Extraction

Feature extraction is computationally more expensive than feature selection since it reconstructs a new set of features from the original variables. In general, feature extraction achieves better performances than feature selection except a few particular domains such as sequence analysis, microarray analysis, mass spectrum analysis, single nucleotide polymorphism analysis and text mining. Feature extraction is based on clustering or projection methods.

5.1.2.1 Clustering methods

It replaces a group of similar variables by a centroid cluster. The new set of features consists of the list of centroids. K-means and hierarchical clustering are well-known examples for such algorithms.
5.1.2.2 Projection methods

It creates a lower dimensional subspace by linear projection of original data. Principle Component Analysis (PCA) and Singular Value Decomposition (SVD) are the most common examples for such algorithms.

5.2 Comparison Methodology

The objective is to evaluate the discovered insights from use case 1 (Section 4.1) and use case 2 (Section 4.2). The use case 3 (Section 4.3) is not suitable for evaluation because the data set does not have a label which is a desired variable to be classified or predicted. Meanwhile, the evaluation method follows a data mining process from the data preparation, modeling and validation as shown in Figure 25.

Figure 25: Graphical representation of comparison methodology
RapidMiner is used to implement a procedure for comparison. RapidMiner is a software platform that supports all steps of the data mining process including results visualization, validation and optimization (Hofmann, 2013). The processes used in RapidMiner are presented in Figure 26 and Figure 27.

Figure 26: Evaluation process of case 1 in RapidMiner

Figure 27: Evaluation process of case 2 in RapidMiner
5.3 Evaluation of Case 1

5.3.1 Experiment Variables

In a scientific experiment, there are two types of variables- dependent and independent. The independent variable is the variable that is manipulated by a researcher to measure the effects of a dependent variable. The dependent variable is the response by the manipulated independent variable. Table 29 classifies the experiment variables used.

<table>
<thead>
<tr>
<th>Experiment variable</th>
<th>DM phase (Section 1.3)</th>
<th>Sub-steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent</td>
<td>Data Preparation</td>
<td>Filter Example Range (70:30)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Date to Number</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Map</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Replace Missing Values</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nominal to Numerical</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Normalize</td>
</tr>
<tr>
<td>Dependent</td>
<td>Modeling</td>
<td>Dimensional Reduction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Modeling</td>
</tr>
<tr>
<td>Independent</td>
<td>Evaluation</td>
<td>MAPE</td>
</tr>
</tbody>
</table>

Table 29: Experiment variable of use case 1

5.3.2 Data Preparation

The Filter Example Range splits the data range of the entire date set. The proportion of data is divided in the ratio 70:30. When the training set is not sufficient, the prediction performance is poor. As the training set increases, the prediction performance increases correspondingly. However, if the proportion of training data approaches 100%, the performance decreases due to the overfitting. Therefore, a sufficient test data is also critical for data mining. (Andrew, 2012)

The next three functions (e.g. Data to Numerical, Map and Nominal to Numerical) transform values into the numerical type in this experiment. Replace Missing Value inserts an average value when there is a null value. Normalization is a scaling process that casts the range of data into a specific boundary such as [0, 1] or [-1, 1]. Additionally, non-numerical variables should be removed before applying the method of PCA or MRMR.
5.3.3 Modeling

Dimensional reduction is a part of modeling phase. PCA, MRMR and TDA are used for the dimensional reduction. These dimensional reduction techniques calculate a weight value corresponding to the correlation with label variable. Then, Select Attribute function selects the attribute if the normalized weights is greater than 0.5. Note that the result of performance is not only dependent on the chosen reduction algorithm, but also on the chosen modeling algorithm. Therefore, this experiment creates two models, in order to prove the result is not dependent on the predictive model. The used algorithms are Neural Network and Support Vector Machine (SVM).

5.3.4 Evaluation

The models trained with the data set (70%), are applied on the remaining data set (30%) for the evaluation. Mean absolute percentage error (MAPE) is used as a measure of prediction accuracy. The better the modeling approaches the lower the error rate. R is additionally used to implement MAPE. The implementation of R code is available in Appendix B.

- MAPE = \( \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \), At: an actual value, Ft: a forecast value

Table 30 depicts the MAPE of each model depending on the chosen reduction technique. In general, neural network model produces better performances than the use of SVM. While TDA can reduce irrelevant variables as well as perform the second best prediction, the best performance is produced by neural network model with no dimensional reduction. If the input data set is particularly small and simple, modeling with no dimensional reduction can performs a decent prediction.

<table>
<thead>
<tr>
<th></th>
<th>Dimensional reduction</th>
<th>All</th>
<th>PCA</th>
<th>MRMR</th>
<th>TDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modeling</td>
<td>Reduction rate</td>
<td>0 %</td>
<td>66.67 %</td>
<td>88.89 %</td>
<td>77.78 %</td>
</tr>
<tr>
<td></td>
<td>Selected features</td>
<td>all</td>
<td>winter, solar_rad, temp</td>
<td>day_type</td>
<td>day_type, sch_holiday</td>
</tr>
<tr>
<td>Evaluation</td>
<td>Model by Neural Network</td>
<td>3.0546 %</td>
<td>11.1026 %</td>
<td>5.7003 %</td>
<td>3.6406 %</td>
</tr>
<tr>
<td></td>
<td>Model by SVM</td>
<td>10.9843 %</td>
<td>11.0649 %</td>
<td>10.6166 %</td>
<td>10.7778 %</td>
</tr>
</tbody>
</table>

Table 30: Performance comparison of use case 1
5.4 Evaluation of Case 2

Unlike previous evaluation of case 1, the data used in case 2 is placed in a high dimensional space. Compare to the simple data set, one can distinguish how dimensional reduction technique influences the prediction performance.

5.4.1 Experiment Variables

The definition of dependent and independent variables is introduced at previous Chapter 5.3.1. The experiment variables used in case 2 are classified in Table 31.

<table>
<thead>
<tr>
<th>Experiment variable</th>
<th>Phase of DM (Section 1.3)</th>
<th>Sub-steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent</td>
<td>Data Preparation</td>
<td>Filter Example Range (70:30) Set Role Replace Missing Values Normalize</td>
</tr>
<tr>
<td>Dependent</td>
<td>Modeling</td>
<td>Dimensional Reduction Modeling</td>
</tr>
<tr>
<td>Independent</td>
<td>Evaluation</td>
<td>F₁ Score</td>
</tr>
</tbody>
</table>

Table 31: Experiment variable of use case 2

5.4.2 Data Preparation

The proportion of data is divided in the ratio 70:30 by the reason described in subsection 5.3.2. The function of Set Role determines a label variable. The label in this experiment is the churn variable. Replace Missing Value replaces zero values by a null value. Normalization is a scaling process that casts the range of data into a specific boundary such as [0, 1] or [-1, 1]. Additionally, non-numerical variables should be removed before applying the method of PCA or MRMR.

5.4.3 Modeling

Similar to the evaluation of applied case 1, the modeling phase includes a dimensional reduction. The calculated weight values are filtered if they are greater than 0.5 by the Select Attribute function. Like the previous experiment, two models are created in order to prove
the result is independent on the predictive model. Unlike the modeling algorithms in case 1, this experiment utilizes Naïve Bayes and Decision Tree, which demand less computational costs. The model creation by Neural Network or Support Vector Machine requires the decent hardware resources, which are beyond the one available for this thesis (CPU: Inter Xeon E5420 @ 2.5 GHz, RAM: 8 GB).

5.4.4 Evaluation

Two trained models are applied on the test data set for the evaluation. This experiment uses a F$_1$ Score as a measure of prediction accuracy. The F$_1$ Score considers both the precision $p$ and the recall $r$ of the test to compute the score. $p$ is the number of correct positive results divided by the number of all positive results, and $r$ is the number of correct positive results divided by the number of positive results that should have been returned. F$_1$ Score reaches its best at 1 and worst at 0 (“F1 score,” 2015). F$_1$ Score is desired to use when the class distribution is unbalanced. The class distribution is the ratio between positive and negative values in the data. The class distribution of use case 2 is 10/90 (positive/negative) which is a biased distribution.

- F$_1$ Score = \[ 2 \frac{PR}{P+R} \]

Table 32 represents the F$_1$ Score of each model corresponding to the chosen reduction technique. In general, the model created with Naïve Bayes performs better prediction than one by Decision tree. Each method reduces the unrelated variables. Note that the highest reduction rate does not mean the best reduction method. If it reduces too much variables like PCA in this experiment, prediction performance is poor. The best performance of F$_1$ Score is produced by following two condition: Naïve Bayes model with TDA and no dimensional reduction. It is remarkable that TDA provides the same performance although it reduces 83.26 % of original data set.

<table>
<thead>
<tr>
<th>Modeling</th>
<th>Dimensional reduction</th>
<th>All</th>
<th>PCA</th>
<th>MRMR</th>
<th>TDA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reduction rate (no. of selected features)</td>
<td>0 % (0)</td>
<td>92.70 % (17)</td>
<td>57.08 % (100/default)</td>
<td><strong>83.26 % (39)</strong></td>
</tr>
<tr>
<td>Evaluation (F$_1$ Score)</td>
<td>Model by Naïve Bayes</td>
<td>0.147</td>
<td>0.005</td>
<td>0.146</td>
<td><strong>0.147</strong></td>
</tr>
<tr>
<td>Evaluation (F$_1$ Score)</td>
<td>Model by Decision tree</td>
<td>0.016</td>
<td>0.002</td>
<td>0.023</td>
<td><strong>0.036</strong></td>
</tr>
</tbody>
</table>

Table 32: Performance comparison of customer data
5.5 Lesson Learned

Interesting enough, the model with fewer features can predict better results than the one with all variables. The differences in prediction performance are more obvious when it comes to the data set in high dimensional space. Perhaps Ayasdi’s TDA does not achieve the greatest reduction, but the prediction accuracies are outstanding compared to other techniques. One may insist that other techniques (particularly MRMR) can deliver a similar prediction as good as TDA can do. However, one thing you should take into account is the efficiency. The other techniques are sensitive at each parameter selection. Users have to iterate parameters repetitively, while TDA does not need any prior assumption to achieve a more than decent result.
Chapter 6

Conclusion and Future Research

6.1 Lessons Learned

6.1.1 Topological Network Creation

Topological Data Analysis is known for unsupervised data analysis, which demands no prior assumptions. Furthermore, Ayasdi’s TDA even automates the output network creation. Therefore, one would not be convinced to research the mathematical logic behind. However, treating the Ayasdi’s TDA as a black box is not recommended. Understanding the mathematical background enriches the quality of data analysis. It helps to determine which metrics and filters to use for a specific purpose of analysis. That is particularly critical since chosen metrics and filters can determine the shape of topological networks.

6.1.2 Insights Discovery

The output of topological networks does not deliver much information, because insights discovery is not automated. Qualitative insights are delivered by varying color spectrums or defining and comparing subgroups. This is a critical step to determine improved insights. In fact, the insights discovery in Ayasdi is intuitive and interesting since users can play with shapes and colors compared to statistical methods. This lowers the entry barrier for the user who does not have any statistic or domain knowledge.
6.2 Limitation

Ayasdi accepts only tabular formatted data as input. This can be tackled by using other tools and expansion packs. A more pressing problem is the isolation from other tools. For instance, it does not have any driver to access Hadoop or any other databases. This does not allow Ayasdi to analyze streaming data. Furthermore, it interrupts the data analysis procedure. Data preparation should be processed outside of Ayasdi. Once a new data set is prepared, Ayasdi is able to import and analyze it. Finally, discovered insights by Ayasdi should be exported to another tool in order to build a prediction model and validate it. The complicated procedure to achieve this is introduced in Section 4.3.

6.3 Future Research

6.3.1 Implement an Open Source TDA

In order to overcome the isolation, an open source version of TDA called "Python Mapper" (Müllner, 2005) can be implemented in the Hadoop environment. The developer used to be a developer of the TDA project at Stanford University with the cofounders of Ayasdi. The “Mapper” covers the key part and delivers qualitative results similar to the one that Ayasdi covers. The complete data mining procedure can be carried out in Hadoop without any interruption.

6.3.2 Apply TDA on a Solved Problem

Another interesting opportunity is applying TDA on a solved problem. Following paper (Shen et al., 2013) is a neural science research paper published by a brain research department of Max Planck Institute. The entire datasets and Matlab codes are also available to the public. Since feature selection is a long-standing tradition in bioinformatics, it is promising to apply TDA and compare the insights.
Bibliography

ACE Science Center, 2015a. Data set of solar wind proton [WWW Document]. URL http://www.srl.caltech.edu/cgi-bin/dib/rundibviewmagl2/ACE/ASC/DATA/level2/mag?mag_data_1day.hdf%21hdfref;tag=1962,ref=6,s=0

ACE Science Center, 2015b. Data set of interplanetary magnetic field [WWW Document]. URL http://www.srl.caltech.edu/cgi-bin/dib/rundibviewmagl2/ACE/ASC/DATA/level2/mag?mag_data_1day.hdf%21hdfref;tag=1962,ref=6,s=0

Andrew, 2012. Why split data in the ratio 70:30?


F1 score, 2015. Wikipedia, the free encyclopedia.


KDnuggets, 2002. Polls: What main methodology are you using for data mining?


Müllner, D., 2005. Mapper, open source TDA.


Training data of energy consumption, 2015.


Appendix A
Pig Scripts

A.1. USED SYNTAX

Start Pig

$ pig

Execute Pig script with nohup option

[nohup] pig ‘file name’.pig [&]

Register jar file

REGISTER ‘jar directory’;

Load a data set

alias = LOAD ‘data path’
[USING function]
AS (schema)

Reconstruct a schema

FOREACH...GENERATE block used with a relation (outer bag). Use this syntax:
alias = FOREACH alias GENERATE expression [AS schema] [expression [AS
schema]...];

Join operation

alias = JOIN alias BY {expression|'('expression [, expression ...]')'}
(,alias BY {expression|'('expression [, expression ...]')'}...)
[USING 'replicated' | 'skewed' | 'merge' | 'merge-sparse']
[PARTITION BY partitioner]
[PARALLEL n];

Union operation
alias = UNION [ONSCHEMA] alias, alias [, alias …];

Store an output include header
STORE alias INTO 'directory' [USING function];

A.2. CONSTRUCT DATA

A.2.1. Earthquake

-- Register jar file
REGISTER '/opt/cloudera/parcels/CDH-5.2.0-
1.cdh5.2.0.p0.36/lib/pig/piggybank.jar';

-- Load the data set except header
rawEarthquakes = LOAD '/user/user1/datalake/Earthquakes/1904_2015.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (time:chararray, latitude:float, longitude:float, depth:float,
mag:float);

-- Reconstruct the schema
earthquakes = FOREACH rawEarthquakes
GENERATE
   SUBSTRING(time,0,4) AS year
   , DaysBetween(
      ToDate(SUBSTRING(time,0,10),'yyyy-MM-dd'),
      ToDate(CONCAT(SUBSTRING(time,0,4),'0101'),'yyyyMMdd')
   )+1 AS doy
   , latitude, longitude, depth, mag;

-- Store the output including header
STORE earthquakes INTO 'processed/earthquakes'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
('','NO_MULTILINE', 'WINDOWS', 'WRITE_OUTPUT_HEADER');

A.2.2.  Active Fire

REGISTER '/opt/cloudera/parcels/CDH-5.2.0-1.cdh5.2.0.p0.36/lib/pig/piggybank.jar';

F1 = LOAD '/user/user1/datalake/Fire-2001-2015/20001101-20021031.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
('','NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (latitude:float, longitude:float, brightness:float, scan:float,
track:float, acq_date:chararray, acq_time:int);

F2 = LOAD '/user/user1/datalake/Fire-2001-2015/20021101-20030731.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
('','NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (latitude:float, longitude:float, brightness:float, scan:float,
track:float, acq_date:chararray, acq_time:int);

F3 = LOAD '/user/user1/datalake/Fire-2001-2015/20030801-20040731.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
('','NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (latitude:float, longitude:float, brightness:float, scan:float,
track:float, acq_date:chararray, acq_time:int);

F4 = LOAD '/user/user1/datalake/Fire-2001-2015/20040801-20050730.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
('','NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (latitude:float, longitude:float, brightness:float, scan:float,
track:float, acq_date:chararray, acq_time:int);

F5 = LOAD '/user/user1/datalake/Fire-2001-2015/20050731-20060730.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
('','NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (latitude:float, longitude:float, brightness:float, scan:float,
track:float, acq_date:chararray, acq_time:int);
F6 = LOAD '/user/user1/datalake/Fire-2001-2015/20060731-20070731.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (latitude:float, longitude:float, brightness:float, scan:float,
track:float, acq_date:chararray, acq_time:int);

F7 = LOAD '/user/user1/datalake/Fire-2001-2015/20070731-20080730.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (latitude:float, longitude:float, brightness:float, scan:float,
track:float, acq_date:chararray, acq_time:int);

F8 = LOAD '/user/user1/datalake/Fire-2001-2015/20080731-20090730.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (latitude:float, longitude:float, brightness:float, scan:float,
track:float, acq_date:chararray, acq_time:int);

F9 = LOAD '/user/user1/datalake/Fire-2001-2015/20090731-20100730.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (latitude:float, longitude:float, brightness:float, scan:float,
track:float, acq_date:chararray, acq_time:int);

F10 = LOAD '/user/user1/datalake/Fire-2001-2015/20100731-20110730.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (latitude:float, longitude:float, brightness:float, scan:float,
track:float, acq_date:chararray, acq_time:int);

F11 = LOAD '/user/user1/datalake/Fire-2001-2015/20110731-20120730.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(',', 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (latitude:float, longitude:float, brightness:float, scan:float,
track:float, acq_date:chararray, acq_time:int);
F12 = LOAD '/user/user1/datalake/Fire-2001-2015/20120731-20130730.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(',' , 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (latitude:float, longitude:float, brightness:float, scan:float,
   track:float, acq_date:chararray, acq_time:int);

F13 = LOAD '/user/user1/datalake/Fire-2001-2015/20130731-20140731.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(',' , 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (latitude:float, longitude:float, brightness:float, scan:float,
   track:float, acq_date:chararray, acq_time:int);

F14 = LOAD '/user/user1/datalake/Fire-2001-2015/20140731-20150228.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(',' , 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (latitude:float, longitude:float, brightness:float, scan:float,
   track:float, acq_date:chararray, acq_time:int);

F15 = LOAD '/user/user1/datalake/Fire-2001-2015/20150301-20150708.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(',' , 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (latitude:float, longitude:float, brightness:float, scan:float,
   track:float, acq_date:chararray, acq_time:int);

rawFires = UNION F1, F2, F3, F4, F5, F6, F7, F8, F9, F10, F11, F12, F13,
   F14, F15;
fires = FOREACH rawFires GENERATE
   SUBSTRING(acq_date,0,4) AS year
   , DaysBetween(
    ToDate(SUBSTRING(acq_date,0,10),'yyyy-MM-dd'),
    ToDate(CONCAT(SUBSTRING(acq_date,0,4),'0101'),'yyyyMMdd')
   )+1 AS doy
   , latitude, longitude, brightness
;
STORE fires INTO 'processed/fires'
USING PigStorage(',');

**A.2.3. Volcano**

REGISTER '/opt/cloudera/parcels/CDH-5.2.0-
1.cdh5.2.0.p0.36/lib/pig/piggybank.jar';

rawVolcanos = LOAD '/user/user1/datalake/Volcano/volcano1904-2015.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(',' , 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (year:chararray, Mo:chararray, Dy:chararray, tsu:int, Eq:int,
Name:chararray, Location:chararray, Country:chararray, latitude:float,
longitude:float, Elevation:int, Type:chararray);

volcanos = FOREACH rawVolcanos GENERATE
    year,
    , DaysBetween(
        ToDate(CONCAT(CONCAT(year, Mo), Dy), 'yyyyMMdd'),
        ToDate(CONCAT(year, '0101'),'yyyyMMdd')
    )+1 AS doy
    , tsu, latitude, longitude;

STORE volcanos INTO 'processed/volcanos'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(',' , 'NO_MULTILINE', 'WINDOWS', 'WRITE_OUTPUT_HEADER');

**A.2.4. Heat Flow**

REGISTER '/opt/cloudera/parcels/CDH-5.2.0-
1.cdh5.2.0.p0.36/lib/pig/piggybank.jar';

rawThermal = LOAD '/user/user1/datalake/thermal-2010-.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(',' , 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (latitude:float, longitude:float, elevation:int, hearFlow:int);

STORE rawThermal INTO 'processed/thermal'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(',', 'NO_MULTILINE', 'WINDOWS', 'WRITE_OUTPUT_HEADER');

A.3. **INTEGRATE DATA**

A.3.1. **Earthquake, Solar wind and Magnetic**

-- Register jar file
REGISTER '/opt/cloudera/parcels/CDH-5.2.0-
1.cdh5.2.0.p0.36/lib/pig/piggybank.jar';

-- Load three data sets except header
solars = LOAD '/user/user1/datalake/Solar/solar_wind_1998_2015.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(’,’, 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (year:chararray, doy:chararray, hr:float, min:float, nH:float, vH:float);

magnetics = LOAD '/user/user1/datalake/Magnetic/magnetic_1997_2015.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(’,’, 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (year:chararray, doy:chararray, mag:float);

earthquakes = LOAD '/user/user1/processed/earthquakes/earthquakes.csv'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(’,’, 'NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (year:chararray, doy:chararray, lat:float, lon:float, depth:float, mag:float);

-- Filter
Fsolars = FILTER solars BY (year > 2014);
Fmagnetics = FILTER magnetics BY (year > 2014);
Fearthquakes = FILTER earthquakes BY (year > 2014);
-- join operation
DStime = JOIN Fsolars BY (year, doy), Fmagnetics BY (year, doy),
Fearthquakes BY (year, doy);

-- Store the output include header
STORE DStime INTO '/user/user1/processed/DStime/2014-
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(', ', NO_MULTILINE', 'WINDOWS', 'WRITE_OUTPUT_HEADER');

A.3.2. Earthquake, Volcano and Heat flow

REGISTER '/opt/cloudera/parcels/CDH-5.2.0-
1.cdh5.2.0.p0.36/lib/pig/piggybank.jar';

earthquakes = LOAD '/user/user1/processed/earthquakes/part-m-00000'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(', ', NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (year:int, doy:chararray, latitude:float, longitude:float, depth:float, 
   mag:float);

volcanos = LOAD '/user/user1/processed/volcanos/part-m-00000'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(', ', NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (year:chararray, doy:chararray, tsu:int, latitude:float, 
   longitude:float);

thermals = LOAD '/user/user1/processed/thermal/part-m-00000'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(', ', NO_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER')
AS (latitude:float, longitude:float, elevation:int, hearFlow:int);

EVT = UNION ONSCHEMA earthquakes, volcanos, thermals;
STORE EVT INTO '/user/user1/processed/EVT'
USING org.apache.pig.piggybank.storage.CSVExcelStorage
(', ', NO_MULTILINE', 'WINDOWS', 'WRITE_OUTPUT_HEADER');
Appendix B

MAPE in R

basePath <- "C:/Users/Kimi/Documents/05 nPower/Result"
setwd(basePath)
filename <- "result.csv"
alldata <- read.csv(filename, header = TRUE, sep =",", dec = ".")

# determine the number of columns
attLength <- length(alldata[1,])
datasize <- length(alldata[,1])

labels <- alldata[0:datasize, attLength-1]
results <- alldata[0:datasize, attLength]

# print absolute error time series
error <- abs(results - labels) / labels
mape <- mean(error)
print(mape)

# mape <- 1/length(labels)  * sum(abs(labels - results) / labels)
print(paste("MAPE = ", mape))