A Survey of Predictive Analytics in Data Mining with Big Data

BY

DICKSON WAI HEI LAM

A project submitted in partial fulfillment

of the requirements for the degree of

MASTER OF SCIENCE in INFORMATION SYSTEMS

Athabasca, Alberta

July, 2014

© Dickson Wai Hei Lam, 2014
DEDICATION

I dedicate this essay to my parents Lai Hoi Yin and Lam Bing Fat, my brother Lam Tung Hay, my sister Tracy Lam and my brother-in-law Matthew Chao. Last but not least, my niece and my two nephews, Kasey, Evan and Gabriel. For their understandings, supports and encouragements throughout the entire MScIS program in the past four years.

I am especially thankful to my wife, Tracy Low, for her love, patience and support during the past four years. It is with her encouragement that I found the confidence to have embarked on this journey. It is also with her perseverance that we are able to conclude this journey with this essay.
ABSTRACT

This paper explores the area of Predictive Analytics in combination of Data Mining and Big Data. The survey indicates an accelerated adoption in the aforementioned technologies in recent years. Businesses and researchers alike take great interests in furthering the use of Predictive Analytics in enhancing Business Intelligence and forecasting ability across a wide range of applications. This research essay explained some of the underpinnings in enabling predictive capabilities in data analysis and Data Mining. Also, it incorporated the characteristics of Big Data as the supplementary enabler to augment the way we perceive Data Mining. Predictive analytics is the next frontier for innovation that is built based on century old concepts and techniques such as mathematical analysis and statistical analysis.

Keywords: Predictive Analytics, Data Mining, Big Data, Analytics, Statistical Analysis, Machine Learning
ACKNOWLEDGMENTS

I would like to express my deepest gratitude to the faculty members of Master of Science in Information Systems (MScIS). I would also like to express my sincere appreciation to the MScIS Graduate Program Director, Dr. Larbi Esmahi, for his acceptance of being my essay supervisor. I would like to thank him for his advice and endless support throughout the development of this essay.
# TABLE OF CONTENT

DEDICATION .................................................................................................................. 2

ABSTRACT ....................................................................................................................... 3

ACKNOWLEDGMENTS ...................................................................................................... 4

TABLE OF CONTENT ................................................................................................... 5

LIST OF TABLES AND FIGURES .................................................................................... 9

CHAPTER I ....................................................................................................................... 12

INTRODUCTION .............................................................................................................. 12

Statement of the Purpose .............................................................................................. 12

Research Problem ......................................................................................................... 14

Background ..................................................................................................................... 14

The Problems .................................................................................................................. 16

The Definitions of Predictive Analytics and Big Data .................................................... 17

Organization of the Remaining Chapters ...................................................................... 18

CHAPTER II .................................................................................................................... 20

REVIEW OF RELATED LITERATURE ............................................................................ 20

Introduction ................................................................................................................... 20

Explanatory versus Predictive Modeling ....................................................................... 23

The Role of Predictive Model ......................................................................................... 24

Underlying Models Differences ..................................................................................... 24

Data Mining and Big Data ............................................................................................. 28

The Big Data Problems ................................................................................................... 29

The NoSQL Solution ........................................................................................................ 31

The Apache Hadoop Platform ......................................................................................... 33

NoSQL Data Model versus Relational Data Model ......................................................... 35
Predictive Analytics Versus Other Forecasting and Mining Methods ........................................ 36
The Predictive Model Markup Language (PMML) ................................................................. 40
PMML Adoption ......................................................................................................................... 40
PMML Document Structure ....................................................................................................... 41
PMML Interoperability and Application ..................................................................................... 44
PMML Enabled Architecture ..................................................................................................... 46
Cloud Computing ...................................................................................................................... 48
Dichotomy of Predictive Analytics Skillsets ............................................................................. 50
Predictive Analytics Performance Optimization ......................................................................... 52

CHAPTER III .............................................................................................................................. 57

PREDICTIVE ANALYTICS APPLICATIONS ........................................................................... 57

Social Computing ....................................................................................................................... 58
Large-Scale Machine Learning at Twitter Inc. ........................................................................... 58
Network Relationship ................................................................................................................. 62

Education Support .................................................................................................................... 64
Predicting At-Risk Students ....................................................................................................... 64
Predict Course Success .............................................................................................................. 66

Video Gaming ........................................................................................................................ 67

Law Enforcement ..................................................................................................................... 70

Business Applications ............................................................................................................. 76

Financial Engineering ............................................................................................................. 78

Summary ................................................................................................................................... 81

CHAPTER IV ............................................................................................................................. 84

METHODOLOGY ....................................................................................................................... 84

Preliminary Discussion ............................................................................................................. 86

Preliminary Analysis of the Topic Data Using Google Trends .................................................... 88

Discussion ............................................................................................................................... 90

Context-Awareness and Big Data ............................................................................................. 90
APPENDIX B – RESEARCH TOOLS

PRODUCTIVITY SOFTWARE
INTERNET BROWSERS
OPEN SOURCE PREDICTIVE ANALYTICS AND DATA MINING TOOLS
PYTHON RELATED STATISTICAL LIBRARIES
LITERATURE SEARCH ENGINES
RESEARCH PAPER ONLINE DATABASES
RESEARCH MANAGEMENT TOOLS AND SERVICES
ONLINE COMMUNITIES
# LIST OF TABLES AND FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Business Intelligence Taxonomy</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>Decision Support Process</td>
<td>21</td>
</tr>
<tr>
<td>3</td>
<td>Analytics Processing Taxonomy</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>Explanatory Model</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>Predictive Model</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>Descriptive Analytics Taxonomy</td>
<td>27</td>
</tr>
<tr>
<td>7</td>
<td>Data mining adopts techniques from many domains (Han, Kamber, &amp; Pei, Data Mining: Concepts and Techniques, Third Edition, 2011)</td>
<td>29</td>
</tr>
<tr>
<td>8</td>
<td>HDFS Architecture (HDFS Architecture Guide, 2014)</td>
<td>33</td>
</tr>
<tr>
<td>9</td>
<td>Predictive Analytics in Data Mining</td>
<td>38</td>
</tr>
<tr>
<td>10</td>
<td>Predictive Analytics Taxonomy</td>
<td>39</td>
</tr>
<tr>
<td>11</td>
<td>PMML Schema (src: <a href="http://www.dmg.org">http://www.dmg.org</a>)</td>
<td>42</td>
</tr>
<tr>
<td>12</td>
<td>Life Cycle of Data Mining Project using ADAPA (Guazzelli, Stathatos, &amp; Zeller, 2009)</td>
<td>45</td>
</tr>
<tr>
<td>13</td>
<td>EMC Greenplum MPP database architecture (Das, Fratkin, Gorajek, Stathatos, &amp; Gajjar, 2011)</td>
<td>47</td>
</tr>
<tr>
<td>14</td>
<td>Automated modeling workflow (Kridel &amp; Dolk, 2013)</td>
<td>52</td>
</tr>
<tr>
<td>15</td>
<td>General Forecasting Process (Fischer, et al., 2013)</td>
<td>54</td>
</tr>
<tr>
<td>16</td>
<td>Extension of 3-layer schema architecture (Fischer, et al., 2013)</td>
<td>56</td>
</tr>
<tr>
<td>17</td>
<td>Bayesian Net Model of Intent to Proliferate (Sanfilippo, et al., 2011)</td>
<td>69</td>
</tr>
</tbody>
</table>
Figure 18: Group Recidivism Rates (Jennings & M.C.J, 2006). Note: n=number of individual per category (scale) and per risk class, %=percentage of the individual in a specific class were re-arrested

Figure 19: Bayes' theorem formula

Figure 20: "Predictive Analytics" Search Term - Google Trends Chart, February 2014

Figure 21: "Data Mining" Search Term - Google Trends Chart, February 2014

Figure 22: "Big Data" Search Term - Google Trends Chart, February 2014

Figure 23: Combined Search Terms - Google Trends Chart, February 2014. Blue color = Predictive Analytics, Red Color = Data Mining, Orange Color = Big Data

Figure 24: The evolution of context definition (Kiseleva, 2013)

Figure 25: Layered conceptual framework for context-aware systems (Baldauf, Dustdar, & Rosenberg, 2007)

Figure 26: Meta-Learning Architecture (Singh & Rao, 2013)

Figure 27: An example of context-aware system design (Kiseleva, 2013)

Figure 28: Context managing framework architecture (Baldauf, Dustdar, & Rosenberg, 2007)

Figure 29: An example of a normal distribution "Bell Curve"

Figure 30: Variance formula

Figure 31: Standard Deviation formula 1

Figure 32: Standard Deviation formula 2

Figure 33: Correlation Coefficient formula

Figure 34: A Bayes net for the medical diagnosis example (Patterns of Inference, 2014)

Figure 35: A visualized example of linear regression (Natural Resources Canada, 2012)
Figure 36: Clustering of a set of objects using the k-means method; for (b) update cluster centers and reassign objects accordingly (the mean of each cluster is marked by a C) (Han, Kamber, & Pei, Data Mining: Concepts and Techniques, Third Edition, 2011) .......................................................... 106

Figure 37: Multi-layer feed forward neural network (Han, Kamber, & Pei, Data Mining: Concepts and Techniques, Third Edition, 2011) .................................................................................................................. 108

Figure 38: Binary Threshold Neuron formula ............................................................................................. 109

Figure 39: An example of binary response dependent variable (x=income, y=gender) ..................114

Figure 40: Value of knowledge about event (Fülöp, et al., 2012)................................................................. 120

Figure 41: CEP-PA conceptual framework .............................................................................................. 121
CHAPTER I

INTRODUCTION

Statement of the Purpose

The cumulating amount of data has pressured researchers and practitioners to devise new techniques and data processing models to tap into the invaluable source of Big Data. One such usage in extracting knowledge from the vast amount of data is in Predictive Analytics which allows us to gain insights in predicting unknown events and future activities. Within the context of Data Mining, Predictive Analytics pairs with statistical analysis to provide a very interesting combination of techniques for knowledge discovery.

The predictive nature within the domain of statistical analysis paves the way for a wide variety of real life applications. From clinical analytics in Clinical Decision Support System (CDSS) to business analytics in Operations Research (OR), Predictive Analytics aids decision makers to make choices and solve problems that have long lasting impacts. The foundation that brings new opportunities and possibilities of continuous improvement came from the centuries old disciplines of mathematics and statistics. From data collection to data analysis, mathematics and statistics interweave into every aspect of scientific data research endeavors. Predictive Analytics is fundamentally dependent on the disciplines of mathematics and statistics. In fact, the roles that the two disciplines play are crucial to the evolutionary development process of Data Mining and Big Data, which is still shaping our collective understandings of Predictive Analytics.

The intent of the selected research topic is to survey the current landscape of Predictive Analytics and Data Mining within the context of Big Data. Although many underlying components of Data Mining and Big Data have been the subject of discussions and research
interests for many years, the demand for better data management and data analysis tools continue to accelerate. One of these trends in this regard is the rise of NoSQL databases which challenge even the *relational data model*, a prevalent and dominant database model design since the second half of twentieth century (Codd, 1970).

The purpose of this paper is to explore Predictive Analytics in conjunction with Data Mining and Big Data. This research will also touch on the concerns related to *cloud computing*, *mobile computing* and *social computing* to the extent that helps to solidify the discussion points on the application of Predictive Analytics. During the initial phrase of the research paper development, it was apparent that Predictive Analytics is still an emerging term. While a simple internet search on the terms “Big Data” and “Data Mining” generated many hundreds of peer-reviewed academic papers from reputable online libraries such as IEEE Xplore and ACM DL, a query on “Predictive Analytics” yielded only marginal amount of search result in terms of volume and subject variety.

With this in mind, the aim of this paper is to solidify our understanding by surveying the current landscape of *Predictive Analytics in Data Mining with Big Data*. The research result would contribute to the knowledgebase of the evolving field of Predictive Analytics. In doing so, this essay attempts to support the survey with contemporary best practices, empirical experiments and case studies to shed light on this embryonic discipline. Also, the research result would indicate theories, methodologies, models and tools that are commonly adopted by researchers and frequently used by businesses. Thus, the research covers the spectrums in both general application and specialized academic research domains.
Research Problem

Background

The need to predict future or to explain pattern of natural phenomena bear many implications to human understanding of epistemology. To anticipate an event, is by preparing a response to a possible outcome, which is to say, judging from what have been known and infer them to the unknown. The advantages of the ability to look ahead to yet-to-occur events are plentiful, the following are a few examples of how the application of prediction allows us to avoid unfavorable impending outcome:

- Improves customer satisfaction with personalized purchasing recommendations;
- Improves business competitiveness by being able to foresee customer behavioral changes;
- Avoids systematic economic downfall by assessing consumer credit risk scores;
- Predicts earthquake by detecting and analyzing seismic activities;
- Supports agricultural planning by forecasting weather;
- Predicts seasonal influenza virus strain.

For consumers, a recommender system based on association rules analysis, advises on merchandise purchase suggestions that might be of interest to online customers. The goal is to improve overall customer satisfactions in reducing the time to manually locate interesting items. This might seems trivial, however, when Predictive Analytics improves the online purchasing experience of a consumer, the ripple effect could encourage more people to make more online purchases rather than purchasing from brick and mortar stores. Thereby, the online shopping environment eliminates the need for consumers to travel with transportation. Collectively, this could lead to the reduction in traffic accidents and decreases CO2 emission footprint due to
lower overall usage of automobiles. Therefore, Predictive Analytics can exert impactful cascading changes on individuals that can lead to societal changes. The positive environmental impact as a result of human behavioral change is only one of the many examples that Predictive Analytics can deliver.

Predictive Analytics, strictly speaking, is a subset of Data Mining field which is part of the Data Science discipline. The term Analytics itself derives from the science of data analysis that is commonly associated with another term Business Intelligence to describe the provisioning of decision support in businesses. Predictive Analytics is supported by applying mathematical and statistical techniques to derive meaning from data and systematically find patterns in data for decision making directives. The applications of Predictive Analytics range across both the academia and the industries. The relationship between Business Intelligence and Data Mining is depicted in Figure 1.
To use Predictive Analytics, is to apply mathematics, statistics and probability theory in conjunction with the overarching computer science discipline of machine learning, data modeling and algorithm development.

**The Problems**

Predictive Analytics has a broad scope and wide application. Predictive Analytics cannot be performed in isolation and should be considered a systematic approach for systems to work in unison to derive result. To that end, Predictive Analytics at its core is realized by machine learning and substantiated by statistical techniques to model, analyze and deduce knowledge.

The extracted data is expected to contain actionable information from seemingly uninteresting dataset. Predictive Analytics mandates the use of database systems to maintain a collection of...

---

1. The diagram depicts the relationships between various components of Data Mining and Data Warehousing in relation to Business Intelligence. Each arrow represents a specific hierarchical relationship between a set of given two entities. The entity receiving the head of an arrow is the sub-entity of its parent entity which contains the tail of an arrow.
data for processing and uses analytical modeling to transform raw data into actionable knowledge.

To support the initiative of applying predictive analytics, a number of issues in relation to data management and computational resources must be addressed. From an infrastructure point of view, what network components\(^2\) and standards\(^3\) are required to support Predictive Analytics? From an architecture point of view, what software and services\(^4\) are needed to construct and support Predictive Analytics? From a conceptual standpoint, what are some of the principles, issues and risks\(^5\) involved in Predictive Analytics, especially with regards to ethical concerns? Within the context of Data Mining, what are some of the best practices\(^6\) in integrating Predictive Analytics in both academia and business settings? Finally, how does Big Data fit into the equation and how Big Data augments the infrastructural, architectural and conceptual paradigm in our collective understanding of information management?

While Cloud Computing is part and partial to Big Data application, it will not be the focus of this paper and will only be briefly discussed within the domain of Predictive Analytics.

**The Definitions of Predictive Analytics and Big Data**

Gartner Research’s definition of Big Data is widely adopted; the three Vs of Big Data consists of Volume, Variety and Velocity (Gartner.com, 2014). While the three Vs definition of Big Data is prevalent and distinguishes the unique aspects of Big Data, it does not serve a suitable description for Big Data within the context of Predictive Analytics. As such, this paper

---

\(^2\) Refer to the Cloud Computing section for more information.
\(^3\) Refer to The Predictive Model Markup Language (PMML section for more information.
\(^4\) Refer to Predictive Analytics Performance Optimization section for more information.
\(^5\) Refer to CHAPTER V for more information.
\(^6\) Refer to CHAPTER III.
A Survey of Predictive Analytics in Data Mining with Big Data

attempts to provide a definition of Big Data that adequately describe its properties within Predictive Analytics.

**Big Data:** The available and accessible set of historical data and metadata that capture seeming unrelated multidimensional facts and events based on the result of human-created and machine-generated actions.

Given the above definition of Big Data, the definition of Predictive Analytics is defined below.

**Predictive Analytics:** To maximize the signal-to-noise ratio through the analysis of Big Data. To use the result of such analysis in combination of the advanced techniques of statistical modeling and the assistance of high performance computing devices, to derive meaningful information that provide a higher-than-guessing accuracy and precision. The derived information is capable of predicting trends and the validated result of each prediction will be used in updating the underlying statistical model continuously and perpetually.

**Organization of the Remaining Chapters**

The approach in this paper will take the path of conducting literature review in the context of Predictive Analytics, Data Mining and Big Data independently and collectively as a single problem domain. Meta-analysis\(^7\) will be employed throughout this paper to contrast conceptual constructs proposed by different researchers and to identify overlapping areas of concerns.

Starting with CHAPTER II, this essay attempts to provide an in breadth coverage of the current landscape of innovation in research and development with regards to the subject matter.

---

\(^7\) Meta-analysis is a type of research method for producing new information by contrasting the results from previously related researches. Conducting meta-analysis does not require designing and performing an actual experiment.
CHAPTER II contains mainly literature reviews of existing research papers that discussed the subject matter of Predictive Analytics in Data Mining with Big Data. The chapter would provide readers with a rudimentary understanding of the subject matter. In the later part of CHAPTER II, it progresses onto more specifically focus research areas such as intra DBMS modeling\(^8\).

CHAPTER III highlights the practical use of Predictive Analytics across many domains.

CHAPTER IV deals with methodologies, methods, best practices that are employed by researchers and practitioners within the domain of the subject matter. CHAPTER V discusses issues, challenges and recent trends in both academic and commercial spaces. Finally, CHAPTER VI summarizes the conclusions and recommendations resulted from this study. Some appendixes are also attached at the end of this report to provide additional material that supports our analysis of the different aspects presented in different chapters.

\(^8\) Refer to the Predictive Analytics Performance Optimization section for more information.
CHAPTER II

REVIEW OF RELATED LITERATURE

Introduction

The research topic in question is an area of intense interest from researchers and practitioners. The expectation of Predictive Analytics to deliver value is a promising proposition. Being able to forecast trends and to predict future behaviors has many implications ranging from disease control to credit risk scoring as they relate to a particular population or individual. To understand Predictive Analytics, one must first understand the concepts and usage of Data Mining as it pertains to the seven steps of *Knowledge Discovery From Data* (KDD). They are, *data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation* (Han, Kamber, & Pei, *Data Mining: Concepts and Techniques*, Third Edition, 2011). As such, the promise of Predictive Analytics hinges on Data Mining to provide meaningful dataset in order to fully appreciate its predictive prowess.

As more people are starting to understand the benefits and values that Predictive Analytics can bring, the trend to higher adoption of Predictive Analytics is becoming evident. The statistics shown in (Waller & Fawcett, 2013) suggested an increasing popularity in the space driven by the higher business competition level in recent years. The recent uptake in the interest of Big Data also fused the interest in Predictive Analytics as the two domains are closely related.

Predictive Analytics and Data Mining are part of the data science field (Waller & Fawcett, 2013); Predictive Analytics is considered a subset of Data Mining field due to the logical order of KDD operation. Data Mining takes precedence in this respect as it identifies existing patterns and trends in seemingly unrelated and uncorrelated data. Performing Data Mining is done in a descriptive way.
The paper by (Haas, Maglio, Selinger, & Tan, 2011) explained descriptive analytics in relation to Data Mining as a last resort for decision-making. This is because descriptive analytics only describes the present conditions, whereas predictive analysis is a model-driven and data-driven approach for generating what-if scenarios which fully exploits the nuances of underlying data. Given this description, the logical order of decision support process is show in Figure 2.

Descriptive $\rightarrow$ Explanatory $\rightarrow$ Exploratory $\rightarrow$ Predictive $\rightarrow$ Prescriptive (Decision)

Figure 2: Decision Support Process

The relationship between the different models is shown in Figure 3. In this sense, Descriptive Analytics takes precedence over Predictive Analytics. Predictive Analytics relies on Descriptive Analytics to provide the descriptive information as well as a foundational framework in order for it to function and be effective.

The paper by (Haas, Maglio, Selinger, & Tan, 2011) introduced the role of Prescriptive Analytics in decision making. Prescriptive Analytics is supported by Predictive Analytics which in turns supported by the deterministic and stochastic optimization techniques such as Decision Tree method and Monte Carlo method, respectively. To that end, Predictive Analytics produces what-if scenarios for Prescriptive Analytics to derive objectives focus and constraints balanced decisions. Thus, Prescriptive Analytics is logically separated from Predictive Analytics and Prescriptive Analytics is positioned as a post-process of Predictive Analytics for final decision support. In other words, the order of operation is described as $X \rightarrow Y \rightarrow Z$ where $X$ is Descriptive Analytics, $Y$ is Predictive Analytics and $Z$ is Prescriptive Analytics.
Relying on only Descriptive Analytics in Data Mining to extrapolate future outcome can be disastrous as illustrated in the example provided in (Haas, Maglio, Selinger, & Tan, 2011). The example of “Extrapolation of 1970-2006 median U.S. housing prices” chart demonstrated the effect of a shallow prediction (i.e. extrapolation) versus deep predictive analytics where the extrapolated housing prices beyond year 2006 was widely different than the actual prices recorded between the year 2006 and 2010.

While Descriptive Analytics provides a simple picture of linear relationship between past and future events, it is however not a solid predictor for the future. That is to say, Descriptive Analytics can help in situations where extrapolation and interpolation are acceptable but prediction using extrapolation is destined to fail as the margins of probable errors are not factored in the equation. The main differences between Predictive Analytics and Descriptive

---

9 Each arrow represents the logical order between each component. For instance, Predictive Analytics depends on the result from both the Exploratory Modeling and Predictive Modeling components.
Analytics lie in the application of advanced techniques, some of which are, machine learning, advanced modeling, causality approximation, root-cause analysis, sensitivity analysis, model validation and many other probability driven model designs. As such, Haas et al. concluded that the focus for Business Intelligence should operate more with deep Predictive Analytics rather than shallow Descriptive Analytics. The standpoints of model and data need to have a balanced perspective and must not overly emphasize solely on the part of either the model or the data when performing analysis.

**Explanatory versus Predictive Modeling**

The concerns brought by (Haas, Maglio, Selinger, & Tan, 2011) are shared by Shmueli et al. in (Shmueli & Koppius, 2010). The authors explained how the common approach of empirical models for explanation (i.e. explanatory statistical model) is fundamentally distinct from the empirical models for prediction (i.e. empirical predictive model). One of the key distinctions between the aforementioned two models is the difference in analysis goal. Explanatory model is used for testing causal hypotheses while predictive model is ideal for predicting new observations and assessing predictability levels.

Explanatory model and predictive model can be complementary in certain regards; however they differ from each other in terms of achieving the end goal at an empirical level, because they are used for different purposes and in different contexts. The notion of causality between theoretical constructs is deeply rooted in the explanatory model, whereas predictive model depends on the associations between measurable variables (Shmueli & Koppius, 2010).

In the context of empirical modeling, the selection of statistical modeling methodology and the expectation of application determine the techniques employed. Principally, the dichotomy of explanatory modeling and predictive modeling leads to a contrasting outcome that
affects model selection, model techniques, model evaluation and validation. The result of these differences was summarized in Table 1.

The Role of Predictive Model

The role of Predictive Analytics in scientific research differs from Explanatory Analytics as well. Predictive modeling shares the principles of grounded theory in research design, in which, the main purpose is to generate theory from both qualitative and quantitative data. In fact, predictive modeling is theory generative in nature rather than theory derivative. Like grounded theory, predictive modeling attempts to produce a theory (i.e. a prediction) based on observations (i.e. data).

Another role of the predictive model is to assess the predictive power of any predictive model based on the measure of distance between theory and practice (Shmueli & Koppius, 2010). Explanatory models are usually constructed with in-sample dataset that serves as a training set and are validated using out-of-sample dataset from the same sample set. Models created under this methodology can be rigid and bias towards the sample dataset and unable to account for rapid shift in data context due to a possible model overfitting, which diminishes their predictive power. Therefore, the theory generative properties of predictive modeling are more suitable to be used to infer to probable outcomes through measureable variables where causality plays a lesser role to probability.

Underlying Models Differences

In fact, the result of applying explanatory techniques to predictive modeling would inevitably defeat the original goal and diminishes the predictive power of a chosen model. The key point in the two modeling approaches is the type of relationships being measured as well as the external factors themselves. The author in (Shmueli, 2010) defined the terms as follow:
• **Explanatory Modeling:** A retrospective approach to explain the underlying causal relationships between theoretical constructs that is both descriptive and explanatory. This definition is depicted in Figure 4.

• **Predictive Modeling:** A prospective approach to predict unknown outcome based on the association relationships between measurable variables. This definition is depicted in Figure 5.

![Explanatory Model](image1)

*Figure 4: Explanatory Model*

![Predictive Model](image2)

*Figure 5: Predictive Model*

Figure 4 depicts the basic premise of an explanatory model, in that, the cause and effect constructs are linked by a unidirectional cause-effect relationship. In statistical terms, the cause construct can be expressed as an independent variable (i.e. factor) and the effect construct can be realized as dependent variable (i.e. outcome). Empirically, it is important to point out that
**correlation does not equate to causation.** Variables that are correlated, either positively or negatively correlated, can be affected by externally latent variables rather than the observable variables. For example, if variable $X$ and $Y$ are correlated with a correlation coefficient (i.e. $R_{(X,Y)}$) value of 0.9, it can be both influenced by a third variable $Z$ that directly manipulates variable $X$ and $Y$. However, $Z$ was not detected because $Z$ is not an observable variable. In other words, had the effect of variable $Z$ not been present, the correlation between variable $X$ and $Y$ will no longer be observed and the perceived correlation between the two variables $X$ and $Y$ is a direct result of a third variable $Z$. Therefore, only variable $Z$ has a direct causal effect on variable $X$ and $Y$. However, without the knowledge of $Z$, the illusion was perceived as if variable $X$ and $Y$ are causally related.

Explanatory Model allows researchers to measure the model validity of a given set of hypotheses. Methods for measuring model *internal validity* can be achieved through *R*-squared and *p*-value statistical calculations that validate and evaluate variables causality and statistical significance. The result can be used to either accept or reject a particular hypothesis for a quantitatively explanation on a given phenomenon. However, the power to explain within a historical context for which explanatory model was designed to perform, does not extend to predictive application which operates with little regard for causality, but it operates with a higher focus on measurable outputs.

Predictive Model, on the other hand, is determined by a loose association between measurable variables within an observable relationship that does not denote any direct causal meaning. This relationship is shown in Figure 5. This key difference enables predictive model to be highly adaptive as it meets the goal to achieve high predictive power rather than explanatory power.
Explanatory modeling methods such as *simple regression*-type methods are not suitable for predictive tasks as described in (Shmueli & Koppius, 2010) and (Shmueli, 2010). A collection of Descriptive Analytics techniques are presented in Figure 6 based on the survey result. To improve predictive power, methods involving machine learning and probability calculation often yield better results, some of the techniques are shown in Figure 10. Predictive model is not intended for past-conformity as there exists a risk of model overfitting for model that is tightly coupled to historical events. For predictive model to be able to foretell an event, a level of tolerance must be built into the model in order to adequately handle unforeseen events.

![Descriptive Analytics Taxonomy](image-url)

*Figure 6: Descriptive Analytics Taxonomy*
The following table summarized the arguments presented in (Shmueli & Koppius, 2010) and in (Shmueli, 2010):

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Explanatory Model</th>
<th>Predictive Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targeted Relationship</td>
<td>Descriptive and Explanatory</td>
<td>Forward-looking</td>
</tr>
<tr>
<td>Targeted Relationship</td>
<td>Causation</td>
<td>Association</td>
</tr>
<tr>
<td>Targeted Relationship</td>
<td>Theoretical Constructs</td>
<td>Measurable Variables</td>
</tr>
<tr>
<td>Perspective</td>
<td>Retrospective</td>
<td>Prospective</td>
</tr>
<tr>
<td>Data Continuity</td>
<td>Continuous changes in data</td>
<td>Continuous and discontinuous changes in data</td>
</tr>
<tr>
<td>Data Bias</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Data Volume</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Data Cleansing</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Requirement</td>
<td>Undesirable</td>
<td>Expected</td>
</tr>
<tr>
<td>Data Partition</td>
<td>Less Common</td>
<td>Common</td>
</tr>
<tr>
<td>Exploratory Domain</td>
<td>Limited</td>
<td>Wide</td>
</tr>
<tr>
<td>Exploration</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Interactivity</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Data Dimension</td>
<td>Simple regression-Type Models</td>
<td>Algorithmic Modeling such as neural networks</td>
</tr>
<tr>
<td>Reduction</td>
<td>High</td>
<td>and k-nearest-neighbors</td>
</tr>
<tr>
<td>Popular Analysis Models</td>
<td>Simple regression-Type Models</td>
<td>Algorithmic Modeling such as neural networks</td>
</tr>
<tr>
<td>Model Transparency</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Power Assessment</td>
<td>R Square and F-Test</td>
<td>Backtesting</td>
</tr>
<tr>
<td>Techniques</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Data Mining and Big Data**

Data Mining is an all-encompassing term to describe the methodology, strategy and the common approaches to extract knowledge out of raw data using techniques from many domains shown in Figure 7. Performing Data Mining is to put data into context, to produce information that is relevant and useful for solving specific problems. Very often, these problems are concerns that crosscut many academic and business areas, making the multidisciplinary field of Data Mining an indispensable one for solving problems using an evidence (i.e. data) based approach. With the improved understanding of data through Data Mining, the need to predict future
occurrences based on historical references prompted the development of Predictive Analytics. Predictive Analytics forecasts trends and behavior patterns augmented by the three properties of Big Data (i.e. volume, variety and velocity).

![Data Mining Techniques Diagram](image)

*Figure 7: Data mining adopts techniques from many domains (Han, Kamber, & Pei, Data Mining: Concepts and Techniques, Third Edition, 2011)*

**The Big Data Problems**

The challenge in the understanding of Big Data stems from the nebula term *Big Data* coined by many researchers to describe the accelerating data volume since the dawn of the internet age.

The publicly accessible internet data have been increasing in size at an exponential rate as far back as 1995 as shown in (Leiner, et al., n.d.) and in (Lesk, n.d.). Between the year 1999 and 2001, the estimated number of internet web pages has grown from less than 1 billion to more than 4 billion pages in a period of 2 years (Murray & Moore, 2000). This rate continued to accelerate, the estimated size of the index-able internet as of 2011 is approximately 5 million terabytes or 5
exabyte of data (Clair, 2011). The ever increasing data volume managed by organizations is putting pressure on storage capacity. In order to meet the demand, a different strategy is in order to deal with many of these challenges.

The European Bioinformatics Institute (EBI), part of the European Molecular Biology Laboratory (EMBL), had gathered approximately 28 petabytes worth of bioinformatics data as of this writing (About Us Background, 2014) to assist tasks such as DNA sequencing, drug resistance testing and drug R&D. The large dataset spans many different types, spanning genes, proteins expressions, small molecules and protein structures.

The European Organization for Nuclear Research (CERN) faces the same storage capacity and data volume challenge with an annual growth of 15 petabytes worth of data for physicists to sift through (Computing, 2014). The data are generated by the Large Hadron Collider (LHC) at a rate of 600 million times per second in an effort to recreate the moment immediately following the big bang event based on the Big Bang Theory. Each particles collision set forth a chain reaction that led to a series of complex events. The exponential growth of each collision and the subsequent reactions created massive amount of data to be persisted in CERN Data Centre (DC) for physicists to analyze.

Furthermore, the Low Frequency Array (LOFAR); a radio interferometer for detecting low radio frequencies between the range of 10 MHz and 240MHz, collects data in petabytes scale per year with each single file approaching terabyte sizes (Begemana, et al., 2010) for geoscience and agricultural applications.

The sheer volume of internet data also gave raise to data complexity and variety in terms of data structure, data sources and data types. Researchers have devised various strategies to collect and process such tremendous amount of data in both data volume and data variety
aspects. Some of the strategies involve storing unstructured data in distributed NoSQL databases and use MapReduce method for job-based distributed data computation. In this domain, the Apache Hadoop Platform (Welcome to Apache™ Hadoop®, 2014) is currently dominating the research and commercial space.

The NoSQL Solution

The statistical algorithms that underpin many of these strategies will be discussed in CHAPTER IV. Many of these strategies built on the traditional disciplines of mathematics and statistics, created an overall interdisciplinary approach to tackle the challenges brought by Big Data. As such, the computer science discipline plays an important role as researchers continue to innovate using an interdisciplinary approach that combines mathematics, statistics and software engineering. The advancements in database technologies and distributed computing are some of the great examples that combat the three Vs problem of Big Data based on the centuries-old principle of divide-and-conquer.

Applying Data Mining in Big Data is a complex process that involves many procedures and is poised to uproot our collective understandings of the most fundamental component in information management, the relational database technology. As such, this gave rise of the NoSQL database technology (Russom, 2011) (Menegaz, 2012) (Chen, Chiang, & Storey, 2012) that was designed to better handle the problem of the three Vs of Big Data. NoSQL is a non-relational data model approach designed to face the challenges presented by Big Data, that are, volume, variety and velocity.

In short, NoSQL is designed with Big Data in mind that have made trade-offs to satisfy the requirements on performance, size, transactional support and features. One such design criteria led to the handling of semi-structured and unstructured data in a distributed environment
for data management. The various implementations of NoSQL on the market today include CouchDB, MongoDB, FlockDB and Apache Casandra, to name a few. Many of these implementations rely on the Hadoop (The Apache Software Foundation, 2013) software library which includes common NoSQL utilities, distributed file system, job scheduler and cluster management as well as the centerpiece component *MapReduce*.

The MapReduce method simply describes the procedure of work distribution (i.e. map) to computing nodes and result aggregation (i.e. reduce) from the nodes. The MapReduce method was designed for distributed and parallel data processing, an important strategy to handle the overwhelming volume of data and rapid data creation that Big Data imposes.

Note that there is no single or unified way to implement NoSQL solution. However, many existing implementations can be categorized into the following three NoSQL database types: Key-Value Store, Graph Database and Document Store (Han, Kamber, & Pei, 2011). Each implementation of these database types had made trade-offs to balance the various concerns. These concerns are characterized by their respective advantages and disadvantages. For instance, Graph Database is optimized for acyclic and cyclic graph objects while Key-Value Store represents data in an unstructured key-value pairs for flexible data representation.

The common theme is that they are based on non-relational data model and distributed computing focus. Also, most of the NoSQL solutions today are unstructured data focus and they are not full ACID (i.e. Atomicity, Consistency, Isolation and Durability) compliant (Han, Kamber, & Pei, 2011) which is the one of the most important features in today dominated relational data model.
The Apache Hadoop Platform

The Apache Hadoop Platform comprises of the two key components: a distributed file system called HDFS for distributed data management and data storage, as well as, the MapReduce method for distributed data querying and data computation. There are other supplementary technologies in the Apache Hadoop Platform including Apache Sqoop. However, they are not the focus of our discussion in relation to our research topic of Big Data.

HDFS stands for Hadoop Distributed File System and it was designed to run on commodity hardware. This design goal allowed the operation of HDFS to be highly scalable and available. To the end, HDFS provides built-in support for data fault-tolerance (i.e. data replication) and load-balancing (i.e. MapReduce). In other words, HDFS is a single logical file system distributed across many data servers and it is able to scale on demand based on required capacity. A high-level HDFS architecture is shown in Figure 8.

Figure 8: HDFS Architecture (HDFS Architecture Guide, 2014)
Given the architecture shown in Figure 8, MapReduce provides a means to query data across the disparate data servers. MapReduce, in a nutshell, consists of the following two functions:

- The Map function is performed by the master node to partition and distribute function input into multiple small inputs to be handled by downstream worker nodes. That is, \( f(x, y) \rightarrow \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots \end{bmatrix} \) where \((x, y)\) is the key-value pair input of the Map function and the matrix is the output of the breakdown of original input.

- The Reduce function is also performed by the master node to aggregate and collect outputs from the disparate worker nodes. That is, \( f(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix}, \begin{bmatrix} y_1 \\ y_2 \\ \vdots \end{bmatrix}) \rightarrow z \) where \( z \) is the result of the aggregation of the individual outputs from the worker nodes.

Both HDFS and MapReduce are low level components of Hadoop. Of course, performing operations directly against low level APIs are time-consuming and unproductive. Without the introduction of some higher level software layer, it would be difficult to drive Hadoop adoption. For this reason, the burgeoning NoSQL based DMBS that are designed to operate on top of the HDFS becomes an indispensable component of any Hadoop implementation.

Recent Hadoop development includes subprojects such as Apache Accumulo, Apache Bigtop, Apache Chukwa, Apache Sqoop and Apache Flume to improve the overall capability of Hadoop platform. Most notably, the Hadoop YARN project is a MapReduce alternative to data processing.
**NoSQL Data Model versus Relational Data Model**

The object structure in relational data model is rigid. For instance, a relation (i.e. table) is defined with a set number of attributes (i.e. columns) as well as a collection of tuples (i.e. rows) representing each instance of an object with matching attributes. Relationship, in this case, is represented by an additional link attribute (e.g. foreign key) introduced in a derived relation (e.g. child table). A rigid model such as relational model has many merits, one of which is a higher level of entity and referential integrity insurance. Since each tuple in a relation is guaranteed with a set number of attributes, constraints (e.g. not nullable and foreign key constraint) can be set in accordance to the level of data integrity required to ensure that data at each tuple level will adhere to the defined constraints.

The NoSQL model rivals the traditional relational data model in that trade-offs have been established in favor of higher performance and scalability. Some of the NoSQL implementation includes, but not limited to: Apache Cassandra, Apache CouchDB, Berkeley DB, FlockDB, MongoDB, Neo4j and Object DB. All of the abovementioned NoSQL implementations fall under one of the following three model categories:

- **Key-Value Stores**: The content of the data is represented as a collection of individual key and value pairs. Dynamo (DeCandia, et al., 2007) is one of the many implementations under this NoSQL model.

- **Document Store**: The content of the data is organized in a per document container object. A single document (e.g. XML and JSON) encapsulates all attribute data for a given object. Couchbase Server (Couchbase.com, 2012) is one of the many implementations under this NoSQL model.
• **Graph Database**: The content of the data are represented in graph objects based on Graph Theory. The object attributes are described in nodes and edges that are interconnected with other object attributes. AllegroGraph (Shimpi & Chaudhari, 2012) is one of the many implementations under this NoSQL model.

Contrary to relational data model, the above NoSQL model structures are not rigid. As such, they are often referred as unstructured data models. The flexibility exhibited in NoSQL data model is highly desirable in Big Data application, due to the fact that, Big Data is inherently unstructured and diverse. NoSQL data model provides a foundational framework for managing Big Data where Predictive Analytics can efficiently perform actions against disparate and distributed databases.

**Predictive Analytics versus Other Forecasting and Mining Methods**

The fields of Big Data and Data Mining have been evolving to bring new ideas and innovations to the field of information management. Though, the term and application of Predictive Analytics in Data Mining came before the term Big Data. Performing analytics, predictive or not, predated both Big Data and Data Mining. This is partly due to the fact that the underpinnings of analytics stem from some of the traditional disciplines, some of which are, mathematics and statistics (Han, Kamber, & Pei, 2011).

The application of analytics has far reaching impact to business bottom-line. Businesses often need to find patterns in their customer behavioral data in order to derive strategies to improve business prospect as they better their understanding of the customers. A multidisciplinary and interdisciplinary approach to uncover business intelligence began to emerge. This translates into tangible practices of risk management, outlier detection, web analytics, logistics management and business optimization. As Big Data gains popularity, the
underlying promise of Predictive Analytics remains the same. However, given the vast amount of data and wide variety of data available at disposal, combined with the fact that data are arriving at near real-time speed, approaching knowledge extraction with novel methods is in order.

Human has long devised numerous techniques to predict and anticipate different outcomes. Industries such as the insurance industry are highly dependent on advanced predictive techniques for their business. This gave rise to an entire discipline dedicated to the study of risk and uncertainty, the actuarial science. Insurance companies employ actuaries to calculate insurance premium based on various factors as to forecast population and assign risk scores to groups of individuals. This deterministic model based on the predetermined segregation of population sample is achieved through the same underpinning techniques as Predictive Analytics. However they vastly differ in methodology and approach, which highlighted one of the major differences between actuarial forecasting and Predictive Analytics.

Actuarial forecasting approaches problems from a top-down fashion and addresses questions that can be answered for a predetermined chance and probability. Predictive Analytics, on the other hand, answers questions through a bottom-up approach at individual level (Siegel, 2013). For example, retail business would be interested in calculating the likelihood of customer A to purchase item Y when item X was purchased during T hours. Determining $Y$ in $\{X, Y\}$ for every A and T leads to actionable outcome for which basket analysis is done via association rules (Han, Kamber, & Pei, 2011). Predictive Analytics provides actionable decision-support information that can benefit targeted marketing by looking at data gathered at individual level – a bottom-up theory generative approach. It is in fact the granularity of data itself that became the enabler for Predictive Analytics.
Conceptually, Data Mining is a superset of Predictive Analytics as shown in Figure 9. From a methodology perspective, Predictive Analytics is a term describing the principles and techniques used in Data Mining specifically for predictive analysis. The Data Mining methodologies includes *characterization and discrimination, frequent patterns, associations, correlations mining, classification analysis, regression analysis, cluster analysis* and *outlier analysis*, as shown in Figure 9. A detail taxonomy for Predictive Analytics is shown in Figure 10.

*Figure 9: Predictive Analytics in Data Mining*
Figure 10: Predictive Analytics Taxonomy

The techniques in Data Mining shown in Figure 9 are applicable to Predictive Analytics for predictive analysis shown in Figure 10. However, Predictive Analytics has a strong preference in employing the classification and regression analysis methodologies which will be
discussed in CHAPTER IV. Many popular implementations such as decision tree and artificial neural network fall under classification and regression analysis category. The rule-based classification methodology is particularly suited for certain Predictive Analytics tasks since the class/label identification and prediction-to-class/label mapping processes are initiative and easily explicable. The regression methodology is common in numerical based analysis to produce a mathematical function that best describes the data. Other techniques in Data Mining are also used during predictive analysis, however, they typically points to usages involving data preprocessing that deals with descriptive and explanatory type analysis.

*The Predictive Model Markup Language (PMML)*

PMML is a standard markup language for Data Analytics. The plethora of techniques used in Data Mining and Machine Learning exert pressures on model interoperability between applications and services. The XML based Predictive Model Markup Language (PMML) was created by the independent and vendor-led consortium called the Data Mining Group (DMG). The DMG standardizes model file format containing model definition (Data Mining Group, 2014). Since the inception of PMML version 0.7 in July 1997, the DMG group has been continuously improving the model definition coverage defined by the PMML schema. The most up-to-date version of PMML is version 4.1 published on December 2011 with enhanced post-processing capabilities and new model elements amongst other updates. An example of the generated PMML code can be found under the PMML Code Example section of APPENDIX A.

*PMML Adoption*

As of this writing, there are over 30 members from data analytics industry and government organizations have adopted the PMML model. Also, many modeling applications have adopted PMML. Some notable modeling applications are: IBM SPSS (IBM, 2014),
KNIME (KNIME, 2014), R (R, 2014), RapidMiner - Extension (RapidMiner, 2014) and Weka (Weka, 2014). Many dataset repositories are also adopting PMML and are offering sample datasets in PMML format. In fact, the UCI Machine Learning Repository currently hosts approximately 300 datasets (UCI Machine Learning Repository, 2014) for model development and testing, many of which are in PMML standard format.

PMML is gaining adoption in both academia and business domains. The reflection in PMML adoption suggests a convergence of modeling techniques and the stabilization of competing predictive modeling methodologies for Predictive Analytics. In fact, the Model section of the PMML schema implements some of the standard modeling methods such as Support Vector Machine (i.e. SupportVectorMachineModel) and Decision Tree (i.e.TreeModel). In order to define a XML schema that is interoperable, the elements in XSD that describe the underlying models must contain the model structure and the composite parts. The PMML schema thus echoes a general consensus amongst applications that support the common predictive models, further hinting to the maturity of the Predictive Analytics field.

As noted in (Guazzelli, Jena, Lin, & Zeller, 2011), the main goal of DMG is the development of PMML, and now aims to make PMML a de facto standard to represent Data Mining models. Many researchers share the same view of embracing PMML and understand its benefits to the research community (Guazzelli, Stathatos, & Zeller, 2009). As a standard, PMML benefits also the Cloud Computing community through promoting interoperability and openness.

**PMML Document Structure**

Under the PMML specification, the components that made up the model are defined under the following major sections (DMG, 2014) as shown in Figure 11. The individual parts of a PMML document are defined to specifically describe a subset of an overall model. The PMML
document structure ensures the interoperability between supporting modeling applications by means of a well-defined schema and use of XML human readable open data format.

Figure 11: PMML Schema (DMG, 2014)

- **Header Section**: The header section of the PMML schema contains model metadata of copyright information, model description, generator application, name, version, annotation and timestamp. An example of a simple header section can be found at the PMML Code Example - Header Section.

- **DataDictionary Section**: The DataDictionary section contains the number of data fields and field definitions such as data type, data range, available data values and data field category. An example of DataDictionary section can be found at the PMML Code Example - DataDictionary Section.

- **TransformationDictionary Section**: The TransformationDictionary section contains general model preprocessing conditions that describe how data will be
transformed from the original state into a desired state. The types of data
transformation PMML supports are normalization, discretization, value mapping,
functions and aggregation. An example of TransformationDictionary section for
value mapping transformation can be found at PMML Code Example -
TransformationDictionary Section.

- **Model Section:** The Model section is represented by MODEL-ELEMENT group
under the PMML 4.1 schema with the main purpose of describing various
modeling techniques used in research and analysis. This is the core section of the
PMML 4.1 standard as most intra-model information is contained within the
model section. The information included within this section varies greatly due to
the differing terminologies and underlying concepts across modeling techniques.
The latest version supports the following models: AssociationModel,
BaselineModel, ClusteringModel, GeneralRegressionModel, MiningModel,
NaiveBayesModel, NearestNeighborModel, NeuralNetwork, RegressionModel,
RuleSetModel, SequenceModel, Scorecard, SupportVectorMachineModel,
TextModel, TimeSeriesModel and TreeModel. The element name unambiguously
represents the names of the modeling techniques. For instance,
SupportVectorMachineModel element is meant to capture model information for
Support Vector Machine (SVM) model. An example of Model section for SVM
can be found at PMML Code Example - Model Section – Support Vector
Machine.
**PMML Interoperability and Application**

The Service Oriented Architecture (SOA) provides the foundational framework for web services to interoperate based on common HTTP communication standards such as Simple Object Access Protocol (SOAP) and Representational State Transfer (REST). Thus, the standards achieve interoperability amongst participating services. The same holds true for predictive modeling in a cloud computing environment where PMML is one of the key enablers to maintain interoperability amongst modeling applications.

The PMML standard brings tangible benefits in modeling application interoperability, improve collaboration amongst researchers and streamline workflow that involves the multistep process of predictive modeling. In this regard, cloud computing such as Software as a Service (SaaS) model (i.e. Google Apps, Amazon EC2) benefits greatly from predictive modeling with PMML, for reasons that software interoperability and integration are critical in cloud computing.

The Amazon EC2 (i.e. Amazon Elastic Compute Cloud) enabled ADAPA scoring engine is one such example that fully took advantages of PMML and SaaS model. The discussion in (Guazzelli, Stathatos, & Zeller, 2009) used ADAPA as the key example. In the study, ADAPA was deployed as a *model verification system* that operates within a cloud environment. The predictive model was built on the PMML standard for describing *model definition* that aids the modeling process from *model design to verification* as shown in Figure 12. The predictive algorithms were expressed in PMML to be scored by multi-instances ADAPA engine hosted on Amazon EC2 infrastructure.
The many benefits of utilizing PMML in conjunction with the ADAPA engine within the Amazon EC2 environment had been illustrated by the El Niño Neural Network modeling example in (Guazzelli, Stathatos, & Zeller, 2009). The inherit benefits (e.g. low startup cost, distributed computing power, streamline management and robust APIs, etc.) of cloud computing combined with the interoperability of PMML, streamlines the process of predicting modeling to a much greater degree. This applies to every step within the process, from model verification to modeling testing.

Of course, the application of PMML in a cloud platform does not limit only to model verification as in the previously discussed ADAPA study. Note that PMML was designed to be multi-purpose in the predictive modeling domain and many researchers have performed model execution based on model described in PMML to derive predictions. So to consider PMML merely as a model data persistence protocol is an understatement to its potential.

In (Das, Fratkin, Gorajek, Stathatos, & Gajjar, 2011), Das et al reaffirmed the view that the portability and interoperability of PMML are bridging the gaps between all participants whom involved in the data mining process. PMML is a conduit that links between cross-teams.
and stimulates inter-organization communication. Consequently, PMML reaches far beyond other direct means of team collaboration. PMML fosters team communication and collaboration that were lacking in the predictive modeling practice domain.

**PMML Enabled Architecture**

A clear example of how PMML bolsters team communication and collaboration is the experiment conducted in (Das, Fratkin, Gorajek, Stathatos, & Gajjar, 2011). The experiment involves an EMC Greenplum database which is a derivative of PostgreSQL database, to act as the backbone for the *in-database processing* experiment. EMC Greenplum was selected in the experiment for its Massively Parallel Processing (MPP) share-nothing database architecture that supports SQL and MapReduce parallel processing.

Under the MPP architecture, a typical configuration of Greenplum database is a collection of servers divided into two roles: the *master* host and the *segment* host. The master host is often setup with redundant servers as its role is critical in that the *master host* is responsible for listening to client queries, optimally allocating the queries to *segment hosts* based on *parallel query plan* and returning processed results to the client application. The segment hosts are responsible for the actual performance of the query allotted by the master host. The above process marks the archetypal MapReduce operation. This architecture is depicted in Figure 13.
The MPP database architecture is modular and distributed in nature as it was designed to support cloud computing and Big Data processing. The MPP architecture demands an open and interoperable model information exchange format to fully exploit its parallel processing potential. This led to the experiment in (Das, Fratkin, Gorajek, Statthatos, & Gajjar, 2011) where the researchers deployed the El Niño regression model to Greenplum database for predictive data processing tasks. The El Niño regression model was created in R and the model was encapsulated and deployed in PMML format.

One of the key aspects of how PMML aided the deployment process involved creating SQL functions as the query execution language for Greenplum and the dynamic mapping of SQL function to a section of PMML definition. The researchers highlighted that the information
contained within the PMML DataDictionary and MiningSchema sections can be effortlessly mapped to the SQL function specification. The active mining fields in PMML can be mapped to the SQL function input parameters while the predicted mining field in PMML can be mapped to the SQL function output parameter. This mapping process involving PMML-to-SQL-functions conversion can be automated as the PMML schema and the SQL function specification are clearly defined to support the model execution. In (Das, Fratkin, Gorajek, Stathatos, & Gajjar, 2011), the SQL functions took the form of User Defined Functions (UDF) and they were created to support the massively in-database parallel processing design by facilitating high performance data selection and data querying.

Once the UDFs are created, the results from the execution of those UDFs are then returned to the client application for the performance of the actual predictive analysis. This architectural design maintained a high performance benchmark score. The authors in this experiment not only highlighted the apparent advantage of using PMML in a natively PMML compliant applications such as R, but also illuminated how the open format of XML and PMML helped in binding and automating seemingly incompatible systems such as Greenplum and R. The demonstration by the authors encourages more researchers and practitioners to adopt the PMML standards across the entire predictive modeling lifecycle.

Cloud Computing

Cloud Computing is a term denoted to a collection of hardware and software services supported by organizations to provide on-demand access to these resources on the internet. According to (Vouk, 2008), Cloud Computing build on the success of preceding technologies including Virtualization, Distributed Computing, Grid Computing, Utility Computing, Web and Software Services.
Many businesses have taken notice of Cloud Computing platform in recent years as a way to take advantage of utility computing paradigm exemplified by the pay-as-you go payment model and virtualized computing infrastructure. Cloud Computing primarily exists in one of the following service levels: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). Each progression from one level to the next built on the previous level and cumulatively adds layers of computing services. Given this layered separation of concerns, businesses can purchase computing services on demand rather than as a lump sum investment, such as paying for a package software solution that often adds costs to integration and service management without utilizing the full solution capability.

The Cloud Computing paradigm brings many advantages in the context of statistical modeling with Big Data. Model execution often demands high power data processing servers and high capacity databases to sustain the strenuous operations imposed by Big Data mining algorithms. These requirements would also levy a substantial upfront investment in hardware and infrastructure. Cloud Computing makes a good economical option for predictive modeling in this case.

In (Kridel & Dolk, 2013), the authors took Cloud Computing to the next logical level in SOA: The Predictive Analytics as a Service (PAaaS). The authors recognized PAaaS and Models as a Service (MaaS) are a subset of Software as a Service (SaaS) model under the Service Oriented Architecture (SOA). PAaaS offers Predictive Analytics modeling services that are tailored to common modeling needs by considering two layers of analytics service, they are called: deep structure layer and surface structure layer. Depending on the resources availability within an organization, one of the two aforementioned layers in performing predictive analytics
can be externalized to the Cloud Computing platform while retaining other analytics functions within the organization.

In addition, the ADAPA engine described under the PMML Interoperability and Application section of this essay illustrated a Predictive Analytics and Cloud Computing use case. The interaction between the user and the cloud-enabled ADAPA system is supported by both the web console mode and the web service mode. The former allows the end user to remotely operate a GUI console through a HTTPS connection while the latter allows web service API communication between programs and ADAPA. Using web service API also allows batch processing to automate model verification process - from loading sample dataset to reporting model score result. The cloud-enabled ADAPA system addressed the important concern of security as well. The connections between all endpoints are secured via HTTPS and the dynamic instancing nature of ADAPA means no residual information retained once the instances have been terminated.

**Dichotomy of Predictive Analytics Skillsets**

The motivation behind the separation of the two layers of concerns stems from the deeply divided human resource skillsets (e.g. Information management versus statistical modeling) and their availability within an organizations. For many businesses, a successful implementation of Predictive Analytics project is a great challenge. This is due to the dichotomy of key skills between IT professional and modeling analyst.

The domain specific skillsets for IT professional and modeling analyst are distinct. The different skillsets calls for very different qualifications. IT professionals are often responsible for ensuring data and system availability (i.e. surface structure layer) while modeling analysts focus on statistical modeling and predictive algorithm development (i.e. deep structure layer). This
dichotomy creates a challenging proposition for small businesses to gain entry into the Predictive Analytics space without the help of PAaaS to set the stage for two key cornerstones of predictive analytics: infrastructure platform and modeling framework.

The chance of acquiring individuals with the proper skillsets becomes increasingly rare as the available human resources decreases within an organization. Therefore, talent acquisition is challenging for small organizations with respect to initiating Predictive Analytics project compared to large organizations. This is why it is important to understand the logical separation between skillsets that are required to deal with deep structure layer functions versus the skillsets that are needed to maintain surface structure layer functions. To clarify, the deep structure layer refers to a stable and foundational schema that rarely changes (Kridel & Dolk, 2013). The surface structure layer, on the other hand, is considered as various instantiations of the deep structure layer that exists in malleable and volatile forms (Kridel & Dolk, 2013).

To put this in a concrete context, a deep structure layer encapsulates model schema and model execution related items and activities whereas the surface structure layer houses the dataset that feed the model as well as the supporting infrastructure for providing the dataset. For instance, IT professionals focus on data acquisition and data access provisioning tasks which operates within a problem domain of surface level layer. The IT professionals support data scientists to concentrate their effort in the modeling process where data modeling is in a problem domain of deep level layer.

Given the abovementioned separation of concerns, a set of discrete predictive analytics functions can be laid out and automated by organizations as they see fit. Predictive modeling workflow requires model creation, model execution, model verification and model data to form a complete predictive modeling lifecycle. A consumer of a PAaaS service can be those who do not
possess the modeling skillset (i.e. data mining, model scoring, etc.) but well-versed in business problem domain (i.e. domain data acquisition and provisioning). This allows business to ignore details of modeling workflow and deals only with the needed dataset provisioning as shown in Figure 14.

Figure 14: Automated modeling workflow (Kridel & Dolk, 2013)

Figure 14 describes an automated self-service modeling system workflow reference model for PAaaS. This approach removes the burden of knowledge from the organization to implement an all-encompassing modeling workflow from the ground up.

**Predictive Analytics Performance Optimization**

While the performance in computing system improves and the latency decreases between connected computing services, it is understandable to expect higher performance when the modeling services are physically nearer the DBMS where the data were queried. However, geographical distance imposes a physical limitation and negatively affects network latency
between computing nodes which has been a continuing challenge for researchers. Therefore, the benefits in amalgamating the modeling services and the data services into a single platform would improve system performance.

This proposition of a single platform approach was described in (Fischer, et al., 2013). The authors presented a solution in which performing forecasting time series data directly within DBMS ensures consistency between data and models in addition to a measurable increase in efficiency. This is largely due to reducing data transfer that made possible by in-database optimization techniques. As the authors described, there has been a growing trend of integrating statistical services, predictive and other data analytics into DMBS given the above reasons. For example, using time series analysis can help finding out the parallel between the forecasting applications involving sales data, energy data and traffic data.

Time series forecasting has been conventionally used in many of these applications. A typical time series forecasting process consists of the following consecutive parts: model identification, model estimation, model usage, model evaluation and model adaptation as shown in Figure 15. Note that the aforementioned process is cyclical where the final sub-process (i.e. model adaption) will in turn provide feedback to the model identification and model estimation sub-processes.
The authors in (Fischer, et al., 2013) pointed out that, although there are evident that the in-database forecasting process is gaining momentum, in-database forecasting is still lacking in functionality. The limited forecasting capacities of the major commercial DMBS such as Microsoft SQL Server and Oracle Database are examples of such concerns. To that end, many researches have proposed ways to further optimize *intra DBMS modeling* and commonly break down into the following items:

1. **Integrated Declarative Forecast Queries**: Additive keyword such as FORECAST can be used with a SQL Query statement.

2. **Deep Relation Query Processing Integration**: Build-in forecast operators in conjunction with standard SQL operators and seamless data joins with current and historical records.

3. **Transparent and Automatic Query Process**: Model attributes can be added automatically to ad-hoc models for multi-dimensional time series model processing.
4. **Query Optimization**: Using I/O-conscious skip list data structure for very large time series data and advance techniques in models reuse in pre-computed multi-dimensional data.

5. **Efficient Update Processing**: Model-aware and model-specific mechanisms are developed to continuously update the underlying model through a stream of incoming time series values.

To bring to light the current issues of effective model process, the proposed ways include the abilities to execute forecast queries with optimization built into DBMS. First and foremost, the suggestion in (Fischer, et al., 2013) is the inclusion of *Time Series View* alongside with the traditional view to add support for time series data representation. The Time Series View is essentially a materialized view with time series attributes incorporated where the date attribute is treated as first class attribute. It also contains built-in considerations of statistical information such as standard deviation and confidence intervals. Time Series View combines the support of continuous data integration where the predicted values from the Time Series View are replaced with current data as they become available.

Besides the external schema (i.e. Traditional View and Time Series View) discussed above, a generic *forecasting system architecture* (Figure 16) based on the traditional ANSI/SPARC architecture was also proposed in (Fischer, et al., 2013). The architecture involves a *conceptual schema* (i.e. model composition) and *internal schema* (i.e. logical access and physical access path). This was designed within a traditional relational DBMS environment containing two separated virtual domains: *relation data* and *forecast models*.
The forecasting system architecture as shown in Figure 16 contains the important components to augment the existing capacities of a relational DBMS. The vertical line in Figure 16 separates the relational model (on the left) and the forecast model (on the right). Each schema layer in the forecast model incrementally adds native support for the Time Series View. This design demonstrated the feasibility of extending a relational DBMS architecture to support forecasting capability without a complete redesign to accomplish a forecast-capable DBMS.
CHAPTER III
PREDICTIVE ANALYTICS APPLICATIONS

As discussed in the Explanatory versus Predictive Modeling section of CHAPTER II, we highlighted the differences between Descriptive Analytics and Predictive Analytics. Descriptive Analytics is a derivative of all characteristics in an explanatory model, which, has a long history of research and practical applications that help to describe causality and explanatory statistical based phenomena. Data Mining has been mostly associated with Descriptive Analytics for knowledge discovery. Data Mining, in the context of Descriptive Analytics, surfaces the facts and causal links between variables leading to the discovery of trends and insights. As such, Data Mining allows one to survey the present situation by pruning away noisy data and ambiguity in a given complex situation that exhibits a set of convoluted cause-and-effect relationships. These relationships are often entangled with assumptions, perceptions and potential false information. While both models have the power to inform, however, only predictive analytics can transform the practice of decision-making.

The Data Mining techniques have been well researched and understood by researchers and practitioners. In (Han, Kamber, & Pei, 2011), Han et al. explained in details of the various Data Mining techniques such as pattern mining, classification, associations, clustering as well as the advanced techniques involving machine learning methods such as Bayesian belief network and support vector machines. The aforementioned methods involving supervised machine learner is particularly suited for Predictive Analytics. One of the many examples is the Backpropagation artificial neural network. Backpropagation artificial neural network was invented to solve complex problem in a way that mimic a part of the central nervous systems of a human brain. The function of a human brain has the ability to learn from experience and
recognizes new pattern which is the basis for prediction and decision making. This technique will be discussed in detail under the Artificial Neural Network section of CHAPTER IV.

To translate the model, methodologies and techniques into practical terms, this chapter sets the stage for illustration of real-life Predictive Analytics applications across industries. The discussion thus far has been mostly focusing on the fundamentals of Predictive Analytics. In this chapter, we will be focusing on the research effort in the practical domain instead of focusing on the theoretical discussion as we did in CHAPTER II. The varying techniques and methods will be covered in depth in CHAPTER IV while this section will be focusing on the real-life applications in Predictive Analytics.

**Social Computing**

*Large-Scale Machine Learning at Twitter Inc.*

Twitter, Inc., a well-known online social networking and microblogging service, with hundreds of millions of active users sending primarily short text messages and search queries, employ both Descriptive Analytics and Predictive Analytics for sentiment analysis and trend prediction respectively.

The example of Twitter Inc. touches many aspects of the essay’s subject matter. From a Big Data perspective, a high volume of messages and events represented data captured through 200 million users at a velocity measured at 400 million daily tweets and a variety of content type such as text, images and videos (Moore, 2013).

The challenges are evidently shown by the accumulating stream of raw data at a daily rate of 100 terabyte (Lin & Ryaboy, 2013). To navigate through the mountain of data at petabyte-scale, the research in (Lin & Ryaboy, 2013) described the Big Data disruption to Twitter Inc. and
the novel means to address those problems. The authors identified the problem domains in the areas of data capturing and data processing.

Firstly, classifying captured data is an important step. There are two classes of data: the *business centric data* such as customers and contracts that are part and partial to the business. This class of data has always been maintained in an organization and it is critical to the day-to-day business operation. However, there exists a second class of data that represents *user behaviors*. This class of data is equally important but often overlooked. This particular kind of data is crucial to many businesses and especially important to a social media company like Twitter Inc.

Secondly, the analysis portion of the Big Data equation is the Descriptive Statistics that includes the use of Online Analytical Processing (OLAP) and Extract-Transform-Load (ETL). They pervade the majority of many IT operations and they are synonymous to what commonly known as *Business Intelligence* (BI). The authors in (Lin & Ryaboy, 2013) recognized the need to perform Predictive Analytics that cannot be done effectively by traditional methodologies and tools but it is achievable with advanced modeling techniques such as *machine learning*.

Thirdly, the open-source disruption, largely credited to the *Hadoop* open-source implementation of MapReduce and surrounding technologies that cemented the infrastructure for Big Data mining.

The contrasting concepts are clear in terms of solving problems that are well-known and those that are latent. Solving unknown problem requires a new kind of modeling that is designed for turning vague directives into concrete and solvable problems. The authors also explained that typical preprocesses of Data Mining, such as data cleansing and data normalization, are still important for avoiding skewing of predictions as a result of missing data, outlier data and
generally incorrect data. The arduous data cleansing process at Twitter Inc. still requires a large amount of manual intervention and have yet to be fully automated. This is largely due to the current system complexity as a result of rapid business growth as well as the loose coupling system architecture design (Lin & Ryaboy, 2013).

Upon the completion of the first step of Data Mining, the authors proceeded to tackle a business concern of user retention, which is, a *stochastic classification problem* of what attributes and behaviors of an active user can be used to predict the future activity of another given user. That is, the probability of a user becoming inactive is predicted through the shared attributes of other users who previously became inactive. This problem can be represented as \( P(Y|X) \) where \( X \) is a given user whose inactivity was previously recorded in a set of attributes represented in \( X = \{x_0, x_1, x_2, \ldots\} \) and \( Y = \{y_0, y_1, y_2, \ldots\} \) is a set of user attributes and behaviors for a predictable case of a user \( Y \).

Once a predictive model is developed, the predictive accuracy of a given model can be assessed through the use of *backtesting* method. In backtesting, historical records are used in both model training as well as in model validation. Suppose the entire dataset consists of five years’ worth of user activity data, the data in the first four years can be partitioned off into a model training dataset and to be used to train the model. Once the model has been trained, the model then generates predictions on future user activity. The result of the prediction can be validated and assessed by the data in the final year of the historical dataset. Thus, backtesting allows a full assessment of the predictive power of a given model based only on historical data.

Backtesting is achievable because both variables, the prediction outcome and the prediction criteria are both known facts in the chosen dataset. This allows researchers to compare the known result (i.e. today’s data) and the predicted result (i.e. computed value) in order to
tweak the model parameters to improve overall predictive power. The backtesting method was the basic form of model validation that Twitter Inc. currently employs.

Outside the internal concerns of Twitter Inc. business operation, the public Twitter.com API provides a means to externalize some of the massive internal data. Many researchers have conducted experiments on Twitter Inc. data for various novel measures of hypotheses. The machine learning approach that underpins the Twitter Inc. platform was discussed in (Lin & Kolcz, 2012) where integration between the Hadoop Pig platform (Hardoop, 2014) and the Twitter Inc. massive machine learning engine was discussed in great length.

Another example that leverages Twitter Inc. data was presented in (ARIAS, ARRATIA, & XURIGUERA, 2013) which described the various ways one can correlate the stock market volatility, movie box office revenue and presidential polls result by performing sentiment analysis on Twitter Inc. data. In particular, the daily Twitter Sentiment Index, coupled with the general sentiment time series data, had been used in a few studies referenced in (ARIAS, ARRATIA, & XURIGUERA, 2013). These studies showed the correlations between the sentiment of Twitter users and the forecast targets such as Dow Jones Industrial Index (DJIA) and US influenza rates.

The result of analysis in (ARIAS, ARRATIA, & XURIGUERA, 2013) was deduced into a decision classification tree called a summary tree. The summary tree was then built based on the experimental results with self-tuning capabilities to increase its predictive ability. The summary tree was not used in the prediction execution process itself but it was used as a supplementary tool to capture experimental observations. It is important to mention that the study was conducted in an experimental design where a control group was used to determine whether the Twitter Sentiment Index alone helped the predictive model to prove the stated
hypothesis. The resulting predictions were overwhelmingly positive. The study concluded the nonlinear models (e.g. support vector machine model and artificial neural network) yielded the highest successful prediction rate as compared to the result produced by linear models (e.g. simple regression model).

Like many social networking services, Twitter Inc. suffers from the peril of spam messages. In (Wang, et al., 2013), the authors employed a Random Tree algorithm for the spam classification model based on click traffic data and the shorten URLs generated by Bitly.com. The model produced predictive result with 90.81% accuracy and 0.913 F1 measure value. A fully developed spam classification model can be used as a predictive model for future spam detection. This is a classic binary classification problem of \( Y = f(X) \) where \( X = \{x_0, x_1, x_2, \ldots \} \). The model was constructed to classify an email message as either a spam or non-spam message by predicting a binary target variable of \( Y \). The study showcased the practical use of classification model to solve tangible business problem of spam traffics in a predictable manner.

The above discussed Twitter.com case studies underline the convergence of Big Data, Data Mining and Predictive Analytics in a practical business domain. They exemplified the application of predictive modeling techniques in solving both known and unknown business problems.

**Network Relationship**

In many predictive applications, using only one predictor (e.g. age) to predict a particular outcome (e.g. education level) is an inexact measure that is problematic due to the lack of consideration of other important variables. Adding other attributes (e.g. race, height, income level, etc.) of an entity can improve the validity and accuracy of a measure but the model would
still lacks breadth. The reason is that many relationship data between the entities themselves were not accounted for and very often, they are very important in improving predictability.

The social element in network connections is one aspect where it can enhance data quality, which is to say, taking into account a person’s social role reveals more data about the individual’s behaviors. These data would otherwise be hidden had the network connections not been considered. In (Nankani & Simoff), the authors incorporated the network relations component in their Classification and Regression Tree (CART) predictive model to enrich the entity dataset. The authors took into account the network relationships between actors such as co-authorship, co-participant in a given academic domain to predict and forecast the two outcomes below:

1. A given research project will be funded or not.

2. The predominant category of personal publication.

As it was noted, the target variables are both a binary label representing funding decision and a discrete label that defines the publication categories such as book, conference paper and journal articles.

The methodology depends on the measure of actor centrality in graph theory. Actor centrality is often used in network analysis which refers to the degree (i.e. indegree or outdegree), closeness and betweenness of actors (i.e. student, instructor, administrative staff, etc). By incorporating the network relationship data and by using the Salford systems CART tool (Nankani & Simoff), the authors were able to enrich existing data to improve the predictive power of their predictive model that leads to an overall enhanced predictive model. Proving that the hypothesis that information about network structure can improve the predictive accuracy of a given model.
Education Support

Predicting At-Risk Students

In the realm of pedagogical analytics, maximizing student participation rate often fosters a positive education environment which is the goal of many educators. In (Annika Wolff, 2013), the authors presented an interesting correlation between clicking behavior of students and the course outcome of students in an Virtual Learning Environment (VLE) conducted at The Open University Institute. The result of the analytics work was used to predict at-risk students so interventions can be administered in advance.

Generally speaking, the VLE activity information (e.g. hyperlinks click frequency) was considered a strong predictor in many pedagogical researches in an online learning environment. This is because, a number of problematic obstacles exist exclusively in distance learning environment that do not exist in canonical educational institutions such as geographic and time zone differences, lack of in-person face-to-face learning environment, imbalance of educational background among students. The authors employed a number of independent variables to predict the dependent variable of at-risk students whom tend to either fail a particular course or dropping out a program entirely.

Besides the activity data in VLE, the authors also incorporated financing, demographic, course subject area and general course information as independent variables in the study. The authors pointed out the danger of misinterpreted correlation such as click frequency, which represent only an aspect of student activity measure, does not always correlate with the course outcome. The perception of students with high click frequency leads one to presume an engaged student in terms of online access frequency. Conversely, a low click frequency would imply a low engagement student and consequently lead to a negative course outcome. In reality, student
preference on learning material should also be taken into account. Preference is a latent factor that cannot be measured by activity data as some students prefer printed materials over online materials and therefore it would skew the abovementioned model result.

The subsequence steps demand the consideration of outcome classifiers which are the classification labels for what outcomes constitute an at-risk student. The authors defined such labels as *performance drop* and the binary response of *course outcome* (i.e. pass or fail). The execution of model was done in a multiple combination of predictor variables, such as, standalone VLE Activity data, standalone Tutor Marked Assessments (TMA) data, VLE and TMA data, etc. The resulting predictor performance on performance drop outcome favors a combined data categories (i.e. VLE and TMA) based on decision tree algorithm. The course outcome dependent variable (i.e. pass or fail) showed a different picture where VLE only data (i.e. total clicks, delta in clicks, clicks relative to historical context, etc.) yielded the best predictive performance based on the precision, recall and f-measure scores.

The other permutations of the model involved the inclusion of demographic data which had been confirmed to improve overall predictive power. Also, over the course of the assessment, the predictors performed in varying degree of accuracy and precision at each point in time. Exclusive VLE data yielded better score during the early stage while other forms of measures produced better overall result at the later stage of the experiment.

The authors concluded the paper with a confirmed hypothesis based on the above predictive model that, *click frequency in online learning environment do correlate positively with course outcome* as long as the measure takes into account the overall timeline and clicks volatility. For instance, if click frequency of a student decreases abruptly during a course, this
correlates positively to a negative course outcome. Postulating such hypotheses and confirming them by analyzing and predicting student performance in a VLE is important to many educators.

**Predict Course Success**

In (Barber & Sharkey, 2012), the authors conducted an empirical experiment at the University of Phoenix where multiple predictive models were created based on a dataset containing 340,000 student records. The authors proposed three models with each containing different independent variables as predictors for course success prediction.

Model 1 is a *Logistic Regression* model targeting an ordinal dependent risk level variable, the values are, high, neutral and low risk. While model 1 performed admirably at 90% accuracy on pass (i.e. low risk) and fail (i.e. high risk) with \( p < 0.5 \), the authors expressed concerns over the model. The data that supported model 1 originated from multiple databases with varying level of data quality that required significant data cleaning and perpetration in order to bring the data to an acceptable level. Also, the researchers had reservation over the level of granularity of the risk level with the concern of levels being over generalized.

Model 2 was built using *Naïve Bayes algorithm* in RapidMiner and it included variables that were not present in model 1 such as military status and financial status as well as the inclusion of the *discussion board posting count*. In model 2, the researchers omitted a few independent variables that are insignificant and lack predictive power through previous observations in model 1. These include the attributes of gender and age, amongst others. Model 2 showed improvement over model 1 for the weekly predictive accuracy measure reaching 95% accuracy by week 3. The most influential independent variables in model 2 are *credits earned versus credits attempted ratio*, *previous financial status* and the most powerful predictor being *cumulative points earned*. 
Model 3 was built with the hypothesis that student engagement is positively correlated with course pass grade; this is in alignment with the hypothesis discussed in the previous Predicting At-Risk Students section. The independent variables under consideration included a lower form of discussion board activity data that could suggest engagement level (i.e. time since last course, public post versus private post to instructor).

The authors concluded the paper with the emphasis on model accuracy and utility which are the two topmost considerations. Utility refers to whether the resulting model result is actionable and can yield fruitful outcome for student success. The experiment also highlighted that the combination of differing predictor variables with the varying predictive models, would yield a very different prediction outcome.

**Video Gaming**

The video gaming industry has long been engaged with machine learning. The video gaming industry is one of the early adopters in Artificial Intelligence (AI) research and development as far back as 1959 (Kaur, et al., 2013). Video gaming industry, also called interactive entertainment industry, is intrinsically tied to the feedback mechanism of the anticipation of human responses by establishing mutual and reciprocal human-machine relationship. An enjoyable video generates human emotional responses which is a direct result of machine learning techniques. Some of these emotional responses are amusement, excitement, contentment, wonderment and surprise.

In (Geisler, 2002), a great amount of details went into explaining enemy opponent AI design in First Person Shooter (FPS) video game. The feature vector representing player decisions (i.e. accelerates, changes movement, changes facing, or jumps actions) are captured for supervised learning through the learning algorithms of ID3 Decision Tree, Naïve Bayes and
Artificial Neural Network. The learning algorithms reinforce the game’s AI agents such as emery opponents.

More recently, researchers have devised techniques to create supervised machine learning models by capturing human gameplay activities and behaviors through video game to create a human-like AI model. This is an effective way of using human actions to build model that embodies human behaviors.

One such example is to use machine learning models to solve real-world problems through video gaming as explained in (Sanfilippo, et al., 2011). The authors proposed and developed a prototype system for generating predictive outcome in assessing the propensity for state and non-state actors in participating illicit nuclear trafficking through Bayesian belief network and agent-based simulations. The goal of the model design proposed by the authors was to capture real-life human interaction behaviors through video game simulation. The data are then analyzed through machine learning techniques based on the Technosocial Predictive Analytics (TPA) framework. The application for the model in (Sanfilippo, et al., 2011) has a very strong focus on the deployment of Predictive Analytics in the field of government and military surveillance, which was used to predict illicit activities for the purpose of resource allocation and strategic resource deployment to high risk regions.

The simulation itself was constructed as a multiplayer game where players took the roles of actors in a staged illicit trafficking scenario based on the following conceptual components: Technosocial Modelling, Knowledge Encapsulation Framework and Analytical Gaming. The authors described the aforementioned components in great depth but the main thrust of the concept as it relates to Predictive Analytics, was the inclusion of System Dynamics and Bayesian
Belief Network methods through supervised learning. These methods are supported by a training dataset that was created by multiplayer gaming sessions simulating real world events.

The Bayesian Network model as shown in Figure 17 was selected in modeling the illicit trafficking game experiment as the learning model. This is because the stochastic nature of player reactions to random events can lead to a number of possibilities that are non-deterministic in nature. Also, System Dynamics modeling were used to map model parameters (e.g. intent to establish alliance) to game parameters (e.g. initiate communication) to control elements of game environment. Using Bayesian Network and System Dynamics are a two-prong approach in model selection to tailor specialized aspects of the game engine.

Creating strategic and tactical human responses is a difficult proposition for unsupervised machine learning techniques, especially, given the fact that human contextual information is
acquired through *experience*. Contextual information is a necessity in devising strategic and tactical decisions. Information arrive at human through a number of random and established channels. Structure channels such as television news is very much different from other unstructured information sources such as social interaction or word-of-mouth communication. Particularly, in offline communications involving verbal and non-verbal information are not commonly captured to be analyzed by machines. Thus, these types of contextual information often exist solely in human brains. To that end, a paper by (Riensche & Whitney, 2012) fused the two concepts of predictive analytics and gaming succinctly by proposing *knowledge transfer* from human to model through gaming. This maximizes the information input to machine including contextual information. The paper focused on war game simulation which is type of *Analytical Gaming*, a form of *Serious Gaming* that dealt with players’ ability and facilitates knowledge extraction.

**Law Enforcement**

In criminology, the study of criminal behavior of individuals as relates to social science is an interdisciplinary field that deals with, at the very least, behavioral science and law enforcement theory. Many results of the criminology study have been directly influencing lawmakers as well as legislations of the criminal justice systems. The use of quantitative statistical methods in criminology in identifying chronic offenders have long been used and the practice is dated back to 1972 (Jennings & M.C.J, 2006) in order to predict high-risk offenders.

The predictive model led to the construction of prediction instruments for measuring criminal behavioral risks. In the article by (Jennings & M.C.J, 2006), the authors discussed the practice of criminal classification by means of *risk assessment instruments*. The risk assessment instruments draw on data from a wide area such as psychological data, socioeconomic data,
demographic characteristics and conditions. The result of such measure had led to the intelligence-led policing such as proactive policing, problem-oriented policing, community-based policing and knowledge-based policing. They are the constituents of an overarching forward-looking model of crime prevention that relies on risk assessment techniques for both new offenders and reoffenders.

The paper by (Jennings & M.C.J, 2006) discussed the result of an empirical experiment with a randomized sample of offenders whom have a history of arrests within the past three years. The sampled offenders were studied over the course of six months and the observed behaviors were recorded including the number of re-arrests made during the six months period. The independent variables for the prediction model were selected as following: property crimes scale, person crimes scale, drug/alcohol scale, crime severity scale, repeat offender scale and the violence scale. The dependent variable in this case was the recidivism rates of a prior offender.

The result was calculated using principal components analysis method and the result suggested a strong indication of a positive correlation amongst the aforementioned six factors. The final result from the chi-square tests led to the following table of figures for the three risk levels of recidivism, displaying a strong positive correlation between risk levels and classes as shown in Figure 18.
Figure 18: Group Recidivism Rates (Jennings & M.C.J, 2006). Note: \( n \)=number of individual per category (scale) and per risk class, \( \% \)=percentage of the individual in a specific class were re-arrested.

Before predictive analytics became a recognized term, many researchers in criminology had used the same statistical approach in identifying high risk individuals with the intent to balance individual rights and public safety, particularly in the area of reoffending and recidivism risk assessments. According to the article by (Greengard, 2012), a growing trend amongst law enforcement and modern policing is to use data and model driven mathematical and statistical techniques to better direct crime prevention. Like other predictive modeling applications, predictive policing makes use of high dimensionality dataset spanning across space (e.g. locations) and time (e.g. time of day). They are made up of past historical crime records that can be deduced into geographical, temporal and distribution correlated patterns.

Human behaviors often carry a certain pattern in a form of habit which applies to criminal activities as well. For instance, it is statistically proven that the probability of a past crime event would increase the likelihood of higher future crime activities occurring within the same vicinity of a past crime. This can lead to a concentrated occurrence of crimes due to a collective social presumption of tolerable activities. This phenomenon is called a Broken Windows Effect. The effect can be observed through the forming of crime hotspot clusters where a prolonged deterrent action has not been taken in those areas with high criminal activities.
Generally speaking, law enforcement has policies in place to review reports of past crimes stored in databases for the purpose of crime mapping. The move from reactive policing to proactive policing is not a new trend in law enforcement practices, but the shift to a stronger reliance on proactive and predictive approach is in fact gaining momentum. The proactive policing translates into patrols deployment in areas that require crime deterrent. One of the many methods in proactive policing can be done through crime pattern analysis.

Predictive policing, on the other hand, requires advanced analytical tools and techniques to increase crime predictability for the required accuracy. One of the key examples in predictive policing was demonstrated by the Memphis Police Department’s intelligent crime fighting solution, called the Blue CRUSH (Criminal Reduction Utilizing Statistical History). The design goal of Blue CRUSH is to surface actionable insights from two commercial analysis support solutions: IBM SPSS Statistics and ESRI ArcGIS. Blue CRUSH resulted in a 30% reduction in overall crimes and 15% reduction in violent crime in the four-years span and directly responsible for 50 arrests on drug related crimes since the Blue CRUSH deployment.

Another example of predictive policing application was presented in (Hollywood, Smith, Price, McInnis, & Perry, 2012) by the National Law Enforcement and Corrections Technology Center. The authors were clear on correcting the perceptions held by many people in regards to the effectiveness of predictive policing. The authors stated that predictive policing does not equate to divination and that we should perceive predictive policing as a decision support model. The long practice of crime mapping is considered a basic form of predictive policing to anticipate criminal activity. Crime mapping is an assistive and informative tool for crime prevention. The predictive policing process mirrors many iterative-based processes where forming a feedback loop is necessary. The feedback loop allows the result of a single process
iteration to enrich the input of subsequent iterations. The predictive policing process (Hollywood, Smith, Price, McInnis, & Perry, 2012) comprises of four high level sub processes: data collection, data analysis, police operations and criminal response.

In fact, many crime hotspot identification methods involve a combination of multiple techniques such as near repeat methods, risk terrain modeling, regression and other common data mining methods as well as spatiotemporal methods. These mathematical methods have been studied in great details in academia as presented in (Short, D’Orsogna, Brantingham, & Tita, 2009). They are not exclusive to predictive policing but the methods apply to Predictive Analytics in the most fundamental ways. These methods allow the deduction of past crime data to produce insights that represent crime patterns. Many of which are very often temporal and spatiotemporal relative, in terms of the day/night cycles, weekend versus weekday, paydays, sporting events, concert events and time of year. These independent variables are used in correlation analysis to help predict future crime activities.

The potential of predictive policing is worthy of continuous research within the realm of Predictive Analytics. An assessment was conducted in (Yang, Wong, & Coid, 2010) to assess the efficacy of violence prediction led by predictive policing. The authors highlighted the difference between high-frequency and low-frequency crimes where low-frequency crimes such as serial killing and school shooting tend to generate many false positive type I errors. In those cases, Type I errors are costly and socially prohibitive to act upon.

The simple model of assessment-to-prediction-to-intervention is a common model for actuarial risk assessment and Predictive Analytics alike. The actuarial risk assessment model is an accepted standard of forensic risk assessment practice which draws in multiple constructs ranging from clinical (e.g. personality disorder) to situational (e.g. community support). In
actuary, the selection of varying constructs changes based on the subjects of assessment. The practice of actuary derives predictions based on empirical evident and professional judgment. The method is subjected to repeated empirical validation while keeping both static (e.g. ethnicity) and dynamic (e.g. received treatment) predictors under consideration.

The authors discussed in depth of the relationships established on violence and psychopathy with respect to Predictive Analytics. The comparison of predictive efficacy was performed on 9 actuarial assessment instruments: Hare Psychopathy Checklist-Revised (PCL-R), Violence Risk Appraisal Guide (VRAG), Violence Risk Assessment Scheme (HCR-20), Level of Service Inventory-Revised (LSI-R), Psychopathy Check List: Screening Version (PCL:SV), Lifestyle Criminality Screening Form (LCSF), General Statistical Information on Recidivism (GSIR), Sexual Violence Risk-20 (SVR-20) and Static 99. The predictive efficacy of abovementioned instruments were studied and concluded with similar performance given the same context. The majority of the variance in efficacy was due to methodological features such as age and length of follow-up, which is to say, the conclusion was such that none of the instruments studied can produce significant standalone advantage that are able to differentiate itself from the other instruments in terms of predictive power, given the same methodological features.

The result suggested that the efficacy of the simple tools (e.g. summing) might have reached a plateau and the needs for novel means of identifying and combining risk predictors is in order. This serves as a great remainder to consider the significance of the role of data in Predictive Analytics.
Business Applications

Many business applications of descriptive analytics are customer centric, meaning that the goal of many businesses is to improve customer service experience in exchange of business loyalty. Companies add tangible business values to their businesses by having a better understanding of their target customers, leading to an increased demand of analytics that aids decision makers to better drive business decisions to thrive amongst competitors. Traditional Business Intelligence (BI) answers standard queries that are descriptive analytics driven, which is to say, queries that answer only in historical context based on descriptive and explanatory models. The models are designed to work with past dataset. While predictive analytics also depends on historical data, and that data is in fact historical by definition, the key difference is the timeliness of the data and the type of model used.

In (Nauck, Ruta, Spott, & Azvine, 2006), the authors explained the common approaches for data analysis and highlighted the inefficiency in many linear and nonlinear regression based modeling methods. The regression model often lacks depth in terms of revealing multidimensional variables dependencies and it is not designed to reveal latent variables while factor analysis excels at that. The authors proceeded to base their analysis in Bayesian network, citing a superior alternative to regression based analysis. Bayesian network, based on Bayesian Theorem, is a way to discover information about the structure of statistical dependencies. The authors employed a Bayesian networks modeling tool called Intelligent Customer Satisfaction Analysis Tool (iCSat), to perform overall sensitivity analysis and what-if analysis in supporting customer satisfaction analysis. The goal of the analysis was to identify customers whom are in jeopardy (e.g. potential high churn rate), customer satisfaction target setting and field force performance.
Bayesian network deals with classification problem based on Bayes’ Theorem. Given a class label $Y$ and the dataset $X$, the probability of $Y$ occurrences in $X$ is determined by $P(Y|X)$. That is, the probability of $Y$ happens given $X$. The resulting formula for a single condition is determined by the formula shown in Figure 19.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

*Figure 19: Bayes’ theorem formula*

When dealing with multiple conditions, the results of the formula (Figure 19) are multiplied per each condition to produce the overall probability value. Bayesian Belief Network can be seen as a way to visualize a collection of naive Bayesian rules and allows for a way to determine probability values based on the combined relational structure between dependent and independent variables. The dependent and independent variables are also directionally significant. A non-bias and a priori way to calculate probability based on a nested set of variables. This concept is deeply rooted in Bayes’ theorem.

Given the complexity involved in a large size Bayesian Network model, the authors relied on iCSat to build models and to perform data analysis that deals with what-if scenarios. This is possible because of the nature of structural relationships of a directed acyclic graph. Bayesian Network can handle noise and partial information using local and distributed algorithms for inference and learning (Pearl & Russell, 2000).

Another example of Business Application was illustrated in (Waller & Fawcett, 2013), the authors discussed the key concepts where the exploitation of domain knowledge, as evidently expressed in the variety of skillsets required by predictive analytics research, including statistics, forecasting, optimization, discrete event simulation, applied probability, data mining and analytical mathematical modeling, are a reflection of market demand for specialized skillsets to
solve problems within a specific domain. The predictive power of predictive analytics in combination of Big Data augments the field of Supply Chain Management. Optimizing logistics with respects to customer, sales, carrier, manufacturer, retailer, inventory, location and time, is curial and it is key to balancing cost and service level, which are of primal concern to many businesses.

The result from business optimization through Predictive Analytics directly adds values to businesses and optimization necessitates the higher *volume, variety* and *velocity* properties of Big Data as the foundation for analytics. Predictive Analytics helps researchers and practitioners to analyze the multifaceted field of Supply Chain Management in that it approximates the relationships between variables with the use of statistical and mathematical deduction methods, to draw predictions based on the result from data mining historical data in both quantitative and qualitative means.

**Financial Engineering**

The survey discussion thus far has identified a number of applications where Predictive Analytics can directly or indirectly influences the mode of established operations to further enhance existing capacities and horizontally scale across industries and organizations. It is evident that Predictive Analytics possess a crosscutting solution that deals with issues and challenges faced by many academic institutions, business corporations and organizations with specialized functions. However, none other has generated the same level of controversial public discourse that took place in financial service sector. The emphasis has been on finance investment where *financial engineering* is a strong focus. The field of financial engineering is very much grounded in the notion of prediction.
The machine learning application in investment finance is prevalent in the industry. This is because the multifactor market performance variables are vital to investment decisions and business strategy. Machine learning methods used in predicting security price patterns are a prominent example of Predictive Analytics for stock traders and portfolio managers. This view is evidently supported in (Soulas & Shasha, 2013). The authors employed the Statstream software to correlate securities using a *sliding window technique* and to predict foreign exchange rate changes.

In financial investment services sector, the institutional investment firms such as hedge fund companies, pension fund companies, mutual fund companies and insurance businesses, collectively known as the *buy-side* firms, are innately dependent on prediction. They often make investment decisions based on recommendations from the *sell-side* firms such as market research companies and brokerage firms.

Traditionally, the data input coming from the sell-side firms arrive through casual channels such as conversations that occur during meeting and between phone calls. Quantifying information that comes from casual conversations is problematic as the information obtained does not get propagated throughout the rest of the system. This reduces the chance to capture information that lead to a traceable and actionable trading recommendation. Many institutional investment firms have developed in-house algorithmic trading platform to automate trading executions as well as for trading decision support. These algorithmic trading systems often rely on a constant stream of data from multiple sources in order to derive trading decisions that leads to a binary decision of *buy or sell* of a particular security.

Hedge fund companies in particular, have developed such proprietary algorithmic trading systems to provide Predictive Analytics functions. To that end, many trading algorithms sift
through massive volume of market information based on predictive modeling to reach at a prospective decision based on the potential performance of any given security and market conditions. Rapid data feed come into the system that varies in scope and in kind, characteristics that are shared by Big Data.

In many ways, the market information resemble the definition of Big Data of unstructured dataset where variety means data from blogs, published news, interviews, phone conversations and other forms of unstructured communication that are in high velocity and volume. They fused together with advanced modeling techniques involving the calculation of alpha (α) and beta (β) coefficients based on regression model (e.g. Capital Asset Pricing Model) to access risk level on expected return. Therefore, the application of algorithmic trading is the prime example of integrating Predictive Analytics in Data Mining with Big Data.

One such system is called Alpha Capture System (ACS) as discussed in (Thomas, 2011). ACS is a collaborative and quantitative research platform for buy-side firms to track, rank and audit as well as to reward sell-side firms for their investment analysis idea inputs and trade recommendations. The ACS provides a single point of contact for the submission of trade ideas which came from semi-structure data format that combine a binary recommendation (i.e. buy or sell) and the textual analysis report. Prior to downstream processing, the data will be cleansed and data-mined and subjected to ideas classification and prioritization based on securities attributes such as firm type and investment dollar amount.

Marshall Wace hedge fund is one of the first fund investment firms to design and develop a homegrown Alpha Capture and Generation System (ACGS) called Trade Optimized Portfolio System (TOPS) as early as 2001 (Thomas, 2011). TOPS ranks trade ideas offered by sell-side researchers through Data Mining and makes trade decisions algorithmically based on Predictive
Analytics. These practices of ACS have garnered endorsement from the Financial Services Authority (FSA) of UK for its traceability, auditability and build-in accountability for both reward and penalty. The decisions from ACS are transparent in terms of system dataflow. In the report by (Financial Services Authority (FSA) UK, 2006), the authors clearly stated that ACS promotes good practices that are policy driven, risk assessed, compliance and alert capable, audit focus and generally considered a transparent system. The FSA regards such system a major step forward that can avoid insider abuse and other criminal offences as compared to the traditional means of communications.

In this section, we highlighted the integration of Big Data and Predictive Analytics in the most practical sense through the illustration of financial engineering. The background of financial engineering exemplified Big Data in which it unifies social media data and expert opinions to derive automated trading decisions. All of which exploit the massive volume, rapid velocity and ranging variety properties of Big Data.

Summary

The practice of Data Mining had undergone a tremendous shift in adoption and application in both the academia setting and business setting. As pointed out by (Piatetsky-Shapiro, 2007), the value of applying Data Mining in many industries, had been realized and cemented in our collective understanding. The meta-analysis approach of this essay is to contrast different studies across many industries done by researchers and practitioners involved in actuary, forecasting, statistical modeling, computer science, machine learning and business intelligence. These are only a subset of a small sample of the overarching reach that Predictive Analytics can extend to.
There are evident that shows the pervasiveness of Predictive Analytics in our society. As a researcher, this essay has illustrated a number of research articles related to the subject matter, as well as discussion in a variety of applications. As a consumer, online recommender systems by companies such as Amazon and Netflix are prime examples of our direct reliance on Predictive Analytics. As a citizen, predictive policing techniques employed by law enforcement agencies directly improve the safety of the people they serve. As a patient, pharmaceutical companies rely on Predictive Analytics to assist drug research and development as well as genetic research, which directly affect us in terms of medical advancement and disease control. The examples are numerous. The effects of Predictive Analytics have made significant impacts on our lives, sometimes surreptitiously, including a lesser known application in bioterrorism surveillance (Berndt, Bhat, Fisher, Hevner, & Studnicki, 2004). The industries of social computing, general business applications, education and pedagogy support, video gaming, law enforcement and financial engineering are just some of the applications discussed in this essay. Evidently, at a minimum, Predictive Analytics can enrich existing decision support models given the discussion in this chapter. The effects of Predictive Analytics go beyond the applications and industries that we have discussed thus far.

To conclude, we discussed the twitter case study and concluded that marrying Big Data with Data Mining to transform machine learning techniques into measurable and actionable predictions for spam classification using cutting-edge infrastructure and open source toolset. Also, we discussed how Predictive Analytics can support educators in helping students early-on in the learning process by predicting at-risk students whom show signs of distress and negative behaviors correlated with course dropout. We discussed video gaming industry, which is, one of the early adopters in machine learning research with the goal to create highly human-like
behaviors of Artificial Intelligence agent design. Predictive Analytics helps game designers to make video games interesting and engaging to appeal to video gamers by evoking human emotions. In areas where Predictive Analytics can be a lifesaving supplement, law enforcement continues to use Predictive Analytics to identify crime hotspots with systems that alert police for violent crimes prevention. Health industry applies Predicative Analytics on biometrics information to help physician to monitor life threatening vital signs.

The aforementioned industries capitalize on the methodologies and methods of Predictive Analytics to extract the telltale signs of the solutions to the problems they are to solve. These signs are often buried deep within their data. The facts are often captured but scattered across multiple domains in a form of distributed dataset, calling for Data Mining methods to reconcile the knowledge embedded within.
CHAPTER IV
METHODOLOGY

The format of this essay took a qualitative approach of meta-analysis to study and survey the current landscape of Predictive Analytics in Data Mining with Big Data. As such, the methodology employed in the creation of this essay is neither observational nor experimental. The development of this essay relied greatly on previously peer-reviewed literatures. They are to allow inferences made from formerly conducted observational or experimental exercises to support the arguments made herein. The experiences derived from various online research outlets such as IEEE, ACM and SpringerLink are the main sources of information in answering the proposed questions stated in CHAPTER I.

The approach to the research subject is done qualitatively with a specific focus on the subject matter based on grounded theory, so the approach taken for this research is bottom-up and inductive in nature, through the ideas and information embodied by the literatures reviewed. As such, in alignment with the grounded theory principles, which are to help us to understand complex problems through a comprehensive, systematic and inductive approach to developing theory. This research does not provide a hypothesis but attempts to generate a theory for the research constructed through existing research results published by other researchers. The research result of this paper also includes a set of taxonomy diagrams (shown in Figure 1, Figure 3 and Figure 6), identifying the terms, methodologies, techniques and methods used in Predictive Analytics as it relates to Data Mining. This is in alignment with the research objective to survey the technological landscape of the subject matter.

The questions stated under the Research Problem section define the limits of the scope of discussion in order to avoid scope creep. While the proposed problems are not definitive, as they
are meant to set a tone for the paper, they are meant to take the first step in reconciling some of the overlaps and misunderstanding between the well-known disciplines and practices. Also, they define the categories of research discussion as a way to delimit the research periphery in the presence of voluminous research papers available on the subject matter. The research process does not only rely on academic papers but also other forms of publications such as technical articles, information technology magazines, conference proceedings, textbooks and other online materials that would help the research effort. However, the focus of source information references and citations come from peer-reviewed academic papers obtainable through reputable outlets such as IEEE and ACM online databases to ascertain information trustworthiness.

The diverse sources of information helped to develop the research constructs and the progressive improvement in attaining construct validity. The multiple information sources provide the basis for triangulation of facts, statistics and events that are relatable and traceable through previously peer-reviewed observational and experimental studies. Since the purpose of the research paper is to survey the information technology landscape, conducting an experimental research is not required. As such, aspects of the research will be quantifiable through online datasets and the result generated from previous research studies.

The primary research tool is a personal computer with Windows 8.1 as the operating system. The personal computer has unrestricted internet access and has sufficient hardware power to support the applications and tools needed to aid in the research. The software and online services for the research includes, but not limited to, the software and services listed on APPENDIX B under the following sections: Productivity Software, Internet Browsers, Open Source Predictive Analytics and Data Mining Tools, Python Related Statistical Libraries,
Preliminary Discussion

The common theme in the topic of Predictive Analytics has been identified as an advanced technique to forecast future outcome in both population domain and individual basis. The majority of the research papers referenced in this essay have utilized Data Mining techniques where Predictive Analytics was mentioned. This indicated some of the techniques in Data Mining are treated as prerequisite to Predictive Analytics. Particularly, in the paper by (Nyce, 2007), stated by the author that, using Data Mining techniques to avoid the “garbage in, garbage out” modeling conundrum is first and foremost to any kind of data analysis including Predictive Analytics.

Some of the research papers indicated the use of Big Data in the context of Data Mining, however Big Data is not prescriptive within the domain of Predictive Analytics in Data Mining. While certain industries can benefit from Big Data such as insurance industry and finance industry in credit scoring (Nyce, 2007) and SCM (Waller & Fawcett, 2013) to optimize business operation and logistics performance, as well as retail business for targeted marketing and advertising. Not all Data Mining and Predictive Analytics applications mandate the use of Big Data. However, Big Data do represent a major enabler in certain applications and industries.

Data Mining and Predictive Analytics are inseparable in this discussion within the context of academia while Big Data is supplementary to the discussion. Big Data adds an additional dimension to the already complex multidimensional field of Data Mining. As previously mentioned in CHAPTER I, employing Big Data means uprooting our collective understanding of the most fundamental database technologies. In that, the commonly employed approach to the
management of Big Data is the non-relational NoSQL model, as well as the distributed and parallel processing in computing model. Regardless, the trade-offs exist in employing Big Data and it can be complementary in most research subjects and fields of study as indicated in CHAPTER II and CHAPTER III.

The benefits Big Data brought to the research community are the simple fact that it provides an additional data dimension that houses the latent variables, which, allows for uncovering correlations that are normally unobserved. For example, the system behind Google Flu Trends (Google Inc, n.d.) has the ability to forecast influenza outbreak faster than U.S. Centers for Disease Control and Prevention (CDC) in previous years (CBC, n.d.). In this particular case, the two distinct approaches employed by two very different organizations (i.e. Google and CDC) show prediction accuracy favoring those that include Big Data (i.e. Google). A recent study by (Preis, Moat, & Stanley) predicted financial market volatility using Google Trends and produced accurate market performance predictions. The underlying premises of Predictive Analytics are correlation and causation of factors, incorporating Big Data helps to broaden our view of research and analysis to account for unseen constructs. Thus, Big Data can enhance the predictive power of Predictive Analytics provided that we have accurate and reliable data through Data Mining and well defined models through statistical modeling.

Another observation made in Predictive Analytics is that it is often temporal dependent. For instance, the paper by (Maciejewski, et al., 2011) linked spatial and temporal views to analysis and forecasting which led to the use of time series analysis and prediction as a means to discover hotspots. The aforementioned examples of Google Trends are also temporal dependent. Time is one factor that exists as an independent variable in almost all research papers encountered thus far. In fact, many research methods such as cross-sectional study and
longitudinal study are inherently time and location dependent. As such, putting data in the context of *time* is an important aspect of ensuring research accuracy and precision.

**Preliminary Analysis of the Topic Data Using Google Trends**

The data collection for this essay is primarily done through literature review. Google Trends was used to uncover some of the interesting phenomena. As a preliminary exercise, one of the approaches is to correlate the point in time with the search terms where there had been an increase in interests of the subject matter. The three search terms (i.e. Data Mining, Predictive Analytics and Big Data) were submitted to Google Trends as shown in Figure 20, Figure 21, Figure 22 and Figure 23, with an emphasis given to Figure 23 where a convergence between the red line (i.e. data mining) and the orange line (i.e. big data) can be observed in year 2013, hinting to a correlation of interests amongst the three terms. The charts also reflected the history of the terms used within academia with the term “Data Mining” leading in search volume since 2005. This is due to the relative early establishment of research interest and adequate understanding in the concepts of Data Mining dating back to 1996. Note that the highlighted (in yellow) dotted line represents the *predicted* trends in volume.

*Figure 20: "Predictive Analytics" Search Term - Google Trends Chart, February 2014*
To quantify the collected data, the first step is data selection and noise reduction, that is, remove unnecessary data to produce a higher signal to noise ratio. In the case of this research, the noise data can be represented by research discussions that are irrelevant to answering the research questions. Regardless, the step-by-step approach in the entire process spectrum of KDD,
from collection to visualization, will be carried out qualitatively and quantitatively using some of the recommended tools and services as detailed in this essay.

Discussion

The study of information systems technology enables the understanding of how human interact with computer systems. In the broadest sense, human provide instructions as inputs to computer systems for execution, the result of the execution are presented in audiovisual responses. This encapsulates the basic form of human-machine communication. Unlike human-to-human communication, human-to-machine communication lacks richness in context. Human communicates with each other in ways beyond the spoken and written language. Human communication includes contextual rich information such as body language (e.g. hand gesture and facial expressions) and entity relationship (e.g. friendship, kinship and acquaintance). Human-machine communication, however, have yet to reach the fullness and richness of information exchange that inter-human communication provides.

Context-Awareness and Big Data

Many researches have been done to expand on the idea of context-aware computing and indicated that the definition of the term “context” changes meanings overtime. The definition of the term “context” changes parallel to the evolution of computing systems. The definition also changes parallel to the increasing reliance human has on machines for data analysis. From the location-centric definition in 1992 to the history of previous interaction in 2008 and the 2012 definition focus on context discovery (Kiseleva, 2013), researchers have continuously supplement computer systems with contextual information in hope to increase machine intelligence as it relates to human cognition and communication. The goal is to increase the overall analytical power of computer information processing by incorporating contextual data
that are dynamic and relevant in any given situation. The evolution of context definition between 1992 and 2012 is shown in Figure 24.

![Figure 24: The evolution of context definition (Kiseleva, 2013)](image)

The following definition of context-aware computing was given by (Dey, 2001):

“Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves.”

The above definition still hold true today in the face of the dynamic nature of what we consider as contextual information. At a high level, contextual information is entity-associated situational information that is relevant to the interaction between user and application. The example of an entity is clear, a person (e.g. user), place (e.g. location) and object (e.g. mobile device). However, what constitute context remains an abstract definition as it is dynamic in nature. It is important to recognize contextual information as variables that exist outside the normative independent variables for a given analysis. For instance, spatial information is not an
absolute piece of contextual information for all problems. Rather, it can be supplementary to problems that are not inherent to spatial analysis. For instance, crime mapping as illustrated in the Law Enforcement section of this essay is clearly a location-bound model and therefore spatial data will not be considered as contextual but is treated as the one of the core independent variables to predict the dependent variable of crime rate.

The author in (Dey, 2001) proposed a high level abstract architecture for context-aware application design called the *Context Toolkit*. It consists of three abstractions of *context widget*, *context interpreter*, and *aggregator component* to provide features of *capture and access of context, storage and distribution* and *independent execution* for context-aware applications. A more recent paper published by (Baldauf, Dustdar, & Rosenberg, 2007) described the layered conceptual framework for context-aware design that resembles the Context Toolkit by (Dey, 2001). The layered conceptual framework is as shown in Figure 25.

![Layered conceptual framework for context-aware systems](image)

*Figure 25: Layered conceptual framework for context-aware systems (Baldauf, Dustdar, & Rosenberg, 2007)*
In the world of internet data, the novel and simple way of *tagging* is one of the many ways to provide context to data, which is, a simple way to annotate data with metadata that can help to classify information.

For instance, tagging on internet articles adds significantly richer and broader information using keywords such as semantics keywords. The approach of tagging often produces a richer semantic context to the text than techniques relies only on words extraction from text. This is because the act of words extraction is done *after the fact*, which is to say, automated algorithms are being put in a position to make inferences based on supervised learning technique in order to make sense out of the text. A common technique in this case is a decision tree which enforces conformity to a rigid, restrictive and hierarchical categorization format. On the other hand, a tagging exercise is a *human activity*, which is to say, the contextual awareness and contextual information assigned by the person who created the article, and that can be directly translated into metadata. To that end, metadata information adds a new dimension to data that lack context.

Folksonomy is one way to produce collaborative *tags* for information classification and categorization. Tags add contextual information to content data to support a greater level of human-application communication. The simple means of folksonomy is a great way to allow for a more accurate information retrieval and response system.

We have established that context is important to information processing within the realm of information systems. Big Data embodies the ideas and practices of context-aware computing where the information in silos are not expressive as correlated information in masses. Big Data extents the notion of what context consists of. Reasonably, for computer systems to predict an imminent event with high accuracy requires high resolution dataset and context-aware algorithms. Reasoning and arguing within interpersonal communication requires facts. Ideally,
only those facts that are relevant to the discussion at hand. Same is true for computer systems to reason and argue with data effectively. We are to analyze data that is pertinent and applicable only to answer the problem in question. Since relevance, pertinence and applicability in facts/data are subjective measures, Big Data is a means by which we can embody contextual information. To reasonably derive relevant information associated with analytical directives that require rich context to be effective, Big Data does provide context for computer systems as the foundation for context-aware analysis. As such, Data Mining and Predictive Analytics both operate within the realm of Big Data that can take information processing to the next level. This is because of Big Data, for its ability to be conceptually closer to the interpersonal form of communication.

To bridge the gap between Predictive Analytics and Data Mining, the paper by (Kiseleva, 2013) clearly depicted the feedback loop dependency between Predictive Analytics and Data Mining. In that, the dependency exists in context mining, context modeling and context integration processes.

Recommendation system is one of the examples in the direct application of context mining. Many online retailers use recommendation system to make merchandise suggestions to online shoppers. Context in this case takes many forms and often context is hidden. For example, an online travel agent company sends marketing emails to potential customers. The company would target those customers with the aligned purchase intent given the temporal context and hidden context. The temporal context could be an upcoming public holiday which suggests the travel availability of the potential customers. The hidden context is the motivation behind the purchase intent which could be the workplace stress experienced by the customers. Mining both
the temporal context and hidden context produce the purchase intent of potential customers whom might be interested in a leisure trip.

To that end, an architectural description to context mining is depicted in Figure 27 where the variable set of $C$ represents the contextual categories, variable $L$ represents a set of individual learning procedures with the key component of contextual feature set $F$, as well as the two mapping functions of $G$ and $H$. In short, once the contextual features have been identified, the function of $C_i = G(F_i)$ can map features to contextual categories and the function of $L_i = H(C_i)$ can map contextual categories to one or more individual learners. This is a high level framework of what a context-aware application architecture would consist of. This framework was designed to anticipate the dynamic nature of contextual information.

The design of a context-aware application architecture (Figure 27) shows resemblance of the Meta-Learning Architecture (Figure 26) described in the paper by (Singh & Rao, 2013) as well as the context managing framework architecture (Figure 28) by (Baldauf, Dustdar, & Rosenberg, 2007). The design similarities between the abovementioned architectures further prove the validity of a fundamental context-aware system design. The context-aware design architecture also fall under the category of ensemble machine learner or bagging with the individual learners trained on different datasets with varying levels of techniques and data biases.
Figure 26: Meta-Learning Architecture (Singh & Rao, 2013)

Figure 27: An example of context-aware system design (Kiseleva, 2013)
Context discovery and context integration as described by (Kiseleva, 2013) is necessary to build context-aware predictive models so as to create context-aware application. While temporal and spatial information are pertinent to many situations, it is important to clarify that temporal and spatial based variables are not classified as contextual information. For instance, time series analysis perceives time as one of the independent variables and crime mapping perceive location as a key variable to predict crime. However, temporal and spatial data in the aforementioned examples do not constitute as context element and should be treated as first-class variables during modeling.

**Basic Statistical Methods and Techniques**

Many statistical methods and techniques involve searching and measuring the central tendency of a population as a reference point for other derivative techniques. The central tendency can take the forms of mean, median and mode to represent arithmetic average, middle value(s) and highest frequency values, respectively. Most commonly, mean measure is often used. The central tendency is always required to visualize certain characteristics of a population at a high level. For instance, the measure of standard deviation (represented in the Greek
alphabet sigma $\sigma$) in a demographic distribution can display a degree of population skewness in reference to the Gaussian distribution (i.e. normal distribution), which is based on the Central Limit Theorem. A Gaussian distribution is a distribution where the mean, median and mode measures are all equal to one another. It can be visualized as a symmetrical curve called the Bell Curve as shown in Figure 29 where the population mean (represented in the Greek alphabet $\mu$) equals to 200.

![Number of People Per Age Group](image)

*Figure 29: An example of a normal distribution "Bell Curve"

The measure of standard deviation and variance ($\sigma^2$) provide a mathematical way to depict the degree of deviation of a particular sample population using the sample mean ($\bar{x}$) from a normal distribution. To calculate $\sigma^2$ of a given sample population, the calculation is done based on the square of the sum of the differences between each data point and the mean, then divide the result by the number of data points. The mathematical formula is shown in Figure 30.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

*Figure 30: Variance formula*
Once $\sigma^2$ is calculated, to calculate $\sigma$ is simply the square root of $\sigma^2$ which are shown in Figure 31 and Figure 32.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

*Figure 31: Standard Deviation formula 1*

$$\sigma = \sqrt{\sigma^2}$$

*Figure 32: Standard Deviation formula 2*

The example shown in Figure 29 is an example of *univariate analysis* where there is only one independent variable called “age group”. On the other hand, *multivariate analysis* allows the use of two or more independent variables. A *bivariate analysis* deals with analysis of two variables and measuring the correlation in bivariate analysis is essentially the cornerstone of all Predictive Analytics methods.

Linear correlation coefficient (represented in $r$) is a mathematical way to discover hidden relationship and detect any correlation between two variables. $r$ is calculated by the sum of each product of the differences between two variables and the respective means, divided by the product of the standard deviation of the two variables and multiple the result by the number of data points. The formula is shown in Figure 33.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})^2}{n\sigma_x \sigma_y}$$

*Figure 33: Correlation Coefficient formula*

The value of $r$ exists between the range of -1 to 1 where 0 represents absolute non-correlation, positive number refers to positively correlated and negative number refers to negatively correlated. The direction (i.e. positive or negative) of the correlation refers to the directional movement of the two variables. If variable $x$ increases while $y$ decreases, there is a
negative correlation exists between $x$ and $y$. Similarly, if both $x$ and $y$ increases or decreases create a corresponding change against each other, then it is said that there is a positive correlation between $x$ and $y$. Therefore, the result of $-1 < r < 0$ represents the degree of negative correlation and $0 < r < 1$ represents the degree of positive correlation of the two measured variables.

There are many versions of the $r$ mathematical formula but the underlying promise is the same. It is used to compute and output a representative numerical value that corresponds to the level of correlation between two variables. Many statistical modeling techniques rely on this concept. A simple linear regression modeling based on least squared method is one such example.

All of the predictive statistics used in statistics is inferential statistics for modeling data that are high in randomness and uncertainty, a measure that accounts for entropy which is inherent to the data being modeled. A common way in scientific and academic research for statistical hypothesis test is to confirm hypothesis based on experimental analysis involving statistical methods. Covariance, correlation and the measure of standard deviation are methods that generate information apposite to be inferred to a hypothesis or established theory. All models are subjected to model validation to ensure high predictability and inference-making ability. A general approach to model validation is to train model with a randomized subset of the sample data and use the remainder sample data to validate. Splitting sample data for both purposes in model training and validation allows researchers to maximize the value of the available data. This dual-purpose practice can take the forms of holdout and subsampling, cross-validation and bootstrap methods.
Data Mining Methods and Techniques in Predictive Analytics

Under the Predictive Analytics section of CHAPTER II, we briefly discussed the various methodologies in Data Mining and where do Predictive Analytics fit in within the Data Mining archetype. The strong focus on data visualization methods in Predictive Analytics is evidently reflected on the popularity of scatterplot charts across many industries. The commonplace scatterplot chart is a visualization tool for linear and non-linear regression-type analysis. Classification is also another highly regarded methodology in Predictive Analytics for its ability to distinguish data classes by labelling them for future inferences. This process is often called concept classification. Both regression analysis and classification analysis are prevalent methodologies in machine learning. Many techniques derived from them such as neural network based methods (Mathewos, Carvalho, & Ham, 2011), clustering methods (Zeng & Huang, 2011) and dimensionality reduction methods (Vlachos, Domeniconi, Gunopulos, Kollios, & Koudas, 2002). For example, the simple decision tree model consists of both classification and regression variants called classification tree and regression tree, respectively. Regardless of the chosen methodology and method, the historical context in Data Mining is crucial to the development of any predictive model. Mining historical information of known facts allows a method to find similar and probable outcome for the unknown. This concept underpins the fields of Data Mining and Predictive Analytics as well as all of their derived methodologies and methods.

Classification

To understand how useful classification methodology is in Predictive Analytics, one must presume inference is a requisite of prediction and therefore any technique that is grounded in logical inference can be used for Predictive Analytics. Classification at its core is an inference engine because it follows a logical order of operations based on learnt rules to segment data into
common constituents. Classification models that use training data to generate model rules are a type of *supervised learning*. Most classification techniques are supervised learning based and can be modelled based on techniques such as decision tree classification, naïve Bayesian classification, Bayesian belief networks classification, support vector machines (e.g. kernel approximation) classification, nearest neighbors classification, stochastic gradient descent (SGD) classification, multiclass and multilabel classification.

Many of the aforementioned classification techniques are also used in solving regression problems, they are models that can be applied to both classification and regression problems. The major difference between classification and regression is that classification deals with *discrete* and *categorical* dataset while regression deals with *continuous* and *numerical* dataset.

Bayesian belief networks is a classification technique that can be used to recover information about the *structure of statistical dependencies* among variables, quantifying probabilistic conditions such as $P(Y|X)$. The probabilistic conditions can be represented in a graph network of nodes, marking the connections between events and belief variables that are either *conditionally dependent* or *structurally independent*. The result can be observed and quantified as shown in Figure 34 and illustrated below:

- Given $X \rightarrow Z \rightarrow Y$, variable $Y$ is conditionally dependent of $X$ through $Z$ such that $P(Y|Z,X)$.

- Given $X \rightarrow Y, X \rightarrow Z$, variable $Y$ is conditionally dependent only on $X$ and structurally independent of $Z$ such that $P(Y|X)$.

In Bayesian belief network, the probabilistic based conditional rules are captured within a *conditional probability table* (CPT) for each variable that presents a given $P(Y|X)$ condition. Given that, a conjoin condition can be calculated mathematically through the multiplication of
\[ P(Y|X) \text{ conditions to arrive at a joint probability distribution under the following Bayesian inference formula, for the condition } X \rightarrow Y: P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \] where \( P(X) \) is the product of the conditional dependence variable set (including the parents of \( P(X) \)) that \( Y \) dependents on.

Support Vector Machine (SVM) is also a frequently used classification technique. SVM deals with both the linear and non-linear classification problems. SVM concerns with delineating data points with the use of hyperplanes while maximizing the margins (i.e. the distance between the two hyperplanes). To that end, SVM shares the same principle of clustering method which will be discussed in the Clustering section.

**Regression**

Regression analysis shares many predictive properties of classification which is ideal for Predictive Analytics for its ability to normalize variables through function approximation.
Visually speaking, regression is a way to describe the dots-and-lines spatial relationships where the dots represent the data points and the lines represent the relationships amongst the data points. An example of visualized linear regression method is shown in Figure 35. In short, regression analysis allows linear and non-linear fitting of scattered independent variables through mathematical techniques. The result of regression analysis is often visualized via scatterplot chart. Regression methods such as non-linear regression, logistic regression, support vector machines regression, stochastic gradient descent regression, Gaussian processes regression, decision tree regression and isotonic regression can provide forecasting capability visually and mathematically.

![Figure 35: A visualized example of linear regression (Natural Resources Canada, 2012)](cm_above_chart_diam.png)

In the world of Predictive Analytics, logistic regression is a form of regression analysis that is a highly regarded method for prediction due to the probability based sigmoid function of
\[ P = \frac{1}{1 + e^{-\beta_k}} \] where \( P \) is the calculated probability and the \( \beta \) is the matrix vector of \( \beta_k = a + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_kx_k + e \) which comprises a linear equation and error (\( e \)) parts. The \( P \) produces a logistic “S” shape curve visually between the value of 0 and 1. Rather than producing a continuous numerical dependent variable as linear regression does with ordinary least squares method, logistic regression produces a binary response that is best to describe any dichotomous phenomenon and thus it is suitable for classification problems.

**Clustering**

In most cases, the classification and regression methodologies involve supervised learning methods. Supervised learning refers to model building with pre-labeled training dataset where manual intervention of data classification was performed on the training dataset, which is to say, supervised learning models learn from examples prepared by human.

Clustering is an unsupervised learning methodology without the need for human intervention for training dataset, which means that, clustering methods operate directly on live data. The unsupervised learning nature of clustering is both advantageous and limiting. Unsupervised learning techniques generally produce inferior predictive power than supervised learning techniques. However, unsupervised learning such as clustering benefits from autonomous operation, relying entirely in intrinsic data properties such as central tendency and density to filter and assemble data based on the relative relationships (e.g. Euclidean distance). Very often, clustering serves as a great first step for data analysis. Clustering methods such as k-Means clustering, Gaussian mixtures, hierarchical clustering, spectral clustering, mean-shift clustering, DBSCAN clustering and affinity propagation clustering, all depend on the pre-existing relationships between data points.
K-Means clustering is one of the most common approaches in clustering. K-means clustering algorithm expects an input of $k$ where $k$ is the expected number of cluster groups which the algorithm would produce. A group of data sharing the nearest mean value is algorithmically determined to form a single cluster relative to the expected number of clusters. The k-means clustering method would continue to form clusters and rebalance previously formed clusters until $k$ number of cluster groups have been created. This iterative process is shown in Figure 36. This clustering technique has strong dependence on the statistical techniques discussed in the Basic Statistical Methods and Techniques section.

![Figure 36: Clustering of a set of objects using the k-means method; for (b) update cluster centers and reassign objects accordingly (the mean of each cluster is marked by a C) (Han, Kamber, & Pei, Data Mining: Concepts and Techniques, Third Edition, 2011)](image)

**Artificial Neural Network**

Artificial neural network mirrors the biological design of a human brain in which neuron is emulated as perceptron containing a function called *activation function* or *transfer function*. The functions are encapsulated within a multi-layer topology of interconnected perceptrons consisting of *input layer, hidden layer* and *output layer*. This simple definition allows for artificial neural network to spread across the entire range of machine learning discipline.

Artificial neural network can be used in areas that require non-linear prediction, prediction for phenomenon with covariate relationships, data classification, feature extraction.
(e.g. PCA and Factor analysis), data compression and general application in image processing (e.g. de-noising and recognition). Consider the most basic form of artificial neural network in non-linear machine learning called the backpropagation neural network. The backpropagation neural network is based on the multilayer feed-forward neural network topology (Figure 37) with the supplemental ability to back propagate across sub-layers within the hidden layer, as a means to update previously learnt weights and bias. Backpropagation neural network is a supervised learning method which means that training dataset with targeted labels are provided to train the perceptrons’ connection (i.e. edge) weights and unit (i.e. node) bias. Since the neural network topology allows for many-to-many connections between perceptrons, it structurally defines each perceptron from one layer passing output value to the perceptron in the next layer as an input value in order to form a complete network. Each perceptron makes a decision based on output from the previous perceptron and adds their connection weight value and bias value to derive a binary response of 1 or 0. A collection of these perceptrons forms a network of nodes that produce a collection of binary output values similar to the way human brain adapts and learns new knowledge based on synapses that occurs between axon and cortical neuron.
One way to implement linear perceptron is to use *binary threshold neuron* formula as shown in Figure 38 to produce a binary response (i.e. 0 or 1, true or false) for all outputs of other input neurons based on *weighted sum*, where $w_i$ is the weight per neuron connection, $x_i$ is the input value, $b$ is the bias value and $y$ is the neuron output based on the calculated weighted sum of $z$. The value of $y$ can be used to determine if an output will be directed to the next perceptron for a given calculation. For instance, if $y = 1$ then perceptron will send output to the next perceptron; if $y = 0$ then perceptron halts output. A collection of these small decisions thread together to form a neural network with zones (i.e. areas of selected perceptrons). The zones within a neural network represents pathways based on learnt data which would directly influence...
future outputs based on similar inputs (i.e. training dataset). This representation and mechanism work much like the synapses of a human brain. The synapses become stronger overtime when a particular part of memory has been exercised more. Other implementations include sigmoid neuron and stochastic binary neuron and both implement logistic regression for binary output.

\[ z = b + \sum_{i=1}^{e} x_i w_i \text{ and } y = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases} \]

*Figure 38: Binary Threshold Neuron formula*

Backpropagation artificial neural network is an inverse function of a typical feed-forward artificial neural network model for supervised learning purpose. To train a backpropagation model, the error derivative is derived from the delta of the predicted value and the actual (i.e. labelled) value which is used to update the model. This process occurs continuously and iteratively.

Researchers have made many improvements to artificial neural network in hope of enhancing the model to better solve real world problems. One such method is to hybridize with fuzzy logic to create neural fuzzy network system (Nirkhi, 2010). Contrary to the standard sigmoid function that produces only binary response, the neural fuzzy network system produces a degree of measure based on fuzzy rules and fuzzy sets.

**Conclusion**

In concluding this section, we have discussed the basic premise of Predictive Analytics in which Predictive Analytics, as part of Data Mining, is realized by the algorithms that implement the machine learning models. The machine learning models are underpinned by the century old disciplines of statistics and mathematics. The importance of statistical analysis and the role it plays in machine learning are also discussed.
The learn-to-predict paradigm of Data Mining dominated most of practical applications in the form of supervised learning techniques and methods, the algorithms that are derived from these machine learning methods has the capacity to infer based on data. Inferences made by these algorithms are logical and evidence based and the result is justified empirically given the deduced, features extracted, dimensionally reduced and correlated dataset. This lead to the development of the overarching methodologies of classification and regression which provide data labeling and concepts grouping.

Conceptually, regression methods such as the simplest form of regression, ordinary least regression, is identical to the classification methods such as decision tree, in which, the data are normalized and categorically divided by a single label (e.g. classification) or a line (e.g. regression). The label or line abstracts away the complex relationships between the independent variables as well as the conditional expressions and mathematical calculations that best describe the data being analyzed.

It is when we transpose the inference into prediction based on learnt data, that we make the leap from descriptive and explanatory analysis to prescriptive forms of analysis. Another way to look at Predictive Analytics through the lens of statistics is the transformation of independent variable into predictor variable. Finding the most fitting independent variables to infer to a dependent variable, and through which to derive a predicted outcome. This is the reason that the modeling techniques are similar between retrospective and prospective ends of the analytics spectrum.

The consensus amongst researchers at the current state of Predictive Analytics is that the differentiator in predictive performance came from the quantity and quality of the data during supervised learning rather than favoring the algorithms and models themselves. This is where
Big Data becomes a significant interest in future research. The pinnacle of predictive modeling presently arrived at the *Ensemble* modeling methods which will be discussed in the next section.

We conclude this section with the understanding that Predictive Analytics problem is a statistical problem and that statistical problem can be solved by many of the aforementioned techniques.
CHAPTER V
ISSUES, CHALLENGES, AND TRENDS

Introduction

It is evident that the techniques and methods described in CHAPTER IV are an integral part of Predictive Analytics. The design of models and advanced techniques are important to improving predictive accuracy. However, predictive modeling will inevitably be optimized to reach a pinnacle where the models and methods themselves will no longer provide a measurable lift against competing models, resulting in a predictive performance plateau. In such cases, no gain in predictive power can be achieved simply by improving the model itself or by applying other advanced techniques. The tipping point for better predictive performance tends to give way to the data that feed the models. In which case, the characteristics of Big Data provide an abundant source of training data for predictive modeling.

The quality of the data certainly plays a critical role in the predictive power of a given model. Data quality is multidimensional where data recency, data volume, data variety and data veracity are the basis to provide a timely, abundant, unbiased and accurate training dataset for any given model. If the training dataset is of poor quality, no predictive model can excel and the result could even be inferior than a random guess. If the data are bias, in a sense that they represent only a minority of a subset of the common view, no predictive model could predict any result beyond the boundaries set by the data that inherently limited the model. This is the classic “garbage in garbage out” saying or the “the output is only as good as the input” principle of any computing systems. Human and computing systems are bound by the same principles, predicting acceptably without bias and flaws equate to learning without knowing, a paradoxical
conundrum that no human or any computing system can solve because we can only process within certain confines that are restrained by the presently available and accessible information.

Uncommon events are rare events that are difficult to predict. Through the discussion thus far, we underscored the difficulty to predict without inference and to infer without evidence. The learning models reinforce our collective reliance on data as a single means to provide evidence for inference that leads to prediction. Committing a type I error (i.e. false positive) on predictions might not be an issue if the error results in only minor inconvenience. For instance, incorrectly forecasted meteorological conditions such as rainfall in an urban setting might not be impactful to the lives of the individuals living in the area. However, generating a false negative on rare but catastrophic event, thereby committing a type II error, could have disastrous consequences such as major natural disaster (e.g. earthquake), terrorist attack and major economic downfall. Balancing between a tolerable level of false positives and discovering momentous false negatives marks the major obstacle that statistician and mathematician have been battling with since.

Performing statistical modelling will inevitably lead to the conundrum of goodness of fit. An important measure used by researchers to assess if the developed model is describing the observations (i.e. data) between the dependent variable and the independent variable adequately. This is done by measuring what we expect to see (i.e. hypothesis) versus what we actually saw (i.e. observations). Some of the methods for goodness of fit measure include Pearson's chi-squared test and p-value method.

When a model is unable to sufficiently describe the data, it is said the model is underfitting the data. For instance, using regression analysis as an example, describing a dependent variable with a binary response using linear regression is undesirable and would
grossly underfit the data due to high residuals (Lesson 19: Logistic Regression, 2014). Using Figure 39 as an example, suppose there is a significant income disparity between female and male where the income of male is higher than female. This would result in a binary response dependent variable (i.e. gender) for any given income. A linear regression in this case is not a descriptive method and would result in a straight line penetrating two clusters of data points as shown in Figure 39. A good model in this case would be the logistic regression model as discussed in the Regression section.

Gender (1 = Male and 0 = Female)

Figure 39: An example of binary response dependent variable (x=income, y=gender)
On the contradictory, in the predictive realm of statistics where a model can overfit the data rather than underfit which can cause diminishing predictive power of a model for unseen events. Reasonably, fitting the model fully and perfectly is great for explanatory and descriptive work. If researchers can impeccably identify the variables for a given outcome, fitting the model exactly to the data is desirable. However, the opposite is true for predictive modeling. Predictive model should not fully and perfectly fit the data. This is because an overfitted model is a model that lacks agility and therefore possesses minimal predictive ability. In this case, the margin of error in predictive model represents the degree of flexibility in dealing with unseen events.

To that end, a functional predictive model should strike a balance between overfitting and underfitting of the training data because certain level of rigorous and nimbleness must be present for prediction. While the majority of the model designs are to make parallel to the training data, it is also important to consider outliers and odd cases, a built-in lenience for exceptions is needed to build capacity for prediction.

**Big Data Issues, Challenges and Trends in Predictive Analytics**

Using the Google Flu Trends example illustrated in the Preliminary Discussion section of CHAPTER IV. The example illustrated the achievement of a higher level of predictability when Big Data was introduced into the equation. However, a recent study by (Lazer, Kennedy, King, & Vespignani, 2014) suggested the underlying Google Flu Trends prediction mechanism led to an overestimation of the prevalence of flu between year 2011 and 2013. The study is a reminder for us of how Big Data might not actually deliver what it has promised.

Realizing the association between prediction and statistics and deriving predictions means solving classification problem, the Big Data paradigm shift within the domain of Predictive Analytics becomes an apparent and significant factor. This is because, the Big Data
characteristics provide an abundant sources of already classified data that are available and accessible on the internet. Supervised machine learning with Big Data reaps the benefits from the attributes and characteristics of Big Data, that is, the volume, velocity and variety aspects. The volume aspect of Big Data represents size, scale, dimension and the amount of data collected which is an important enabler for prediction. Prediction performs intrinsically better with more data, which is to say, the high data volume corresponds to more granular class labels for classification or high data points for regression function approximation.

In the broadest sense, more data is always better for prediction since data adds context to the inductive nature of all supervised training methods. Furthermore, more data adds extra dimensionality to the set of independent variables which allows for high dimensional data analysis to account for multi-factors and latent variables. While additional independent variables do not naturally become high value predictor variables, the probability of discovering predictor variables increases as more independent variables are available through Big Data.

Machine learning methods such as backpropagation artificial neural network requires vast amount of training dataset to be effective due to the recursive process of the update of weight values throughout the network topology. The argument for how the high volume aspect of Big Data can benefit Predictive Analytics is identical to that of scientific experiment. Scientific experiment requires sufficient number of observations in sample data to be representable in a population. Having insufficient observations likely decrease external validity which impair inference-made outside of the sample, that is to say, the model is unable to sufficiently make prediction beyond its sample data. Small data makes the case for biased data; big data increases the chance for an evenly distributed randomized sample and reduces the risks for data bias. Big Data can avoid the overrepresentation and underrepresentation of any given dataset due to the
fact that, the sample size is approaching the population size. Therefore, Predictive Analytics within the context of Big Data, at the very least, can benefit from the unprecedented high volume of training data. This is because, more is not just more, *more is different* (Anderson, 1972).

Consider the high velocity aspect of Big Data, prediction demands real-time or near real-time data to account for rapid shift in data context. Prediction model becomes uninformed when operating on antiquated data that lacks data *recency*. The future will always remain a degree of uncertainty notwithstanding the advanced analytics techniques we have discussed this far. However, uncaptured events and uncollected data are missed opportunities. The real-time or near real-time data feeds enable the prediction model to sift through data *into the future* relative to model that relies on hours or days old data. The benefits are two-folds. Firstly, model operating on timely dataset avoids making prediction on events previously occurred, thus maximizing computing resources on the most relevant context. Secondly, achieving real-time data processing means processing against the current rather than the past, an important distinction for applications such as algorithmic trading.

Finally, the variety aspect of Big Data fuels the arguments made on the volume aspect of Big Data in which wide-ranging data types and data sources add data volume to the model training data. Also, the data type and data source variety in Big Data, in and of itself, are metadata that can be contextualized for Predictive Analytics. This is because significant differences exist between different data types as well as the data sources which are contextual self-descriptive, to a certain degree. For instance, consider the law of evidence, significant differences exist between a witness written deposition and a video evidence that captured an event of a crime. Admissible video evidence would prove to be a powerful piece of evident in a
court case. This is because video evidence generally provides a richer context that is rarely available in written deposition based on malleable human memory. In that sense, the metadata for document and video data type, is the fact that video would carry a higher degree of evidential weight value than a written deposition in the case above. Furthermore, news information that came from a personal website generally carries less weight of trustworthiness than a reputable news outlet such as the Canadian Broadcasting Corporation (CBC). Therefore, the variety aspect of Big Data not only provides a diverse data types and sources for data, but also provides the needed context that is often overlooked during predictive modeling.

**Trend in Big Data Application**

Processing textural data and visual data require vastly different techniques to perform. Processing a written document would require text analytics (Kimbrough, Chou, Chen, & Lin, 2013) such as natural language processing techniques that involves deep linguistic processing methods. However, processing images of objects in photo requires object recognition techniques that involve edge detection methods while processing video data requires motion detection techniques. All of which are some forms of features extraction techniques to address very different problem domains that very often involve data dimension reduction techniques for information processing. The diversity in data types unquestionably created challenges for researchers, on one hand, more is different; on the other hand, more is complex and difficult.

The assorted data processing techniques mentioned in previous chapter existed prior to the advent of the Big Data era. However, the existence of Big Data accelerated the growth in the application of the aforementioned techniques which are heavily used in predictive modeling. As in the study by (Jarrold, et al., 2010), using data mining, machine learning and text analytics techniques to predict future patient brain health in the areas of: cognitive impairment, depression
and pre-symptomatic Alzheimer’s disease. The authors of the study applied multiple techniques on patient language expressions, one of which was supported by the Linguistic Inquiry Word Count (LIWC) software (James W. Pennebaker; Roger J. Booth; Martha E. Francis, 2014) by performing text analytics on transcribed audio interview data. The authors concluded the study with the suggestion to further exploit web intelligence (i.e. Big Data) to further research the application of text analytics in the disease prevention domain. The potential was highlighted by the success prediction rate between 73% and 97% in the study.

In another study by (Schwegmann, Matzner, & Janiesch, 2013), the authors facilitated predictive event-driven process analytics under the Complex Event Processing (CEP) technology by exploiting event log data to improve business functions, an architectural approach to integrate Predictive Analytics with CEP. CEP is a broad term where event data is described as “anything that happens, or is contemplated as happening” (Llaves & Maué, 2011). CEP thus refers to the processing aspect of the ubiquitous event data. The nature of event data ties directly to the discussion of Big Data.

Log event data are often overlooked and underutilized, however, data embedded in machine generated log files are often context rich that can help other aspects of a given operation. To combine Predictive Analytics with CEP is to use Predictive Analytics with Big Data to improve processes, especially the velocity aspect of Big Data.
As shown in Figure 40, the value of an event depreciates over time on the application of Predictive Analytics. Ideally, proactive actions can be administered on an impending event. Beyond that point, it would become a reactive action that is still high in practical value but it is limited within a sub-second period; a period of time immediately follows an event. To that end, the authors in (Fülöp, et al., 2012) proposed a conceptual framework for combining CEP and Predictive Analytics as depicted in Figure 41. In that, the framework consists of predictive event processing agents within a predictive event processing network to synergize both aspects of Predictive Analytics and real-time processing of event data.
Beyond the realm of complex event processing, the multi-purpose application of Predictive Analytics was also researched within the context aviation surveillance Big Data. Specifically, the Aircraft Situation Display to Industry (ASDI) data generated over the years ranging from flight arrival information to oceanic report, from plane maintenance history to flight route information. The study was discussed in (Ayhan, Pesce, Comitz, Gerberick, & Bliesner, 2012). The authors applied Predictive Analytics in the aforementioned massive surveillance Big Data accumulating at rate of approximately 1GB per day. The collected information combined with the developed system illustrated by the authors, can answer practical concerns such as flight route optimization. The author concluded the paper with plan to include meteorological conditions information to further enhance the underlying model supported by SPSS Modeler software.
**Predictive Analytics Issues, Challenges and Trends**

A study by (Niculescu-Mizil & Caruana, 2005), the authors quantitatively identified 10 common supervised machine learning methods, that are, boosted tree, support vector machine, boosted stumps, naïve bayes, artificial neural network, bagged tree, random forests, logistic regression, decision tree and memory-based learning. Each method possesses a degree of bias within its structured definition of the methods, resulting in a distortion in prediction. The authors proposed two methods (i.e. platt calibration and isotonic regression) for supervised machine learning *calibration* in an attempt to reduce the distortion produced by the learning methods.

The result suggested that the calibration methods do in fact produced a measurable improvement in the learning methods (e.g. SVM, boosted stumps) while bagged trees and artificial neural network did not produce any observable improvement. The best learning methods in terms of minimizing errors are: boosted tree with platt scaling calibration, random forest with platt scaling calibration, non-calibrated bagged tree and non-calibrated artificial neural network. The study benchmarked the 10 most common methods and the result do suggest a plateau in supervised learning model performance, particularly, both bagged tree and artificial neural network performed *worse* after calibration. Although not mentioned in the study, a few of these models fall under the category of *ensemble methods* which further divided into two subcategories: bootstrap aggregating (e.g. random forests) and boosting (e.g. boosted tree) methods.

**The Ensemble Approach**

To appreciate the ensemble approach is to embrace diversity in aggregate where collective intelligence can be achieved. Ensemble based supervised machine learning methods are gaining acceptance and exemplified by the one million dollars Netflix Prize competition. The
goal of the event was to predict movie rating of Netflix users, a form of collaborative filtering problem. The top performing team in the event had all employed ensemble based models (Siegel, 2013). An interesting series of events occurred during the competition where competing teams began forming alliances amongst each other to compete with other rival alliance teams. Drawing on the expertise of the models created by other teams to create an ensemble model with meta-learning capability which allows automated model selection based on the strengths and weaknesses existed within the individual models.

The event concluded with BellKor’s Pragmatic Chaos ensemble team as the winner of the Netflix Prize event (Koren, 2009) which was an alliance team formed amongst BellKor, BigChaos and PragmaticTheory teams. The ensemble model developed by the team produced a 10% improvement in movie rating prediction over the Netflix’s internal established method. The winning model employs a Gradient Boosted Decision Trees (GBDT) method as discussed in the related literature by (Ye, Chow, Chen, & Zheng, 2009). GBDT is a boosting based ensemble decision tree method consisting of multiple decision trees.

Ensemble methods are gaining momentum, it is because the underpinning meta-learning idea of combining weak leaners to make a strong leaner means existing model performance can be improved by aggregation. For instance, using AdaBoost (Freund & Schapire, 1999) as an example, the algorithm maintains a set of weight values over a set of base learners and the weight values were adjusted over time through supervised learning. The AdaBoost operation not only smooth out the biases exhibited within the constituent leaners, but also matches the model with any given domain problem (i.e. predictor) that has the best empirical evident (i.e. in terms of previous predictions) as the best performer. Therefore, boosting based method such as AdaBoost, can learn from the result of the base leaners to play to their strengths in an assembly
of leaners. The strength of ensemble based method is evidently established by the empirical evident in (Ye, Chow, Chen, & Zheng, 2009) and in (Siegel, 2013).

**The Concept Drift**

Another advantage to boosting method is the dynamic nature of the technique itself which help to alleviate the challenges posed by concept drift (Venkatesan, Krishnan, & Panchanathan, 2010). Concept drift defines the phenomena involving changes in the predictive nature of the independent variables used in the underlying data and model. Concept drift describes data that exhibits a shift in variable relationship in concept, which is different from the training data used during the model’s supervised learning process. In other words, the training data used to train a model no longer representing the current data being processed by the model.

This is understandable, since the training data is always a subset of the population data and therefore poses a chance for concept drift. As well, the population data is dynamic in nature and not static. Consequently, a model that was trained with sample data must relearn at some point to be in alignment with the new sample data. Many techniques have been developed by researchers to detect concept drift as discussed in (Masud, Gao, Khan, Han, & Thuraisingham, 2009). AdaBoost in this case, to a certain degree, possesses inherited properties to guard against concept drift due to the dynamic weight value which represents past performance of a given learner. In this case, if a base leaner exhibits a sudden drop in predictive performance relative to other base leaners, a concept drift might have occurred and thus requires model rebalancing and possible model retrain.

**Trends and Advancements**

Beyond the consortium, pooled or ensemble methods, other advances in academia and scientific communities are thriving in the machine learning domain. Particularly, the *Python*
based open source communities (e.g. SciPy and Scikit). A list of Python based libraries were identified in the Python Related Statistical Libraries section under APPEDIX B. Notably, the scikit-learn python machine learning library which offers a comprehensive collection of ready-to-use machine learning algorithms as identified in the Data Mining Methods and Techniques in Predictive Analytics section of CHAPTER II.

On the image processing side, scikit-image python image processing library was created to support specific object and facial recognition applications. Other non-Python based library includes the java based Apache Mahout (What is Apache Mahout?, 2014) and the C++ based Dlib library (Machine Learning, 2014). Also, the R platform (The R Project for Statistical Computing, 2014) and the STATISTICA products (StatSoft, 2014) by StatSoft have a long history in academia and commercial space. They are used by many researchers and data scientists for statistical modeling and analysis work. Other open source software supporting Predictive Analytics application was identified in the Open Source Predictive Analytics and Data Mining Tools section in APPEDIX B.

On the Big Data side, Apache Hadoop is the most recognizable and dominant opensource solution to the Big Data problem as discussed in The Apache Hadoop Platform section of CHAPTER II. This leads to many commercial adaptations of Apache Hadoop platform, some of the notable Hadoop distributions are: Cloudera, Cloudspace, EMC Greenplum, Hortonworks, IBM BigInsights Enterprise Edition and Think Big Analytics. In the non-Hadoop proprietary commercial space, includes Oracle Database (Oracle Inc., 2013) and a host of companies listed in (Gartner, 2014) for Business Intelligence and in (Gartner, 2013) for DBMS.

On the infrastructure side of discussion, traditional DBMS was designed with transactional processing in mind and does not scale well to support the modern enterprise data
warehouse, as discussed in The NoSQL Solution section of CHAPTER II. Executing these functions in the database, closer to the data, can avoid data movement and thus accelerate analytics performance and increase throughput. Further, centralizing analytics in the database facilitates version control, reduces duplication, and extends the DBMS’s management, security, and auditing infrastructure to analytical data and functions.

The direction to in-database analytics architecture was realized by the discussion on EMC Greenplum (EMC Inc., 2012) study under the PMML Enabled Architecture section of CHAPTER II. Other competing commercial products in the in-database analytics space are IBM Netezza (IBM Inc., 2012), SAS (SAS Inc., 2007), Teradata Database (Grimes, 2012) and Oracle Database (Oracle Inc., 2011). As previously mentioned, the R platform has a long history and it is pervasive amongst practitioners, which attracted many commercial and open source products to integrate their products with the R platform such as Teradata Database (Teradata , 2013) for in-database analytics using R.

IBM Watson famously outdid top human contestants in the game of Jeopardy! (Siegel, 2013). The event serves as a practical example of Predictive Analytics with Big Data in action. IBM Watson utilized text analytics and ensemble models along with internet data to analyze question (in a form of answer in Jeopardy!) phrased in human language, to predict the most likely answer by ranking the top candidate answers based on their probability scores. Of course, the Jeopardy! challenge was meant to be a showcase of IBM Watson’s ability to process massive information. One of the first commercial applications of IBM Watson was to support healthcare preapproval decisions based on clinical and patient data (IBM Inc., 2013), applying the same techniques and methods used on Jeopardy! on practical application of management decision support.
Another notable trend in predictive modeling is the *Uplift Model* (Siegel, 2013). Uplift model has the practical properties of finding the differentials in class/label when an intervention is introduced, which is to say, solving the problem of \( \Delta P(Y|X) = P(Y_i|X_i) - P(Y_{i+1}|X_{i+1}) \) where \( \Delta P(Y|X) \) is the probability of delta of the effect between the original class/label and the intervention. Also, to measure a combined effect of a treatment, we can perform \( P(Y|X) = P(Y_i|X_i) + \Delta P(Y|X) \).

For instance, a popular example amongst uplift modeling business application was that, marketing professionals often interest in knowing the *persuadable* individuals to concentrate targeted marketing efforts then to spend resources on the unlikely individuals who are *unpersuadable or less-than-effective-persuadable*. In that case, discovering the persuadable individuals is also a specialized classification problem because focuses would be given to those who are classified as persuadable based on commonly shared predictors. Since uplift modeling is dealing with classification problem that can be solved by classification methodology. Association rules such as basket analysis is certainly a logical choice given the problem domain, however, decision tree method is one of the simplest form of classification techniques that is simple to understand and implement in most cases.

In (Rzepakowski & Jaroszewicz, 2011), clinical trials often involve *control group* in order to observe the isolated effect of a treatment. Uplift modeling is also applicable in this case. Uplift modeling not only helps to predict the differences between treatment and non-treatment groups, but also allows researchers to predict the *lift* from single or multiple treatments applied to treatment group(s). Furthermore, uplifting modeling can model the effects on each action (i.e. treatment) applied as well as measuring the *lift* in the *degree of effectiveness* per treatment or per a series of treatments. This is an example of how uplift modeling can be incorporated into
clinical trials to better support and to more accurately predict clinical trial successes (Jaskowski & Jaroszewicz, 2012).

**Ethical Concerns and Issues**

The ethical concerns overlap both the problem domains of Big Data and Predictive Analytics. In this context, the concern for Big Data stems from the data collection and storage aspect of the problem domain while Predictive Analytics matters in the implementation problem domain. That is, we must look at the Big Data side of our concern prior to investigating in Predictive Analytics. Big Data, by definition, increases the data resolution and thus contains details that are of concerns to some.

Privacy is a basic human right enforced by laws in most industrialized countries including Canada. Canada established the *Privacy Act* (Government of Canada, 1985) to complement the *Access to Information Act* (Government of Canada, 1985) in order to balance the right of information access and privacy protection of Canadians. Canada also has the *Personal Information Protection and Electronic Documents Act* (PIPEDA) to enforce the use and collection of personal information in the private sectors (The Office of the Privacy Commissioner of Canada, 2009). The province of Alberta has *Freedom of Information and Protection of Privacy Act* (FOIP) (Alberta Legislature, 2000) and *Personal Information Protection Act* (PIPA) (Alberta Legislature, 2003), defining the boundary of limits to which public bodies (i.e. government) collect, use and disclose of personal information, a framework to protect the privacy of Albertans for information held within public bodies.

In terms of legislated entities, the *Office of the Privacy Commissioner of Canada* (Government of Canada, 2012) and the *Office of the Information and Privacy Commissioner of
Alberta (OIPC, 2012) represent residents on issues and concerns related to privacy in Canada and Alberta, respectively.

Respecting privacy of individuals often results in conflict with the intrinsic values of Big Data. Sensor data that came from surveillance video, mobile phone GPS data, biometrics data from implanted devices. Other electronic communication data such as email messages, text messages, digitalized voice and video messages as well as web traffic log. They are all examples of what constitute Big Data that are advantageous to Predictive Analytics for reasons already discussed in this essay. However, these data sources suggest an incompatible view of personal privacy.

The fear of information misuse is one aspect of the overall privacy concern. Take GPS data as an example, location-aware applications and devices that track the physical location of an individual. This would bring convenience from photo geotagging to navigation support, but they also brings the concern of information exploitation. Consider a geotagged picture taken inside of an individual’s own home which had been inadvertently circulated to the public internet by a third party. The photo might reveal valuable objects within the home that pique the interest of a perpetrator. Together with the geotagged information (i.e. latitude and longitude) embedded within the photo, this will make a compelling case of potential criminal consequence due to privacy invasion.

As much as data, in and of itself, are morally neutral, the perception of data when view under different contexts is what challenge researchers to operate within the confines of personal privacy and professional ethics. Applying Predictive Analytics allows us to make observations and decisions about the future based on presumptions; that is, applying statistical inference to predict into the future state based on data. There is always a level of uncertainty with prediction
and must not be considered as a guaranteed outcome. For instance, as discussed in Law Enforcement section of CHAPTER III, predicting recidivism of offenders is an approximation and not an exact measure. It can be used in supporting parole decisions but certainly cannot be used in conviction of a future crime. Also, consider the social and moral consequences of job candidate selection based entirely on predictive scores deduced from the Big Data. Predictive Analytics application that targets individuals based on their shared attributes inferred by the individuals’ behavioral data can lead to misclassification, discrimination and damage to reputation. In addition, preemption could undermine our traditional models of justice, due process and individual freedoms (Office of The Privacy Commissioner of Canada, 2012). Since there is no certainty in prediction and only accuracy in measures, we cannot determine future event as absolute but only as probable outcome.

**Conclusion**

The challenges remain in developing models that are robust to noise but adaptive to change. These challenges continue to push the ability of researchers to strike a balance between model overfitting and model underfitting, as well as to account for the dynamic nature of data we used in prediction.

The issues, challenges and trends discussed in this essay represent only a subset of the subject matter, albeit important ones, in understanding the limitations of Predictive Analytics in the context of Big Data. Certainly, the incorporation of Big Data has been proven to advance to the application of Predictive Analytics in both theoretical and pragmatic senses. The IBM Watson Jeopardy! challenge captured the interests of the general public in what Predictive Analytics with Big Data can achieve. The enablers behind IBM Watson’s ability to answer natural language questions comprises of multilayers of advanced technologies.
Predictive Analytics applies to a wide-ranging applications as previously discussed, novel means of predictive application will only become more widespread in the near future. The trend in predictive application is as evidenced by the recent patient filed by Amazon Inc. for the *anticipatory package shipping process* (Amazon Inc., 2013).

The fusion of Big Data and ensemble methods employed by IBM Watson marked a milestone in Predictive Analytics as the Netflix Prize winner team did in exploiting collective wisdom in deriving predictions. Mining unstructured data with complex data types will continue to challenge researchers to invent novel means to tackle problems. However, in doing so, navigating the fine line between accurate predictions and the privacy of individuals, remains a great challenge. Researchers and practitioners must abide by the ethical principles with professional attitudes to respect with rights of individuals.
CHAPTER VI

CONCLUSIONS AND RECOMMENDATIONS

Conclusions

This essay paper took the position that Predictive Analytics is a specialized interest area within the Data Mining domain. It focuses on prediction that makes inferences to unknowns and for that reason, it is considered as a subset of Data Mining field. Big Data and Predictive Analytics might have begun as marketing terms with the intent to attract attentions and excite people into the discussion of the subject matter. However, the current landscape is such that the terms are now synonymous to the ideas, methodologies, methods and techniques that the terms symbolized.

For these reasons, this essay attempts to elucidate the confusion by starting the discussion with an overview of the current landscape of the subject matter in CHAPTER I and CHAPTER II. Then, cross-industry case studies were discussed in CHAPTER III as well as the commonly employed methodologies were discussed in CHAPTER IV. This lead to CHAPTER V on the discussion of the issues, challenges and trends of the subject matter. At various points during the discussion, the juxtaposition between Data Mining and Predictive Analytics was provided and the symbiotic relationship between Predictive Analytics and Big Data was demonstrated throughout the essay. Illustrations using figures and tables were created or referenced and cited to provide visualization in areas that best illustrate the concepts embedded within.

During the course of the construction of this essay as well as the copious amount of literatures reviewed, a pattern began to emerge in terms of shared practices between the different practice domains. The result of this realization was captured in the content of this essay as well as the developed taxonomy diagrams as shown in Figure 1, Figure 3, Figure 6 and Figure 10. The
taxonomy diagrams were an attempt to differentiate the constructs in the convoluted field of analytics, a visualization tool created towards the goal of solidifying the understanding of the subject matter in the readers’ minds of this essay.

Researchers and practitioners should not blindly believe that Big Data or Predictive Analytics as a one-size-fits-all solution to our problems. It is however, a solid decision support tool that further broadens our reach on data that results in the previously unattainable rapid knowledge acquisition.

**Suggestions for Further Research**

The torrent of Big Data came with tremendous opportunities and also costs for materializing the potentials for data analytics. This research indicated that the Big Data evolution is currently at the early stage, the full potential has yet to be realized. To put this in perspective, a direct quote from Eric Schmidt, CEO of Google Inc. in year 2010: “There were 5 Exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days.”. Almost four years after the statement, we saw the advancement of NoSQL databases, Cloud Computing and the ecosystems that are built around open source platforms such as OpenStack and Hadoop.

**SOA**

From a software architectural perspective, we have NoSQL databases and Hadoop platform to handle the challenges brought by Big Data. From an infrastructural perspective, we have Cloud Computing platform (i.e. IaaS, PaaS and SaaS) with hardware visualization to bring operating cost down (i.e. utility computing) to support the many applications of Predictive Analytics with Big Data. To that end, the infrastructure and the toolsets are now in place to handle Big Data. The industry as a whole will continue to evolve around the Software Oriented
Architecture (SOA) to bring Data-as-a-Service (DaaS) to a more mainstream level, especially in the context of Big Data.

Predictive Analytics will also capitalize on SOA to bring organizations the services of Predictive-Analytics-as-a-Service (PAaaS). Combining DaaS and PAaaS, the end result of the full realization and ubiquitous adoption of this synergized technology, is what we can look forward to in the next frontier in the application of *Predictive Analytics in Data Mining with Big Data*.

With that said, the research community thus needs to focus more on the SOA aspect rather than on the modeling methods and algorithms advancements, as we have reached a plateau in those dimensions as discussed in previous chapters. Therefore, our collective attentions should concentrate on making the technologies more available, accessible and affordable to the community at large.

**Real-time analytics**

Another challenge related to Big Data and Predictive Analytics is the ability to process information at real-time or at nearly real-time speed. The early version of the Hadoop Platform did not provide adequate support on this front as the MapReduce method (Vavilapalli, et al.) operates in batch processing mode. Under the batch processing mode, the notion of a job is intrinsically a scheduling-based activity, which is to say, a job based approach is incompatible with real-time application as there always exists a significant delay between job executions.

In addressing these issues with real-time data processing, the Hadoop Platform version 2.0 introduced a number of components including Hadoop YARN and Apache STORM to provide continuous real-time streaming analytics. Other important Apache projects that deal with real-time Big Data problems include Apache Spark, Apache Drill and Apache S4. Reducing the
time gap between recording an observation and making a decision will maximize the effect of any prediction.

Making timely decision is critical to the application of Predictive Analytics. However, even with the accompanying technologies introduced in Hadoop 2.0, the exponential growth of Big Data would eventually pose a risk in spite of the most cutting-edge developments in the Apache Hadoop project. There exists an asymmetrical growth rate between Big Data and Hadoop. Big Data is projected to grow exponential until the year 2020 (Gantz, Reinsel, & Lee, 2013) while the current Hadoop architecture currently scale linearly. The gap will continue to grow between data that are available and the ability to process them in the timely manner.

**NoSQL**

The NoSQL data model offers the following characteristics: Basically Available, Soft State, and Eventual Consistency. These characteristics are commonly known as *BASE*. Contrary to *ACID* (Atomicity, Consistency, Isolation and Durability), which is common amongst relational data model, *BASE* offers a less rigid data model that favors performance over strong data consistency.

The reason that many adopted the NoSQL data model for Big Data is because, NoSQL data model is able to handle the characteristics of Big Data where relational data model fall short. The shortcomings in NoSQL data model as it relates to relational data model becomes the strength in which NoSQL can excel in an era of Big Data.

If ensuring data atomicity and strong data consistency are too difficult to do with Big Data and relational data model, then we would design our applications around these limitations and learn to accept the trade-offs using NoSQL. However, there are situations where we would expect strong transactionality for our operations such as securities trading data and financial
transactions. These situations can also be benefited by Big Data and thus, an improved version of NoSQL with better transactionality support is desirable. Solutions such as FoundationDB (FoundationDB, 2014) have made the first step in improving NoSQL with ACID support. More research is needed to validate the viability and to determine the cost of ACID based NoSQL data model. To that end, a hybridized data model is to be looked forward to that fuses the benefits between the relational data model and the NoSQL data model.
REFERENCES

About Us Background. (2014, 01 29). Retrieved from European Molecular Biology Laboratory - European Bioinformatics Institute: http://www.ebi.ac.uk/about/background


academicindex.net. (2014, February). Retrieved from AcademicIndex.net:
http://www.academicindex.net/


http://pdfpiw.uspto.gov/.piw?docid=08615473&SectionNum=3&IDKey=2809BD982F3D&HomeUrl=http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2%2526Sect2=HITOFF%2526p=1%2526u=%252526Fnetauthhtml%252526FPTO%252526Fsearch-bool.html%2526r=1%2526f=G%2526l=50%2526co1=AND%2


http://www.dmg.org/v4-1/GeneralStructure.html


Google Correlate. (2014, February). Retrieved from Google.com:
http://www.google.com/trends/correlate/

http://www.google.org/flutrends/about/how.html


http://www.priv.gc.ca/au-ans/index_e.asp

doi:10.1145/2093548.2093555


Han, J., Kamber, M., & Pei, J. (2011). In J. Han, M. Kamber, & J. Pei, *Data Mining: Concepts and Techniques, Third Edition*. Morgan Kaufmann. doi:978-0123814791

Han, J., Kamber, M., & Pei, J. (2011). In J. Han, M. Kamber, & J. Pei, *Data Mining: Concepts and Techniques, Third Edition*. Morgan Kaufmann. doi:978-0123814791


*IEEE Xplore Digital Library.* (2014, February). Retrieved from IEEE.org:
http://ieeexplore.ieee.org/Xplore/home.jsp

*IET Inspec.* (2014, Feb). Retrieved from The Institution of Engineering and Technology:
http://www.theiet.org/resources/inspec/


doi:10.1007/BF02885683


Lesk, M. (n.d.). *How Much Information Is There In the World?* Retrieved from Lesk.com:
http://www.lesk.com/mlesk/ksg97/ksg.html

Lesson 19: Logistic Regression. (2014). Retrieved from Professor Courtney Brown's YouTube Video Classes:
http://www.courtneybrown.com/classes/video_classes_Courtney_Brown.html


A Survey of Predictive Analytics in Data Mining with Big Data


Oracle Inc. (2011). *In-Database Analytics: Predictive Analytics, Data Mining, Exadata & Business Intelligence*. Retrieved from


http://www.oracle.com/technetwork/database/options/advanced-analytics/advanced-analytics-wp-12c-1896138.pdf?ssSourceSiteId=ocomen


*Patterns of Inference*. (2014, 2 23). Retrieved from Probabilistic Models of Cognition:

https://probmods.org/patterns-of-inference.html


http://www.nature.com/srep/2013/130425/srep01684/full/srep01684.html


Siegel, E. (2013). *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die.* Wiley.


Statsmodels. (2014, February). Retrieved from Sourceforget.net:

http://statsmodels.sourceforge.net/
http://www.statsoft.com/Products/STATISTICA/Product-Index


APPENDIX A – PMML CODE

PMML CODE EXAMPLE

<DerivedField name="Field2" optype="continuous" dataType="double">
  <NormContinuous field="Field1" mapMissingTo="Field3" outliers="asExtremeValues">
    <LinearNorm orig="Original Value 1" norm="Normalized Value 1"/>
    <LinearNorm orig="Original Value 2" norm="Normalized Value 2"/>
    <LinearNorm orig="Original Value 3" norm="Normalized Value 3"/>
  </NormContinuous>
</DerivedField>

PMML CODE EXAMPLE - HEADER SECTION

<Header copyright="KNIME">
  <Application name="KNIME" version="2.8.0"/>
</Header>

PMML CODE EXAMPLE - DATADICIONARY SECTION

<DataDictionary numberOfFields="10">
  <DataField dataType="integer" name="Age" optype="continuous">
    <Interval closure="closedClosed" leftMargin="17.0" rightMargin="90.0"/>
  </DataField>
  <DataField dataType="string" name="Employment" optype="categorical">
    <Value value="Private"/>
    <Value value="Consultant"/>
    <Value value="SelfEmp"/>
    ...
  </DataField>
  <DataField dataType="string" name="Education" optype="categorical">
    <Value value="College"/>
    ...
  </DataField>
</DataDictionary>

PMML CODE EXAMPLE - TRANSFORMATIONDICTIO

<MapValues outputColumn="longForm">
  <FieldColumnPair field="gender" column="shortForm"/>
</InlineTable>
A Survey of Predictive Analytics in Data Mining with Big Data

<row><shortForm>m</shortForm><longForm>male</longForm>
</row>
</InlineTable>

<row><shortForm>f</shortForm><longForm>female</longForm>
</row>
</MapValues>

PMML CODE EXAMPLE - MODEL SECTION – SUPPORT VECTOR MACHINE

<SupportVectorMachineModel modelName="SVM" functionName="classification" algorithmName="Sequential Minimal Optimization (SMO)" svmRepresentations="SupportVectors">
    <MiningSchema>
        <MiningField name="Age" invalidValueTreatment="asIs"/>
        <MiningField name="Income" invalidValueTreatment="asIs"/>
    </MiningSchema>
    <Targets>
        <Target field="TARGET_Adjusted" optype="categorical">
            <TargetValue value="0"/>
            <TargetValue value="1"/>
        </Target>
    </Targets>
    <LocalTransformations>
        <DerivedField dataType="integer" name="Private_Employment" optype="ordinal">
            <NormDiscrete field="Employment" mapMissingTo="0.0" value="Private"/>
        </DerivedField>
        ...<DerivedField dataType="string" name="Age_binned" optype="categorical">
            <Discretize field="Age">
                <DiscretizeBin binValue="0.92">
                    <Interval closure="openOpen" rightMargin="18.0"/>
                </DiscretizeBin>
                ...</Discretize>
        </DerivedField>
        <DerivedField dataType="double" displayName="Age" name="Age*" optype="continuous">
            <Extension extender="KNIME" name="summary" value="Z-Score (Gaussian) normalization on 4 column(s)"/>
        </DerivedField>
    </LocalTransformations>
</SupportVectorMachineModel>
<NormContinuous field="Age">
  <LinearNorm norm="-2.8430412477523532" orig="0.0"/>
  <LinearNorm norm="-2.76942928421807" orig="1.0"/>
</NormContinuous>
</DerivedField>

<DerivedField dataType="double" displayName="Income*" name="Income*" optype="continuous">
  <Extension extender="KNIME" name="summary" value="Z-Score (Gaussian) normalization on 4 column(s)="/>
  <NormContinuous field="Income">
    <LinearNorm norm="-1.2164099347279616" orig="0.0"/>
    <LinearNorm norm="-1.2163955713786447" orig="1.0"/>
  </NormContinuous>
</DerivedField>

<DerivedField dataType="double" displayName="Deductions*" name="Deductions*" optype="continuous">
  <Extension extender="KNIME" name="summary" value="Z-Score (Gaussian) normalization on 4 column(s)="/>
  <NormContinuous field="Deductions">
    <LinearNorm norm="-0.1983193433770224" orig="0.0"/>
    <LinearNorm norm="-0.19538425035943" orig="1.0"/>
  </NormContinuous>
</DerivedField>

<DerivedField dataType="double" displayName="Hours*" name="Hours*" optype="continuous">
  <Extension extender="KNIME" name="summary" value="Z-Score (Gaussian) normalization on 4 column(s)="/>
  <NormContinuous field="Hours">
    <LinearNorm norm="-3.297304369766391" orig="0.0"/>
    <LinearNorm norm="-3.2150250058375485" orig="1.0"/>
  </NormContinuous>
</DerivedField>

<DerivedField dataType="double" displayName="Age_binned*" name="Age_binned*" optype="continuous">
  <FieldRef field="Age_binned"/>
</DerivedField>

</LocalTransformations>

<PolynomialKernelType coef0="1.0" degree="1.0" gamma="1.0"/>

<VectorDictionary numberOfVectors="741">
  <VectorFields numberOfFields="53">
    <FieldRef field="Age*"/>
    <FieldRef field="Income*"/>
  </VectorFields>
</VectorDictionary>
A collection of services and software aided the research during the construction of this essay:

**PRODUCTIVITY SOFTWARE**

1. Microsoft Office 2013 Suite (Word, Excel and PowerPoint)
2. Microsoft Project 2013
3. Microsoft Visio 2013

**INTERNET BROWSERS**

4. Google Chrome internet browser
5. Internet Explorer internet browser
6. Firefox internet browser

**OPEN SOURCE PREDICTIVE ANALYTICS AND DATA MINING TOOLS**

8. Weka (Weka 3: Data Mining Software in Java, 2014)
9. PSPP (GNU PSPP, 2014)
10. Orange (Orange 2.7 for Windows, 2014)
11. KNIME (KNIME, 2014)
13. ELKI (LKI: Environment for Developing KDD-Applications Supported by Index-Structures, 2014)

**PYTHON RELATED STATISTICAL LIBRARIES**

    a. NumPy library (NumPy, 2014)
b. pandas library (pandas, 2014)

c. SymPy library (SymPy, 2014)

d. IPython library (IPython, 2014)

15. scikit-learn library (scikit-learn: Machine Learning in Python, 2014)


17. Matplotlib library (Matplotlib, 2014)

18. Statsmodels library (Statsmodels, 2014)

19. MpMath library (mpmath, 2014)

LITERATURE SEARCH ENGINES

20. Google (Google, 2014)

   a. Google Correlate (Google Correlate, 2014)

   b. Google Scholar (Google Scholar, 2014)

   c. Google Trends (Google Trends, 2014)

21. WolframAlpha (WolframAlpha, 2014)

22. Dogpile (dogpile, 2014)

23. iSeek at (iSeek Education, 2014)

24. refseek (refseek, 2014)

25. Virtual LRC (Virtual LRC, 2014)

26. AcademicIndex.net at (academicindex.net, 2014)

27. Digital Library of The Commons Repository (Digital Library of the Commons Repository, 2014)

28. Microsoft Academic Research (Microsoft Academic Research, 2014)
RESEARCH PAPER ONLINE DATABASES

29. IEEE Xplore Digital Library (IEEE Xplore Digital Library, 2014)

30. ACM Digital Library (ACM Digital Library, 2014)

31. EBSCO Colleges and Universities Online Resources for Academic Libraries
   (EBSCO Colleges and Universities Online Resources for Academic Libraries, 2014)

32. IET Inspec Database (IET Inspec, 2014)

33. SpringerLink Database (SpringerLink, 2014)

34. Athabasca Online Library at (Athabasca University Library, 2014)

RESEARCH MANAGEMENT TOOLS AND SERVICES

35. Mendeley (Mendeley, 2014)

36. Qiqqa (Qiqqa, 2014)

37. Colwiz (Colwiz, 2014)

38. Evernote (Evernote, 2014)

39. Zotero (Zotero, 2014)

ONLINE COMMUNITIES

40. Academia.edu (Academia.edu, 2014)

41. Khan Academy (KhanAcademy Probability and Statistics, 2014)