A Framework for Knowledge Discovery within Business Intelligence for Decision Support

A thesis submitted for the degree of Doctor of Philosophy.

By

Rajveer Singh Basra

Brunel Business School
Brunel University

Abstract

Business Intelligence (BI) techniques provide the potential to not only efficiently manage but further analyse and apply the collected information in an effective manner. Benefiting from research both within industry and academia, BI provides functionality for accessing, cleansing, transforming, analysing and reporting organisational datasets. This provides further opportunities for the data to be explored and assist organisations in the discovery of correlations, trends and patterns that exist hidden within the data. This hidden information can be employed to provide an insight into opportunities to make an organisation more competitive by allowing manager to make more informed decisions and as a result, corporate resources optimally utilised. This potential insight provides organisations with an unrivalled opportunity to remain abreast of market trends. Consequently, BI techniques provide significant opportunity for integration with Decision Support Systems (DSS). The gap which was identified within the current body of knowledge and motivated this research, revealed that currently no suitable framework for BI, which can be applied at a meta-level and is therefore tool, technology and domain independent, currently exists. To address the identified gap this study proposes a meta-level framework: - 'KDDS-BI', which can be applied at an abstract level and therefore structure a BI investigation, irrespective of the end user. KDDS-BI not only facilitates the selection of suitable techniques for BI investigations, reducing the reliance upon ad-hoc investigative approaches which rely upon 'trial and error', yet further integrates Knowledge Management (KM) principles to ensure the retention and transfer of knowledge due to a structured approach to provide DSS that are based upon the principles of BI.

In order to evaluate and validate the framework, KDDS-BI has been investigated through three distinct case studies. First KDDS-BI facilitates the integration of BI within 'Direct Marketing' to provide innovative solutions for analysis based upon the most suitable BI technique. Secondly, KDDS-BI is investigated within sales promotion, to facilitate the selection of tools and techniques for more focused in store marketing campaigns and increase revenue through the discovery of hidden data, and finally, operations management is analysed within a highly dynamic and unstructured environment of the London Underground Ltd. network through unique a BI solution to organise and manage resources, thereby increasing the efficiency of business processes. The three case studies provide insight into not only how KDDS-BI provides structure to the integration of BI within business process, but additionally the opportunity to analyse the performance of KDDS-BI within three independent environments for distinct purposes provided structure through KDDS-BI thereby validating and corroborating the proposed framework and adding value to business processes.

Key Words

Business Intelligence, Decision Support Systems, Framework, Case Studies, Advanced Analytics, Intelligent Agents.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor and friend; Dr. Kevin Lü (Brunel Business School, Brunel University, UK). His support and guidance has been invaluable throughout my further education. He has been pivotal and truly inspirational in my life.

Any piece of written work, especially a task as significant as a doctoral thesis, requires commitment, dedication and can prove to be an arduous journey. Although stressful for the person undertaking the research, the journey can prove equally long and difficult for those loved ones, family and friends alike, who share this journey. I would therefore, like to thank deeply, my Father, Mother and family, especially my Sisters and Brothers-in-law, for their continued support and patience. Without their support, I could not have undertaken, let alone completed this research. For which I also especially thank my good friend Dr. Simran Grewal (School of Management, University of Bath, UK), who not only provided me with much appreciated advice and guidance from the day I commenced this research, but also undertook the laborious task of proof-reading this thesis.

I would also like to thank my friends, whom I consider my family, who have been there to support me throughout this journey. They have been there for me, to not only provide support and guidance, but most importantly believed in me on the rare occasions on which my own self-belief may have faltered. I would also like to give a special mention to my dog; Guinness. Who not only kept me company whilst writing this document, but also patiently waited in the event I was ever late to feed or walk him.

This has genuinely been an experience which determines ones character and a task which I am happy to have had the opportunity to endure, as it was one which often felt as though it was a truly perpetual endeavour. Yet, with support from those mentioned and anybody I may have accidently omitted, along with the ability and insight provided to me by God, a journey which has finally been completed.

Waheguruji Ka Khalsa, Waheguruji Ki Fateh.

Rajveer Singh Basra.

Dedicated to my Nieces;

Simran, Priya, Anu

&

Preeti.

Publications Arising From This Study

Basra, R. & Lü, K. (2008) 'Enhancing the Capabilities of Marketing Executives with Business Intelligence Techniques', *In publication process*.

Basra, R. & Lü, K. (2008) 'Analysing Organisational Data through Business Intelligence to Enhance the Capabilities of Direct Marketing Strategies', *In publication process*.

Basra, R. & Lü, K. (2008) 'Increasing the Impact of Sales Promotion Strategies through Advanced Analytics and Business Intelligence Techniques, *In publication process*.

Basra, R. & Lü, K. (2008) 'A Framework to Support Knowledge Discovery within Business Intelligence', *In publication process*.

Basra, R. & Lü, K. (2008) 'Analysing Multidimensional Databases using Data Mining and Business Intelligence to Provide Decision Support', *In Proceedings of the 10th International Conference on Enterprise Information Systems*, Barcelona, Spain, 12 - 16 June.

Basra, R. & Lü, K. (2007) 'Investigating A Multi-Agent System For Scheduling And Planning On The London Underground', *International Journal of Intelligent Systems Technologies and Applications (IJISTA),* vol.2 , no. 1, pp. 3-19.

Basra, R., Lü, K., Rzevski, G. & Skoblev, P. (2005) *Resolving Scheduling Issues of the London Underground Using a Multi-Agent System*. Holonic and Multi-Agent Systems for Manufacturing: Volume 3593/2005 (Book Chapter) pp. 188-196.

Basra, R. & Lü, K. (2004) 'A High-Granularity Multi-Agent System for Advanced Planning and Scheduling In Public Transport', *In Proceedings of the IEEE/WIC/ACM Joint International Conference on Intelligent Agent Technology and Web Intelligence*. Beijing, China, 20-24 September, 2004.

Basra, R. & Lü, K. (2004) 'Scheduling London Transport Resources Using a Multi-Agent System', *In Proceedings of the 9th UK Academy for Information Systems International Conference,* Glasgow, Scotland, 5 -7 May.

Table of Contents

List of Figures

List of Tables

List of Code-tables

Chapter 1:

Introduction

This opening chapter provides the background and motivation for this study, in addition to an overview for the requirement to investigate a meta-level framework for Business Intelligence (BI). The framework should encompass the functionality to provide decision makers with a greater level of knowledge upon which to direct future business direction, whilst facilitate knowledge acquisition and retention from the experience of integrating BI within business processes. The chapter will subsequently introduce the proposed solution, scope, contribution to knowledge, aims and objectives of this study.

1.1 Introduction

Reduced cost of storage equipment has facilitated companies to store vast amounts of data not only on their staff and resources but also their business needs and most significantly their customers. Typically, key information is retained and knowledge extracted to enable organisations to discover customer needs and preferences, the process through which customers make decisions, the competition, conditions in the industry, in addition to general economic, technological, and cultural trends. Furthermore, coupled with the growth of the internet and web based services, organisations are able to access previously untapped markets. However, for these opportunities to be fully exploited it is imperative that companies are capable of effectively managing this ever increasing quantity of information. Traditionally, such information and data is stored in data warehouses. Hence, data warehouses facilitate an organisation to store and analyse their data in order to utilise this it to make informed decisions. It is this requirement of the need for informed decisions that has lead to the development and growth of Business Intelligence (BI).

BI amalgamates tools such as reporting, data depositories, on-line analysis tools and data mining, in addition to providing access to data that has been integrated and cleaned, allowing for the data to be analysed, manipulated transformed. The data can then be explored to assist organisations to discover correlations, trends and patterns that may exist between this information, enabling an insight into how to make an organisation more competitive by allowing manager to make more informed decisions and as a result, corporate resources optimally utilised (Simmers, 2004; Zhao et al, 2006). Traditionally, the focus of BI has been upon the analysis of batch data that is updated periodically, be it daily, weekly or monthly. Since, batch updating has previously been 'at best' conducted daily, real-time has been considered anything better than that (Finnie et al, 2005). However, with the internet facilitating a larger quantity of transactions, and blurring the lines between time zones and countries, it is imperative that businesses move toward a model that not only allows for true real-time analysis but in addition facilitate access to as many customers as possible at the least possible cost. Organisations therefore require a model that will facilitate decision maker and managers to not only be able to target previous customers and deal with current needs, but also predict customers and forecast trends. Furthermore, with such a dynamic environment it is essential that BI tools and information be provided to managers enabling them to make informed decisions. Thus BI forms a very well suited platform to be integrated with Decision Support Systems (DSS).

Since their introductions in the 1970s, decision makers in a number of fields have employed DSS to aid in critical decisions. Decisions concluded through DSS have proven to bear a significant effect upon the nature and performance of an organisation (Arnott & Pervan, 2007). DSS are an interactive computer based system that is intended to provide support to the decision makers engaged in solving

various semi- to ill-structured problems, involving multiple attributes, objectives and goals (Rupnik et al, 2006). It should be noted however, that DSS are not autonomous systems and therefore is not intended to replace decision makers; the aim of DSS is to provide support for decision making, thereby endeavouring to improve the effectiveness and quality of decisions (Mladeni et al, 2002).

Hence DSS provide the functionality to aid the process of managing, information and data. It has been identified that the market for systems which provide information/data management advances in (approximately) 20-year cycles. The initial period can be identified from the 1950s. It was during this period in a seminal IBM Journal article in October 1958 that the term 'Business Intelligence' was first proposed (Luhn, 1958). At this time organisations collected data from non-automated sources, yet lacked the computing resources to properly analyse the data, as a result, companies often made decisions primarily on the basis of intuition. During the 1970s to 1990s the data management sector dominated by companies such as SAS, IBI, and IBM, was characterised by production reporting on mainframes. This eventually evolved to the current 'modern era of Business Intelligence'. In the early 1990s, Howard Dressner, then an analyst at the Gartner Group, coined 'Business Intelligence' as a term through which to encapsulate the intelligent management and analysis of data, to increase business efficiency (Burstein & Holsapple, 2008). BI is now widely used, especially in the world of practice, to describe analytic applications (Watson & Wixcom, 2007). The modern era of BI, can be characterised by user friendly client/server-based BI tools underpinned by a number of investigative and analytical techniques, of which the key techniques are known as data mining and intelligent agents. As a result, there has been significant interest to integrate BI technology to provide intelligent DSS (Lawton, 2006).

1.2 Motivation

It is the potential of BI to integrate with DSS that has motivated this research. As highlighted earlier in the chapter, BI amalgamates tools such as reporting, data depositories, on-line analysis tools and data mining, in addition to providing access to data that has been integrated and cleaned. Users are able to analyse, manipulate, transform and combine data to assist organisations to discover correlations, trends and patterns aiding an organisation to become more competitive (Simmers, 2004). Yet, the integration of the global economy, customer data increase rapidly, how to use BI to make strategic decision becomes a critical issue (Zhao et al, 2006), although this has provided significant incentive for the integration of BI and DSS. Due to the increasing quantity of data that is stored, traditional DSS or conventional artificial intelligence techniques for supporting the decision making process, fail to address situations which require a vast amount of data to be analysed from databases. As a result, BI still faces several challenges paramount to which, is the fact that BI remains essentially poorly integrated with the decision support process (Zhang et al, 2007).

The key to addressing the problem of the inadequate integration of BI and DSS is to ensure that the opportunities available within a dataset are full recognised and the correct technique, or often the correct combination of techniques is applied to analyse the data and approach the problem. Data is not just the record of facts. If understood correctly, valuable knowledge can be discovered from these records and future predictions can be made. Consequently, data provides the integral criteria for effective integration of BI with DSS. Furthermore, it has been addressed that organisations are collecting and storing data in a greater quantity than ever before. This data although occasionally collected for specific reason, through surveys, questionnaires etc. may also be collected as part of routine organisational practices, e.g. data collected on the details of customers at the point of purchase. In addition to customers, data may also be collected internally by organisations. As organisations endeavour to compete on a larger global market, with an increasing number of competitors, organisations will accumulate an increasing number of resources, which must be optimally managed to gain a competitive edge. Consequently, organisations aspiring to strive in the global market must manage data, be it on customer and external forces, or the internal resources of an organisation. It is this necessity for integrated BI based DSS that can explore and interrogate data with a view to provide decision makers and mangers with the best possible criteria upon which to base decisions that has motivated this research.

1.3 Proposed Solution

For the effective integration of BI and DSS, it is imperative that the manner through which BI solutions are integrated and utilised within organisations be addressed. In addition, consideration should be given to the methodology employed to realise BI solutions and explore organisational data. Stevenson (2006) has identified three basic functions of an organisation (figure 1.1). Thus, it is data regarding these functions that must be managed and analysed for optimum operations.

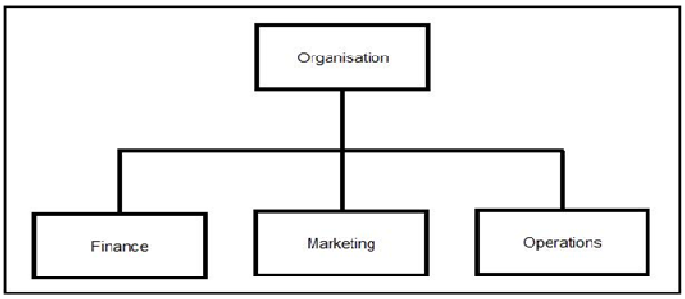


Figure 1.1: Basic functions of an organisation.

Consequently, DSS using BI technology which can explore, analyse and support the three basic function of an organisation will provide an invaluable tool. Currently, the ineffective integration of BI and DSS can be attributed to ineffective assimilation of organisational data and data requirements.

Furthermore, BI solutions are implemented in an ad-hoc manner utilising frameworks which have not been explicitly created for the discovery of BI as a means of decision support. Having investigated the conventional approaches to BI and its related technologies and techniques (explored in greater detail in chapter 4). It is clear that there is no meta-level framework which exists for the investigation of BI within business processes. As a result, to address the dynamic issue that exists when interrogating data for a BI project and meet the needs of an organisation intending to integrate BI technology into their daily operations. Conventional methods such as 'Agile', 'Prometheus', 'Gaia' and even the 'waterfall model', are effective at collecting requirements and specifications. However, there is a distinct weakness within these process models for selecting appropriate techniques that enable the most effective analysis of data. This issue has been addressed by process models such as CRISP-DM, SEMMA and Knowledge Discovery in Databases (KDD). SEMMA however, is devised for explicit use with a particular tool (SAS Enterprise Miner), and neither SEMMA nor CRISP-DM provides effective support for data understanding, preparation and modelling within BI. These shortcomings become even more critical when the knowledge once extracted and analysed is to be used for decision support. In addition, none of these process models, including KDD provide the opportunity to investigate various tools and technologies which can be explored to investigate BI solutions.

CRISP-DM as emphasised on their homepage (www.crispdm.org), does provide an example of how a structured approach to the investigation of a technology and its associated techniques, can increase the effectiveness of solutions. Hence, much like the manner in which the CRISP-DM process model was motivated by the need for an approach that was tailored to data mining, so too does BI require an explicit framework which can facilitate knowledge discovery and retention. The requirement for a structured approach is supported by the Gartner Group. As a result, it has been recommended that organisations take advantage BI functionality in a competitive environment. To ensure that organisations can reap the full benefits of BI integration, research has emphasised the necessity to prepare for changes in technology and products. Furthermore, in order to explore BI to drive business transformation it is advised through studies that, organisations adopt the following practice into organisational operations (Forsling, 2007):

1. Alter the method for the implementation and management of the organisations information architecture and application portfolio.
2. Re-asses and alter the procedure through which BI is integrated into business processes.
3. Increase focus on developing user skills. In addition to instilling a culture for the use and analysis of information. This culture should be considered an integral aspect for achieving business objectives and transformation.
4. Establish a BI Competence Centre. A steering committee has been proven to substantially increase efficiency in many successful companies.

Historically, each department within an organisation has purchased their own BI solutions. This frequently results in a number of different tools being implemented throughout the various departments. It is the limitations and inefficiency of this ad-hoc, distributed approach to BI integration, which has resulted in organisations endeavouring to implement a uniform system throughout the various departments in an organisation. This will ensure that each department not only adopts a similar and rational way of analysing and measuring data, yet additionally, increasing the impact and operational efficiency of the implemented BI solution (Forsling, 2007). However, the effective integration and exploration of BI, requires a structured meta-level framework, especially if BI technology is to be integrated and discovered to provide support for decision makers.

The solution proposed by this research to address the shortcomings of the integration of BI and DSS is to investigate and develop a framework that can support this process. The proposed framework KDDS-BI (explored in greater detail in chapter 5), presents an approach explicitly developed for the exploration of organisational data. Based upon and extending the CRISP-DM and KDD process model, KDDS-BI introduces a structured meta-level approach to BI investigation. This framework provides the facilities through which existing data can be interrogated with BI for novel and innovative analysis. This analysis can then be further explored to enable executive decision making to take place. KDDS-BI is investigated with a view to provide a tailored framework that facilitates BI integration for the exploration of organisational data with a view to providing decision support. Thereby providing a suitable framework, hence, guidelines for an organisation which can be observed when launching their own BI projects and interrogating organisational data for enhancing decision support and gaining a competitive advantage.

1.4 Scope of Thesis and Contribution to Knowledge

This research proposes a framework that addresses the requirement for a structured meta-level approach to BI integration for providing not only decision support but to further enable knowledge retention from BI integration experiences. The investigation will conduct an in-depth analysis of the background theory relating to BI technology and its associated techniques. Given, that a framework can be defined as a system, since it provides a structured approach to resolving a problem (Fowler, 1998) a ‘hard systems engineering approach’ is adopted as a methodological framework within which the research can be conducted and validated. A hard systems engineering approach adopts a scientific approach to problem solving, placing a greater value upon logic and rationality, over intuition and thereby placing less emphasis upon the human element. As a result, related work and conventional methods can be studied and analysed to provide an insight into the requirements and aspects that will provide a complete, robust and valid meta-level framework for knowledge discovery within BI.

Stevenson (2006) has identified three basic functions that define organisational activity (figure 1.1-page 4):

Finance: defines the activities responsible for securing financial resources at the best possible rate. Finance also involves the allocation of financial resources (capital) throughout an organisation, be it for budgeting, analysing investment proposals and/or providing funds for the daily operational activities of the organisation.

Marketing: is the activity that identifies the needs and requirement of target consumers. In addition, marketing is the organisational activity which strives to ensure that consumers are not only aware but further, desire the goods and services provided by an organisation.

Operations: is the organisational activity which is responsible for the production of goods and/or provision of services offered by an organisation. Thus, operations ensures the inputs, such as land, labour, capital etc. can be transformed into goods and services. This transformation implies a process of adding value, which operations activities endeavour to achieve at the lowest cost, whilst maintaining the highest possible standards.

Finance can be considered a function of operations and marketing, as both aim to maximise the finance available to an organisation be it through increased revenue or lower manufacturing cost, furthermore many financial activities refer to the process of allocating resources (capital) optimally throughout an organisation. Consequently, to ensure that the framework proposed through this investigation is robust and valid, the framework will, be explored through three case studies:

1. The initial case study will explore the framework in the domain of 'direct marketing'. Thereby permitting the analysis of the frameworks ability to provide decision makers with the means through which to predict consumer segments that can be targeted through marketing strategies.
2. The second case study will explore the framework in the domain of 'sales promotion'. Thereby permitting the analysis of the frameworks ability to provide decision makers with the means through which to target consumers at the point of purchase, cross-promote products and increase revenue in sub-optimal performing areas.
3. The third case study will explore the framework in the domain of 'resource allocation'. Thereby, permitting the analysis of the frameworks ability to provide decision makers with the means through which to optimally allocate and distribute resources throughout an organisation.

The analysis of the framework through the means of case studies will ensure that the framework can be tested and validated in situations resembling those typical of which, require organisations to investigate BI-based approaches to provide DSS. As discussed, data is often collected routinely by organisations. Thus, to truly maximise the potential of a meta-level framework, existing data will be explored and analysed for innovative and novel objectives. The case studies however, provide duel functionality, since not only do they demonstrate the capability of the framework in a variety of

domains through a meta-level framework. The case studies will permit the analysis of BI techniques and tools. Modern BI applications can be defined into two classes of intelligence tools. The first class of tools are built on database management systems and implemented to interrogate large amounts of operational data. Traditionally held in data warehouses, data is interrogated in order to extract useful information, trends and patterns from otherwise arbitrary data (Zhang et al, 2007). The second class of tools, also referred to as competitive intelligence tools, systematically collect and analyse information from the competitive environments to assist organizational decision making (Chung et al, 2002). Hence, since BI amalgamates a number of technologies, many of which require various tools, selecting the case studies in a variety of organisational activities permits the analysis of the framework using a variety of tools and techniques, in instances which require novel solutions requiring a combinatorial approach of various BI techniques.

Despite vast research into the various applications of the two classes of BI tools, and the techniques that underlie these tools, it can still remain unclear to an organisation integrating BI practices into their daily operations, which techniques are most suitable for their needs. Discovering the most suitable techniques to be integrated can be a costly and time consuming endeavour for companies, yet no framework currently exists to aid in the evaluation and selection of BI techniques for decision support. For this purpose, the case studies permit this aspect of the study to address the requirements of a framework which can not only aid effective integration of BI for decision support, but further ensure that the most appropriate tool and technique is selected. This will ensure that the proposed framework and solutions can be considered a valid and robust contribution.

1.5 Aims and Objectives

The primary aim of this study is to address the requirement for a framework which can explicitly facilitate the investigation and analysis of BI techniques to provide decision support; through the analysis of the various aspects of BI and an in-depth study of conventional approaches to BI integration, a framework will be formulated which can address the short-comings of conventional approaches. Once formulated, the framework will be investigated to provide support for decision makers within a variety of organisational activities. Due to the increasing complexity of the competitive global marketplace, it is essential that BI techniques be explored to provide decision makers with novel and innovative DSS, which can provide key insight into trends hidden within data to help direct future operations through forecasting and prediction. For this to be effective, it is imperative that the correct technique is selected for the explicit task. To facilitate knowledge discovery and retention, the investigation of novel and innovative DSS based upon BI techniques will not only provide a road map detailing the situations in which particular BI techniques are appropriate, but also illustrate the potential benefits which can be reaped from using an approach tailored to BI, to

structure the analysis of data which is routinely collected by an organisation. For this purpose, this study will not only propose a framework which can structure BI investigations at an abstract level, but also validate this framework through the use of three case studies. These case studies will transverse a number of areas in which BI is applicable to demonstrate that the developed framework can be applied at a meta-level. Since, each case study will cover a distinct area, they will illustrate that the framework is tool and technique independent. Rather the framework will structure the investigations within each case study at a meta-level thereby providing a means through which to validate the proposed framework, thereby, verifying and evaluating the findings and conclusions of this study. The objective of this study can be defined by the following hypothesis:

"A meta-level framework for Business Intelligence will facilitate knowledge discovery and decision support."

In order to test and validate the hypothesis, the research aims of this study can be defined as:

- Conduct an in depth critique and review of the literature and the theories which underpin the domain of BI and related strategies. This will subsequently, provide a background to this study. Whilst permitting the identification of a suitable research methodology, which is able to capture and generate data to test the focal theory, via a specific research design.
- Undertake exploratory studies within the research methodology to hypothesis a meta-level framework which can facilitate the selection of suitable BI techniques to develop solutions and address organisational requirements for providing decision support to executives.
- Develop a meta-level framework to structure BI investigations and select suitable techniques.
- Evaluate the framework through three case studies which will assess the proposed frameworks capability to support the process of selecting suitable techniques for BI integration.
- Asses the significance of the framework within each case study and the implications that occur through the development life cycle.

1.6 Structure of the Thesis

Since the scope, aims and objectives of the research have been determined and declared the structure for the thesis can be explicated. The thesis will consist of three sections; the preliminary section, (consisting of Chapters 2-4) will explore and investigate the background theory and literature, in addition to the selected methodology pertaining to this research. The second section (consisting of Chapters 5-8) will define and illustrate the formulation and proposal of a suitable meta-level framework. Furthermore, this framework will be explored through a number of case studies, each requiring novel and innovative solutions based upon organisation data requirement. The final section of the study (Chapter 9) will conclude the findings and performance of the framework. In addition to

providing a conclusion for this research and proposing future recommendations which may be observed to further this study.

Section 1:

The initial section of this study will explore and investigate the background theory, literature and methodology pertaining to this research. In addition to, exploring and scrutinising the related work and conventional approaches for BI investigations.

Chapter 2:

Literature Review

This chapter provides a context to the study by reviewing the published literature on Business Intelligence (BI) and its associated technologies and techniques. BI is further defined to distinguish it from Knowledge Management (KM). Moreover, the intricacies of BI techniques are considered, in addition to the mechanisms for decision making and Decision Support Systems (DSS). Consequently, it is the theories and techniques explored in this chapter which despite remaining unchanged will underpin this research.

2.1 Business Intelligence (BI)

It has been identified that the Information/Data management systems market advances in 20 year cycles. The initial period can be identified from the 1950s, it was during this period, in a seminal IBM Journal article (Luhn, 1958) that the term 'Business Intelligence' (BI) was first coined. Luhn (1958, p. 314) wrote in his paper that:

"Business is a collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defence, et cetera. The communication facility serving the conduct of a business (in the broad sense) may be referred to as an intelligence system. The notion of intelligence is also defined here, in a more general sense, as the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal".

During this initial period organisations collected data from non-automated sources, yet lacked the computing resources to properly analyse the data, thus, companies often made decisions primarily on the basis of intuition. During the 1970s to 1990s the information management sector dominated by companies such as SAS, IBI, and IBM, and characterised by production reporting on mainframes. Given that the information management infrastructure of an organisation can greatly affect its performance and ability to remain competitive, it is nevertheless, (often) a large-scale investment. It is for this reason that, accounting researchers have attempted to accurately analyse the explicit effect of the information management infrastructure upon performance. This will enable the optimum level of investment for an organisation to be determined. In the early 1990s, there were many event and market valuation studies using archival data, which sought to demonstrate the 'payoff' from information management investment (Dehning & Richardson, 2002). By the early 2000s, this research resulted in a focus toward specific information management applications, such as Enterprise Resource Planning (ERP) systems, that can emphasise stronger and more meaningful relationships between adoption and performance effects. Nigel Rayner, research vice president at the Gartner Group has stated (Forsling, 2007 p. 1):

"*Companies that have invested heavily in ERP are now realising they need to invest in BI to extract value from the massive amount of data they are storing as a result of ERP."*

Consequently, the research and focus toward specific information management applications has led to the "modern era of Business Intelligence" (Elbashir et al, 2008).The modern era of BI has witnessed the evolution of query reporting, and OLAP technology being migrated from client/server to Web-based architecture and the development of broad suites of BI tools from vendors such as Business Objects, Cognos, and Hyperion (Lawton, 2006). It was during this period that Howard Dressner, at the time an analyst at the Gartner Group, defined BI as an umbrella term to describe *"concepts and methods to improve business decision making by using fact-based support systems."* (Burstein &

Holsapple, 2008 p. 128). BI is now widely used, especially in the world of practice, to describe analytic applications and technologies which are applied to gather, provide access to and analyse the data and information about an enterprise for business decisions (Wu et al, 2007). This is achieved through the use of standards, automation and specialised software, including analytical tools, allows large volumes of data to be extracted, transformed, loaded and warehoused (Watson & Wixson, 2007; Power, 2007). Consequently, BI amalgamates tools such as reporting, data depositories, on-line analysis tools and data mining, resources planning intelligent agents, amongst others. BI enables users to analyse, manipulate, transform and combine data to assist organisations to discover correlations, trends and patterns, a number of modern BI software allows users to cross-analyze and perform deep data research rapidly (Golfarelli et al, 2004). As a result, in modern applications of BI, managers are able to quickly compile reports from data, for forecasting, analysis, and business decision-making or organise resources to increase the competitive edge of an organisation (Simmer et al, 2004; Zhang et al, 2007).

Advances in hardware, software, communication, and online transactions have led to a truly integrated global economy. These advances together with reductions in the associated cost have facilitated companies to store an ever increasing volume of data (Han & Kamber, 2006). For organisations there is a necessity to rapidly analyse this data to explore opportunities, consequently, how to use BI to make strategic decisions has become a critical issue (Zhoa et al, 2006). As a result, BI models are being implemented in a number of industries, for a variety of purposes; this has led to the development of many BI tools. However, in the past BI tools where prohibitive due to their complex nature requiring a specialist operator, this is changing and in the face of increased demand. BI developers have endeavoured to make their tools more user-friendly and ensure that they are able to integrate more seamlessly with the current applications (Ortiz, 2002). This has led to increased interest in BI applications. As discussed, various forms of BI technology have been available since the fifties, yet, it is only now that this technology has begun to mature and move into the mainstream (Finnie & Barker, 2005). The financial services, telecommunications, and manufacturing are the industries where BI has most actively been adopted (Ortiz, 2002). As BI technology has matured, various industries have also adopted BI into there business model. According to a worldwide forecast for BI platforms for the period 2007-2012, it has been predicted that the market is forecast to reach $5.8 billon (£2.9 billion) during 2008, which is an 11.2 per cent increase from 2007. Worldwide BI platform revenue is forecast to grow at a compound annual growth rate of 8.1 per cent through 2012, thereby, reaching $7.7 billion (£3.9 billion) in 2012 (Forsling 2007). This has been predicted in spite of economic slowdown; in contrast the market for BI platforms is expected to continue to grow at a strong rate. Further to this, in a survey of 1,500 Chief Information Officers (CIOs), studies discovered that BI projects were the number one technology priority for 2008 (Knights, 2008). Thus, BI is

currently the top-most priority of many chief information officers - for the second year running. According to Gartner Groups' research vice president, Andreas Bitterer (Forsling, 2007 p. 1):

"BI has become a strategic initiative and is now recognized by CIOs and business leaders as instrumental in driving business effectiveness and innovation".

2007 witnessed major consolidation among BI vendors. During this period SAP acquired Business Objects for €4.8 billion (£3.3 billion), whilst IBM paid $4.9 billion (£2.5 billion) for Cognos and Oracle purchased Hyperion for $3.3 billion (£1.7 billion). As a result mega vendors' predicted to own more than one third of the global BI market by 2010 (Knights, 2008). Despite such large commercial interest in BI and BI platforms, this curiosity has not been solely from commercial vendors. There has also been significant attention and investigation within the research community, reflected in the large number of open-source BI tools that are available and studied for a variety of research projects (Watson & Wixom, 2007). Irrespective of the area they are applied in, contemporary BI applications can be defined into three classes of intelligence tool; 'resource planning', 'End user QRA (query, reporting and analysis)' and 'advanced analytics' (Chung, 2003).

BI resource planning tools are an evolution of conventional ERP systems that were invested in, during the 1990s. Along with data, contemporary organisations are amassing large quantities of resources. These resources range from physical, tangible resources such as vehicles (e.g. company cars), IT equipment and offices, to human resources, such as employees or clients / customers. Just as telecommunications have facilitated the collection of a greater amount of data to be collected, they have also facilitated the increase in the distribution of organisational resources. Organisations often distribute departments into offices physically located in various locations in a country or even distributed in various countries. Due to the advancements in telecommunication, it is now possible for these various offices can operate as one without finding themselves subject to diseconomies of scale, due to the ability to share information in real-time (Curtis & Cobham, 2008). The core of a conventional ERP system is a set of planning modules that translate the anticipated demand into plans that can be implemented to coordinate the various aspects of managing supply, production and distribution. Other modules in the ERP software aid the organisation to implement the plans and integrate them into daily operations in addition to providing computerised support for purchasing, receiving, sales and various other operations. However, ERP systems are far more difficult to maintain then general software development due to the size, complexity and dynamic nature of the environments within which they are expected to function (Kwon & Lee ,2001; Taylor, 2004). BI resource planning tools integrate artificial intelligence techniques such as intelligent agents to ensure that the short comings of traditional ERP systems can be overcome.

End-user QRA tools include ad-hoc query and multidimensional analysis tools as well as dashboards and production reporting tools. Query and reporting tools are designed specifically to support ad-hoc data access and report building by either IT or business users. Generally data is queried via SQL to investigate and analyse data structures and uncover shallow knowledge. Shallow knowledge, is knowledge that is by nature factual, and can be stored and manipulated easily. Multidimensional analysis tools traditionally include both on-line analytical processing (OLAP) servers and client-side analysis tools that provide a data management environment used for modelling business problems and analysing business data and producing reports. In relational data, each piece of data correlates to one row and one column, each of which can be considered a dimension, thus relational data is considered to be two-dimensional. Multi-dimensional databases however provide a higher-level perspective of the data by providing further dimensions to include core components of your business plan such as; Accounts, Time, Products and Markets. Each dimension consists of individual components known as members. Although the dimension will tend to remain static, members will generally be dynamic, e.g. new customers or products added. Multi-dimensional analysis supports interactive examination of large amounts of data from many perspectives facilitating for the interrogation of data that if configured correctly can quickly provide answers to queries via OLAP and SQL (Pederson & Jenson, 2001; Thompson, 2002).

QRA tools meet the requirement for organisations to produce weekly or monthly reports based upon shallow knowledge to generate aggregated views of data enabling stakeholders and management to view the state of their business such as the value of assets, and distinguish between the assets which are obtaining their goals to the resources that are performing sub-optimally. Often these reports can be provided to the users through Description tools. Description tools employ two techniques Profiling and Reporting, Description tools are designed as analysis applications. Originally intended to support the growing data warehousing market, the embedded methods description has evolved from simplistic sampling and frequency analysis to more complex cross-column dependency and embedded structure analysis techniques (Lee & Park, 2005). Profiling is a BI strategy that entails collecting information often combing information from various databases and presenting it in an organised manner to develop a representation that will facilitate business decisions (Witten & Frank, 2005). Reporting extracts information from data stored in databases, applications and data warehouses into an understandable format for users, thus enabling organisations to gain feedback on issues regarding business operations and resources, facilitating more efficient decisions. Although reporting can be text based it is often a technique that visually ascertains the status and performance of a business enterprise via key business indicators commonly referred to as a ‘Digital Dashboard’. Digital Dashboards provide users with a visual, at-a-glance display of data pulled from disparate business systems to provide warnings, action notices, next steps, and summaries of business conditions by using devices such as red/green/yellow lights, alerts, drill-downs, summaries, graphics such as bar charts, pie charts,

bullet graphs, spark lines and gauges that are usually set in a portal-like environment that is often role-driven and customizable. An example of a digital dashboard by SAS Dashboards is depicted in figure 2.1 (Zhang et al, 2007).

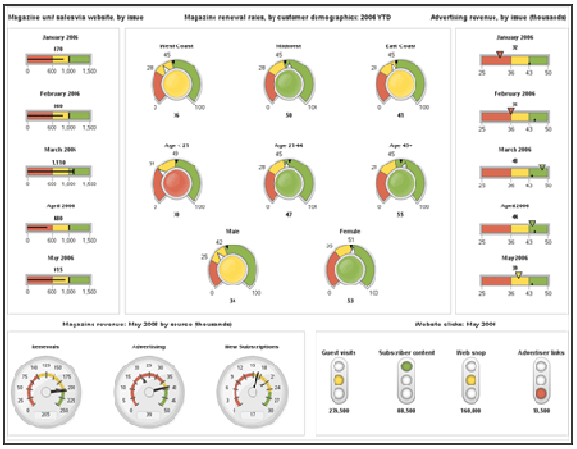


Figure 2.1: SAS digital dashboard.

Advanced analytics tools (which can also be referred to as competitive intelligence tools) are generally built upon database management systems and implemented to interrogate large amounts of operational data at a deeper level through QRA (query, reporting and analysis) tools. The data, traditionally held in data warehouses, is scrutinised in order to extract useful information, trends and patterns from what is often considered arbitrary data. Advanced analytic tools systematically collect and analyse information from the competitive environments to assist organisational decision making as illustrated by figure 2.2. This class of tools includes data mining and statistical software. Data mining, which will be explored in greater depth later in this chapter, is a key BI tool.

According to Mladenic´ et al (2003), data mining is concerned with finding models and patterns from the available data, which can be achieved through the extraction of data from a database by utilising software that can isolate and identify previously unknown patterns or trends in large amounts of data (Han et al, 2006). There are a variety of data mining techniques that reveal different types of patterns such as, predictive data mining algorithms, descriptive data mining algorithms, rule induction, Neural Networks and Clustering. These techniques provide an advanced means of analysing data to discover relationships in data and make predictions that are hidden, not apparent, or too complex to be extracted using QRA software (Lavrac, 2006). It should however, be noted that advanced analytic tools are also currently being actively researched to investigate data that is held in multidimensional database. Multidimensional databases have received significant interest and research from the data mining community. Investigation of multidimensional databases through data mining techniques can often facilitate a much more intricate and deeper analysis for hidden knowledge, which is not possible

through OLAP or SQL alone (Yuefeng, 2007; Pinto et al, 2001; Thompson, 2002).

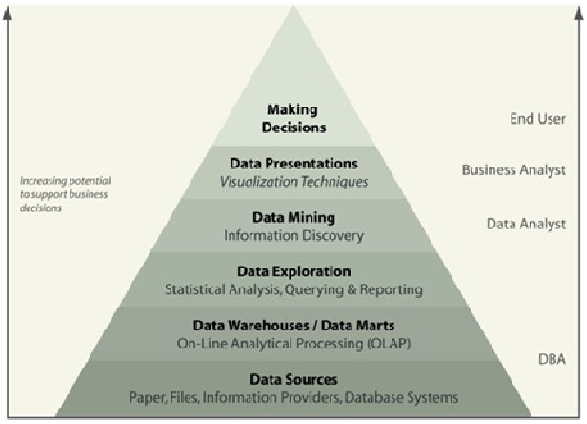


Figure 2.2: Supporting decision making through BI techniques.

2.1.1 Business Intelligence (BI) and Knowledge Management (KM)

As discussed, organisations have been exploring and investing in technology that will enable them to manage information and data for analysis, which will provide them with a competitive edge. Along with BI, knowledge management (KM) has proven to be technique which provides a significant return on investment (Herschel & Jones, 2005). BI can be defined as the process of turning data into information and then into knowledge (Golfarelli et al, (2004). While KM can be defined as a systematic process of finding, selecting, organising, distilling and presenting information in a way that improves an employee's comprehension in a specific area of interest. Given the similarity between the two technologies it is imperative that the distinction between the two be established, in view of the fact that a survey by OTR consultancy revealed that 60 percent of consultants did not understand the difference between the two technologies (Herschel & Jones, 2005; Cody et al, 2002).

Herschel & Jones (2005 p. 46) address the difference between KM and BI, by stating that *"BI focuses on explicit knowledge, but KM encompasses both tacit and explicit knowledge. Both concepts promote learning, decision making, and understanding."* Tacit knowledge can be thought of as the knowledge that an individual encompasses, therefore, is intrinsically linked to the experiences and perspective of an individual, thus can prove cumbersome to formalise. In contrast, explicit knowledge is knowledge which has been or can be formalised, coded and stored. Tacit knowledge is therefore, easier to capture, since effective transfer of tacit knowledge requires extensive personal contact and trust (Mclaughlin, 2007; Debowski, 2006). KM much like BI is a descriptive term which suffers from far more misuse. Like BI, the term KM is frequently applied to resolve problems for which it has never truly been intended. Consequently, this has resulted in organisations implementing large scale KM integration projects that have failed, largely due to a lack of understanding (Cheng et al, 2003; Cody et al; Mclaughlin, 2007; Herschel & Jones, 2002; Debowski, S. 2006). It is therefore, essential that an

organisation ensure that it has clearly defined objectives when deciding upon the most suitable approach. If the aim of an organisation is to ensure that it can gain a competitive edge from information that it has collected then it is wiser to implement a BI strategy. Furthermore, BI has proven to be successful when not only analysing data, but when investigated to uncover trends and patterns that are hidden deep within datasets. These hidden trends and patterns can be investigated to forecast future directions (Watson & Wixon, 2007).

BI provides organisations with a further benefit; easier integration with existing technologies. BI applications can be built to increase the current information management capabilities of an organisation, with less risk than that which is associated with KM (Ortiz, 2002). BI tools can be integrated with organisational strategies such as Customer Relationship Management (CRM). Being one of the leading business strategies CRM integrates sales, marketing and service across multiple business units and customer contact points. In addition to this CRM helps companies understand the value of customers, identify and target their most profitable customers, encourage and maintain high-quality relationships that increase loyalty and profits (Lee & Park, 2005). However, for this to be successful it is imperative that customer profitability precisely evaluated thus, targeting the most profitable customers, consequently, companies will utilise historical data and through BI techniques such as ETL (Extract, Transform and Load) and data mining techniques, to extrapolate this data in order to (as accurately as possible) predict trends and forecast expected growths within their business operations and provide this analysis to managers. This enables organisations to make focused decisions for the future. Forecasting can also be utilised to find correlations between various products, thus forecasting can be explored to investigate whether, the increased consumption of a particular product, will bear an effect upon the performance of another? Consequently, BI can be utilised to view not only current action, but also suggest the most suitable direction an organisation should take, consequently BI can be an invaluable tool for decision–makers and managers (Stein & Dhar, 1997). However, the effectiveness of BI tools is reliant upon the quality of data it utilises. Therefore, rely upon the data being of a high standard and quality information and knowledge extraction are pivotal to a successful BI implementation. Consequently, it is imperative to investigate the techniques that can be implemented to select and analyse organisational data. Knowledge Discovery in Databases (KDD), which will be discussed in the following chapter, is one process, which can be investigated to ensure the highest quality of data is available for BI applications.

2.2 Knowledge Discovery in Databases (KDD)

Information systems have enable users to gather an ever increasing amount of data. Data obtained regarding a particular environment is the basic evidence used to build theories and models of any domain. Moreover, companies use such data to better understand the market, trends and customer behaviour, enabling them to gain a competitive advantage, increase efficiency and provide more

valuable services to customers. The increased volume of data captured and the potential benefits that can be reaped from this data has in turn led to a need for computational techniques which can provide support to help unveil meaningful patterns and structures from these datasets. Knowledge Discovery in Databases (KDD) is one attempt to address a problem that the digital information era has made ever-present; data overload (Roiger & Geatz, 2003). The phrase '*Knowledge Discovery in Databases*' was coined at the first KDD workshop in 1989 (Piatetsky-Shapiro, 1991), to emphasise that knowledge is the end product of a data-driven discovery. It has since been popularised in the field of artificial intelligence and machine-learning. KDD can be defined as the non-trivial extraction of implicit, previously unknown, and potentially useful information from databases. Further to this, in recent years the number and the size of databases across a variety of fields have rapidly increased due to various factors (Witten & Frank, 2005). The falling cost of storage and technological advances in connectivity have resulted in companies being able to capture and store data with ease. This growth, by far exceeds human capacities to analyse the databases in order to find implicit regularities, rules or clusters hidden within the data. Therefore, knowledge discovery becomes more and more important in databases and he applications based upon these concepts (Lazcorreta, 2008).

Knowledge discovery is an area of research that amalgamates several disciplines, including statistics, databases, artificial intelligence, visualisation, high-performance and parallel computing (Wu, 2004). At an abstract level, KDD is concerned with the development of methods and techniques that facilitate the deciphering of data. The basic problem addressed by the KDD process is one of mapping low-level data into other forms that might be more compact and user-friendly in the form of a short report, more abstract such as, a descriptive approximation or model of the process that generated the data, or more useful for example, a predictive model for estimating the value of future cases. At the core of the process is the application of specific data mining methods for pattern discovery and extraction (Han & Kamber, 2006; Geist, 2002). The traditional method of turning data into knowledge relies upon manual analysis and interpretation (Witten & Frank, 2005). An example of this can be cited within the health-care industry, it is not uncommon for specialists to periodically analyse current trends and changes in health-care data on a quarterly basis. This enables the specialists to provide reports detailing the analysis to the sponsoring health-care organization. In turn these reports form the basis for future decision making and planning within health-care management. Another example of this which can be cited is the efforts of planetary geologists. Planetary geologists sift through remotely sensed images of planets and asteroids, carefully locating and cataloguing such geologic objects of interest such as, impact craters (Roiger & Geatz, 2003). Knowledge discovery is apparent in almost any field; science; marketing; finance; health care; retail etc. (Luo, 2008). The classical approach to data analysis relies fundamentally upon analysts becoming intimately familiar with the data, thereby, enabling the analysts to perform as an interface between the data, users and products. However, in many areas this form of manual probing of a dataset can be slow, cumbersome, expensive, and highly

subjective. Further to this as the volume of data captured increases manual data analysis has become impractical in many domains. Consequently, the requirement to scale up human analysis capabilities in order to analyse the large volumes of data is both economic and scientific. The KDD process model provides a means through which analysis capabilities can be up scaled through a scientific method (figure 2.3).

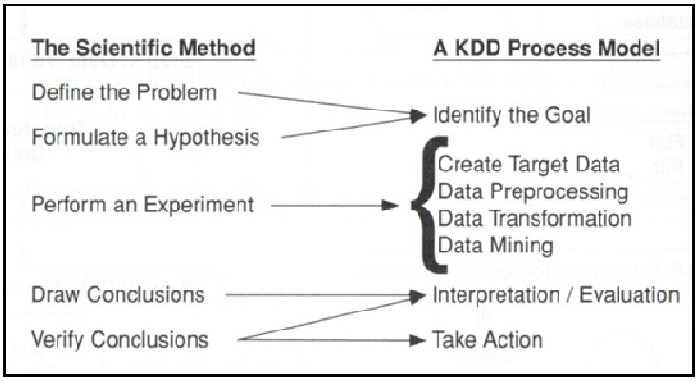


Figure 2.3: Relating scientific method to KDD process model.

2.2.1 KDD Process Model

The KDD process model is an interactive, iterative, procedure that attempts to extract implicit, previously unknown and potentially useful knowledge from data through a scientific method. Involving numerous steps with many decisions made by the user the KDD process model can be divided into a number of steps. As described by Fayyad et al. (1996) and supported by Rioger & Geatz (2003), a popularised approach of the KDD process can be described to consist of several stages:

1. *Data understanding*: establish goals and investigate data.
2. *Data selection*: selecting a dataset on which discovery is performed.
3. *Data pre-processing*: clean data by removing noise and outliers.
4. *Data transformation*: transform dataset into a form suitable for mining.
5. *Data mining*: choose methods and search for patterns.
6. *Interpretation / Evaluation*: use measures to assess patterns and visualise models.

Initially, it is imperative that an understanding of the application domain should be established in addition to identifying the goals of the KDD process from the viewpoint of the user/customer. This will facilitate in creating a target data set; created by selecting a data set or focusing upon a subset of variables or samples of the data from which meaningful information can be extracted. Once the data set is selected it must be pre-processed. This stage involves cleaning the data to remove any noise or outlier (if appropriate), if any noise or outliers are removed the data must be adjusted to compensate for this, in addition to any missing values and account for time sequence information. Next the data must be transformed, this stage requires that the data be normalised, converted and smoothed. This may entail removing and/or adding attributes, often this stage will also require the data to be modelled

or changed into an appropriate format for the knowledge discovery process to take place. Once an appropriate dataset has been selected, pre-processed and transformed, the dataset is ready to be interrogated to find meaningful information, patterns and trends. Thus, the data mining stage involves deciding which models and parameters might be and matching a particular data mining method with the overall goals of the KDD process through the application of one or more algorithms to the data to extract information and create models, this stage will be examined in greater deal in the following section. Finally, the previous stages especially the results of the data mining stage are interpreted and evaluated. It is possible to return to any of the earlier stages at this point should the need arise. Furthermore, the evaluation can also involve visualisation of the extracted patterns and models or visualization of the data given the extracted models (Combes et al, 2007).

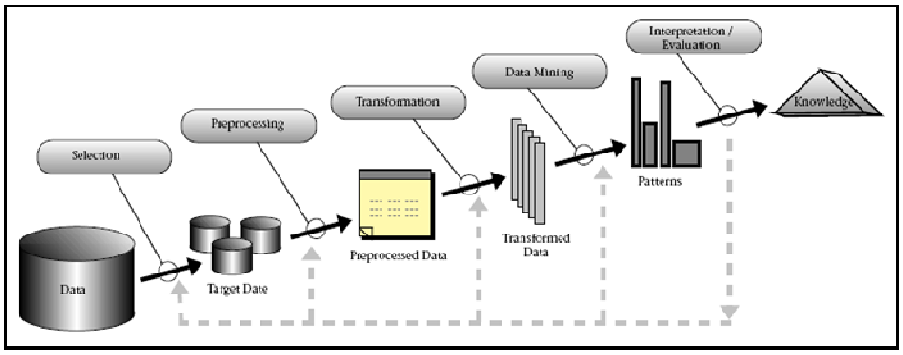


Figure 2.4: An overview of the steps that compose the KDD process.

The KDD process can involve significant iteration and can contain 'loops' within any two stages of the process model. The basic flow of steps is illustrated in figure 2.4. It should be noted that there are a number of variants of this KDD process, such as those that have been published by Adriaans & Zantinge (1996); Brachman & Anand (1996); and Han & Kamber (2006) in addition to others, however, all variants of the KDD process remain close to the stages depicted in figure 2.4. A key concern highlighted through previous research, is the potentially high level of iteration which takes place between the KDD stages, resulting in the process being more arduous and time-consuming than necessary (Roiger & Geatz, 2005). The KDD process provides a framework on which this knowledge can be extracted. Thus, KDD refers to the overall process of discovering useful knowledge from data; however, the data mining stage can consist of a number of techniques each suited to particular circumstances therefore is a stage that although not necessarily more important (to a successful knowledge discovery process) then the other stages, is a key stage within the process should be given careful consideration (Luo, 2008; Peng, 2006).

2.3 Data Mining

Data mining is the nontrivial process of extracting valid, previously unknown, comprehensible, and useful information from large databases and using it. Consequently, it can be considered an exploratory data analysis technique, implemented to discover useful patterns in data that are not obvious to the data user (Mladenic´ et al, 2003; Han & Kamber, 2006). This is achieved via the application of specific algorithms for extracting patterns from data. The distinction between the KDD process and the data mining stage (within the process) is a key issue. The additional steps in the KDD process, such as data preparation, data selection and data cleaning are essential to ensure that useful knowledge is derived from the data. Blind application of data mining methods (data dredging in statistical literature) will often result in the discovery of meaningless and invalid patterns, therefore, is a technique which should be given careful consideration when deployed. Historically, the process of finding useful patterns and trends within data sets has been referred to under a number a guises, such as data mining; knowledge extraction; information discovery; information harvesting; data archaeology; and data pattern processing, amongst others. (Witten & Frank, 2005). The term 'data mining' has mostly been used by statisticians, data analysts, and the management information systems (MIS) communities. Although a stage in the KDD process, the term has gained popularity within the database field, and has become almost interchangeable with KDD (Luo, 2008).

Roiger and Geatz (2005) define data mining as the process of employing one or more computer learning techniques to analyse and extract knowledge from data. Thus, the primary objective of data mining is to search for, identify and reveal trends and patterns contained within the data. This search is automated (or at the very least augmented) by computers (Witten & Frank, 2005). However, searching to identify patterns is not a novel concept, economists, satiations, forecasters and communication engineers amongst others have all sought to uncover hidden meaningful pattern within data. Research in data mining has addressed a broad range of applications as diverse as sales and customer relationship management (Berry & Linoff, 2004; Hung et al, 2006), financial forecasting (Chun & Park, 2006), fraud detection (Fawcett & Provost, 1997), gene mapping (Kantardzic & Zurada, 2005) and mining of health care data (Alonso et al, 2002; Phillips-Wren et al, 2007). In addition to this, a poll conducted by KD nuggets (June 2007) to discover the size of databases that are being mined, found that, nearly 22% of the respondents reported mining databases of 1 terabyte or more, this figure has thus almost doubled since 2006. In addition, it is estimated that the amount of data stored within the world databases doubles every 20 months. With advances in technology, such as the internet and telecommunications, providing organisations access to previously untapped markets and transactions taking place twenty-four hours-a-day in a universal global marketplace. Data mining has been brought to the forefront of business technologies, especially when integrated with BI strategies (Zhang & Zhou, 2004; Tan et al, 2005).

Data mining can be integrated with BI as an advanced analytics tool. These tools are able to increase the capability of organisations to perform deep analysis on the data they collect. Traditionally, this is an objective achieved through QRA tools, OLAP, SQL or statistical packages. QRA tools provide the user with the means through which to conduct descriptive analysis, these tools incorporate various aggregation, allocation, ratio algorithms that provide descriptive modelling functions. Furthermore, database query languages such as SQL, are able to manipulate databases to extract shallow knowledge, which is by nature factual and therefore easier to extract, yet are unable to uncover correlations or extract patterns that are not immediately apparent. The ability to only uncover shallow knowledge is a limitation that also applies to statistical packages, which are generally only applied to small-to-medium sized datasets. Furthermore, statistical packages are ill-suited for analysis involving nominal or structured data-types, which are common to databases. Nor do statistical packages provide any means of incorporating prior domain knowledge as they are completely data driven (Anand & Buchner, 1998). In contrast, data mining techniques such as 'Regression', 'Neural Networks', 'Decision Trees' and 'Clustering' algorithms provide an exploratory modelling function, thereby, interrogating and investigating the data at a deeper level to uncover patterns, trends and correlations that are hidden and would hence remain unobserved through conventional techniques (Adriaans & Zantage, 1996; Thompson, 2002; Mirkin, 2005).

It has been previously discussed, that KDD provides a process model within which data mining forms an integral stage. The KDD process model aims to provide the data mining step with quality data that can be analysed and knowledge extracted. The knowledge gained from data mining is given as a model or synopsis of the data (Peng et al, 2006). For this several data mining models exist (which will be explored in greater detail later in the chapter). A data mining model is can be defined as an algorithm or set of rules that connects a collection of inputs to a specific target or outcome (Berry & Linoff, 2005). All data mining models are founded upon 'induction-based leaning' (Roiger & Geatz, 2005). Induction-based learning is the process of forming general concept definitions via the observation of the particular concepts that are to be learned. Data mining is thus fundamentally connected to learning. Learning, however, is a complex process, Merril & Tennyson, (1977) proposed that four levels of learning can be differentiated:

- *Facts:* are simple statements of truth.
- *Concepts:* are a set of objects, symbols or events, which can be grouped together due to the common characteristic they encompass.
- *Procedures:* detail a sequence of stages or steps that can be followed with a view to achieving a particular objective.

- *Principles:* represent the highest level of learning. Principles define general truths or laws that are basic to other truths.

Computer systems are proficient at learning 'concepts'. Concepts can be equated to the output of data mining. The form of learned concepts is dictated by the data mining technique and the nature of the data/objective. However, pivotal to the ability for data mining tools to learn, is the concept of machine learning. Machine learning is field that has benefited form a vast amount of research. Historical data and inputs can be utilised to provide induction-based models from which computer systems are able to learn and apply to future raw datasets for integration with data mining (Ma, 2007). An advantage of machine learning methods is that they are able to efficiently analyse datasets that contain noisy and missing values. Data mining employs machine learning methods in a practical, rather than theoretical manner for application to real-world problems (Witten & Frank, 2005). Successful data mining, therefore, relies upon the integration of a number of techniques from a variety of disciplines such as database and data warehouse technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualisation information retrieval, image and signal processing and spatial or temporal data analysis (Han & Kamber, 2006).

Data mining strategies can be divided into two broad models Predictive models and Descriptive models (see figure 2.5). Predictive models (not to be confused with the data mining technique prediction) use historical data to analyse current dataset. Descriptive models examine attributes or input variables to identify relationships within the dataset (Dunham, 2003).

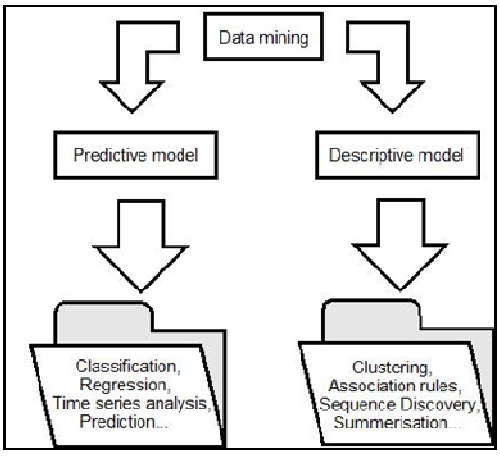


Figure 2.5: Models and tasks of data mining.

Consequently, data mining does not provide a single technique that will accomplish all objectives. Data mining provides a number of techniques whose applicability will depend upon the nature of the task, for example there are algorithms that reveal different types of patterns (Trajkovski et al, 2006).

Each of these algorithms relies upon being able to categorise and represent the knowledge in a manner that will enable the exploration and analysis of data. The technique that is most suited to an objective is dependent upon the requirements of the analysis, in addition to the nature of the raw data (Berry & Linoff, 2004). Typical tasks for data mining are the identification of classes (Clustering), assigning future instances/object to classes (Classification) the prediction of new, unknown objects/instances that are by nature continuous (Estimation), or prediction of future outcomes (Prediction) and the discovery of associations or deviations hidden within databases/datasets (Association Rule Mining) (Geist, 2002). Tasks involving predictive models, such as; Classification, Estimation and Prediction, are examples of directed data mining. In directed data mining, the goal is to find the value of a particular variable. Directed data mining techniques use historical models to build learning models through which to analyse current datasets and are therefore, thought of as 'Supervised' techniques. In contrast, 'Association Rule Mining' and 'Clustering' are examples of undirected data mining where the goal is to find and uncover structures in data without respect to a particular target variable. Undirected data mining is generally considered 'Unsupervised'.

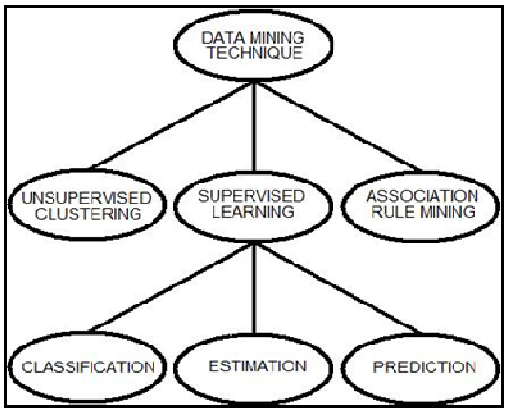


Figure 2.6: Hierarchy of data mining techniques.

This is not an exhaustive list of data mining techniques, however, these techniques are those that are most commonly investigated for the predicaments that organisations face when mining corporate data (Nemati, 2002; Klösgen & Zytkow, 2005). These techniques enable organisations to investigate, analyse and describe their data. These data mining techniques can be categorised according to their objectives in a hierarchy as depicted in figure 2.6 (Roiger & Geatz, 2003; Zhang & Zhou, 2004). These techniques and the algorithms they employ have been explored in greater detail in Appendix A. BI however, does not merely consist of various data mining and analytical models, in contrast there are various other BI techniques such as Intelligent Agents which can be explored to provide models for prediction and resources allocation.

2.4 Intelligent Agents

As discussed, there has been a stark increase in the amount of information organisations are storing. However, along with information and data, due to the advances in telecommunications, organisations are processing an increasing number of transactions in addition to accruing a greater amount of resources. These resources must be efficiently managed, even in the event that they are physically distributed over large distances. Consequently, organisations are discovering the necessity to not only to process a greater volume of commercial transactions over networks, but additionally, the need for co-ordinating and organising a large quantity of resources. Hence, there is an ever-growing requirement for smart *'intelligent agents'* that autonomously perform specific actions.

Intelligent agents are elements of software that perform tasks on behalf of a user without any assistance by making choices. These tasks are often repetitive and mundane. They are performed on behalf of a user to reach a conclusion of provide a suggestion. The conclusions or suggestions are deliberated through the decisions that an agent makes. Decisions are based on rules that software developers have investigated and programmed into the agents design. Another key characteristic of intelligent agents is their ability to learn. Intelligent agents can retain information every time they perform an action enabling them to perform more efficiently every time they perform a function they have previously completed (Curtis & Cobham, 2008). Thus, an intelligent agent is capable of operating independently with well-defined goals, and designated tasks by human operators. Intelligent agents are generally thought of as intelligent, due to their ability to adapt to environments and designated tasks based upon received information. Intelligent agents enhance the capabilities of applications and help users achieve tasks. Intelligent agents can be categorised into two categories, based upon their activities (Abraham et al, 2004):

- *Closed Agent*: A closed agent operates within a specific limited domain. An example of this is a computers 'system agent', which on behalf of a user performs task such as monitor the performance of the CPU, disk fragmentation and security status.
- *Open Agent*: An open agent is one which can interface with other agents and work co-operatively. These agents are capable of performing tasks across wider networks, such as multiple computers or the internet.

Intelligent agents are generally classed through the nature of their actions could be further classified in to several categories:

- *Collaborative*: These agents co-operate or compete with other intelligent agents.
- *Interface*: These agents communicate with a human user. Through some form of Graphical User Interface (GUI), these agents will communicate with users to collect requirements or report results.

- *Mobile*: These agents are able to travel across networks to communicate with other agents and perform actions.
- *Reactive*: These agents are able to interact with their environment and alter the state of the domain. An example of such an agent is anti-virus software that can detect and delete threats as soon as they enter a system.
- *Smart / Hybrid*: These agents contain a combination of traits from the categories that have been defined.

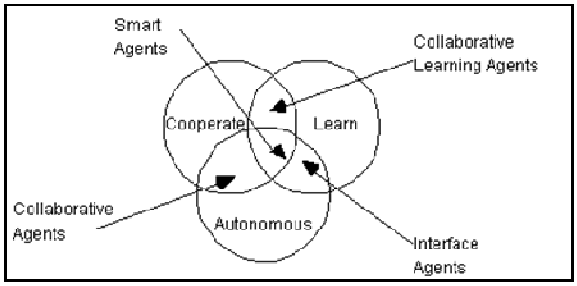


Figure 2.7: A part view of an agent typology.

Figure 2.7 depicts a part view typology of some of these categories and the properties that the various types of agents encapsulate. In the weak notion of agency, agents have their own will (*autonomy*), they are able to interact with each other (*social ability*), they respond to stimulus (*reactivity*), and they take initiative (*pro-activity*). In the strong notion of agency the weak notions of agency are preserved, in addition to which, agents are portable (*mobility*), truthful (*veracity*), follow instructions (*benevolence*), and agents will perform in an optimal manner to achieve goals (*rationality*). Some basic examples of software agents that are currently in widespread use are (Tveit, 2001):

- The animated paperclip agent in Microsoft Office.
- Computer viruses (also referred to as destructive agents).
- Artificial players or actors in computer games and simulations (e.g. Quake, Perfect Dark, Half-Life etc).
- Trading and negotiation agents (e.g. the automatic bid placing agent at E-Bay).
- Web spiders that collect data to build indexes used by a search engines.

Although agents appear in various guises, all intelligent agents share certain characteristics. The predominant of these is autonomy. Despite a certain level of human interaction being necessary, intelligent agents are able to perform tasks, even if limited, on behalf of the user and as a consequence display a certain degree of intelligence. This intelligence may be the ability to make decisions upon which course of action is most suitable to satisfy the agents objectives or be able to respond to a change or alter a characteristic in its environment. The actions of an intelligent agent are defined by its Beliefs, Desires and Intentions (BDI). 'Beliefs' refers to the knowledge that an agent possesses with

regard to its domain and status, 'Desires' encode the agents goals, whereas 'Intentions' define the actions that an agent must perform to satisfy the objectives (Krümpelmann et al, 2008). Furthermore, an intelligent agent should be able to communicate with other agents. Applications that consists of a number of agents each performing limited tasks, yet able to communicate with each other are known as Multi-Agent systems.

2.4.1 Multi-Agent Systems

Multi-Agent systems consist of a collection of intelligent software agents. These agents are capable of accomplishing their tasks through interaction with other intelligent agents or humans. Although individually performing limited tasks, the sum of the interactions amongst the agents is a highly intelligent system. Multi-Agent systems are considered a system with autonomous agents functioning simultaneously and attempting to achieve a goal or to fulfil/satisfy a function or condition. As a result, intelligent agents possess a high-level communication protocols and mechanisms that enable the agent to interact with its environment in addition to other intelligent agents (Farber, 1999). A Multi-Agent system resolves specific issues or specified tasks through negotiation among agents. Consequently, it is the effectiveness of this interaction that determines the effectiveness of the system, rather than sole performance of each agent (Jars et al, 2004). Certain characteristics which define Multi-Agent systems, can be found in expert systems, since expert systems are largely autonomous, and encapsulate the knowledge of a domain. However, expert systems typically, do not co-operate or encompass the ability to learn through experience (Kendall et al, 2000). Just as Multi-Agent systems shares characteristics with an expert system, parallels can be drawn between Multi-Agent systems and object-oriented programming. Thus, it is imperative to investigate the features that distinguish Multi-Agent systems and object-oriented programs.

Multi-Agent systems differ from object-oriented programs, due to the fact that in an object- oriented program an object must perform its function when invoked. Yet, within a Multi-Agent system an agent can refuse as their objectives are based upon their Beliefs, Desires and Intentions (BDI). Hence, the decision of whether to execute an action lies with object that invokes the method (action) in an object-oriented program, in contrast within a Multi-Agent system the decision of whether to act lies with the agent, as a result the agent is able to exhibit a degree of control over their behaviour (Lomuscio & Raimondi, 2006). Table 2.1 proposed by Lin (2007) summarises major differences between an agent-oriented approach and object-oriented approach.

Basic unit	Agent	Object
Unit state	Mental components	Unconstrained
Communication paradigm	Peer-to-peer	Client-server
Communication mode	Message passing	Message passing

Communication type	Local (mobile) + remote (static)	Mostly remote
API	Uniform method call	Unconstrained
Method constraints	Honesty, consistency, etc.	None
Message Type	Speech Acts	Unconstrained
Mobility	Autonomy and mobility-related meta data	No autonomy or mobility related meta data
Inheritance	Metal states	Methods and attributes
intelligence	Intelligent operations	Not always present

Table 2.1: Differences between Agent-oriented programming and Object-oriented programming.

Intelligent agents rely upon an 'ontology' to encode their BDI. The main function of an ontology is to enable communication amongst different systems in a way that is independent from the individual systems technology, information architecture and application domain. The key component of an ontology is a vocabulary of basic terms and precise definitions. However, an ontology is more than an agreed vocabulary; it provides fundamental constructs that are leveraged to build higher-level knowledge. The component terms must be selected with great care, ensuring that the most basic foundational concepts are defined and specified. In addition to a taxonomy or classification of terms and contributing to the semantics, an ontology also includes the relationships between these terms. It is the relationships that enable the expression of domain-specific knowledge. An ontology enables agent-based systems to simultaneously 'interoperate without misunderstanding' and 'retain a high degree of autonomy, flexibility and agility' (Wooldridge, 2002).

Multi-Agent systems have been successfully investigated for computer games (Kobti & Sharma, 2007), co-ordinated defence system (Reichel, 2008), transportation (Miao et al, 2006), geographic information systems (Gervais, et al, 2007), in addition to various other application domains. The dynamic and adaptive capabilities of Multi-Agent systems, has made them an ideal technology for investigation within BI. As a result, there has been significant research into the business applications of Multi-Agent systems. Consequently, it is widely being advocated for agent systems to be investigated to achieve automatic and dynamic load balancing, high scalability, error detection and recovery (especially self healing networks), manufacturing, supply chain management and logistics amongst other applications (Lin, 2007). A major problem facing manufacturing organisations is how to provide efficient and cost-effective responses to the unpredictable changes taking place in a global market. This problem is further convoluted by the supply chain networks. Conventional solutions such as manufacturing execution systems, supply chain management systems and enterprise resource planning (ERP) systems do not provide adequate facilities for addressing this problem. However, augmenting conventional approaches to planning/management systems, with Multi-Agent systems and BI techniques can provide an innovative technique for resolving dynamic problems such as this (Zhang et al, 2006).

However, to successfully resolve dynamic problems an intelligent agent within a Multi-Agent system must be capable of autonomous, flexible action in order to meet its design objectives. As a result, intelligent agents are (Singh & Huhns, 2005):

- *Pro-activeness*: Intelligent agents are goal orientated. In addition they are capable of 'taking the initiative' to direct their actions in order to satisfy these goals.
- *Reactivity*: Intelligent agents are aware of their environment and capable of responding to any changes in order to satisfy their goals and objectives.
- *Social ability*: Intelligent agents are capable of interaction. This interaction can be amongst agents, enabling them to collaborate, compete or learn in order to satisfy objectives. Or this interaction may take place between agents and users.

Within Multi-Agent systems, each intelligent agent performs a limited task, with the sum of these tasks providing an intelligent system. The performance of a Multi-Agent system, therefore, relies upon the sum of the performance of the constituent agents. Thus, efficient communication between the agents is imperative.

2.4.2 Intelligent Agent Communication

Communication forms the basis upon which interactions in a social organisation are able to take place. Multi-Agent systems can be considered as social organisations, as it is the collective sum of the agent interactions that enables the system to satisfy its objectives. Consequently, within a Multi-Agent system there will be a degree of communication (bargaining) between agents, in addition to the communication which occurs between agents and users. This communication process can be summarised by applying a template to agent negotiation architecture (Rzevski, 2002):

- Resource Agents greet 'resource owners' and help them register or log in. Collect data on their preferences, to discover and store any knowledge useful for satisfying expectations.
- Client Agents greet 'clients' and help them register or log in. Collect data on client preferences to discover and store knowledge that may be useful for satisfying expectations.
- Client Agents, send message to Demand Agents, informing them of resources that are required and the amount of notional monetary units they are prepared to pay for resources.
- Demand Agents send messages to Resource Agents describing requirements, stating the amount of notional monetary units they are prepared to pay for resources.
- Resource Agents send messages to Demand Agents describing the resources that are on offer and the amount the resource costs.
- Demand Agents and Resource Agents negotiate offers and make deals.

- In cases where a full demand-resource matching is not possible, agents accept a partial matching after consulting clients and/or resources owners.
- Payment for partial matching is less than the payment for full matching.
- When new demand or resource arrive, agents try to re-negotiate previously agreed partial matching, offering compensation to agents agreeing to release purchases to another bidder.
- A deal is only agreed if the total value of transactions increases.
- In stable situations agents negotiate and re-negotiate deals until the optimal distribution of resources to demands is achieved (when further increase in the value of transactions is unachievable). In situations categorised by frequent changes in demand/supply conditions, when there is no time to work out the optimum: agents' attempt to achieving a satisfactory level of allocation.
- Demand Agents compete among themselves, as do Resource Agents. However, agents can switch from competition to collaboration, if such an alteration would improve performance.

The described template represents a typical structure of the agent communication process. Multi-Agent systems are thus able to accomplish task due to constituent agents encapsulating the ability to communicate. Agents are able to cooperate, coordinate their actions and carry out tasks jointly, displaying emergent behaviour, else the agent would become locked into a perception-deliberation-action loop. Communication, therefore, forms the basis for the operation of a Multi-Agent system, and is generally expressed as a form of interaction. The dynamic relationship between agents is expressed through the intermediary of mediators (i.e. signals) once interpreted inform the agents of the required actions (Farber, 1999). There are a number of theories of communication. Many of these theories have been formed upon variations of the 'theory of communication'. The theory of communication emerged from telecommunications research Shannon & Weaver in the 1940s. This model consists of the exchange of information between the sender and the receiver, known as an addressee. The information is transmitted via a channel and is encoded upon being transmitted, through a language and decoded by the addressee upon receipt. In agent communication this model can be adopted as agent communication is defined as the intentional exchange of information based upon a shared system of signs (Weiss, 1999). However, for Multi-Agent systems to operate efficiently it is paramount that they have a basis upon which these agents can understand and encode messages, send messages, and decode these signs and messages.

2.4.3 The Foundation for Intelligent Physical Agents (FIPA)

In order to produce a set of standards for heterogeneous and interacting agents and agent-based systems, the Foundation for Intelligent Physical Agents (FIPA) was formed in 1996. FIPA is an Institute of Electrical and Electronic Engineers (IEEE) standards committee that promotes agent-based

technology and interoperability between agent-based systems developed independently by different companies, organisations and researchers. To enhance interoperability FIPA provides specifications for agent platform architectures to support communicating agents, communication languages and content languages for expressing messages and interaction protocols for not only individual messages but complete transactions (Lin, 2007). The 'FIPA 97 standards' were the first set of FIPA specifications, which set out to promote commercial implementation and limited interoperability of agent-based applications. The FIPA 97 standards addressed three major areas:

- Agent Management.
- Agent Communications.
- Agent Software Integration.

In order to be implemented the FIPA 97 specifications required the COBRA IIOP (Common Object Request Broker Architecture Internet Inter-ORB Protocols) as the communication protocol, a speech act based agent-specific communication language, and a set of interaction protocols. These protocols enable dialogues and negotiations to be defined among agents. The FIPA 97 specifications contained several weak areas that effected reliable communication. Furthermore, FIPA 97 did not fulfil the necessity for technology-independent specifications. This was essential due to the rise of Java RMI and HTTP as alternatives to COBRAs' IIOP communication protocol (Bigus & Bigus, 2001).

To address the weaknesses of FIPA 97, a revised set of specifications were proposed in 2000. FIPA 2000 specifications embody several years of work, and have been contributed to from major technology companies including Sun, IBM, Fajitsu, HP, Toshiba and BT to name a few, in addition to contributions from major universities the world over. The major inclusion in FIPA 2000 standards is the definition of an abstract architecture that permits alternative implementations that interoperate. FIPA has also published agent standards and official reference documents, which have been checked through practical application and validated by the FACTS (FIPA Agent Communication Technologies and Services) EU project. The centrepiece of FIPAs efforts has been the development of a standardised agent communication language (Wooldridge, 2002).

2.4.4 FIPA-Agent Communication Language (FIPA-ACL)

Virtually any language can be used to create a Multi-Agent system, such as Java, Telescript, or C++, however these languages are not recognised as an agent communication language (ACL) since their primary function is not to facilitate agent communication. An ACL has the primary function of describing and supporting communication between communicative entities such as intelligent agents (Feng & Lu, 2003). Agent-based applications developed in conventional object-oriented languages must implement a specific ACL for communication. For this purpose, FIPA has investigated and

developed the FIPA-ACL in an effort to standardise the language that agents use to encode the messages that they exchange, thereby, increasing the interoperability of agent systems (Fornora & Colombetti, 2002).

Based upon the speech acts theory developed by Searle in 1969, FIPA-ACL defines an 'outer language'. Since the messages are communicative acts (actions), the exchange of these messages is intended to perform some action (Charlton et al, 2000). The syntax and structure for FIPA-ACL messages resembles that of KQML (Knowledge Query and Manipulation Language: a language and protocol for exchanging information and knowledge proposed by the ARPA Knowledge Sharing Effort[1]). The typical structure element of a FIPA-ACL message is illustrated in code-table 2.1(Wooldridge, 2002):

	ACL message format	Description
1	(<performative>:	Name of communicative act
2	:sender	Name of agent sending message
3	:receiver	Name of agent receiving message
4	:content	Proposition, action or combination of two
5	:language	Language used in content (e.g. SL, KIF, Prolog, ...)
6	:ontology	Ontology used in content (e.g. fipa-pta)
7	:reply-with	Subject
8	:in-reply-to	Re: Subject
9	:conversation-id	Identification of current dialogue
10	:reply-by	Deadline for latest reply
11	:protocol	Interaction protocol used
12	:envelope	Requirements on message transport layer
13	)	

Code-table 2.1: FIPA-ACL message structure.

The collection of performatives that are provided is the major characteristic that distinguishes KQML and FIPA-ACL. FIPA-ACL specifies more twenty performatives, these performatives can be classified into five categories (Fornora & Colombetti, 2002) as illustrated in table 2.2. The most frequent criticism of KQML is the lack of adequate semantics. The lack of adequate semantics was a factor that motivated the development FIPA-ACL. Consequently, the FIPA-ACL has been specified with a comprehensive list of formal semantics. The approach adopted during the development of FIPA-ACL was founded in Cohen & Levesques' (1990) theory of speech acts as relational action, and Bretier & Sadek's (1997) research and development of this theory. The semantics of FIPA-ACL are

[1] http://www.cs.umbc.edu/kqml/: Accessed August, 2008.

based upon a formal language 'SL'. SL enables the actions, beliefs, desires and uncertain beliefs of agents to be represented. This is achieved via the semantics mapping of all FIPA-ACL messages to a SL formula. This SL formula is referred to by FIPA as the 'feasibility condition'.

Passing Information	Confirm, Unconfirm, Inform, Inform-If, Inform-Ref.
Requesting Information	Cancel, Query-If, Query-Ref, Subscribe.
Negotiation	Accept-Proposal, Cfp, Proposal, Reject-Proposal.
Performing action	Agree, Cancel, Refuse, Request, Request-When, Request-Whenever.
Error-handling	Failure, Not-Understood.

Table 2.2: FIPA-ACL performatives.

The feasibility condition is a constraint that the sender of the message must satisfy. In addition to this, the semantics of FIPA-ACL further map each message to an SL-formula that defines the 'rational effect', specifying the purpose of the action taken by an agent. Given that an agent is regarded as an autonomous entity, the relational effect of the message is not regarded as guaranteed, since the purpose can be subject to change. As a result FIPA-ACL requires that the recipient only acknowledge the feasibility condition aspect of the message (Wooldridge, 2002).Within FIPA-ACL, 'inform' and 'request' are the two most important communication primitives. These communication primitives form the basis for all performatives within FIPA-ACL (Helin, 2003). Through the introduction of a set of standards and specifications for agent communication FIPA endeavours to increase interoperability between agent-based systems that are independently developed, enabling agent technology to gain more widespread appeal and attain a true commercial status. Furthermore, this will provide an additional benefit for BI applications. If agent-based BI systems are developed to FIPA-compliant specifications these systems will not only be able resolve internal issues, but interoperate with Multi-Agent systems which have been deployed by clients and/or suppliers. However, Multi-Agent systems,

in addition to data mining and other BI techniques, are only effective if they can provide valuable support for decision makers. It is the ability to make exceptional and valuable decisions augmented by BI techniques that will ultimately provide organisations with BI tools that can increase the competitive performance of an organisation.

2.5 Decision Making

This chapter has reviewed a number of models and techniques for extracting and analysing data, in addition to intelligent agent systems that can be investigated to assign resources. However, for these technologies to be effectively integrated into business processes the BI tools must be able to provide this information to the appropriate type of user or department, be it a manager, HRM department, or any other user within an organisation in a format they are able to understand. This requires accurate feedback, and the capability to make effective decisions. Consequently, decision making plays an integral role in many business processes. Decisions whether *strategic*, *structured* or *dependent* can aid management in determine courses of actions that will provide the greatest benefit to an organisation (Papadakis & Barwise, 2002; Hitt & Collins, 2007):

- *Strategic decisions* are those that dictate long term planning. When making strategic decisions it is essential the decision maker considers all variables, such as environmental, technological, and socio-economic, since these variables are subject to change when planning long-term objectives.
- *Structured decisions* are those that are derived on the basis of a specific business objective.
- *Dependent decisions* are those that are taken based upon the indicators of past business performance or historical data.

Thus, decision making can be defined as *"the ability to process and analyse information and knowledge in order to select and implement the appropriate course of action"* (Nemati, 2002). Consequently, decision making is not only essential for business processes, rather can often determine the success of or failure of business ventures. Herbert Simon and associates identify a number of similarities between decision-making and problem solving, since not only are both processes integral to business activities, but share common a common process model (Turban & Aronson, 2001). In problem solving or decision making it is essential to, observe, recognise and understand the problems or objectives. Once this initial phase has been completed the decision maker or problem-solver must determine the options that are available, evaluate the options and any alternatives, prior to choosing and implanting an option that satisfies the objectives or resolves the dilemma. These phases are depicted in figure 2.8. This model can aid decision makers to reach conclusions or solve problems, however represents a broad guidance to the process, without supplemental details on selecting the most suitable course of action (Goa et al, 2007).

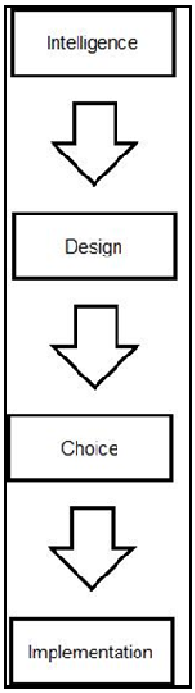


Figure 2.8: Herbert Simon's decision process model.

The conventional approach that is adopted for selecting the most suitable action is the SWOT analysis. SWOT is an acronym for Strengths, Weakness, Opportunities and Threats. The model developed by Ken Andrews in the early 1970s has proven to be successful for decision makers in business processes (Baker, 2000). The SWOT model supports the decision process by enabling the decision maker to identify the factors that must be considered. The SWOT matrix (figure 2.9) can be investigated to identify (Hay & Castilla, 2006):

- *Strengths*: The factors internal to the organisation, which bear a positive effect and can, support the course of action.
- *Weakness*: The factors internal to the organisation, which bear a negative effect and may act as an obstacle.
- *Opportunities*: Factors external to the organisation, which bear a positive effect and can, support the course of action.
- *Threats*: Factors external to the organisation, which bear a negative effect and may act as an obstacle.

	Positive effects	**Negative effects**
Internal Factors	Strengths	Weakness
External factors	Opportunities	Threats

Figure 2.9: SWOT matrix

SWOT models can, therefore, support a decision maker in identifying all key factors that must be considered in the decision process and formulate decisions that can be beneficial to the performance of an organisation. Another acronym that can be utilised is 'USED' (Use, Stop, Exploit and Defend). *Strengths* must be 'Used', *Weaknesses* must be 'Stopped', whilst *Opportunities* should be 'Exploited', and care taken to ensure *Threats* are 'Defended' against by the organisation (Hay & Castilla, 2006). Although models such as SWOT can be investigated to support the capabilities of decision makers, the capabilities can also be augmented with automation or integration with Decision Support Systems (DSS).

2.6 Decision Support Systems (DSS)

Since the introduction of DSS in the 1970s, decision makers in a number of fields have employed DSS to aide in critical decisions (Arnott & Pervan, 2008). However, DSS are becoming increasingly more critical in the daily operation of organisations (Nemati et al, 2002). DSS are interactive, computer-based systems intended to provide support to decision makers within an organisation who are engaged in solving various complex problems that involve multiple attributes, objectives and goals. Thus DSS can be defined as an interactive computer-based system to aide decision makers utilise data and models to identify problems, solve problems and reach the most efficient decision. DSS integrate both data and business models to assist the decision making processes (Rupnik et al, 2006). It should be noted that DSS are not automated systems; rather they are intended to provide support to the decision maker, and as a result, not intended to replace decision-makers, yet rather, improve the effectiveness and quality of decisions (Mladeni et al, 2002). The early DSS that information technologists investigated in the mid-1970s were designed to augment the capabilities of transaction processing systems that 'crunched numbers' and kept accurate accounts of data and resources. During this period organisations became more knowledge oriented. As a result greater emphasis was placed upon the benefits that could be reaped by the analysis and interpretation of information, vicariously creating a demand for interactive DSS capable of providing interactive support to the user (Dhar & Stein, 1997).

The design of early DSS was by nature rudimentary, providing users with software based upon (although themselves, novel at the time) spreadsheet software. These early systems were followed by DSS based upon models from Management Information Science (MIS), Operations Research (OR) and Management Science (MS). Figure 2.10 depicts how disciplines such as MIS, MS and OR were incorporated into Hubert Simons decision model (figure 2.9) (Sprague et al, 1996). DSS integrated these disciplines by incorporating linear programming techniques with user friendly interfaces, these early DSS enabled the user to investigate various 'what-if' options, rather then providing a single 'best' solution, thus, allowing for the exploration of a greater number of possibilities, whilst providing

far greater levels of interactivity and decision support (Turban & Aronson, 2001; Campbell et al, 2006).

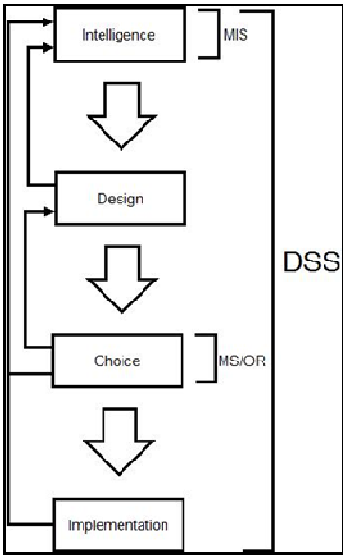


Figure 2.10: Herbert Simon's decision process model extended for DSS.

The development of DSS, although gradual, has moved from the early spreadsheet based DSS toward the more conventional DSS found today that are based upon artificial intelligence techniques. There are various design taxonomies that can be investigated to classify DSS. However, not all DSS can be classed as under a particular model, in contrast they are often a hybrid of models. DSS are generally classed according to the method of assistance they provide to the user. The categories under which DSS can be classed as are (Power, 2002):

- *Model-driven DSS*: emphasise access to and manipulation of a statistical, financial, optimization, or simulation model. Model-driven DSS use data and parameters provided by users to assist the decision maker to analyse a situation; they are not necessarily data-intensive systems. Dicodess is an example of an open source model-driven DSS generator (Gachet, 2004).
- *Communication-driven DSS*: support more than one person working on a shared task; examples include integrated tools such as Microsoft's NetMeeting or Groove (Stanhope, 2002).
- *Data-driven DSS* or *Data-oriented DSS*: emphasise access to and manipulation of a time series of internal company data the systems however can be extended to support external data.
- *Document-driven DSS*: manages, retrieves, and manipulates unstructured information in a variety of electronic formats providing the user with access to these sources.
- *Knowledge-driven DSS*: provide specialised problem-solving expertise stored as facts, rules, procedures, or in similar structures, these systems are similar to 'expert systems'.

According to Michalewicz, et al (2005); Guerlain et al (2000); Zhang, et al (2007); and Zhou et al (2008), DSS even those based upon artificial intelligence models and BI tools, considered intelligent decision support systems have failed to integrate effectively and require far greater research for effective integration. Considering the capabilities of both data mining and DSS, the two approaches can be integrated to better solve data analysis and decision-making. In knowledge management, such integration is interesting for several reasons (Wei et al, 2001). For example, in data mining it is often unclear which algorithm is best suited for the problem. This decision support process could be extended to the initial phase of data collection, thus when data is collected decisions models could be developed to describe the data and ensure that the correct data is collected, thereby improving the quality of results when data is analysed, furthermore, this data could be presented via a digital dashboard. BI tools that can analyse data can be used for predictive maintenance and ultimately within a DSS which could identify fault patterns before the physical fault occurs, alert suitable human operators and have degrading components replaced before a critical failure which would result in unit downtime. These are two examples of how data mining and DSS can be integrated to improve the quality of BI systems (Mladenic´ et al, 2003), it would be of even greater significance if such models, could de developed to handle extremely large datasets, with the results provided via visual BI models to support the decision maker. Academics and practitioners have investigated the architecture of DSS in terms of four major components (Power, 2002):

- The user interface.
- The database.
- The model and analytical tools.
- The DSS architecture and network.

The integration of the DSS architecture with BI systems will increase the capabilities and support offered to the user, in addition to providing organisations with the competitive edge that necessitates such systems. However, it is imperative that there be further investigation in to the frameworks that can aid the integration, design and deployment of these systems. Such a framework will facilitate the integration of these two technologies. Allowing decision-makers access too, in addition to tools that can analyse these vast stores of data that companies are amassing. Furthermore, such DSS will be able to provide users with intelligent decision support (Wei, 2001; Fong et al, 2002; Campbell et al, 2006).

2.7 Summary

This chapter has investigated the tools and techniques utilised to capture, extract and analyse data. As discussed increasing standards, automation, and technologies have led to vast amounts of data

becoming available. Data warehouse technologies have facilitated the establishment of repositories to store this data. Improved BI tools have increased the speed of collecting the data. QRA tools enable faster generation of new reports and analysis of data at a shallow level. Advanced analytic tools and techniques such as data mining have taken these analysis capabilities to the next level enabling organisations to investigate data for hidden patterns, trends and correlations. This chapter has also reviewed some of the data mining techniques in addition to systems based upon intelligent agents that enable these benefits to be exploited and provide the organisations with a true competitive edge. However, for this to be of any real value it is essential that this analysis is optimal and can be utilised, and provided to the correct departments or individuals within an organisation in a format that can provide this competitive edge. Thus for BI tools to be of real value to an organisation it is imperative that they be integrated with DSS and provide organisations with the means through which to analyse and use data or resources efficiently.

The literature which has been explored within this chapter provide a background to BI, DSS and the decision making process. The motivation for exploring these topics is to provide an understanding of the inherent complexities which are involved within BI. As a result, the integration of BI within DSS is a complex procedure which can result in ineffective integration. Bearing in mind the complexities of the techniques which underpin BI and due to multiple techniques which can be applied with a particular situation, it is imperative to research a suitable meta-level framework which can not only be explored to provide structure to the process of integrating BI within business processes, yet furthermore provide a systematic process for selecting the most suitable technique thereby providing the most efficient solution. Since, the integration of BI tools with DSS has until now been of a sub-optimal quality (Michalewicz, et al, 2005; Guerlain et al, 2000; Wei et al, 2001; Zhang, et al, 2007; Zhou et al 2008). It has been discussed that the integration of BI within the business model is a major area of interest and investment for many organisations. Thus, it is imperative that the framework proposed by this research be capable of not only facilitating knowledge discovery within BI with a view to providing intelligent decision support, yet further provide a structured approach, consequently permitting the experiences and knowledge gained from a particular integration project to be transferrable.

Chapter 3:

Methodology

The previous chapters have highlighted many of the related issues that need to be addressed by this study. It will be the aim of this chapter to provide an outline to the methods that will be used to conduct this research and ensure the integrity of the conclusions. This chapter will commence with an investigation into the research model that will be adopted for this research, in addition to reviewing the approaches that were considered, however deemed unsuitable for the requirements of this research. The chapter will then conclude with an examination of the epistemology and methodology that will substantiate the results and contribution to knowledge of this research.

3.1 Methodology: Systems Engineering

In the previous chapter it has been discussed that Business intelligence (BI) is a major area of investment for organisations in number of industries (Ortiz, 2002; Forsling, 2007; Watson & Wixom, 2007). Amalgamating a number of disciplines, the focus of BI solutions has been on analysing information within a company and presenting this information to users to provide a competitive edge. However, there are key problems within BI, primary of which is ensuring the availability of information for BI tools. This issue is the result of a lack of incentive for the sharing of information throughout an organisation. Furthermore, the required information may not always be instantly evident. Another key issue is the poor integration of BI and Decision Support Systems (DSS). Both these issues are the result of correct, valid, integrated and in-time data being unavailable, in addition to the means through which data can be transformed for decision information failing to perform optimally. These issues can however, be resolved through the investigation of a framework which can facilitate an organisation in exploring an optimal BI solution (Xu et al, 2008). Currently however, for BI there is no structured framework for investigating BI based solutions, in particular one that can provide intelligent decision support. Exploration of BI solutions is an ad-hoc process with consultants investigating and providing organisations with these BI solutions on a case-by-case basis. If BI is to reach its true potential then it is imperative that a framework be investigated (Landqvist & Pessi, 2004; Xu et al, 2008).

A framework (within research) can be defined as a means through which to outline possible courses of action or present an approach to a systems analysis project. A common approach to frameworks is an amalgamation of planned or existing behaviours, functions, relationships and objects that address specific application domains and support the development of end-user applications and products directly (Abi-Antoun, 2007). More comprehensive approaches also provide procedure models, in which all developing and maintaining activities are defined (Leist & Zellner, 2006). In their study to understand frameworks, Schwarz et al (2007) identify that the purpose of a framework is to:

1. *Integrate previous research studies*: The ‘data’ for framework articles is discovered in the form of previously published research articles. After collecting the data, a researcher identifies distinct, but related research streams, subsequently integrating these streams. As a result, an output of this process is a cohesive model or table that unifies the separate research streams in a domain, based upon a selective bibliography or list of previous studies.
2. *Assist in the development and testing of theories*: A framework assists researchers to theorise about a phenomenon. Using the ‘data’ a framework raises issues found within the literature and highlights questions that conventional theory has not yet answered. In doing so, a framework becomes an input to the development and testing of a theory.

Consequently, the fundamental objective of a framework is to formalise the approach and in doing so, identify new methods and research opportunities, often drawing from experiences in similar or related fields. Thus a parallel can be drawn between a framework and relational model or system (Curtis & Cobham, 2008). The oxford dictionary defines a system as (Delahunty & McDonald, 2007, p. 699):

"System /'sistəm/ n. complex whole, set of connected things or parts, organised body of things…method, considered principles of procedure or classification; orderliness."

Hence, a framework can be thought of as a system (Fowler, 1998). However, just as a system must adhere to a methodology to ensure that it is valid and of sound design, it is essential that a framework does too. A methodology is a system or plan that is adhered to achieve an objective, since it provides structure and integrity to the process. It is thus, imperative that all research follows a research methodology; thereby enabling the researcher to structure the research (Anderson & Rönnbom, 1999). One methodology that can be adopted for this research is that of systems engineering. The International Council on Systems Engineering Homepage (http://www.incose.org) defines systems engineering as:

"…a discipline whose responsibility is creating and executing an interdisciplinary process to ensure that the customer and stakeholder's needs are satisfied in a high quality, trustworthy, cost efficient and schedule compliant manner throughout a system's entire life cycle. The process is usually comprised of the following seven tasks: State the problem, Investigate alternatives, Model the systems, Integrate, Launch the system, Assess performance and Re-evaluate."

Engineering can, therefore, be employed as a metaphor within software development (Patel, 2005), accordingly systems engineering can provide an effective methodology through which to investigate a framework. In the context of systems engineering, a system can be divided into two categories a; 'Hard systems approach' or 'Soft systems approach'. The soft systems approach should not be confused with 'Soft computing', which is a term that refers to a collection of computational techniques in computer science, machine learning and some engineering disciplines. In contrast the soft systems approach is a popular methodology for examining 'ill defined' systems. Ill defined systems do not necessarily have obvious boundaries or interaction between the elements that compose the system (Yinghong, 2007). The soft systems approach, formalised by Peter Checkland (1999), places emphasis upon user interactions and the human element of a system, they are therefore suitable for problems that are less technical and more reliant upon the role or knowledge that is contained within individuals which cannot be easily defined, thus useful for problems which involve tacit knowledge. In contrast, a hard systems approach aims to find the best solution for optimisation for a problem that is structured with clearly defined boundaries (Yinghong, 2007). A hard systems approach

is, therefore, an approach more suitable for investigating areas in which the problem can be defined through explicit knowledge such as BI.

3.1.1 Hard Systems Approach

Hard Systems approaches also known as structured systems approach is generally applied to explicit knowledge problems, where the problem can be defined within given boundaries, and a single optimal solution proposed (Hitchins, 2007). Hard systems approaches have been investigated and successfully applied to information technology and computer/software systems within a number of organisations. The hard systems approach has evolved from traditional approaches not being able to fully address project control failures, difficulty in designing systems that are highly integrated and the significance placed upon data models within databases. There are a number of features that result in the identifiably of a hard systems approach (Curtis & Cobham, 2008):

- A common assumption that there is an identifiable existing system that can be investigated.
- There is a statement of clear and agreed objectives of systems analysis and design. Thus the problems associated with the current system are well-defined.
- The end result shall incorporate a technological solution. Thus the assumption is made that the short-comings of the current system and the objectives of design can be met by a technical solution.
- There is a single, optimum solution. Consequently, similar to the manner in which the objectives can defined, it is possible to determine whether the designed system will be able to fulfil the objectives.
- The process of analysis and design requires an expert. This stems from the belief that the solution will be technical in nature. The client-expert dichotomy is central to a hard systems approach.

The hard systems approaches has, furthermore, spawned various methodologies, such as systems analysis, structured methods, object-oriented approaches, SSADM (Structured Systems Analysis and Design Method), operations research, data analysis, amongst others. Whilst adhering to the features of hard systems, these methodologies share common characteristics in the approach they adopt (Kock, 2006; Broardman & Sauser, 2008):

- *Top-Down design scheme*: initially emphasis is upon investigating general process and functions, and those that are deemed most important. The initial analysis can then be further investigated and each stage fine-tuned to encapsulate intricate details.
- *Logical design issues take precedent over physical design issues*: emphasising the top-down scheme processes, functions and data is initially investigated logically, ensuring the proposed solution is fully investigated and explored prior to any physical implementation.

- *Modelling functions and processes*: analysts concentrate on investigating the fundamental aspects of the functions and processes that are fulfilled by the existing system or approach. This enables the clarification of the current systems benefits and the shortcomings that the new system must address.
- *Data models*: generally an entity-relationship representation schema is adopted for data modelling. The entity-relationship schema describes information needs or the type of information that is to be stored, which is then mapped to an organisations 'logical data model' (representation of an organisations data). The data models ensure that the data within an organisation can be fully investigated and explored.
- *Documentation*: emphasis is placed upon the importance of documenting the stages of the system. Early In development this documentation is often in the form of pictorial diagrams often produced via CASE tools.

Hard systems approaches are similar to a 'general engineering approach to problem solving'. In addition to systems rooted in information technology/management, hard approaches have been successfully investigated within engineering and military applications. Various authors have identified a wide variety of instances in which a hard systems approach is adopted has been adopted (Boardman & Sauser, 2008; Curtis & Cobham, 2008; Hitchins, 2007; Kock, 2006; Kudyba, 2004; Lawrence et al, 2008; Monahan, 2000):

- Critical path analysis or project planning: identifying those processes in a complex project which affect the overall duration of the project.
- Designing the layout of a factory for efficient flow of materials.
- Constructing a telecommunications network at low cost while still guaranteeing QoS (Quality of Service) or QoE (Quality of Experience) if particular connections become very busy or get damaged.
- Road traffic management and 'one way' street allocations i.e. allocation problems.
- Determining the routes of buses so that as few buses are needed as possible.
- Designing the layout of a computer chip to reduce manufacturing time (therefore reducing cost).
- Managing the flow of raw materials and products in a supply chain based on uncertain demand for the finished products.
- Efficient messaging and customer response tactics.
- Automating human-driven operations processes.
- Globalising operations processes in order to take advantage of cheaper materials, labour, land or other productivity inputs.
- Managing freight transportation and delivery systems.

- Scheduling.
- Personnel staffing.
- Manufacturing steps.
- Project tasks.
- Network data traffic: known as queuing models or queuing systems.

Consequently, a hard systems approach has been successfully investigated in various domains for numerous applications. This is primarily as a hard systems approach is defined by clear boundaries. The current approach and expected outcome can be defined as a result the solution can be considered in order to fill the gap created by the current and desired state logically. The process of attaining the desired state can be described by four distinct phases (Curtis & Cobham, 2008):

1. Investigating the existing system in order to gather information and identify the problem.
2. Analysing the existing system and proposing a solution. The proposal can be a systems designed on paper or via CASE tools that can model the system and processes.
3. Implement the solution.
4. Evaluate and assess and the effectiveness of the solution.

Hence, hard systems adopt a scientific approach to problem solving, placing a greater value upon logic and rationality, over intuition and thereby placing less emphasis upon the human element then soft systems methodology. The hard systems approach is suited to the investigation of BI, since just as BI does; the hard systems approach also places a greater emphasis upon explicit knowledge then tacit knowledge. Although not detached from the human element of the system, the hard systems approach considers the human users simply as an entity that uses the system. As a result, it is imperative that the user and context of use be given careful consideration to ensure the proposed solution is able to operate efficiently, yet individual opinion may introduce bias into the design process. This is not to state that the process is devoid of any human element, rather that the approach is far more data centric, considering the needs of the user from a technical rather then individual perspective. This is a key issue when investigating the applicability of a hard systems approach toward a framework for BI. BI is a technology that is focused toward providing organisations with a competitive edge. However, organisations can be very dynamic environments, with human employees frequently changing, be this due to promotion, retirement, new employee. For this reason, BI solutions should be investigated with key objectives defined clearly, rather then on the opinion of a user or tacit knowledge as the user is subject to change. A rational hard systems approach toward investigating a system is preferable since, in the event a new human fulfilling a role changes, the user role from a system perspective will not. Emphasis upon individual users or designing the system based upon tacit knowledge can often result

in difficulties if the individual user changes, and is a criticism of soft systems approaches and knowledge management (Curtis & Cobham, 2008).

Given, BI is a technique that is exploited to investigate, explore and extract explicit information especially that which is hidden. A scientific method such as a hard systems approach is ideal for investigating a framework for enhancing BI capabilities.

3.2 Research Model

A hard systems approach to systems engineering will provide a methodology through which a systematic framework can be developed for enhancing knowledge discovery within BI. Consequently, this research will amalgamate the disciplines of computer science, artificial intelligence and business management amongst others in order to investigate and enable intelligent decision support to be provided to organisations so that the data they collect can be better utilised for a competitive edge.

A research model is an abstraction that can be applied to understand observations and experiences (Osbourne, 2002). The model of an intelligent systems containing its own representation of knowledge about the world (or domain within which they are to operate) has been the lynch pin in the development of the field of artificial intelligence in particular, in the technological development of knowledge based systems (Dasgupta, 1991). Similarly, BI through data contains models and representations of an organisation. However, within BI the focus of research has been on ad-hoc development of tools with little research to address the issues of practical models for designing and investigating BI. (Williams & Williams, 2003). This is the consequence of the inherent complexity that is involved in investigating BI tools, especially through conventional knowledge discovery paradigms such as the KDD process model. The complexity of using these traditional approaches that lack the high-level abstraction mechanisms required for the engineering of BI, has resulted in ad hoc solutions and paradigms being utilised. Despite there being no clear model for the investigation of BI development, it is possible to provide this research with direction through Vidgen & Braa's (1997) triangle. The original model proposed by Vidgen & Braa was focused toward information systems (Johansson, 2004); nevertheless, it is also be applicable to BI. Similar to information systems, BI is also an interdisciplinary research field. Both, BI and information systems necessitate that a number of disciplines be collectively investigated in a systematic manner through computer based systems.

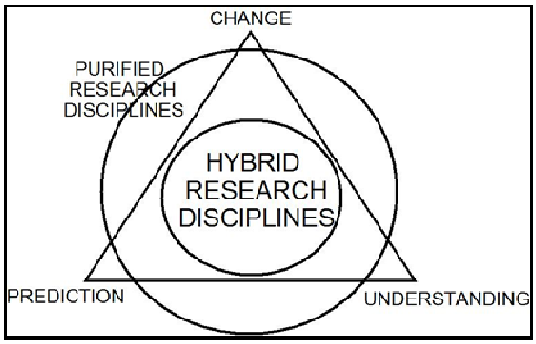


Figure 3.1: Vidgen & Braa's (1997) triangle.

Each apex of Vidgen & Braa's (1997) triangle (figure 3.1) represents a research objective. *Purified research disciplines* apply only one of the research objectives, whereas *hybrid research disciplines* investigate at least two research objectives:

- Change is considered the objective if the investigation is to examine a problem in order to change or alter a situation.
- Understanding enables the researcher to understand the technology and application domain in detail, so that practical problems can be solved accurately.
- Prediction allows the researcher to state a hypothesis that will be confirmed.

Since BI is an interdisciplinary technique and this research aims to investigate BI-based solutions through an innovative approach, the proposed investigation will be a hybrid of 'understanding' and 'prediction' research disciplines (figure 3.1). The research objectives are an understanding of BI, its tools techniques and current methods. This understanding will permit the formulation of a framework. The framework will not be formulated for an investigation in an individual scenario/application domain, but rather, the framework proposed can be investigated in a variety of scenarios, thereby structuring the approach that is adopted. In accordance with Vidgen and Braa's (1997) triangle (figure 3.1) 'prediction' will also underpin the research objectives, since it is predicted that, via a set of objectives, the framework will provide a greater level of support to the user to not only reach business decisions, but more importantly, select the most suitable BI technique. As earlier established a hard systems approach contains four distinct phases; Investigating; Analysing; Implementation; and Evaluation. The format of this research will be as follows:

1. This research will perform an in-depth investigation of the theory underpinning BI and DSS, to identify the opportunities available to facilitate knowledge discovery through a focused meta-level framework.
2. Analysing the existing approaches that are adopted identify any shortcomings.
3. Propose a solution via CASE tools such as UML. Investigating a top-down view of the framework.

4. An implementation of the solution will be carried out through the investigation of case studies.
5. Assess and evaluate the effectiveness of the solution. Consequently, the framework will be tested with a number of datasets of that the performance can be validated.

The four phases of a hard systems approach will form the basis for the research model. Figure 3.2 depicts the research design, and the stages or objectives which must be realised prior to the successful completion of each phase of the research. Initially, a review of the literature and background theory will be conducted to identify the need for a framework, this in turn facilitates the formation of the hypothesises and objectives which underpin this research.

During phase 2, the conventional approaches and the short coming of these approaches will be subject to investigation. This investigation shall permit the exploration of a framework to facilitate knowledge discovery within BI. This framework will be developed with a top-down approach, enabling the broad criteria to be examined within each stage, which can then be further refined and elaborated. Description will form a key component of this phase with the framework design being encapsulated through UML.

Phase 3, will consist of experimenting with the framework with case studies. Case studies are a suitable research method for problems that face complexity and uncertainty in their field of work (Johansson, 2004). These case studies will permit the evaluation and analysis of the framework in phase 4. This research model will facilitate a valid in-depth study for the investigation of a framework that will provide structure to knowledge discovery and technique selection within BI.

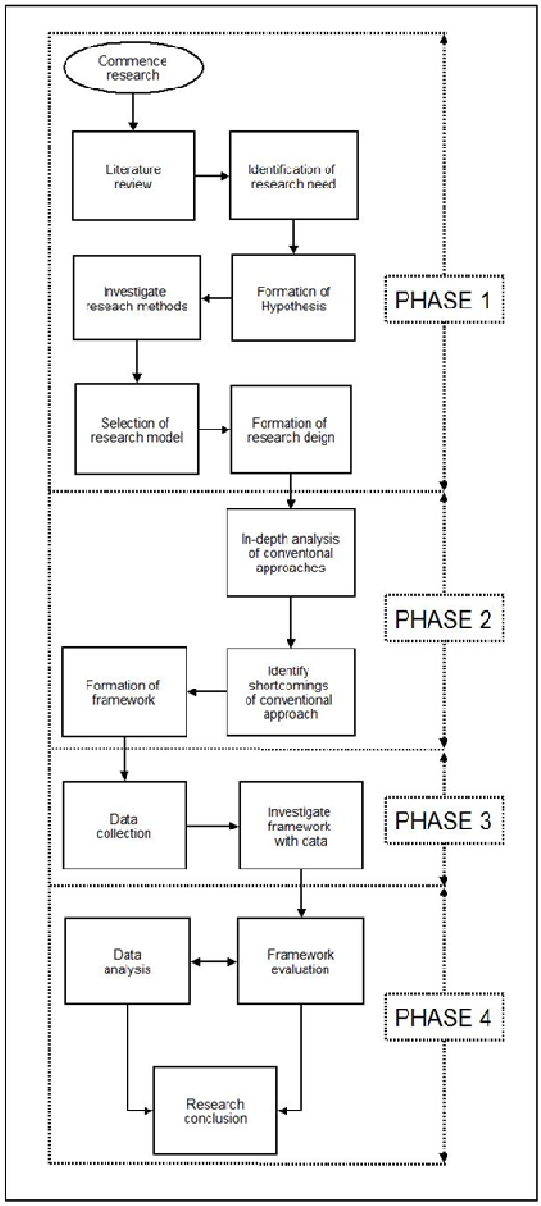


Figure 3.2: Research design.

3.3 Epistemology

As with all research, a model alone is insufficient to scientifically substantiate results and ensure a valid contribution to knowledge. The research process must conform to a philosophical-methodology and epistemology to justify the philosophical standpoint that has been adopted and the assumptions that have applied to the research. Thereby, validating the process and means through which the researcher can ensure that the contribution to knowledge will be scientifically and philosophically defensible (Bradshaw, 1997).

Thus, if the research process is to be scientifically substantiated, both an epistemology and methodology must be adopted by the researcher. Due to there being a number of definitions for epistemology and methodology (especially as this research will draw upon knowledge from a variety of disciplines), a distinction should be made of the definition that will be used throughout this research. Epistemology is the philosophy of knowledge or of how we come to an understanding of the world. Whereas, a methodology is focused on the specific ways, the methods, which we can use to try and understand the world better (Khazanchi & Munkvold, 2003).

Within methodological literature there are two main research paradigms 'positivism' and 'phenomenology' (Esterby-Smith, Thorpe & Lowe, 1999). Since, positivistic theory has been founded from 'natural science' and the belief that there is only one 'true reality', thus the reality can be observed objectively and knowledge gained through observations. In contrast phenomenology attempts to understand human behaviour from the participants' personal frame of reference (Chalmers, 1999). As the research shall reject metaphysics, this research will be grounded from a p*ositivistic* epistemology, the philosophical-methodology of the research that will be rooted in *falsification*. Positivistic theory assumes that the researcher has the ability to study the problem with a clear distinction between the researcher and the subject. Positivistic theory further strives to control all known uncertainty factors to be able to collect as objective and reliable research results as possible enabling the most objective results possible to be obtained (Khazanchi & Munkvold, 2003). This shall be paralleled in this research through the investigation of the framework via the utilisation of case studies (Johansson, 2004). Thus, allowing the technology to be applied directly to the problem irrespective of the programmers' beliefs as these cannot bear upon the performance of the platform, with emphasis upon the explicit knowledge contained within the data. Since, the hardware, programming/BI techniques and data conform to strict regulations and cannot be influenced to any great extent, by the beliefs of any external entities, thereby eliminating any bias. Positivistic theory aims to control any factors of uncertainty to ensure as objective and reliable results as possible (Vidgen & Braa, 1997). The purpose of the system is to provide the positivistic instrument through which the researcher can predict and control the reality of the external world (Hanfling, 1981; Khazanchi & Munkvold, 2003). In this way the research can be thought take a modernist view on positivism: The enlightenment-humanist rejection of tradition and authority in favour of reason and natural science (Keep, McLaughlin & Parmar, 2000).

Positivist theorists employ deductions to reach research objectives i.e. they try out existing theories or assumptions in their research (Anderson & Rönnbom, 1999). This is further exemplified by the theories of Kuhn, who although himself not completely positivistic, accepts that once a science becomes established and structured (Chalmers, 1999), much like BI, artificial intelligence and many of the mathematical techniques that underpin data mining (Bigus & Bigus, 2002; Curtis & Cobham,

2008; Witten & Frank, 2005), it becomes subject to theoretical frameworks and assumptions that are accepted within that community (Chalmers, 1999). Therefore, any assumptions, such as the validity of BI techniques or definitions, which are accepted by the BI community, will remain true to those principles when applied to the research. An objective of this research is to investigate a meta-level framework which can provide support to the process of selecting a suitable technique and structure an investigation within BI for knowledge discovery. Since there is no precedent for this research, it becomes increasingly difficult for an effective evaluation to take place. To prevent this from affecting the validity of the research, the research will investigate the framework through case studies and a variety of datasets.

To substantiate the contribution to knowledge, this research will not merely investigate the theoretical aspect of applying this technology but, as stated, include a number of exploratory investigations through case studies based upon 'real-world' data. The framework will be subjected to extensive testing/evaluation prior to validation. The objective of this testing will be to invoke failure and falsify the framework. This will permit a thorough investigation of the robustness and validity of the research. Accordingly, in the event that it is not possible to invoke failure through a systematic evaluation, the framework can be considered to have fulfilled the objectives, or in the event of failure identify the instances in which the framework is not valid. Falsification is an essential concept in the philosophy of science. Falsification puts forth the hypothesis that a theory cannot be considered scientifically valid if it does not admit to the consideration of the possibility of being false (Magee, 1985). Conversely, if the theory cannot be proven false within reasonable bounds then it can be substantiated as a genuine contribution to knowledge (Bhatia & Frazzoli, 2007). Conforming to the views expressed by Falsification theorists such as Karl Popper; the Baconian and Newtonian instances based upon the primacy of 'pure' observations as the initial step in the formation of theories, is misguided. Rather, all observation is selective and theory-laden (Chalmers, 1999). Thus, to ensure that the system is free from bias, as stated the framework will be tested under various conditions and the nature of the data sets dynamically altered, with a view to evoking the system to fail. Thus, ensuring that the research method can overcome any criticism of positivism which is rooted in the belief that science cannot proceed by corroboration alone (Warnick & Johnson, 2005).

3.4 Summary

This research presents a framework for knowledge discovery within BI to provide a greater level of decision support. Consequently, it has been the objective of this chapter has been to investigate, review and define the methodology, model and epistemology that will underpin and validate the contribution to knowledge.

A framework (within research) can be defined as a means through which to outline possible courses of action or present an approach to a systems analysis project. Since a framework extends further then a sole piece of software, yet provides a systematic approach to finding a solution, a general systems engineering process can be adopted. Given the technical nature of BI, with the emphasis upon explicit knowledge that must be extracted from software and data contained within an organisation, a hard systems approach can be adopted. The concept of explicit knowledge is pivotal to BI and the suitability of a hard systems approach. It is imperative to consider user requirement and objectives, however, the primary objective of the framework should be the investigation of a solution that can fulfil its objectives based upon organisational requirements, rather then those of individual users, as is the case with ad-hoc designs.

The rationale for the selection of a hard systems approach, is due to a hard systems approach having been investigated and successfully applied to information technology and computer/software systems within a number of organisations and application area. Furthermore, a hard systems approach is considered more suitable than a soft systems approach for this study. This is due to the requirements of this study, since the study would require specific input from experts as opposed to general feedback or requirements an approach such as that of a hard system is more suitable since it relies more upon the technical assessment, which in turn is made by the developer or individual conducting the investigation. Consequently, a hard systems approach that can be examined as the underpinning methodology will facilitate through technical understanding the design and examination of a more robust framework. Thus, following a hard systems approach, the research will take a top-down approach to design, ensuring that the logical processes, functions and entity-relationships within an organisation and its data are modelled correctly, with a documented structure. A hard systems approach contains four distinct phases Investigating; Analysing; Implementation; and Evaluation, thus the process of this research will be as follows:

1. This research will perform an in-depth investigation of the theory underpinning BI and DSS, to identify the opportunities available to provide structure to the process of technique selection and knowledge discovery through a focused meta-level framework.
2. Analysing the existing approaches that are adopted identify any shortcomings.
3. Propose a solution via CASE tools such as UML. Investigating a top-down view of the framework.
4. An implementation of the solution will be carried out through the investigation of case studies. This is a suitable research method for problems that face complexity and uncertainty in their field of work (Johansson, 2004).
5. Assess and evaluate the effectiveness of the solution. Consequently, the framework will be tested with a number of datasets of that the performance can be validated.

This research model will facilitate a valid in-depth study for the investigation of a framework that will structure the process of technique selection and knowledge discovery within BI, with a view to providing intelligent decision support.

Since a model alone is insufficient to scientifically substantiate results and ensure a valid contribution to knowledge. The research will follow a positivistic epistemology and ensure any criticisms of positivism can be addressed through falsification. The rationale for selecting this method is due to there being no real precedent for such a study. As a result, to ensure the validity of the research, the research will investigate the framework through case studies and evaluate the framework with the aid of a variety of datasets. After substantial consideration it was reasoned that the case studies (based upon live industrial data) would provide a suitable benchmark upon which the performance of the framework could be scrutinised. The objective of this testing will be to invoke failure and falsify the framework. This will allow a through investigation of the robustness and validity of the research, if it is not possible to invoke failure through a systematic evaluation, the framework can be considered to have fulfilled the objectives, or draw underline those instances in which the framework is not valid. By following this approach a systematic framework for knowledge discovery for decision support within BI can be investigated and evaluated. Whilst ensuring that any analysis, findings and contribution to knowledge is scientifically valid.

Chapter 4:

Related Work: Technical Review

This chapter will analyse the conventional approaches that are employed for the investigation and integration of BI techniques within organisational operations. Exploring and analysing conventional approaches will not only provide the opportunity to scrutinise and study the work related to this study. But will additionally provide the opportunity to facilitate the discovery of the key issues that the framework is to address. Ultimately, it is the aim of this chapter to provide insight into conventional approaches and identify the strengths and weakness of conventional approaches which can be addressed to formulate a meta-level framework explicitly for BI.

4.1 Conventional Approaches to Business Intelligence (BI)

It has been established that BI is becoming increasing crucial to business processes. As new technologies emerge, with increasing storage capabilities, advances in hardware, there has been a shift integrating information technology into many aspects of an organisation. BI amalgamates a number of disciplines such as artificial Intelligence, data mining, intelligent agents, KDD, QRA, DSS, enterprise resource planning, customer relationship management, amongst others for a variety of purpose. Integrating these technologies, together with the various information resources of an organisation provide a competitive edge. It has been this necessity to surpass rival organisations which has been the fundamental motivation for BI integration. Currently, no specific approach for BI integration or development exists, in contrast, the conventional approach to investigating BI solutions is the exploration of an ad-hoc methodology that can be tailored to specific requirements (Golfarelli et al, 2004; Wasserman et al, 2004; Konstantinos et al, 2008; Trestian et al, 2008; Xu et al, 2007). Consequently, the investigation of a solution frequently requires an expert and can, therefore, be a costly and complicated endeavour (Howson, 2008). However, a systematic approach for investigating BI solutions to provide knowledge discovery and retention for decision support, is imperative if the technology is to be effectively and seamlessly integrated into business practices. Two approaches which have gained prominence within the BI community for the exploration of BI solutions are; 'Rapid Application Development' (RAD); and 'Agile' (Shvertner, 2003; Turban & Aronson, 2001; Cutirs& Cobham, 2008; Howson, 2008). These approaches and their corresponding limitations can be further investigated.

4.1.1 Rapid Application Development (RAD)

Rapid Application Development (RAD) is a term originally used to describe a software engineering systems development life cycle (SDLC) introduced by James Martin in 1991. Based upon concepts proposed by Brian Gallagher, Barry Boehm and Scott Shultz, James Martin developed the Rapid Application Development approach during the 1980s at IBM. RAD involves iterative development through the investigation of prototypes (Beynon-Davies et al, 2000). RAD was motivated by the inflexibility of processes developed in the 1970s and 1980s, such as SSADM (Structured Systems Analysis and Design Method). It was considered that conventional approaches, where not dynamic enough due to requirements changing before a suitable solution had been developed, thereby often resulting in inadequate or even unusable systems. Another drawback of conventional methods has been that a methodical requirements analysis phase alone is often insufficient to identify all critical requirements.

The dynamic nature through which requirement of a system can be collected and altered have resulted in RAD becoming a popular approach for investigating BI solutions. RAD is a phased development methodology, thus, the system is developed sequentially through a number of iterations, with each subsequent version providing greater functionality. An advantage of this approach is that a working model can be provided to the user and implemented; whilst further refinement of the system is conducted simultaneously. Figure 4.1, illustrates the RAD process model. Although, the RAD model provides users with an implementation early in the development cycle, that can be investigated for accurate feedback. The implementation of an incomplete solution has also been a criticism of this approach (Turban & Aronson, 2001; Shvertner, 2003).

Central to the concept of RAD is the role of clearly defined workshops. The fundamental concept of these workshops is (Cobham & Curtis, 2008):

- Involve business and information systems personnel.
- Run for a pre-determined duration. This duration is typically 1-5 days.
- Take place in 'clean rooms'. Clean rooms are rooms that have been set aside for the explicit purpose of the workshop. As a result, the workshops are detached from daily operations, thereby, allowing each workshop to be conducted without interruption.
- Be presided over by a 'facilitator'. It is the role of a facilitator, to ensure that content of the workshop, the agenda, and ensure that the workshops remain focused toward deliverables.
- The workshops should be documented by a scribe to ensure that the details can be investigated throughout the SDLC.

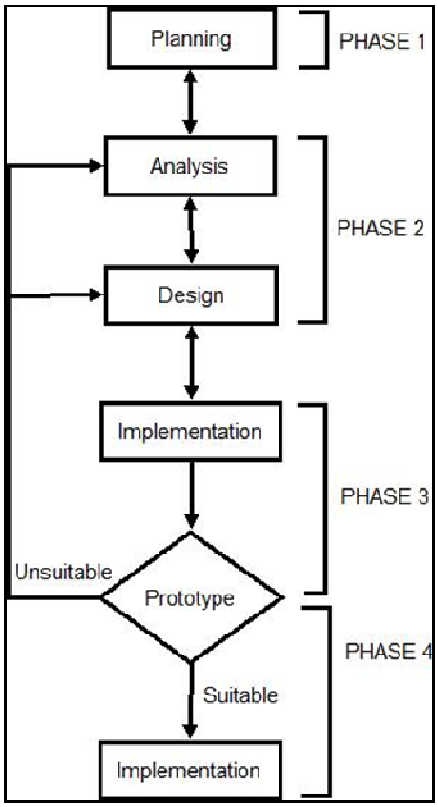


Figure 4.1: RAD process model.

The RAD workshops are intended to reduce the time taken to gather requirements necessitated by traditional SDLCs. A conventional SDLC often involves lengthy investigation of the data models and documentation of an organisation. RAD consists of 4 phases (Curtis & Cobham, 2008). These phases are illustrated on figure 4.1 and detail the objectives and expected deliverables of the stages that comprise the process model:

1. The initial phase involves the collection of requirements to establish the high-level management and strategic objectives of an organisation. These requirements will be collected via workshops to establish a high-level business perspective that can be investigated to drive the development of a solution.
2. The second phase will investigate the requirements gathered in phase 1, and investigate them from a more technical perspective. Following a top-down approach this phase may involve further workshops that assist the design of a system through CASE tools.
3. Phase 3, involves the realisation of the designs compiled in phase 2. Thus an implementation of the system is developed. This development can take the form of a number of prototypes, each of which can offer differing functionality. These prototypes can be assessed by the user and feedback gathered for further prototypes. The various prototype systems are developed and investigated by small teams known as SWAT (Skilled With Advanced Tools) teams.

Emphasis is placed upon the development of the core system, permitting a central system to be rapidly designed and implemented with further functionality provided or removed with successive iterations.

4. The final phase involves comprehensive testing of the system to investigate whether the solution provided all required functionality or further prototypes are required.

In RAD, if the development of a solution falls behind schedule, the functionality will be reduced rather than the deadline extended. RAD developed BI solutions can be readily integrated into business processes, and the dynamic approach is suitable for the dynamic nature of organisations. However, many BI practitioners believe that the RAD approach can result in a system which may fail to meet all requirements (Cobham & Curtis, 2008). Further to this, an incomplete system can result in a negative perception of a solution. For successful BI integration it is imperative that users not only be able to integrate solutions into business processes, but develop BI solutions that are capable of providing a high level of support to the users. The level of support will naturally be reduced, if early implementations with limited functionality result in a negative impression upon the user. In addition, the stages of RAD are extremely unstructured, this often results in prototypes that are developed, but must be subsequently discarded. These limitations have resulted in the RAD process model often overcomplicating the investigation of a solution (Agarwal, 2000).

4.1.2 Agile

The Agile development methodology is a SDLC that has been developed upon the dynamic principles introduced by RAD. Agile software development refers to a group of software development methodologies such as eXtreme Programming (XP) that promote development iterations, open collaboration, and process adaptability throughout the life-cycle of the project (Beck & Andres, 2005). Agile processes involve objectives that can be achieved in small increments, with minimal planning. This is motivated by the requirement to minimise the overall risk, and allows the project to adapt to changes more readily. Agile is an emerging approach that has begun to gain prominence within the BI community (Howson, 2008).

The concept of Agile software development materialised from a gathering of software engineers in 2001. The constituent members of this gathering represented respected individuals from a number of software engineering disciplines motivated by the requirement for an alternative to documentation driven, heavyweight software development processes. The consequence of this gathering was the Agile manifesto[2]. This manifesto is compiled of twelve principles that underpin the concept of the Agile development methodology:

[2] http://www.agilemanifesto.org: Accessed September, 2008.

1. Welcome changing requirements, even late in development. Agile processes harness change for the customer's competitive advantage.
2. Deliver working software frequently, from a couple of weeks to a couple of months, with a preference to the shorter timescale.
3. Business people and developers must work together daily throughout the project.
4. Build projects around motivated individuals. Give them the environment and support they need, and trust them to get the job done.
5. The most efficient and effective method of conveying information to and within a development team is face-to-face conversation.
6. Working software is the primary measure of progress.
7. Agile processes promote sustainable development.
8. The sponsors, developers, and users should be able to maintain a constant pace indefinitely.
9. Continuous attention to technical excellence and good design enhances agility.
10. Simplicity, the art of maximising the amount of work not done is essential.
11. The best architectures, requirements, and designs emerge from self-organizing teams.
12. At regular intervals, the team reflects on how to become more effective, then tunes and adjusts its behaviour accordingly.

The adoption of Agile, methodologies has increased within the BI development community as it provides a SDLC that is not confined by a strict and precise list of requirements. In contrast, Agile encourages a top-down approach, through the identification of a broad set of requirements or required capabilities. This broad set of requirements is further investigated and focused though the development process through prototyping. Prototyping in Agile can range from spreadsheets, mock-screens, reports, to coded applications (Stamelos & Sfetsos, 2007). Agile methodologies consist of five main phases:

- *Planning*: During this phase short ‘user stories’ are developed for each of the requirements of the solution. User stories can be compiled of a few lines of text describing the requirements.
- *Design*: In Agile, emphasis is placed upon simplicity with in the system. Functionality is gradually increased. In general, there is little documentation produced. In contrast, this phase will involve the small development teams attempting to gain a greater insight into the user stories and how to implement solutions.
- *Coding*: Code is produced in stages. Initiating with test code to give direction and focus to development prior to the development of the main code. A common development technique that is utilised in this stage of Agile methodologies is ‘pair programming.’ Pair programming implies that aspects of design are distributed amongst teams of two programmers who will develop their module of code in unison. Seated together at one station, one programmer will

focus upon the detailed coding, with control of the mouse and keyboard, whilst the other will focus upon observation, error detection, and strategic issues regarding the code on a larger scale within the context of the complete system. As code is completed a module can be added to a central repository. The sum of the modules will result in a complete system.

- *Testing*: Unlike traditional systems where the code the systems or solution is tested at the end of the development cycle. In Agile methodologies, testing occurs alongside development / prototyping. This facilitates changes to the system as requirements are subject to change. As a result code is tested prior to integration into the central repository.
- *Implementation*: The process does not end with implementation. Given the dynamic nature of requirements, it may be required to reassess the solution periodically and return to an earlier phase in the event that the requirements alter or new requirements are discovered.

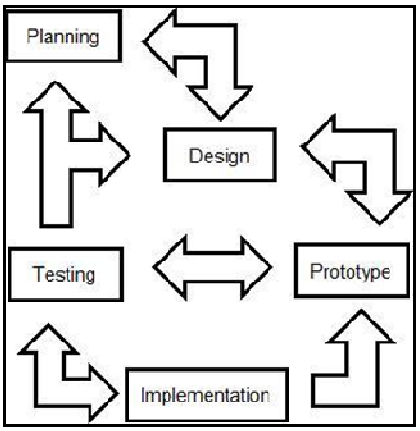


Figure 4.2: Agile process model

Agile methodologies adopt a cyclical approach that emphasises simplicity, modularity, and iteration. The phases and dynamic movement between these phases is illustrated in figure 4.2. The process model, exemplifies the level of iteration that is involved. Although not specifically tailored to BI, in the absence of a specific approach, the dynamic and iterative nature of Agile methodologies has resulted in their emergence to the forefront as suitable process models for BI solutions. Agile methodologies have been investigated for BI solutions in areas such as fraud detection and Business process (Nguyen et al, 2005; Peng et al, 2008). However, Agile methodologies have yet to gain wide spread acceptance within BI. In general, the development of BI solutions consists of ad-hoc methodologies that may integrate aspects of Agile models (Howson, 2008). A major drawback of Agile methodologies is that despite the dynamic nature of requirement gathering, they do not providing sufficient support for the knowledge discovery process. Furthermore, the emphasis of Agile methodologies upon simplicity and limited documentation, which can although result in more rapid development cycles, can also result in unnecessary iterations (Subramaniam & Hunt, 2006). Consequently, in order to gain further insight in to the approaches that are explored when

investigating BI solutions it is imperative to further investigate the conventional approaches to the key technologies that underpin BI solutions.

4.2 Conventional Approaches to Multi-Agent Systems

Multi-Agent systems are based upon the principles of artificial intelligence. As discussed within Chapter 2, Multi-Agent systems consist of a number of intelligent agents that complete tasks on behalf of a user. Furthermore, various factors that distinguish agent-oriented programs form object-oriented programs were reviewed. Due to these factors, standard object-oriented development paradigms are not suitable for designing agent based systems. Consequently, there have been multiple architectures and frameworks that have been investigated for agent design. These frameworks have been proposed in order to provide an underlying infrastructure that can support agent identity, autonomy, co-existence, communication, mobility, security and life-cycle management (Lin, 2007). As discussed in Chapter 2, FIPA has attempted to standardise agent communication with the FIPA-ACL. The FIPA-ACL language has been developed with a view to facilitate various agent applications which have been independently developed, to communicate. For this purpose, FIPA provides specifications for an abstract architecture for a Multi-Agent system. The objective of the of the FIPA 'Abstract Architecture Specifications' is to present an conceptual structural design to accommodate a wide range of commonly used mechanisms such as the various message transports, directory services and popular commercial platforms. The FIPA Abstract Architecture consequently, describes the semantics for each element followed by the relationships between this element and other elements. In addition to the technology-oriented features of its abstract architecture FIPA provides abstract entities for platform management (white and yellow pages) (FIPA, 2008).

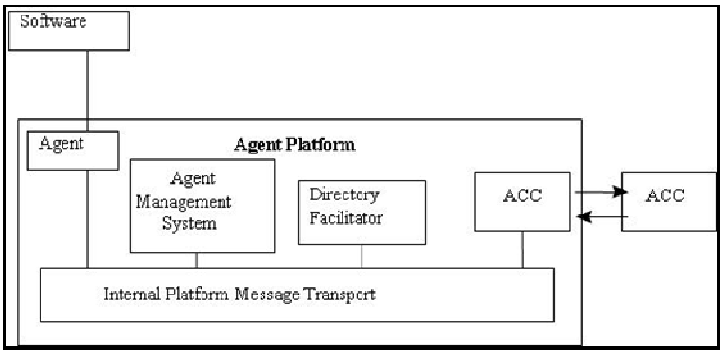


Figure 4.3: FIPA-compliant agent platform.

A Multi-Agent system which conforms to FIPA specifications are known as FIPA-compliant. Agent platforms provide the framework within which agents exist and operate. FIPA-compliant agent platforms (figure 4.3) consist of three 'management agents' (Jade, 2008):

- *DF (Directory Facilitator)*: The DF provides a 'yellow pages' service. The DF is a mandatory, normative agent which acts as a custodian of the directory within a single domain. Facilitating agents to find agents that provide specific services.
- *AMS (Agent Management System)*: A mandatory component of the platform. The AMS exerts supervisory control over the platform managing the agents, by registering and authenticating them. As a result the AMS provides the 'white pages' service by maintaining a directory of logical agent names and their associated transport addresses. Furthermore, the AMS manages the use of the ACC (Agent Communication Channel).
- *ACC (Agent Communication Channel)*: The ACC is the default communication agent that connects all agents within the platform by supporting the storage, retrieval and forwarding of messages. In addition, the ACC is the agent through which multiple agent platforms can be connected.

The FIPA abstract architecture provides a means through which a Multi-Agent system can be modelled. These specifications however, do not provide support for the discovery of requirements and objectives of an explicit users needs. Consequently, agent-oriented methodologies have been researched. These methodologies enable an analyst to systematically investigate a statement of specifications to a design that can be directly implemented. Two agent-oriented methodologies that have been significantly researched are *Prometheus* and *Gaia* (Wooldridge, 2002; Pagham & Winkoff, 2004).

4.2.1 Prometheus

The Prometheus methodology defines a detailed process for specifying, designing, implementing, testing and debugging Multi-Agent systems. Prometheus has been developed over several years at RMIT University, Melbourne, Australia, in collaboration with Agent Oriented Software (AOS) [3]. The motivation for Prometheus was the investigation of a process with associated deliverables which could be taught to industry practitioners and researchers with little or no background experience with intelligent agents.

Prometheus has been explicitly formulated for intelligent agent systems. Thus, rather than consider an agent as simple software process that interacts to achieve an overall system goal, Prometheus provides support for the development of intelligent agents that use goals, beliefs, plans and events. Providing support for agent investigations as a complete SDLC, Prometheus provides a framework for detailing, requirements and specification, detailed design to implementation. In addition, Prometheus facilitates the construction of 'design artefacts'. Prometheus design artefacts relate to the *precepts*

[3] http://www.agent-software.com: Accessed: September, 2008.

(inputs) and *actions* (outputs) of the system, in addition to the interaction protocols, interaction diagrams, scenario diagrams, capability diagrams and plan descriptors. All artefacts are structured; however, whilst some artefacts will be major components of the system others will provide intermediary support. The Prometheus methodology consists of three phases (Padgham et al, 2007a):

- *System Specification*: Emphasis is placed upon identifying the goals and basic functionalities of the system, specifying the interface between the system and its environment in terms of precepts and actions. Furthermore, 'use case' scenarios are developed to illustrate the system's operation.
- *Architectural design*: This phase encapsulates the system's overall (static) structure through a system overview diagram. The output from the system specification is investigated to determine the agent types, and agent descriptors. In addition to describing the dynamic behaviour of the system using interaction diagrams and interaction protocols.
- *Detailed design*: This phase investigates the internals of each agent and the mechanisms the agent will use to accomplish objectives.

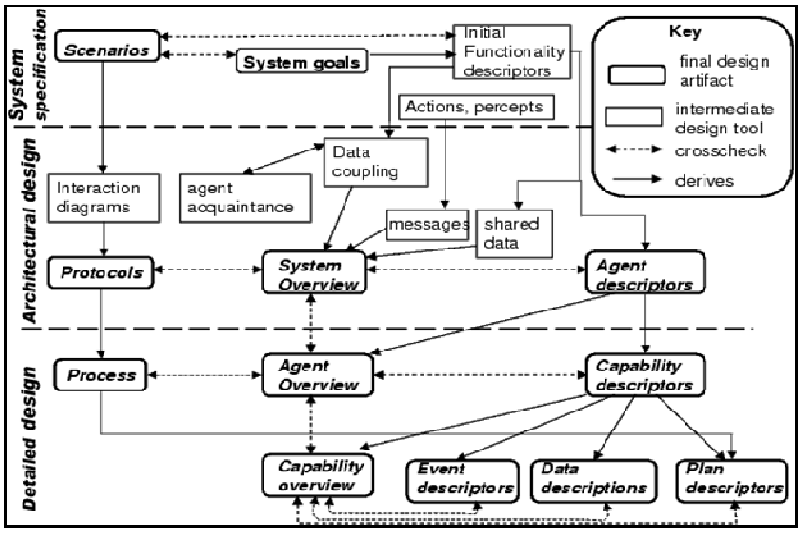


Figure 4.4: Prometheus process model.

The three phases of the Prometheus methodology are illustrated in figure 4.4. The goals that are investigated and determined during the system specification phase should be detailed and encapsulate the objectives of the system as a whole. Investigating goals at a high level ensures that they are more likely to remain constant, even if the required functionalities change. The goals of the system are documented with each goal assigned a name and description, furthermore this documentation will encapsulate the relationships between goals and sub-goals. The documentation will also extend to

creating use case scenarios, which are examples of the system's operation. Use case scenarios provide an effective means through which to investigate goals and ensure any problems that may arise. Use case scenarios consist of a sequence of steps that describe typical uses of the system. Although similar to object-oriented deigns, the details encapsulated in agent-oriented use case scenarios differ. Furthermore, use case scenarios will facilitate the discovery of the functionalities, which are formed through the grouping of related goals, precepts (incoming information from the agents' environment) and actions (the means through which an agent can affect the environment (Padgham & Winkoff, 2004).

Upon determining the functionalities of the system, the systems architectural design can be investigated. Initially, the functionalities can be grouped together to investigate the agent types. The exploration of which functionalities should be grouped together can be investigated through coupling and cohesion. A data coupling diagram can guide this process, whereas an agent acquaintance diagram can be employed to asses whether the couplings are suitable. Once the groupings have been investigated the agent types are described through an agent descriptor form. Once the agents that the system is to consist of have been discovered the interaction between these agents must be established. This will enable a system overview diagram to be discovered. The final stage, builds upon the architectural design to develop the internal structure of the agent. The internal structure of the agent dictates the mechanisms through which the agent will realise its objectives. Each agent is progressively refined by defining its capabilities, internal events plans and detailed data structures. Simultaneous to the internal structure of the agent the interaction protocols are refined to provide process specifications.

The Prometheus methodology is not intended to be followed strictly. The methodology should be interpreted as a set of guidelines, providing the degree of flexibility required for agent-oriented design. To facilitate the stages of the methodology the Prometheus Design Tool[4] (PDT) have been researched at RMIT Univeristy (Phadgham et al, 2007b). The PDT is a graphical editor designed specifically to supports the design tasks specified within the Prometheus methodology. PDT described by Phadgham et al, (2008) can be investigated to facilitate the design and documentation of the systems specification, architectural design and detailed design of an agent-oriented system.

Although 'Implementation' can be considered as a fourth phase, due to implementation being closely dependent upon the platform through which the system is developed. This phase has been omitted from the overall structure of the Prometheus methodology (Phadgham et al, 2007a). Thus, ensuring Prometheus is a tool/platform/application neutral methodology that can facilitate the investigation of

[4] http://www.cs.rmit.edu.au/agents/pdt/: Accessed September, 2008.

agent-oriented systems. Prometheus however, is not the only agent-oriented methodology that has been significantly researched.

4.2.2 Gaia

The Gaia methodology is one of the most researched agent-oriented mythologies. Initially investigated by Wooldridge, Jennings and Kinny (2000), the Gaia methodology is intended to facilitate an analyst to systematically investigate an intelligent agent system from requirements to a design, which contains sufficient details for direct implementation. Emphasising the developer to view the investigation of an agent-system as a process of organisational design, Gaia employs terminology and notation from object-oriented analysis and design (specifically FUSION). Gaia, however, is not simply an application of object-oriented methods. In contrast, the methodology provides an agent-specific set of concepts that a software engineer can explore to understand and model complex systems. The Gaia methodology is intended to provide the developer of an agent-based system with support from inception to a fully detailed design, modelling both the *macro* (social) and *micro* (agent internals) of an agent-based system. The Gaia methodology considers requirements specification independently from the paradigm utilised for analysis and design (figure 4.5).

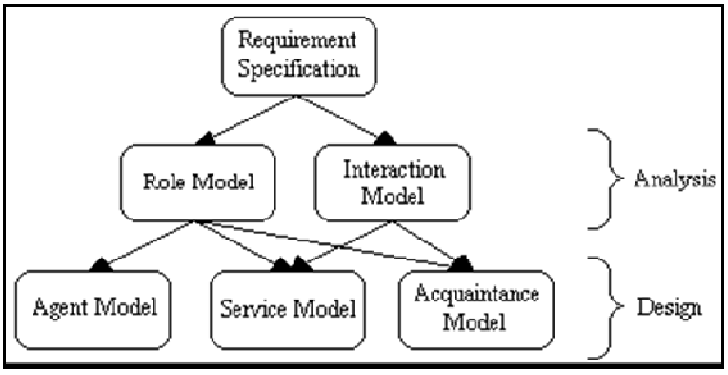


Figure 4.5: The Gaia Methodology.

Adhering to a top-down approach, the Gaia methodology provides a framework which facilitates initial abstract concepts to be encapsulated and investigated in greater detail. Thereby, providing support for the transition from, 'abstract concepts' to 'concrete concepts'. The main concepts within the Gaia methodology are illustrated in table 4.1. Abstract entities conceptualise the system through the analysis phase, yet have no realisation with the final system. Concrete concepts however, are not only explored during the design phase, but will also be present within the final system. As the abstract concepts and original requirements are investigated, implementation bias is introduced to reduce the number of possible solutions that can satisfy the original requirements (Wooldridge, 2002).

Roles	Agent types

Permissions	Services
Responsibilities	Acquaintances
Protocols	
Activities	
Liveness properties	
Safety properties	

Table 4.1: Abstract and concrete concepts within the Gaia methodology.

Once the original requirements have been explored and it has been discovered that an agent approach is the most suitable solution. The developer can conduct an analysis in order to understand and explore a proposed agent-system and its structure that will satisfy the requirements. Within Gaia, the agent-system to be developed is viewed as an organisation. This organisation consists of a collection of roles and the interaction between these roles (figure 4.6).

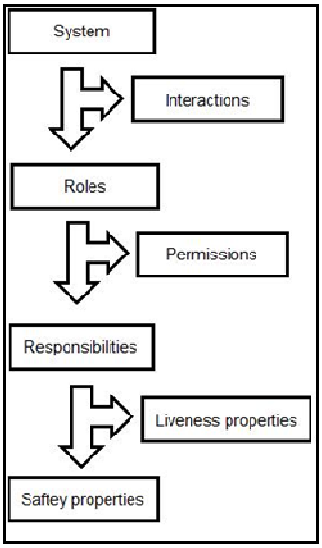


Figure 4.6: Gaia methodology analysis concepts.

Each role within the organisation is defined by four attributes (Wooldridge et al, 2000):

- *Responsibilities*: the key attribute associated with a role. The responsibilities determine the functionality of the role. Responsibilities are typically divided into two types:
    - *Liveness properties*: describe those states of affairs that an agent must bring about, once certain environmental conditions have been met.
    - *Safety properties*: are invariants, and ensure that an acceptable state of affairs is maintained across all states of execution.
- *Permissions*: are the 'rights' that are associated with a role. The permissions thus, identify the resources that are available to that role in order to realise its responsibilities.

- *Activities*: are computations that are associated with the role that the role can carry out, without interacting with other roles.
- *Protocols*: define the manner in the role can interact with other roles.

Role Schema:	*name of role*
Description	*short English description of the role*
Protocols and Activities	*protocols and activities in which the role plays a part*
Permissions	*"rights" associated with the role*
Responsibilities	
Liveness	*liveness responsibilities*
Safety	*safety responsibilities*

Figure 4.7: Template for Gaia schemata for defining roles.

The Gaia methodology specifies that the roles be depicted within a schema (figure 4.7). There a number of notations that must be followed, to specify roles within the Gaia methodology (Sethuraman et al, 2008):

- Protocol names are written in 'Sans'.
- Activities are denoted in a similar font and underlined.
- Liveness expressions have a number of operators:

x . y	=>	x followed by y
x | y	=>	x or y occurs
x*	=>	x occurs 0 or more times
x+	=>	x occurs 1 or more times
x^{w}	=>	x occurs infinitely often
[x]	=>	x is optional
x || y	=>	x and y are interleaved

Once the roles and their attributes have been defined, these abstract concepts can be further investigated. The design phase of the Gaia methodology entails investigating the abstract concepts so that they can be defined as concrete concepts. These concrete concepts will provide details of how the society of agents will operates to realise the goals of the system, in addition to the specific tasks that individual agents will be required to fulfil.

The 'design' phase will initially consist of creating an agent model that will document the various 'agent types' that will be deployed at run time. An agent type is a set of agent roles. A number of roles may be aggregated to increase efficiency, as it will be less processor intensive to have one agent fulfil a number of roles, then a number of agents each with a single role. The agent types and the associated

roles are documented to investigate an agent type hierarchy. Furthermore, the instances of each agent type are documented using instance annotations.

n	=>	There will be exactly n instances.
m..n	=>	There will be between m and n instances.
*	=>	There will be 0 or more instances.
+	=>	There will be 1 or more instances.

In contrast to object-oriented design, there is no 'inheritance' during the design phase. Although, inheritance may be present when the agents are implemented, the Gaia methodology omits this process from the design. This is due to agents being viewed as coarse grained computational systems with an agent system consisting of a comparatively small number of roles and types. In addition, the Gaia methodology is facilitates the investigation of 'what' tasks the agents will do as opposed to 'how', thus a service model for the agent is investigated and documented (Wooldridge et al, 2000). The service model will specify the main activities of an agent. These services are the activities the agent is involved in. Hence, the inputs, outputs, pre-conditions, post conditions, are documented, these will be derived from the 'activities', 'protocols', 'safety' and 'liveness' properties of a role. Once the agent types and services have been modelled the interaction between agent types must be defined. The acquaintance model does not define what or which messages are sent; in contrast it is the links that exists between the agent types that is documented. An agent acquaintance model is documented as a directed graph (Juan et al, 2002). The nodes of the graph correspond to agent types and arcs in the graph correspond to communication pathways.

The analysis and design phases of the Gaia methodology will provide detailed models of the system to be implemented. These models provide sufficient information to be implemented directly through any programming languages, architectures, and techniques that a developer may employ. However, the Gaia methodology has been designed to handle small-scale, closed systems. It has weaknesses that render it inappropriate for engineering complex open systems. Furthermore, although Gaia and Prometheus are suitable for agent-oriented designs, these paradigms do not fully encapsulate a data intensive environment, where an autonomous solution is not always appropriate. Within BI an agent solution may not always provide the most suitable solution, it is essential to investigate BI techniques such as data mining that can facilitate the interrogation of organisational data.

4.3 Conventional Approaches to Data Mining

It has been discussed in Chapter 2 that data mining forms an integral component of the KDD process model. Data mining can be defined as the process of interrogating data in order to identify valid, novel, potentially useful patterns. The advantages of data mining have been explored in a number of

application areas such as, sales and customer relationship management (Berry & Linoff, 2004; Hung et al, 2006), financial forecasting (Chun & Park, 2006), fraud detection (Fawcett & Provost, 1997), gene mapping (Kantardzic & Zurada, 2005) and mining of health care data (Alonso et al, 2002; Phillips-Wren et al, 2007). The versatility and increasing demand for data mining necessitates that this is not a technology that should be applied without first establishing a systematic approach.

A poll conducted by KD Nuggets [5] in August 2007 (figure 4.8), found that a vast number of projects still use ad-hoc methodologies. This however, can be contrasted with the same poll that was conducted in April, 2004 (figure 4.9), and we can see that although a large number of respondent are still using ad-hoc or non-specific methodologies, there has been a fall in this number. Thus, indicating that a greater value is being placed upon the benefits that can be reaped from a methodology.

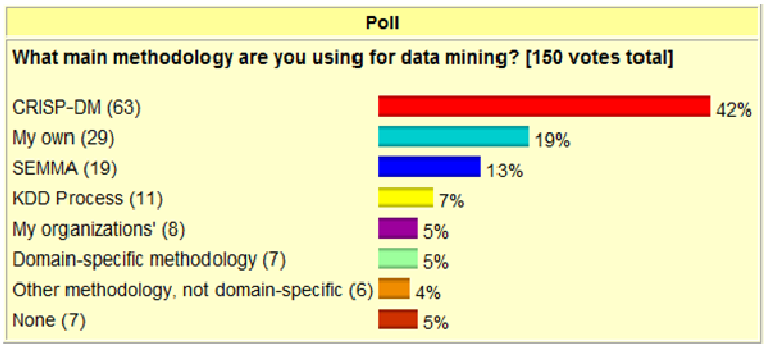


Figure 4.8: KD Nuggets data mining poll August, 2007.

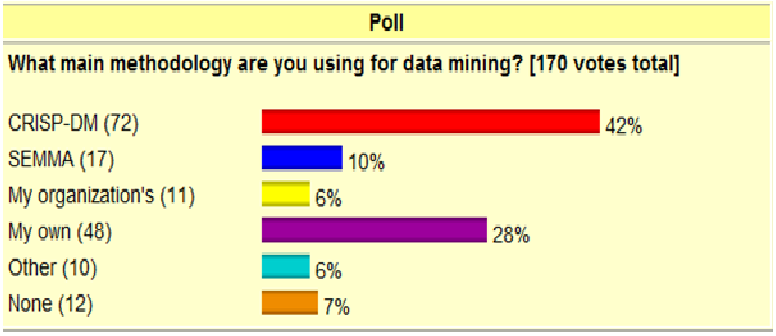


Figure 4.9 KD Nuggets data mining poll April, 2004.

Two approaches that have proven popular are SEMMA developed by SAS, and CRISP-DM, developed under the ESPRIT funding initiative as a European Union project and led by four companies: ISL, NCR, Daimler-Benz and OHRA. These models, their benefits and limitations can be further investigated.

[5] http://www.kdnuggets.com/: Accessed September, 2008.

4.3.1 SAS SEMMA

SEMMA was developed by the SAS Institute, the largest independent vendor in the BI market (Evelson, 2008). Although commonly referred to as a methodology, as stated on the SAS SEMMA homepage[6], SEMMA is not a data mining methodology. In contrast, SEMMA is a logical organisation of the function toolset accessed through the SAS Enterprise Miner application. SAS Enterprise Miner is a data mining tool developed by SAS to facilitate the core tasks of data mining (figure 4.10). Thus, despite generally being considered a data mining methodology, SEMMA has been developed explicitly for use with a specific toolset. However, since the stages of SEMMA can be investigated to examine the considerations that must be observed within a data mining project, it is generally referred to and considered data mining methodology.

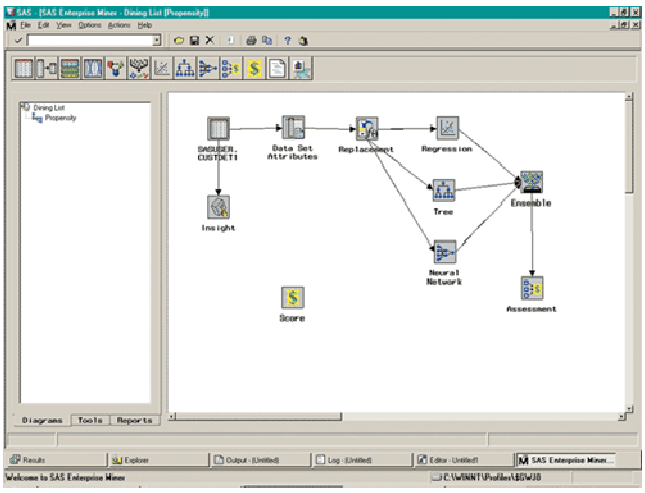


Figure 4.10: SAS Enterprise Miner interface.

The name SEMMA is an acronym for the stages that the process consists of; 'Sample', 'Explore', 'Modify', 'Model' and 'Assess'. These stages can be further explored (Lawrence et al, 2008):

- *Sample*: Although an optional stage, a portion of the data large enough to be significantly representative is selected. The sample data should be small enough to facilitate greater manipulation and faster analysis. The concept underpinning sampling is that mining a representative sample rather than the complete dataset reduces the processing time required to get crucial business information. If general patterns appear in the data as a whole, these will be traceable in a representative sample. However, for those correlations that cannot be discovered with a sample of the data set, may be discovered via summary methods. The sample process exemplifies that SEMMA is a tool dependent process model.
- *Explore*: This stage entails refining the discovery process by searching for anomalies (outliers) and unexpected trends to gain greater insights. Thus a variety of methods, such as

[6] http://www.sas.com/technologies/analytics/datamining/miner/semma.html: Accessed September, 2008.

visual, statistical techniques, factor analysis, correspondence analysis, and clustering can be investigated.

- *Modify*: Creating, selecting and transforming the variables can result in more focused model selection. Developing upon the results of the 'Explore' stage, datasets can be modified to include or remove variables, group variable or variable types. The dynamic and iterative nature of data mining may require modification to take place as new data becomes available.
- *Model*: Modelling involves investigating the dataset with data mining algorithms. This stage may involve investigating the dataset with a variety of data sets. Within SAS Enterprise Miner, aspects of this stage may be automated. As a result the detail of this stage exemplifies the association between SEMMA and a particular toolset.
- *Asses*: Involves the analysis of the modelling stage. Thereby evaluating the outcome and results of the algorithms that have been investigated. There are a variety of methods for assessing data. These methods range from analysing samples to estimate performance on the complete dataset to the use of training sets where the outcome is known, thus results can be compared.

The SEMMA process advocates iteration. Consequently, continuous assessment and evaluation can facilitate new discoveries and a greater understanding of the data. Once the 5 stages of the SEMMA process have been investigated, the result will be a 'Champion model'. This model can be investigated with new and unknown datasets, this is referred to as 'scoring new customer cases' in Enterprise Miner. However, this further exemplifies the key limitation of SEMMA, since within Enterprise Miner the deployment phase is automated by supplying scoring code in SAS, C, Java, and PMML. Furthermore, the automated process not only captures the code for analytic models but also captures the code for pre-processing activities. As a result the applicability of SEMMA is limited if attempting this 'methodology' with a toolset other then SAS Enterprise Miner.

4.3.2 CRISP-DM

Unlike SEMMA, CRISP-DM (CRoss-Industry Standard Process for Data Mining) is a platform independent data mining methodology. The CRISP-DM user Guide documents that the methodology was conceived in late 1996 by Daimler-Benz (now DaimlerChrysler), Integral Solutions Ltd. (ISL), NCR, and OHRA and remains close to the KDD process model. CRISP-DM was motivated by the mounting interest in data mining and the requirement for a clear, widely accepted pure data mining methodology. The four companies that conceived CRISP-DM, were large stakeholders of within the data mining industry:

- *Daimler-Benz* was amongst the industrial and commercial organisations that pioneered the integration of data mining into its business operations.

- *ISL* (acquired by SPSS Inc. in 1998) initially provided services based upon data mining principles. This led to the launch of the first commercial data mining workbench 'Clementine' in 1994.
- *NCR*, aiming to deliver added value to its Teradata data warehouse customers, met its clients' needs with teams of data mining consultants.
- *OHRA*, at the time one of the largest Dutch insurance companies, provided a valuable testing ground for live, large-scale data mining projects.

According to the CRSIP-DM Specifications, the rationale for the development of CRISP-DM was thus to provide a data mining process model that would facilitate organisations in launching their own data mining projects, whilst standardising the industry. It was considered that the development of a non-proprietary, documented, and freely available model could be a means through which to improve the performance of data mining and encourage adoption of data mining strategies, thereby facilitating the industry to develop. In 1997, the founding companies of CRISP-DM had obtained funding from European Commission and initiated the investigation of an industry, tool and application neutral data mining methodology. This process was supported by the founding of the CRISP-DM Special Interest Group (SIG). The aim of Sig was to gather input from a variety of data mining sources. This cumulated in the development of the original CRISP-DM process guide and user manual (CRISP-DM version 1.0) in mid-1999.

The process model for CRISP-DM version 1.0 provides an overview of the life cycle of a data mining project. The process model details the phases and tasks of a data mining project and the relationships between tasks that must be observed. The CRISP-DM process model advocates a top-down approach, thus not requiring the identification of every possible relationship. In contrast, the initial obvious relationships can be identified, prior to further investigation and detailing. More in-depth analysis can take place once further examination of the goals, background and interest of the user, and data has taken place. The life cycle of a data mining project consists of six phases:

- *Business Understanding*: Focuses upon the investigation, discovery and understanding of the project requirements from the perspective of a business. These requirements can then be converted into more technical data mining terminology and a preliminary plan for investigating these objectives established.
- *Data Understanding*: During this phase, the initial data is collected. Once the data has been collected, the data can be further investigated to identify the quality of the data, any discrepancies. Furthermore initial hypothesis may be formed upon this initial investigation.
- *Data preparation*: Consists of converting the raw data to a format that can be investigated with data mining algorithms. Data preparation tasks are likely to be iterated on numerous

occasions throughout the life cycle of a project. Data preparation tasks include; table, record, and attribute selection as well as transformation and cleaning of data for modelling tools.

- *Modelling*: during this phase various data mining algorithms and modelling techniques are investigated. Typically, there are several techniques for the same data mining problem type, thus a number of techniques may need to be investigated with their parameters are calibrated to optimal values. Often data mining techniques have specific requirements on the form of data. Therefore, iteration to the data preparation phase is often required.
- *Evaluation*: By this stage a model (or models) of the data will have been developed. These models should be of a high quality for data analysis prior to final deployment, thus may require iteration amongst the preceding stages. At this stage the models should be thoroughly evaluated to ensure the business objectives have been achieved, this may involve a review of the development stages.
- *Deployment*: Creation of the model should not be considered as the end of a data mining project. The knowledge that has been extracted from the data must be organised and presented to the end-user so that it can be effectively utilised. This process can range from a simple report complex data mining process that can be repeated. In a vast majority of instances, it will not be the data analyst, but rather the end user that will execute the deployment phase.

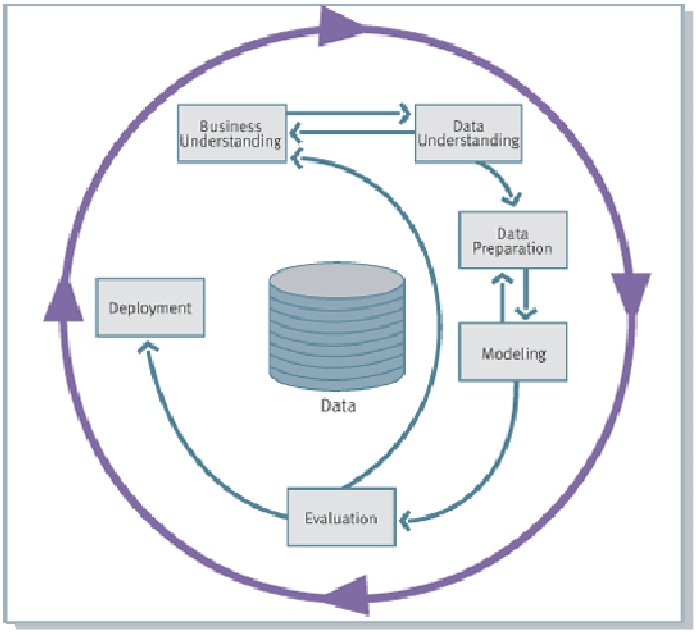


Figure 4.11: CRISP-DM process model.

The phases of CRISP-DM do not have a strict ordering and iteration between phases is advised. Figure 4.11, depicts the most frequent dependencies. The movement between phases should ideally be dictated by the outcomes of each stage, despite the flexibility it is imperative that a project not lose focus as the iterative stages take place. The outer circle in figure 4.11 is representative of the cyclical

and on-going nature of a data mining project. A data mining project does not cease with deployment, rather iteration of phases after deployment is advised, so to integrate discoveries that are made post-deployment, as this stage often results in more focused business questions.

As illustrated by the polls conducted by KD Nuggets in April, 2004 and August, 2007 (page 97; figure 4.8 and figure 4.9) CRISP-DM is the methodology of choice for a large percentage of the respondents for data mining projects. Further to this these projects have been investigated in a number of domains, Forensics (Adderley & Bond, 2007; Venter et al, 2007), Medical, Piatetsky-Shapiro et al, 2003; Kalos & Rey, 2005; Kuo et al, 2007), Telecommunications (Li et al, 2006; Raivio, 2006; Zan et al, 2007), Manufacturing (Abajo, 2004) and Business management (Sharma & Osei-Bryson, 2008) amongst others. Although CRISP-DM has been successful for data mining projects, the methodology is tailored toward data mining, therefore not able to fully support BI projects that often require an integration of multiple technologies. Furthermore, the support offered by CRISP-DM in the deployment phase is limited, as the emphasis is taken from the analyst and placed upon the user. This may not be a critical issue for data mining, however when investigating technologies such as BI, the deployment phase is of fundamental to a successful project. The deployment phase becomes even more critical when the product of the data mining project is to be used as a decision support system as is often the case with BI applications.

4.4 Conventional Approaches to Decision Support Systems (DSS)

If BI technologies are to reach their full potential, then it is imperative that they are capable of providing decision makers with information, which can be understood and efficiently utilised for business processes. DSS are a system that provide one means of achieving this objective, thus for BI to reach its true potential, then the integration of DSS is a vital component. As reviewed in Chapter 2 the development of DSS has moved from the early spreadsheet based systems toward the more conventional DSS found today that are based upon artificial intelligence techniques. There are various design taxonomies that can be investigated to classify DSS; Model-driven, Document-driven, Knowledge-driven, Communication-driven or Data-driven (Data-oriented). DSS are generally classed according to the method of assistance they provide to the user, although many cannot be classed under a particular model, in contrast they are often a hybrid of models. Although classification of DSS is based upon the type of support they provide, the basic approach to development remains the same. The basic approach remains the same for various classes of DSS since the stages of development and requirements investigation are similar. DSS systems are not entirely different from other systems and require a structured approach as the development of a DSS is a very deliberate and orderly approach (Turban & Aronson, 2001).

The development of a DSS, therefore, requires a methodology to provide structure to the process of systems development. There are many 'traditional' SDLCs that can be investigated to provide this structure. Although, the various software engineering SDLCs emphasise slightly differing approaches, they all generally follow certain guidelines and processes. A traditional SDLC will consist of four elementary phases:

- *Planning:* This phase initiates with the identification of a problem. Once the problem has been identified a feasibility analysis is conducted to determine the viability of solutions, the technical feasibility, associated costs and impact upon the organisation. If a project is approved individuals and groups that are to be involved are determined and assigned.
- *Analysis*: This phase investigates the requirements in greater depth, gathering as much information as possible to propose a suitable solution model The requirements are determined from a technical perspective, along with investigations of the current approach and requirements. It is imperative to gather as much detailed information as possible to ensure that the solution model provides an accurate account of what is expected from the solution.
- *Design*: This phase investigates the more technical aspects associated with the solution. Hardware, software, usability, organisational structures are all investigated the fundamental objective of this phase is to gather the system specifications.
- *Implementation*: During this phase the system is constructed, developed tested and deployed.

The four phases of the SDLC may be approached from variety of perspectives dependent upon the overall aims and objectives of the organisation and explicit methodology that is implemented. However, the overall concept remains consistent in any structured approach. Figure 4.12 depicts a popular SDLC known as the waterfall model; which has been investigated for a variety of purposes including the design of DSS.

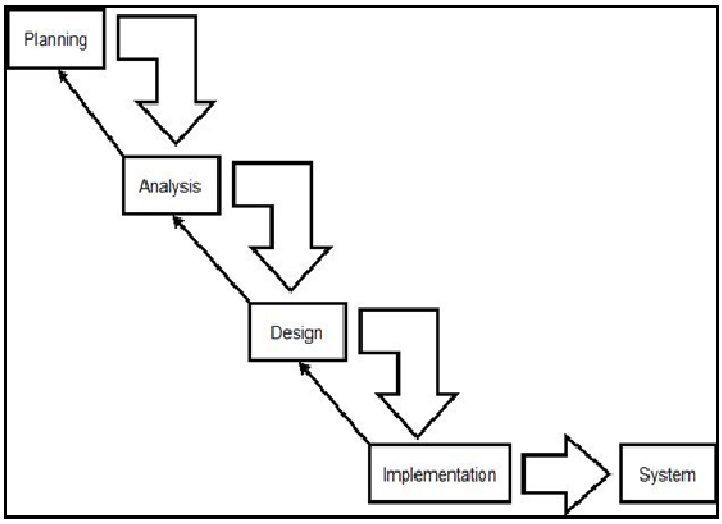


Figure 4.12: Waterfall model.

A software engineering project should follow the right-hand arrows (figure 4.12) until a fully implemented solution (system) has been realised. Although, appearing as a linear process, the left-hand arrows illustrate the cyclical nature of a SDLC. The developmental approach for a DSS should be strongly iterative. This will allow for the application to be changed and redesigned at various intervals. Each phases of a SDLC will consists of a number of processes that are expected to produce deliverables. Dennis & Wixom (2000) and Turban & Aronson (2001) have described these phases and their associated processes and deliverables as illustrated in table 4.2, which can be investigated for the design of DSS. A structured approach such as the waterfall model provides a suitable structure for many information systems, including DSS. Nevertheless, the waterfall model does not provide the explicit details and processes that are required for projects that may involve data mining. As discussed in the previous section the CRISP-DM model is far more suited to the intricate design issues that must be observed for data mining. Since BI is a technology that amalgamates a number of technologies such as data mining and DSS, it too would benefit from a tailored approach that will observe the explicit intricate details of a BI project, since ad-hoc methodologies that are investigated may fail to encapsulate all necessary design issues.

Planning: **(Why build the system?)**	1. Identify business value. 2. Analyse feasibility. 3. Develop work plan. 4. Staff project	System request. Feasibility study Work plan. Staffing plan. Project charter.

	5. Control and direct project	Project Management tools. CASE tool. Standards list. Project "binders" or files. Risk assessment.
Analysis: **(Who, what, when and where will the system be?)**	6. Analyse problem. 7. Gather information. 8. Model process(es). 9. Model data.	Analysis plan. Information Process model. Data model.
Design: **(How will the system work?)**	10. Design physical system. 11. Design architecture. 12. Design interface 13. Design database and files 14. Design program(s)	Design plan. Architecture design. Infrastructure design. Interface design. Data storage design. Program design.
Implementation: **(System Delivery)**	15. Construction 16. Installation	Test plan. Programs. Documentation. Conversion plan. Training plan.

Table 4.2: SDLC phases and their associated processes and deliverables.

4.5 Summary

This chapter has endeavoured to review and analyse the work related to this investigation, thereby providing a technical review of conventional approaches. It has been established that BI is becoming increasing crucial to business processes. As new technologies emerge, with increasing storage capabilities, advances in hardware, there has been a shift integrating information technology into many aspects of an organisation. Nevertheless, no explicit framework for integrating BI within organisations, especially to provide decision support, exists. Hence, the motivation of this chapter has been to not only, analyse these conventional approaches, in their explicit context, but moreover to identify the strengths and weakness of these approaches within the context of BI. It is for this reason that the technical review has been kept distinct to the literature review (Chapter 2), since not only does this chapter review the related work to this study, but further the technical issues which must be considered to ensure that a meta-level framework can be proposed which meets the criteria for BI.

Currently, BI integration is conducted in an ad-hoc method, utilising methodologies which are related only to constituent technologies. As a result, the conventional approach to investigating BI solutions is the exploration of an ad-hoc methodology that can be tailored to specific requirements (Golfarelli et al, 2004; Wasserman et al, 2004; Konstantinos et al, 2008; Trestian et al, 2008; Xu et al, 2007).

Consequently, the investigation of a solution frequently requires an expert and can, therefore, be a costly and complicated endeavour (Howson, 2008). While the conventional approaches considered within this chapter are robust and sufficient to integrate projects that are completely reliant upon the technology they have been designed to integrate, they do not always encompass the flexibility and dynamic requirements of a project that is to fully explore the capabilities of BI. A number of these methodologies, such as RAD; Prometheus; Gaia; SEMMA; CRISP-DM; the waterfall model, have been reviewed in this chapter. Through the analysis of the strengths and weaknesses of these approaches, the following chapter will investigate and formulate a meta-level framework tailored to the knowledge discovery for BI integration to provide decision-makers with a greater level of support and aid in the selection of suitable BI techniques. The following chapter will develop the characteristics of the models defined within his chapter to extrapolate key features which can be further explored to provide a framework which is tailored to technique selection and decision support within BI.

Section 2:

The second section of this study will define and illustrate the formulation and proposal of a suitable meta-level framework. This framework will be explored through a number of case studies each requiring novel and innovative solutions based upon the data and requirements of an organisation. The full case studies can be found in the appendix; however, it is the performance of the framework and its ability to capture requirements and structure technique selection which will form the focus of this section of the study.

Chapter 5:

KDDS-BI: A Framework for Knowledge Discovery and Decision Support through Business Intelligence

The previous chapter reviewed and analysed work related to this study, thereby providing a technical review of conventional approaches. It has been established that BI is becoming increasing crucial to business processes. As new technologies emerge, with increasing storage capabilities, advances in hardware, there has been a shift integrating information technology into many aspects of an organisation. However, no explicit framework for integrating BI within organisations, especially to provide decision support currently exists. Hence, it is the objective of this chapter to analyse BI requirements and formulate a framework that can support knowledge discovery and provide decision makers with a greater level of support. Furthermore, the framework is proposed with a view to structuring BI investigations.

5.1 KDDS-BI

Much like the CRISP-DM process model was motivated by the need for an explicit approach that was tailored to data mining (CRISP-DM Process Guide and User Manual: Available online: http://www.crisp-dm.org/), so to does BI require an explicit framework that can facilitate knowledge discovery. A suitable framework should provide guidelines for organisations, which can be observed when launching their own BI projects and interrogating data. Having investigated the conventional approaches to BI and its related technologies, it is clear that there is no systems-wide framework that exists for the investigation of BI or to address the dynamic issue that exists when interrogating data for a BI project. As a result, this research proposes a tailored framework for knowledge discovery and decision support that will facilitate an organisation wishing to utilise data they have collected, and guide the process of analysing this data to provide a greater level insight for decision support in order to gain a significant competitive advantage. Based upon the conventional approaches analysed in Chapter 4, table 5.1 illustrates a selection (since there is significant overlap) of the advantages and disadvantages to these approaches:

RAD	Iterative development through prototyping. High Level requirement gathering.	Requires repetitive input from non-experts, frequently resulting in conflicting information.
AGILE	Facilitates identification of goals, and permits changes to specifications once development has commenced.	Multiple techniques applied within the system are no explicitly supported.
Prometheus	Testing and design phase is supported. Enables the identification of external factors which can affect the performance of the system.	Does not permit various options for techniques to be explored. Inflexible for system design that does not incorporate Intelligent Agents.
GAIA	Good support for modular systems, which can have functionality added or removed.	Does not facilitate the design of systems which are not naturally modular. Only applicable to Agent-based systems. Restrictive in documentation detailing.
SAS SEMMA	Permits data to be extensively analysed from a top-down view.	Tool dependent
CRISP-DM	Support for documentation, data cleansing and data modelling.	Restive in techniques which can be used
Waterfall Model	Support for requirement gathering. Permits iterative design.	Does not support the process of data preparation.

Table 5.1: Selection of applicable advantages and disadvantages of conventional approaches to BI.

As found in Chapter 4 and illustrated in table 5.1, the various conventional approaches have provided a number of indicators to the benefits and shortcomings that exist when applying technologies that are related to BI. Conventional methods such as Agile, Prometheus, Gaia and even the Waterfall model

are effective at collecting requirements and specifications. However, a key issue for BI projects, which is also a critical weakness of these conventional models, is that the requirements are not always evident from the outset. Moreover, BI projects will often be initiated not due to explicit requirements; in contrast, an organisation may have amassed a store of data that must be interrogated without any prior realisation of deliverables. Hence, BI projects often endeavour to maximise the potential of collected data, rather then satisfy particular unambiguous objectives. In situations such as this the requirements, in addition to any knowledge must be extracted directly from the data. Furthermore, there is a distinct weakness within conventional process models for selecting appropriate techniques that enable the most effective analysis of data. This issue has been addressed by process models such as CRISP-DM and SEMMA (Evelson, 2008). SEMMA however is tool dependent, and neither SEMMA nor CRISP-DM provides effective support for data understanding, preparation and modelling. These shortcomings become even more critical when the knowledge once extracted and analysed is to be used for decision support.

Decision making in business should transfer valuable information into knowledge with a view to providing an organisation with a competitive advantage. This competitive advantage can be discovered through information that can be extracted from the data that an organisation has collected or from the optimum use of resources. Traditional approaches have not been able to address the various issues that must be considered when investigating the integration of BI. In contrast, an approach related to a particular technique is employed, frequently in an ad-hoc manner by a consultant or expert. However, if an organisation is to truly reap the benefits of BI, then it must be feasible for an organisation to launch BI projects, through a structured approach, thereby permitting a uniform approach across all departments. For this reason, a framework is required, which will permit an organisation to investigate and interrogate the data that it collects, in addition to ensuring that the most appropriate technique for the particular objective is selected. In addition, the framework should not only provide support for the knowledge discovery process but also for decisions which influence future business activities. Assessing the finding of Chapter 4, where conventional approaches have been extensively analysed and the advantages and disadvantages of conventional approaches as illustrated by table 5.1, it can be hypothesised that the he key issues that such a framework must address are:

- Identifying goals and functionality of the systems from a technical perspective (technical specifications).
- Facilitate encapsulation of key details, exploration of requirements and use cases.
- Provide support for modelling components of the system.
- Investigating various approaches as a solution to determine which technique will provide optimum results.

- Facilitate a 'Top-down approach' toward investigating specifications, functionalities and requirements, whilst facilitating dynamism.
- Emphasis upon iteration and flexibility toward problem solving.
- Enable clear documentation.
- Ensure that the data can be preparation.
- Support constant assessment of progress and outcomes.
- Facilitate interpreting output and knowledge exploration.
- Clearly presenting the output so that it can be investigated and utilised for decision support.
- Provide support for implementing and deploying a BI solution.
- Provide support for maintaining the system post deployment.

Although not unique in isolation, the concepts investigated to address these key issues, can be applied in a unique and novel manner. Figure 5.1 depicts a framework KDDS-BI (Knowledge Discovery and Decision Support in Business Intelligence) that can be explored to address the shortcomings of conventional approaches which are unable to provide effective support for knowledge discovery, extraction and decision support. Combining various aspects of conventional approaches KDDS-BI was formulated after careful and purposeful deliberation of the requirements, which needed to be addressed for the realisation of a suitable framework that will meet the complex requirements and procedures of a BI project. Prior to formulating KDDS-BI, several iterations of the development process were required. Eventually, KDDS-BI was composed through an amalgamation of concepts related to the conventional approaches discussed. These models have been extended to assimilate these concepts to provide greater support for the various BI techniques and thus, address the weaknesses of the conventional models when applied to a complex BI integration project, especially were the focus is upon knowledge discovery for a greater level of decision support.

Adopting a system engineering approach as discussed in Chapter 3. KDDS-BI provides support for the knowledge discovery process and decision support. Thus, analysis, output and interpretation are key functions thereby providing greater support for decisions. The key issue that must be addressed by a framework for BI can be discovered through 4 stages. Each stage can in turn, each be addressed through 4 sub-stages; Data investigation, Modelling, Development and Decision Support. Although the key denotes a recommended flow of information, the process model is by no-means linear. Conversely, there is a high-level of flexibility between each stage and sub-stages, allowing for a flexible approach for knowledge discovery to be adopted.

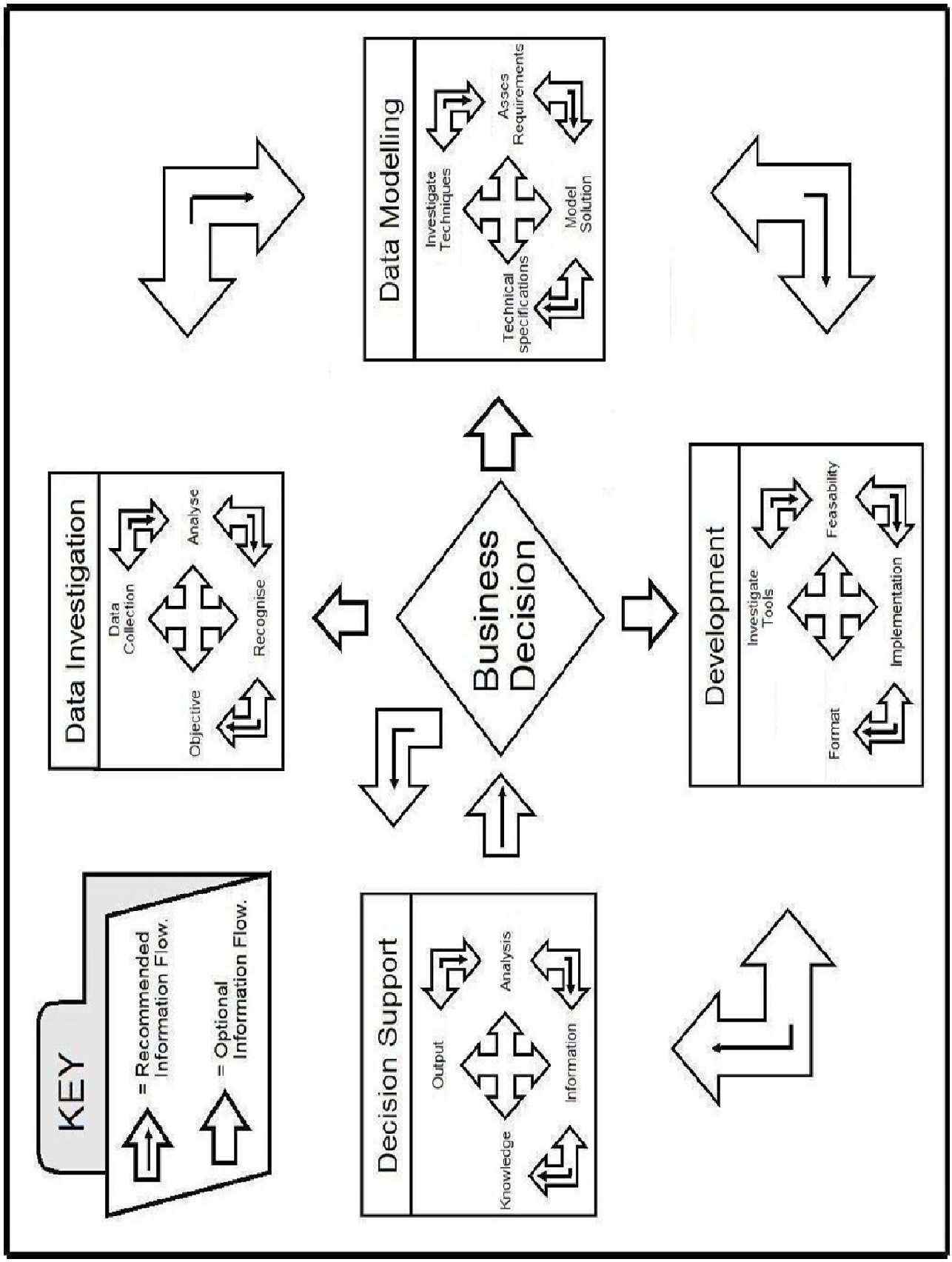


Figure 5.1: KDDS-BI process model.

5.1.1 Data Investigation

Conventional approaches to requirements gathering have been proven to be extremely successful. Approaches such as Agile have been investigated within BI and are gaining popularity (Howson, 2008). Nevertheless, Agile does not provide effective support for investigating a solution explicitly within BI; rather it proposes a generic SDLC. Despite this shortcoming being addressed by CRISP-DM, the process model does not provide a sufficient level of support for the analysis, exploration and extraction of knowledge. Furthermore, as critical to BI projects, CRISP-DM does not directly facilitate the exploration of this knowledge for decision support. Therefore, the limitations of CRISP-DM become apparent when investigating techniques other then data mining.

The initial stage of KDDS-BI; 'Data Investigation' consists of four iterating sub-stages (figure 5.2). It has already been stated that it is often the case in BI investigations that an organisation may not be aware of explicit requirements when initiating the project. Often BI projects originate from a large quantity of data an organisation has collected that they desire to analyse without the explicit knowledge of exactly what the requirements of the project are, or details they expect to find. In such an event not only information and knowledge, but also the requirements must be extracted directly from the data itself. Consequently, once the initial raw data and (if applicable) basic requirements have been collected, the data can be analysed to ascertain relevant objectives.

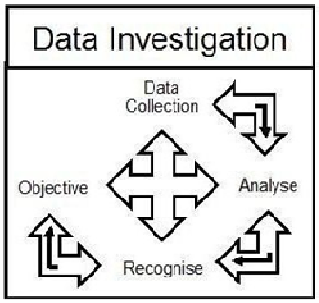


Figure 5.2: Data Investigation stage of KDDS-BI.

Davenport and Harris (2007) identify three aspects that can limit the analytical competitiveness of an organisation:

1. The organisation is plagued by missing or poor quality data, multiple definitions of its data, and poorly integrated systems.
2. The organisation collects data efficiently but often lacks the right data for better decision making.
3. The organisation has a proliferation of BI tools and data marts, however, much of the data remains un-integrated, unstructured and inaccessible.

Davenport and Harris (2007) thus, highlight that ineffective use of organisational data can be a barrier to competitiveness. As a result, within KDDS-BI once data has been collected, the data can be analysed and explored. The 'analyse' sub-stage involves the exploration, profiling and verification of the data. The data is explored to determine the format of the data, for example; number of records and fields in each table, the identities of the fields and any other features of the data. In addition, the data quality must be verified. The data is examined to address whether the data is complete or does it contain missing values, any errors that may be evident and the frequency of these errors.

Once analysed, the opportunities evident within the data can be recognised. This phase will investigate the objectives of the organisation which can be addressed using BI techniques such as querying, visualisation, reporting or intelligent agents. These objectives can then be formalised through a report in the objectives stage. This report will formalise in non-technical terms the objectives and requirements of the project that have been discovered through the initial investigation of the data. Once the objectives report has been compiled it can then be investigated, whether further data is required, in which instance the data investigation stage repeats itself, or if the data and objectives are suitable then the process can move to the second phase: Data Modelling.

5.1.2 Data Modelling

The second stage of KDDS-BI; 'Data Modelling' also consists of four iterative sub-steps (figure 5.3). The initial sub-step of the data modelling stage is to investigate techniques that can be explored to resolve the objectives identified in the 'Data Investigation' phase.

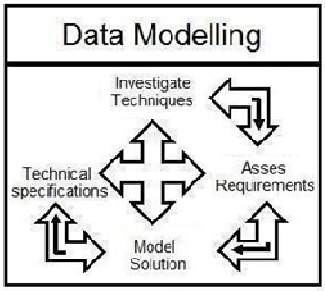


Figure 5.3: Data Modelling stage of KDDS-BI.

This initial sub-step extends the objectives report that was delivered through the data investigation. During this sub-step the various techniques available within BI can be explored and compared to the requirements in order to investigate which technique is most suitable for a particular project. For this to be effectively accomplished any pre-requisites for particular techniques must be considered, such as; does the quality, format, distribution of the particular dataset correspond to the requirements of a technique. Furthermore, it must be determined which BI strategy is most suited to the dataset. Hence,

do the requirements of the dataset result in intelligent agents or advanced analytics being the most suitable BI strategy. These assumptions can be contrasted with the data objectives and requirements.

Once suitable techniques have been explored, the knowledge obtained from this exploration can be drawn upon to assess the requirements and objectives of the investigation. This process will involve converting the objectives defined in business terms into not only more technical specifications, but also more specific objectives that can aid with the conceptual modelling of a solution. Furthermore, the 'Data Investigation' is explored from a more technical perspective. As a result, the technical relationships and discoveries such as the distribution of key attributes, or for example; the target attribute of a prediction task. In addition to the relationships between attributes; results of simple aggregations; properties of significant sub-populations and simple statistical analyses, can be investigated. This assessment of the requirements may identify the BI goals or further support the results of the Data Investigation, through refinement the initial data report, through a more technical description and quality reports. Additionally, this will provide support for future requirements and analysis such as data formatting (transformation and data pre-processing).

This technical and business assessment of the requirements will also provide sufficient information for various BI techniques to be investigated and the most suitable BI strategy to be discovered. Consequently, the processes of KDDS-BI thus far completed will provide sufficient detail for a solution to be modelled. The requirements of the project, business and technical, the relationships and intricacies of the data can all be investigated to ensure that all requirements have been discovered, thereby, providing a conceptual model for the data that will permit the discovery of all associated requirements. The details of the investigation will, however, differ depending upon the technique that has been selected, as the process of creating a conceptual model is dependent upon the requirements the technique and data being modelled. If the data is to be interrogated using advanced analytics the initially the variables (or descriptors) are investigated. If the data is most suited to a supervised approach then the data must be divided into a training set and test set. Figure 5.4, depicts a suitable model for this process. The function of a training and test set has been reviewed in Chapter 2. A percentage of the data will be allocated to a training set and the remaining data constitutes the test set.

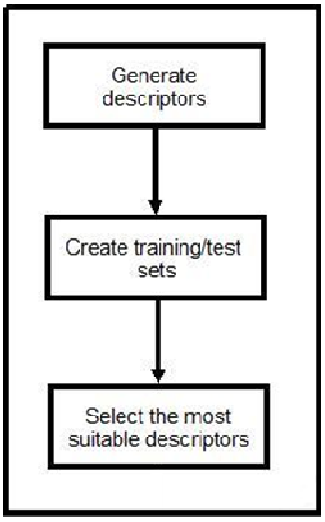


Figure 5.4: Modelling flow chart.

There are a number of methods for dividing the data; one method which has yielded positive result is 'n-fold cross validation'. The data set is divided into 'n' subsets, and the holdout method is repeated n times. The holdout method implies that 'n-1' subsets are used as the training set and the remaining subset is the test set. This process is repeated until each of the subsets has been investigated as the test set. The average of correctly classified instances can then be calculated. The advantage of this method is that the importance of how the data is divided is reduced, since every data item will constitute a test set once and training set 'n-1' times. Dependent upon the objectives the descriptors may consist of a hierarchy of importance or be of equal importance. As explored in Chapter 2, there are a number of supervised algorithms that require a key input variable. This will provide a model that can be further investigated through the aid of a suitable BI tool.

The technique for modelling will vary depending upon the technique that was deemed as most suitable in sub-step 2 of the data modelling phase. If an agent approach has been deemed as more suitable then advanced analytics then it will be essential to model the roles and functions of the agents and the relationships to the data. The roles can be determined by reviewing the requirements and related processes, with a GAIA-schemata (figure 5.5) explored for a formal description of the roles.

Role Schema:	*name of role*
Description	*short English description of the role*
Protocols and Activities	*protocols and activities in which the role plays a part*
Permissions	*"rights" associated with the role*
Responsibilities	
Liveness	*liveness responsibilities*
Safety	*safety responsibilities*

Figure 5.5: Template for schemata for defining roles

Once a suitable technique has been selected and the data models have been explored, be they agent or advanced analytic in nature, a suitable tool can be investigated. However, prior to investigating a tool, it is imperative to examine the technical issue regarding the project. By careful consideration provided to any technical constraints or requirements of the selected technique, the requirements for a suitable tool can be discovered. If necessary the project can return to the data investigation stage or any of its sub-steps. Once a suitable technique and model have been explored and established, in addition to the technical requirements or constraints, such as intended user facilities. A means through which the objectives can be realised can be investigated.

5.1.3 Development

The KDDS-BI process model has so far facilitated the discovery of requirements and suitable models based upon these requirements. It is now, therefore, possible to examine a BI-based tool, which facilitates implementation and exploration of these models. It may be necessary to develop a tool; however, the growth of BI and its related technologies have resulted in a number of commercial and open source tools that have become available. These open-source or commercial solutions can be investigated to provide effective solutions depending upon the requirements of an organisation.

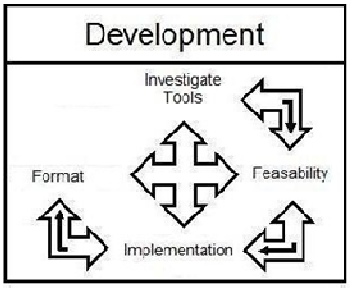


Figure 5.6: Development stage of KDDS-BI.

Due to the availability of BI tools, especially open-source tools and development tools that can be modified to meet the needs of an organisation, the initial sub-stage of the 'Development' stage (figure 5.6) is to investigate these tools. Having previously considered the requirements, technique and created a model, the functional and technical requirements for the tool have been established. There are a number of tools / platforms that are available for both agent implementation and advanced analytics, which can facilitate BI projects and investigation. If there is no tool or platform available through which to investigate the requirements of a project, then it may be necessary to develop a platform for analysis. In such an instance, it is recommended to progress to the 'Implementation' sub-stage. However, if a suitable tool or platform through which to investigate the project exists, it is imperative that the feasibility of said tool/platform be explored and reviewed. This will enable the identification of any potential obstacles, short-comings or may even result in the proposed solution

being unsuitable. if necessary it may be required to return to any of the previous stages of KDDS-BI. Furthermore, as a pre-requisite to the 'Implementation' and 'Format' sub-stages; the investigation of modifications that will be required to integrate the BI-tool and data within an organisation must be conducted. Consequently, the 'Format' sub-step can be further divided into two-sages that can be independently explored; 'Tool feasibility' and 'Data feasibility'.

- *Tool feasibility*: Involves further investigation of the tool or platform that has been selected. It is essential to consider the tool selected is capable of fulfilling all requirements and can be integrated within the organisation. A simple example of tool feasibility is that, many agent development platforms are Java-based. It must, therefore, be established that a java-based tool is suitable within the organisation.
- *Data feasibility*: Involves investigation of the requirements of the data in relation to the tool or platform selected. A simple example of data feasibility is that can the current method of storing data be integrated with the tool. Furthermore, many advanced analytical tools require that the data be formatted into a particular data type, thus it must be established that this is possible.

Once an approach has been deemed feasible, the 'Implementation' sub-stage can be explored. The implementation stage is the integration of the tool or platform. As a result, the model investigated in the 'Data Modelling' stage will be realised. Whilst giving consideration to the output of the tool feasibility study. In the event that a suitable tool or platform has not been discovered, then a technical solution must be devised from a conceptual design. Prior to proceeding to the final sub-step of the development stag, it is recommended that the development in such an event, follow a traditional SDLC, such as those discussed within Chapter 4. Once the tool or platform has been discovered or developed, it must be integrated within the current business processes. Since many of tools are 'stand-alone tools', the implementation stage may only extend to training users once the system has been successfully installed.

Once implemented, to be fully integrated within the organisation, the input of the system must be formatted. This may involve integrating current data stores, or in contrast, formatting data into a suitable data type. The formatting sub-step is essentially the realisation of the issues addressed within the data modelling and feasibility. Once the system and data has been integrated and the data model has been converted from a conceptual design to one that can be implemented and investigated, the system can be explored to provide decision support.

5.1.4 Decision Support

Once the BI tool has been successfully investigated and integrated within business processes along with any data formatting that is required. The system can be investigated with real data with a view to providing a greater level of support for the process of concluding executive decisions (figure 5.7).

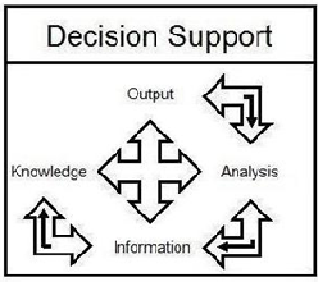


Figure 5.7: Decision Support stage of KDDS-BI.

Initially, the data is run within the system to provide output. This phase is effectively the realisation of the results obtained from the previous stages. As a result, the techniques investigated in the 'Data Modelling' stage, and the modelled solution, formulated in the 'Format' sub-stage of the 'Development' stage are discovered to provide an output via the implemented BI solution. This output will often be in the form of statistics. Additionally, the output can be collected in the form of summaries of information, statistics, graphs, curves and charts amongst others formats, especially in the case of advanced analytics. The output in its raw form represents information which must be analysed, so that useful discoveries can be extracted from this information, and can be more easily understood and transformed into reports to provide business information. Thus it is the objective during the third sub-step 'Information', to extract valuable specifics from the analysis of the output, This information must examine the analysis in detail to ensure that all information can be placed within context and provide feedback relevant to the objectives, in addition to any additional information that may become evident through the analysis, even in the event that this information may not have formed part of the original objectives. The 'Information' sub-step naturally extends to the next sub-step; 'Knowledge'.

During the 'Knowledge' sub-step, the discovered information is explored and converted to a more conventional method that can aid decision support. Thus, the objective is to transform the output and information discovered from the output and analysis into reports and representations that can be more readily understood by decision makers within an organisation. These reports can further aid in the identification of existing problems. Reports taken at regular intervals can be compared to identify any discrepancies in addition to the support that provide in the short-term. This information in turn can be transferred into knowledge that can provide support for decision making. There is no predefined template for the output, rather this will depend upon the output and context, especially with regard to

who will be reviewing the information and discovering knowledge from it. Consequently, during this sub-step, once the output has been obtained, analysed with information extracted that can be converted into a report that facilitates information discovery it provides knowledge that can be transferred into business decisions, thus fulfilling the primary objective of the framework to provide knowledge that can be exploited for effective decision support. Figure 5.8, depicts how the decision support stage of KDDS-BI can be extended to provide business decisions. This step is highly iterative, as once a decision has been taken it is continually reviewed.

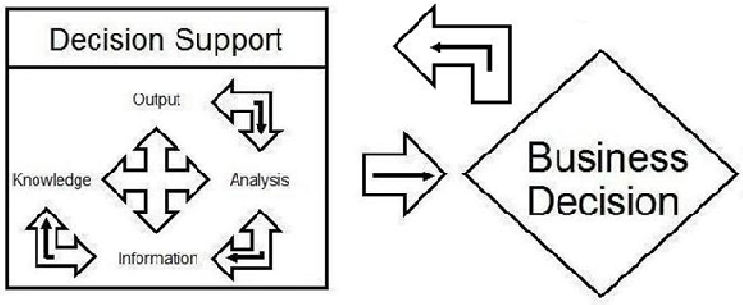


Figure 5.8: Extending the Decision Support stage of KDDS-BI to provide business decisions.

Hence, rather than cease the process model with the identification of a business decision, the 'Decision Support' and 'Business Decision' process repeat continually. Although, this iteration will most commonly take place between the decision support stage and the discovery of business decisions, it is possible to return to any previous stage should the need arise. Hence, as new information becomes available, it is imperative that these discoveries continually provide invaluable knowledge and in turn this knowledge provide effective business decisions. This will continually provide an organisation with a competitive edge.

5.2 Conclusion

This chapter has explored the findings of the previous chapter to investigate the strengths and weaknesses of the identified conventional approaches to formulate a framework explicitly for knowledge discovery to support decision makers through BI strategies. Through the discovery and identification of key issues which are critical to a valid framework for BI a framework was proposed: KDDS-BI. KDDS-BI consists of 4 major phases, with each stage consisting of 4 sub-steps, prior to the final step of providing the support to a decision maker for a decision to be contemplated, (the process of implementing the 'Business Decision' is generally considered a extension of the fourth stage). The design of KDDS-BI extends the basic model of CRISP-DM to permit more dynamic and flexible framework, which permits the investigation of various tools, and techniques, in addition to a combinatorial approach as required by BI investigations. Since BI amalgamates a number of techniques, to fully exploit the potential of BI it is imperative to enable these combinatorial approaches to be explored. In addition, it may not always be clear from the outset which technique is

the most suitable; consequently support has been integrated to facilitate the discovery of these techniques. Furthermore, prominence has been provided to the dynamic nature of BI investigation. Hence, KDDS-BI, suggests a recommended flow of knowledge, however, does not compel the flow to these stages, in contrast, the support for flexibility, enabling stages to be repeated or omitted has been incorporated and is supported.

Initiating the 'Data Investigation' stage, KDDS-BI provides a greater emphasis upon extracting objectives directly from the data, as opposed to the conventional approach of gathering requirements from a user. The rational for such an approach, is that BI is employed to uncover hidden patterns and trends. Consequently, without intimate knowledge of BI techniques, the potential of the data may not be instantly evident. It is therefore, more appropriate to examine the format and type of data to obtain preliminary objectives. The process of obtaining objectives directly from the data is one which is entwined with the 'hard systems engineering' approach, which underpins KDDS-BI. Once the preliminary objectives have been extrapolated, these objectives can be more intimately scrutinised, through the investigation of potential techniques. Once objectives have been determined possible techniques can be explored and a suitable model investigated in the second phase 'Data Modelling'. Since this requires explicit knowledge of BI. KDDS-BI is not proposed as a replacement for an expert or consultant. In contrast, the aim of KDDS-BI is to support and structure at a meta-level the approach adopted for BI investigations. Once suitable techniques or combination of techniques have been determined, the requirements and objectives should be assessed. Since it may have become apparent that the original objectives had been optimistic, or in contrast a deeper level of analysis is possible then previously been considered possible. With the preliminary analysis of techniques and objectives completed, a solution should be modelled; this will not only provide a conceptual template for the solution, but also unveil the exact functionality required from the tool through which to realise the solution. The third phase 'Development' facilitates the investigation of a suitable means through which to realise the conceptual model. However, the feasibility must also be assessed, in the event that the conceptual model is unfeasible, or additionally the inclusion of greater functionality is possible. The platform can then be implemented; this may involve installation or in contrast extend to the development of a suitable solution. In addition to implementation, a degree of formatting may also be required, thereby, enabling the integration of the BI solution with organisation practices. The penultimate phase of KDDS-BI; 'Decision Support', provides support to explore the solution and attain output or results. These finding can subsequently be analysed to extract information and placed into the context of the domain thereby converted into knowledge. This knowledge once attained, can henceforth be provided to decision-makers. Although, the decision concluded, or even if the knowledge provided is utilised is beyond the scope of KDDS-BI. What KDDS-BI will have achieved, is the in-depth interrogation of a provided dataset through an innovative BI strategy.

Chapter 6:

KDDS-BI: Case Studies and Evaluation

In order to investigate the advantages and disadvantages of the proposed framework KDDS-BI, it will be explored with the aid of three case studies. The case studies traverse a number of application domains to demonstrate that the framework is domain neutral and applies to BI investigations at an abstract level. 'Case Study 1' conducted in conjunction with the insurance company: Sentient Insurance and covered in depth in Appendix B, illustrates the functionality and support which can be provided to a BI investigation in the domain of Direct Marketing. 'Case Study 2' conducted in conjunction with Tesco Club Card and covered in depth in Appendix C; examines the performance of KDDS-BI within retail to increase the impact of Sales Promotion strategies. Whilst, 'Case Study 3' conducted in conjunction with London Underground Ltd. (LUL) and covered in depth in Appendix D, examines KDDS-BI within the domain of Managing Organisational Resources. However, it is the aim of this chapter to analyse the contribution of KDDS-BI within the process of conducting these case studies and providing a level of support that has previously been unavailable within BI investigations.

6.1 KDDS-BI: Case Studies

The objective of this study has been to propose a framework which at a meta-level which can provide support and structure for selecting suitable tools and techniques explicitly within BI. Chapter 4 reviewed a number of conventional approaches, which were further analysed in Chapter 5 to propose a framework that can meet the requirements of this study. The proposed framework: KDDS-BI, provides a meta-level framework which can provide structure to the process of investigating BI to refine business processes and performance. Since the framework provides support at a meta-level, KDDS-BI provides explicit support for the selection of tools and techniques, however more importantly, since the approach is structured through a similar methodological approach irrespective of the technique, tool or application domain, KDDS-BI provides opportunities to retain information and experiences. This in turn permits a degree of knowledge management and removes the reliance upon ad-hoc processes, thereby permitting a structured approach to be adopted through which BI can be explored at an abstract level.

Providing support for the process of techniques selection is imperative to a successful BI investigation. Since the underlying objective of BI is to refines business processes and performance, through efficient use of information, which can be achieved by interrogating data at a deep level to unveil hidden and meaningful information. Equally, it is essential that information be provided to the correct departments or individuals within the organisation in a format which is not only easy to understand but can be swiftly integrated within organisational decisions ensuring future activities perform optimally. As discussed in the Introduction (Chapter 1) to this study, Stevenson (2006) has identified three basic functions of an organisation; Finance; Marketing; and Operations.

Finance to some extent can be considered a function of operations and marketing. Although independent in terms of allocating resources, it is marketing and operations were the vast majority of finance is expended and accumulated. Furthermore, the motivation for integrating BI within organisational processes is to refine these processes so that they can generate greater amounts of revenue, through more data awareness/insight, or refine processes so that the associated cost can be reduced. This study will as a result focus upon the potential of KDDS-BI to permit the exploration of BI within ‘Marketing’ and ‘Operations’.

6.1.1 Case Study 1: Direct Marketing (Appendix B)

The initial case study (the full details of which can be found in Appendix B), explored KDDS-BI in the domain of ‘Direct Marketing’. This domain permitted the evaluation of KDDS-BI and the

frameworks ability to provide decision makers with the means through which to predict consumer segments whom are likely to respond to particular marketing strategies, thereby increasing the return on investment.

Although the focus of this case study is upon direct marketing. BI can be extended to analyse a far greater range of marketing strategies. Marketing strategies consist of various approaches such as; branding, advertising, consumer behaviour analysis and distribution amongst other strategies. Many of these are encapsulated within the 'Marketing Mix', which shall be explored in greater detail in the second case study (Section 6.1.2 & Appendix B). However, for the purpose of this case study, direct marketing has been selected as the area of analysis, since this domain provides one which encapsulates the data and dilemmas that are common to many organisations, especially for companies such as Sentient Insurance who are competing with a product such as 'insurance policies', which is difficult to differentiate and distinguish from rival products. Subsequently, a greater emphasis is placed upon an organisations capability to target consumer and increase awareness, thereby coercing the particular product to the forefront within a competitive market.

6.1.2 Case Study 2: Sales Promotion (Appendix C)

For the purposes of examining the performance of KDDS-BI within a dimension of marketing other than direct marketing, the second case study (Appendix C) focused upon 'Sales Promotion'. Sales promotion is a domain which facilitated the analysis of the ability of KDDS-BI to provide decision makers with the means through which to target consumers at the point of purchase, cross-promote products and increase revenue in areas performing sub-optimally. As discussed in greater detail in the introduction for Case Study 2 (Appendix C; Section C.1), the 'Marketing Mix' is far more intricate and consists of various marketing strategies beyond merely sales promotion. Although BI and by extension KDDS-BI could be explored to investigate any of these various strategies. It is the particular problems posed by sales promotion which has resulted in it being determined a prime candidate for this case study. Sales promotion, face several significant challenges; not only is it one of the strategies which accounts for a significant proportion of marketing strategies, but like the domain of 'Direct Marketing', examined in Case Study 1, it is one which an increasing number of companies are employing to differentiate their branding. Through a combination of sales promotion and 'loyalty schemes', large retailers are endeavouring to ensure repeat custom. However, this poses new challenges, in not only ensuring that the correct sales promotion technique is utilised in explicit regions, but also that the sales promotion is bearing a significant impact. As discussed this impact can be difficult to asses; as a result, by 'over-promoting' products, an organisation can diminish profits unnecessarily. Consequently, it is these challenges for which conventional techniques have provided less than desirable results, which has resulted in sales promotion being determined as one of the more suitable areas for evaluating KDDS-BI.

6.1.3 Case Study 3: Managing Organisational Resources (Appendix D)

Both Case Studies 1 & 2 have explored the potential of BI (structured through KDDS-BI) within various aspects of marketing. However, marketing expresses one facet of organisation activity. As discussed, primary organisational activity further encompasses 'Operations' (Stevenson, 2006). Operations defines the process of not only producing goods and services, yet further the management of resources, distribution of goods, services and resources, relationships with suppliers and consumers, amongst other functions. Operations, thereby extends across the entire supply chain. As organisations endeavour to compete upon a global stage, often providing goods and services across a variety of domains, it is of increasing importance that the operations of an organisation be managed optimally. Albeit, many of these fields provide a suitable domain within which to investigate BI integration. It is this ever-increasing necessity to ensure that organisations can optimally manage the operations process, thereby effectively competing with rivals, which has motivated the evaluation of KDDS-BI within this domain. Hence, Case Study 3 (Appendix D) evaluated the potential of KDDS-BI to structure BI investigation for 'resource allocation'. Case Study 3, therefore permits the analysis of the potential for BI to provide decision makers with the means through which to optimally allocate and distribute resources throughout an organisation through an approach which has been previously explored within two significantly diverse application domains.

Since, it is of substantial interest to examine the applicability of novel and innovative BI strategies discovered through KDDS-BI, within a dynamic, complex and unpredictable environment. Although there have been various attempts to manage the operations process, and ultimately provide decision-makers with preeminent information through which to guide organisational direction. Many of these conventional attempts have centred upon various forms of ERP and APS support systems, however as discussed within Appendix D (Section D.1.1), there are several shortcomings of these conventional systems which can be improved through augmentation with BI. Accordingly, resource management has been determined to be the most suitable domain, thereby, providing the opportunity to investigate the capacity of BI to increase the performance of these support systems within such a convoluted domain.

6.2 KDDS-BI: Results, Analysis and Evaluation

The three case studies have demonstrated the way through which KDDS-BI can be applied to exiting datasets which have been collected through routine operations, yet through BI can be analysed to reveal hidden information and provide decision support for business directions. In each instance KDDS-BI permitted a structured approach to be adopted, and solutions proposed which provided business / data insight for decision support. The motivation however for KDDS-BI has been to provide a meta-level framework which is not tool, technique or domain dependent, the case studies

which provided the means through which to demonstrate KDDS-BI can be found in Appendices B, C and D. However, to validate the framework it is vital that the performance of KDDS-BI in each case study be contrasted.

6.2.1 Data Investigation

The initial step of KDDS-BI is 'Data Investigation'. This provides the opportunity to inspect and scrutinise the data set which will be used for analysis. Since KDDS-BI is a meta-level framework, it does not assume any dependence of the data set. Unlike traditional methodologies KDDS-BI does not require data be collected specifically for the project, rather, since the aim of KDDS-BI is to facilitate the integration of BI, the frameworks primary aim is to further extend the use of existing information and facilitate a deep level analysis.

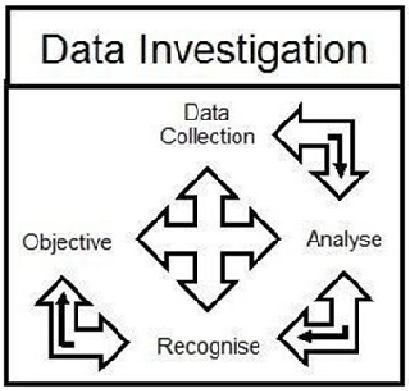


Figure 6.1: Data Investigation stage of KDDS-BI.

For each of the datasets in the case studies the data had not been specifically collected for a deep level analysis. The data set for Case Study 1: Sentient Insurance; consisted of information which is routinely collected as someone purchases a policy. The data set for Case Study 2: Tesco; consisted of information extracted from the details of Tesco's loyalty scheme (Tesco Club Card), and Case Study 3: London Underground Ltd (LUL); the data analysed was based upon the resources, both physical and human which enable the organisation to complete its operations. The primary aim of these data sets is to record customer data, increase brand loyalty or record organisation assets. However, BI can be applied to each dataset to further extract hidden information and provide a greater return on the investment of data collection. KDDS-BI (as illustrated by figure 6.1) provides the mechanisms to guide this process. Once the data had been obtained, it was in each instance, analysed at a meta-level, to discover details specific to that instance and create a list of what details are being categorised, is there any missing or incomplete information and what is the format, such as numerical and/or categorical/nominal, this preliminary analysis enables opportunities to be recognised and a set of objectives to be composed. For each of the case studies the preliminary objectives were:

- Case Study 1:
  - Identify which customers are likely to purchase a particular insurance policy.
  - Identify the attributes that distinguish these customers, so that they can be effectively targeted.
  - Identify any attributes that identify consumers unlikely to purchase a policy.
  - Ensure that the results provided are accurate and can, therefore, be proposed as scientifically valid results.

- Case Study 2:
  - Identify region with lowest sales.
  - Identify the level of impact of promotions on fresh meat sales.
  - Explore opportunities for sales promotions to enhance sales.
  - Identify which products are purchased in together and therefore benefit from joint sales promotions. This information can further be investigated when determining the most effective store/shelf layout, thereby placing products with strong relationships in close proximity.
  - Ensure that the results provided are accurate and can, therefore, be proposed as scientifically valid results.

- Case Study 3:
  - Investigate an intelligent system for optimal allocation of LUL resources.
  - Investigate the capability of an intelligent system to provide decision support.
  - Ensure that the proposed system is effective and robust, thereby capable of providing reliable decision support.

6.2.2 Data Modelling

In each case study the preliminary objectives provided a platform upon which further analysis could take place. Extracted directly from the data, through examination of the data's meta-information, techniques most suited to the data were explored. Since the initial phase of KDDS-BI had provided a number of documented results, namely, the meta-details of the data, objectives specific to that data and details relating to any information which is missing or incomplete, techniques could be objectively considered (figure 6.2).

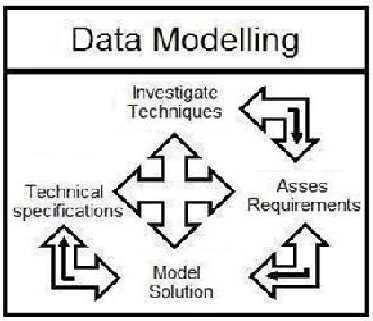


Figure 6.2: Data Modelling stage of KDDS-BI.

In the instance of Case Study 1, at a meta-level the data provided historical information, detailing explicit customer socio-demographic details, in addition to previous purchasing decisions; these details were suitable for a supervised approach. The aptness for this data set to a supervised approach, coupled with the preliminary objectives, provided the required information to delve deeper into specific techniques (a collection of which have been expanded upon in Appendix A). Selecting the correct technique is imperative since BI amalgamates a number of different approaches, of which selecting a few key techniques will enable the most efficient extraction of hidden and meaningful data from the data. For this data set it was determined that the most effective interrogation technique would be a supervised learning approach consisting of 'Bayes Theorem', 'Naïve Bayes', 'Bayesian Networks', 'Decsions Trees' and 'Production Rules', these techniques are explored in depth in Appendix A, however provide a suitable means of identifying explicit characteristics which can be utilised, furthermore cross analysing the performance of these techniques provides significant technique insight, for future data sets that the comp any may wish to analyse.

For Case Study 2, the dataset related to information collected through the Tesco Club Card loyalty scheme. Detailing socio-demographic information collected when an individual enlists for the scheme, in addition to the products purchased and the dates of these purchases. Unlike the data set interrogated in Case Study 1, the information was not provided explicitly for each customer, in contrast the customers, based upon key attributes had been pre-grouped, by both social (life stage) and demographic (location) classifications. Once the preliminary objectives had been considered it was deemed that a combinatorial approach of supervised and unsupervised techniques would provide the most effective insight, since objectives such as assessing the impact of sales promotions and in turn future sales promotion strategies are ideal for exploration through supervised models. In contrast, discovering methods to increase revenue in areas performing sub-optimally requires forecasting and predicting the impact of new sales promotion techniques within the area for which there is no historical data, therefore unsupervised learning is a more robust technique. Whilst association rule

mining provides an ideal means for analysing which products can be cross promoted or sell in conjunction to direct shelving policies.

The third case study analyses the resources of LUL and how the management of these resources can be improved, especially with regard to exploring 'what if' opportunities. Given the meta-information of the data, it is clear that the data describes individual items, which together enable LUL to successfully complete operations. This information, the objectives and the dynamic, flexible and unpredictable nature of LUL operations alludes itself to a solution developed around Intelligent Agents. However, the conclusion of using Intelligent Agents alone is insufficient, of equal importance is the mechanism utilised for decision making. Since KDDS-BI has facilitated the discovery of a preliminary set of objectives and Intelligent Agents as the appropriate technique, the Agent Decision Mechanism cantered around algorithms, such as 'Anytime Algorithms' (see Appendix A, section A.4.1.). Anytime Algorithms posses the ability to deliver results mid calculation can be proposed.

Although each case study in this research has been performed independently for varying organisations, the exploration of various techniques provided the opportunity to only require background documentation once, which can then be used a reference point for future projects. As documented in Appendix A (section A.1), the reference points provided the opportunity to reassess techniques for each case study without having to repeat the documentation process. Once the appropriate technique had been discovered, KDDS-BI provides the opportunity to assess the requirements (figure 6.2). This provides the opportunity to consider what is viable or deduce any further objectives which are possible, thereby providing a full set of viable objectives. Hence the objectives for each of the case studies can be more accurately proposed (for a full description of these objectives can be found in Appendices B, C & D):

- Case Study 1:
  - Interrogate the data set using a variety of supervised learning techniques.
  - Analyse the performance of these supervised learning techniques.
  - Identify which customers are likely to purchase a caravan policy.
  - Identify the attributes that distinguish these customers, so that they can be effectively targeted.
  - Identify any attributes that identify consumers unlikely to purchase a policy.
  - Ensure that the results provided are accurate and can therefore be proposed as scientifically valid results.

- Case Study 2:
  - What is the worst and best performing area in terms of sales?

- In the both the best and worst performing region, which is the largest and smallest consumer and what are the products they purchase the most?
- Is the most product purchased the most / least by the highest / lowest consumer group the same in each area or does this vary for each area, i.e. is it a case of a particular group just being a high consumer or is there a particular trend in an area which can be replicated in others.
- Is this product consistent with the overall, nationwide highest selling product?
- What products can be cross promoted with the highest selling product?
- What is the impact of all sales promotion techniques on this product and is one sales promotion technique more successful?
- What is the overall impact of sales promotion (nationwide) and who is the consumer who sales promotion has the smallest and largest impact on (All products)?
- For both types of sales promotions which consumer responds best and worst to which promotion (lowest performing region)?

➢ Case Study 3:

- Investigate an intelligent agent–based system for optimal allocation of LUL resources, focusing upon the Piccadilly line.
- Include the possibility of duel trains of the same route to mimic, alternative LUL lines.
- Provide an intelligent agent–based system for the scheduling of LUL train drivers.
- Investigate the capability of the proposed systems to provide decision support.
- Ensure that the proposed system is effective and robust, thereby capable of providing reliable decision support.

The stages of KDDS-BI thus far, have enabled a structured approach to be adopted toward the discovery of a complete set of objectives and associated techniques which will enable the analysis of data a deep level to fulfil these objectives. However, in order to ensure that these objectives can be realised and ensure that any major obstacles are pre-empted, KDDS-BI provides the functionality to create a conceptual model for the proposed solution and enable the designer to epitomize how the technique will be used for the explicit objective. A detailed description of each modelled solution can be found in the full case studies contained in the Appendices: B, C and D).

An added benefit which can be reaped from the ‘model solution’ sub-step, is that through extension the technical requirements can be discovered, this can range from certain constraints provided by the system to the hardware specifications. For this study a mid-spec PC was utilised for all case studies:

- Intel Celeron 440 Processor (2 GHz. 800 MHz FSB, 512 KB Cache),

- Windows Vista Home Edition,
- 1 GB RAM,
- 80 GB HDD.

The motivation for the hardware constraint was to represent a platform which did not require any large degree of investment and demonstrated the accessibility of innovative solutions that can be realised through the framework by placing a greater emphasis upon the technique and BI approach. Furthermore, by limiting the hardware specifications, KDDS-BI can not only demonstrate its flexibility, but further the potential to integrate into daily business process through limited requirements or resources.

6.2.3 Development

The third phase of KDDS-BI: 'Development' (figure 6.3), provides the mechanism through which the modelled solution can be realised. This process involved the review of various BI tools, in particular both those pertaining to Advanced Analytics and Intelligent Agents, the full review of these tools can be found in Appendix A, section A.5 and A.6.

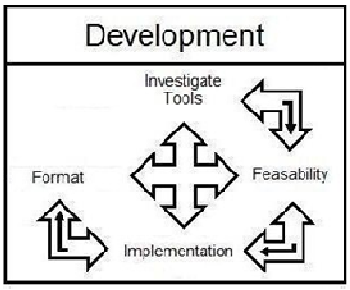


Figure 6.3: Development stage of KDDS-BI.

The review of BI tools in Appendix A, again provides an instance of how KDDS-BI can facilitate the management of knowledge. Since the review of tools, need only be conducted once, for each case study and any future projects the same table could be referred to, in order to discover which tools best served the modelled solution. However, a review of all possible tools alone is insufficient; additionally it is imperative that the technical limitations and modelled solution be considered in-depth to ensure that modelled solution can be matched to the most appropriate tool. After careful deliberation, it was determined that Case Study 1, would employ a combination of an open-source BI platform: Weka; Java and HTML; whereas Case Study 2 employed a combination of Microsoft applications, consisting of: .Net Framework; Business Intelligence Development Studio, Microsoft Excel, SQL Server 2008, SQL Server Data Mining Add-Ins for Excel, SQL Server Management Studio, SQL Server Visual Studio. While the solution of Case Study 3 was based upon the Magent-A application and the Jade platform.

With the solution modelled, tool determined and feasibility established, it is possible to implement the solution. For each of the case studies a detailed step-by-step guide of the process of developing a solution can be found in Appendices B, C and D. However, each of the case studies provides an insight into the abstract nature of KDDS-BI and how the framework can be tailored to varying techniques and platforms. KDDS-BI is applied to provide a meta-level structure, hence is neutral to the BI tool, be it based on open source or commercial programs, or BI techniques, be they Advanced Analytics or Intelligent Agents. As a result, Case Study 1 provides a BI platform based upon open-source software which has been customised through Java and HTML. The solution as demonstrated in Appendix B, can be utilised as either a desktop application or online, in the form of Software as a Service (SaaS). Case Study 2, proposed a solution based upon a combination of various commercial applications. Although requiring a higher cost than the solution for Case Study 1, the use of commercial applications demonstrates how the higher cost can be offset by a reduced implementation time and built upon any legacy commercial applications that may previously be installed by an organisation. Case Study 3 developed the concept of commercial and open-source tools further, through the provision of two solutions, a high granularity Multi-Agent system based upon a commercial platform and a Multi-Agent system based upon course-grained agent developed though an open-source platform. These solutions did nevertheless require further formatting of the datasets prior to analysis. Again, the process of formatting the data for each case study can be found in the fully documented case studies, which can be found in Appendices B, C and D. The 'Format' sub-step encompasses a fundamental process of a BI investigation. Since, BI frequently draws upon a number of technologies, tools and resources, it is imperative that the data be formatted, cleansed and prepared so that it can be effectively interrogated, prior to providing any support for the decision making process. It is in light of this requirement, that KDDS-BI, ensures that for any BI investigation the process of ensuring the data is formatted correctly be given a substantial level of consideration.

6.2.4 Decision Support

The final stage of KDDS-BI represents the cumulative of all previous KDDS-BI stages and the opportunity to gather and analyse the output from the proposed solution for decision support. As illustrated by figure 6.4, this stage is a cyclical one since the process of analysing data is one that can be continually assessed in light of new data and market shifts.

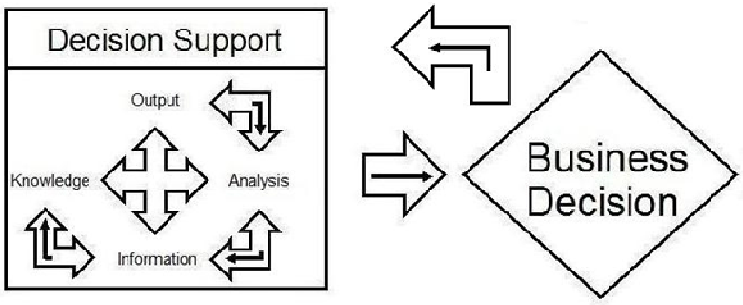


Figure 6.4: Decision Support stage of KDDS-BI.

Case Study 1 provided a unparalleled insight into what characteristics make consumers susceptible to a particular direct marketing campaign, in addition to those characteristics that would result in particular consumers generally being unreceptive to such a strategy. An added insight provided through Case Study 1, is of the performance of individual classifiers when applied to particular data types. This again provides an opportunity to retain knowledge from an investigation which can be documented for further studies.

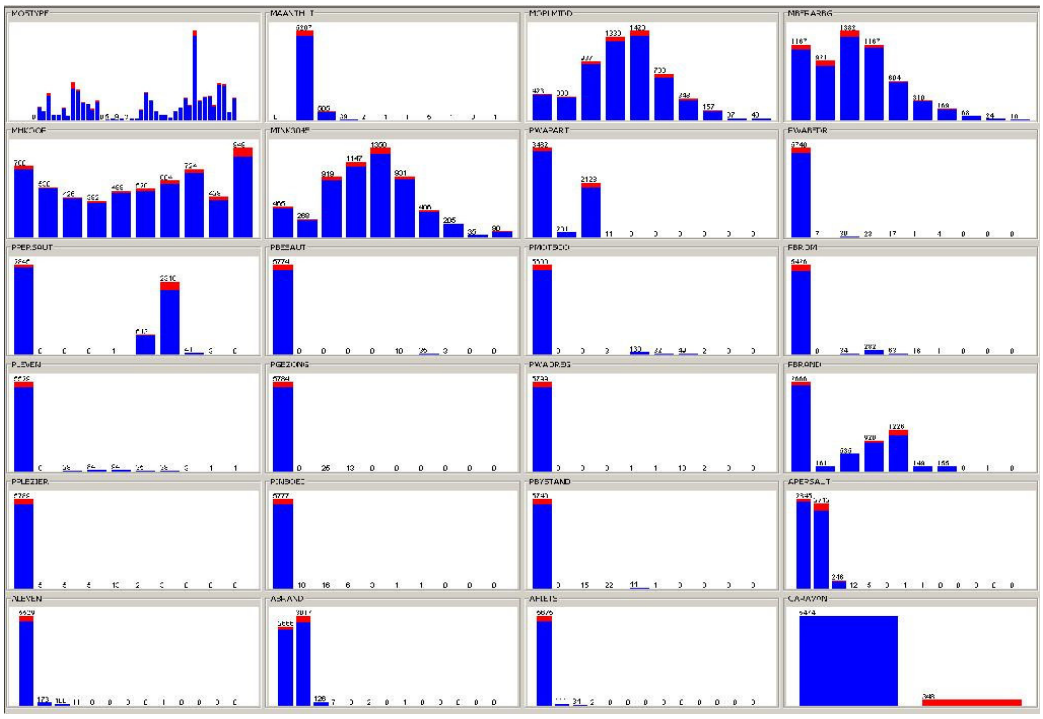


Figure 6.5: Attributes individually visualised once the 'NaïveBayes' model had been investigated.

For Case Study 1, it was Naïve Bayes, which provided the most reliable insight, however the analysis of other classification techniques enabled a level of cross-analysis. Once the models had been re-evaluated the following 7 attributes were amongst the most significant when assessing the criteria for those who purchase caravan insurance policies (figure 6.5):

- PPERSAUT = Contribution to car policies.
- PBRAND = Contribution to fire policies.
- APERSAUT = Number of car policies.
- MOSTYPE = Customer Subtype.
- PPLEZIER = Contribution to boat policies.
- MINKGEM = Average income.
- PWAPART = Contribution to private third party insurance.

The identification of these attributes can be utilised to focus marketing campaigns toward those individuals who will be amongst the largest purchasers of insurance policies, thereby increasing the return on investment.

Case Study 2 examined the impact of sales promotion techniques and endeavoured to discover insight into the purchasing habits of particular consumer segments. Focused upon the lowest performing region, which was discovered to be the 'South West' and through careful analysis of output, compiled through a number of Advanced Analytical techniques, it was possible to gain insight trends within the customer groups, sales promotional techniques and products, to discover the key performance indicators which would ensure that the most appropriate retail strategies can be applied to increase the sales performance of that particular region. As illustrated by figure 6.6, once the relative impact of sales promotions on sales had been determined, it was possible to further analyse the impact of individual techniques.

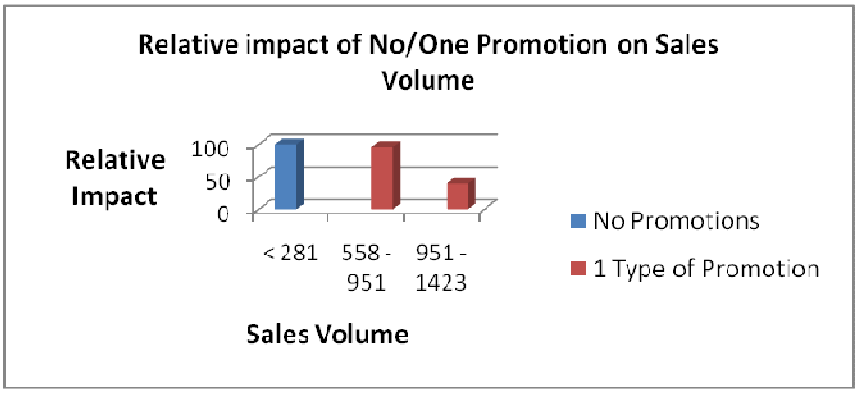


Figure 6.6: Statistical analysis for the impact of sales promotions.

Further analysis provided the insight to discover that the highest purchasing consumer segment in the area was Pensioners, which was the contrary to a nationwide analysis, where Pensioners were amongst the lowest consuming groups.

Analysis was also conducted upon the various products to discover which are liable to particular sales promotional strategies or for cross promotion. As an example of the results (taken from Appendix C), it was discovered that 'temporary price reduction' increased sales more significantly then 'multi-buy' promotions. Thus, if the promotion is motivated by the incentive of increasing sales, temporary price reduction, is a more ideal form of sales promotion. This should not result in multi-buy promotions being disregarded. Multi-buy promotions can be applied to increase the sales of a combination of related products. In order to uncover relationships between products initially Neural Networks were explored to explore the effect that products bear upon one another (figure 6.7).

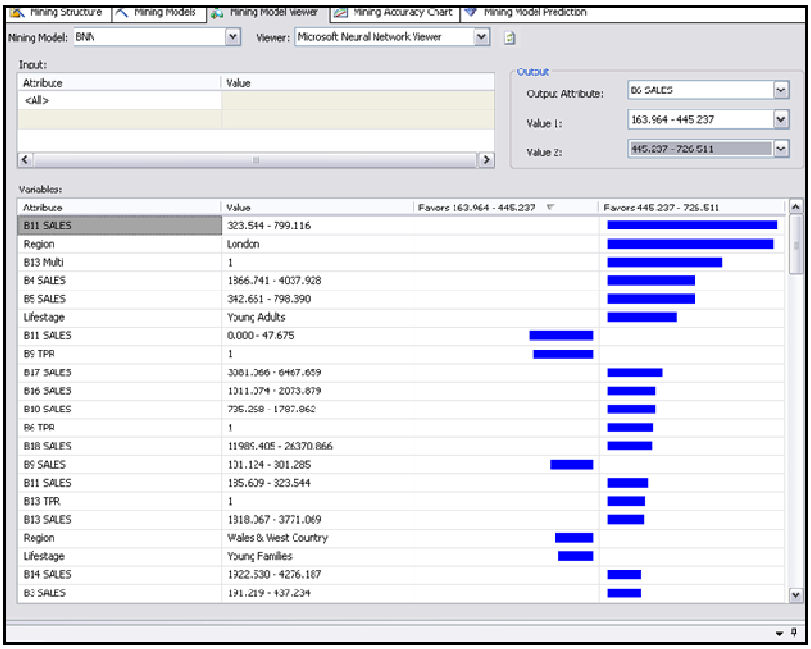


Figure 6.7: Neural network analysis to examine relationship between Beef varieties.

Further analysis through Association Rule Mining that the highest selling variety of Beef (for example); Premium Mince Beef sales, bared a strong correlation to high sales of:

- Premium Fry/Grilling Beef (B2),
- Specially Reared Organic Roasting Beef (B5),
- Healthy Diced Beef (B7),
- Standard Other Fry/Grill Beef (B12),
- Value Fry/Grill Beef (B16),
- Standard Diced Beef (B17),
- Standard Beef Mince (B18).

Hence, these products are appropriate for cross-promotion and can also guide the process of determining shelving policies. An alternate analysis of the highest performing region also provided

insight. By cross-analysing the results of the Decision Tree analysis on the impact of sales promotions in London with that of Neural Network analysis, it was discovered and validated that there is no significant increase in sales figures through sales promotions Hence, sales figures remained consistent, often showing no difference in sales whether there was a sales promotion active or not. Consequently, not only does this determine that stores in London are over discounting by applying sales promotions. Thereby reducing their profit margin, unlike the South West region, which responded with a historical trend of increased sales when sales promotional technique were applied.

Case Study 3, examined the potential of BI to manage the organisational resources of London underground Ltd. (LUL). LUL provided an ideal domain due to it dynamic, flexible and unpredictable nature. To investigate the applicability of KDDS-BI within the domain of resource allocation, two Intelligent Agent based BI solutions: BIDS and BITS were developed. Structured through KDDS-BI there was a degree of analysis as the output was collected, so that information could be extracted and transformed in to knowledge for decision support. BIDS provided a low granularity multi-agent system to schedule drivers to shifts. BIDS agents still posses a great degree of massaging and communicative flexibility, demonstrated by the dynamic ability to search for resources while ensuring the resource acquired results in optimum allocation. The system was tested using a combination of:

- Limited shifts and a high quantity of drivers.
- Limited shifts and a limited quantity of drivers.
- A high number of available shifts and a high quantity of drivers.
- A high number of available shifts and a limited quantity of drivers.

BIDS performed well under these circumstances, displaying the ability to effectively allocate resources. Although explored as a means through which to allocate drivers to shifts, the BIDS can readily be extended to allocate any organisational resource. In contrast to BIDS, BITS provides a more complex intelligent-agent based BI solution. The ability to schedule trains, in multiple events has been a complex problem for which intelligent agents provide a novel solution. The scheduling of trains, due to its complexity can be compared to a supply-chain, hence the use of intelligent agents that are not mobile. In contrast, BITS agents reside in the user's workstation/network; are more secure; developed in highly efficient language; thus enabling them to be considerably more efficient. The number of agents in BITS, as with any Multi-Agent system will vary, depending upon the task. However, since in the potential number of agents simultaneously negotiating, especially in the created scenes, is significantly large, BITS is considered a 'high granularity' solution. In addition, BITS demonstrates the advantages of using a multi-agent approach, while permitting schedules to be generated and calculate usage levels of each of the resources. Furthermore, these details can be exported into an excel file (figure 6.8). It is however, the 'ontology scenes' which provide true insight into the

performance of the system as a decision support tool. These scenes represent simulations of real-world opportunities and dilemmas. In order to demonstrate the capabilities of the system three ontology scenes were constructed:

- *Recovery*: Tests the systems ability to negotiate for alternative resources in the eventuality that current resources become unavailable or superior alternatives become available. Altering the shift of a driver tested the system, so that if no driver was available, nor where any carriages assigned for the return route. The system with absolutely no delay in execution time was able to designate carriages located at the station to the route and report to the user the error of no driver and list alternatives.
- *Rollback in Scheduling*: Evaluated the systems ability to adjust the schedule of a new train to account for a previously defined schedule with a high priority and allow for overtaking at stations where resources allow with minimal delays. The system efficiently created new schedules and allowed for reports to be generated.
- *Dynamic Changes to Schedule*: Demonstrated the ability for a number of previously defined low priority trains to adjust their schedules upon the addition of a high priority train and allow overtaking at relevant points bearing in mind safety implications. The system was able to implement these alterations while ensuring that the trains did not exceed defined speeds and allocated the time segments accurately.

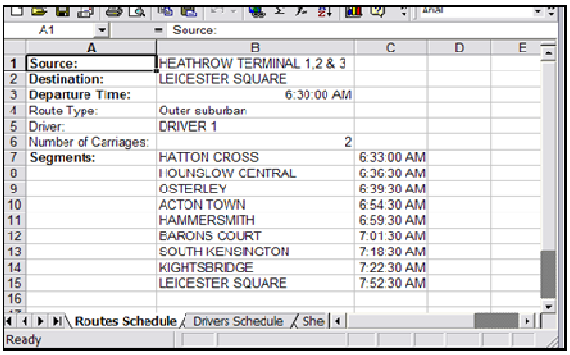


Figure 6.8: Details of the drivers shift exported to a spreadsheet.

Since, the system performed well, whilst providing suitable output and analysis of the scenarios, which can be explored for potential 'what-if' schedules, or even implemented to generate 'active' schedules. It can be deemed that, BITS provides a valuable BI solution through which resources can be allocated optimally within an organisation.

The three case studies have been fully documented in the appendix, however what is illustrated through each of the case studies is that the 'Decision Support' stage of KDDS-BI which necessitates that the output be collected, analysed, information extracted, and transfer to knowledge prior to any

decisions, which further structures the process of gaining insight from BI. Often, within BI a common hindrance is that although useful information is discovered, it is presented in a manner which is difficult tot decipher and in turn not utilised to its full potential. Structuring this process of decision making, ensure that the output, is readable, and more importantly understandable, so that it can be fully explored and provide a greater level of support than that which would be otherwise available.

6.3 Conclusion

This study has presented a domain, tools and technique neutral framework for BI. In order to illustrate these attributes, in addition to evaluate and verify the framework it has been applied using real-world data through three case studies. This chapter has provided an overview of the case studies with direct reference to each stage of KDDS-BI to illustrate how at an abstract level the process of implementing BI. For this reason the application area for the case studies in addition to the explicit solutions, were selected with great care. For each of the case studies:

Case Study 1: Sentient Insurance, (which can be found in full in Appendix B), examined BI in the area of Direct Marketing. This provided an interesting area for investigation, since it is imperative to correctly define and predict customer characteristics.

Case Study 2: Tesco Club Card, (which can be found in full in Appendix C), examined BI in the area of Sales Promotion. This are was selected since it enabled the examination of a number of objectives, ranging from analysing the impact of sales promotions, consumer purchasing habits and sales and cross-sales of products. These objectives all provided substantial opportunities for BI.

Case Study 3: London Underground Ltd., (which can be found in full in Appendix D), examined BI in the area of Managing Organisational Resources. This provided the opportunity to look at applying BI in the organisational area of operations and look at how prediction and forecasting can be utilised to manage resources in complex, dynamic and flexible environments.

It is imperative to clarify that despite each case study being conducted in a different application domain for varying objectives, the structure provided to the investigations remained the same. This structure was provided By KDDS-BI to the analysis of data pertaining to customers or resources which is collected routinely. This permitted a number of objectives to be extracted and developed. These objectives are then refined to provide the basis for the discovery of requirements and constraints of complex projects from an abstract perspective. In each instance, although the individual specifications are distinct, the process and structure remains the same. The objectives themselves provided insight into the requirements for the most suitable technique for analysis. Hence, using the

objectives of each case study as a focal point, these were contrasted against various techniques, structuring the approach in this manner permits flexible but robust method of technique selection.

Since traditional approaches are generally specific to a domain or tool, they cannot, at an abstract level ensure that the requirements of a BI investigation are fully supported through the opportunity to investigate varying approaches. As a result, KDDS-BI not only provides support for the process of selecting appropriate means for realising project objectives, but as demonstrated by the degree of overlap between the case studies, many of these common details can be documented and used as a reference point as illustrated by Appendix A. Consequently, these reference points can be used in future projects and simplify the process for technique selection and tool selection. Once objectives and techniques have been established, these can be used to discover appropriate means for a solution and data analysis, KDDS-BI provides the flexibility to gradually discover the requirements it is possible to look at techniques at an abstract level through modelling solutions to ensure that the correct approach is selected and the decision support is presented in the most appropriate manner, thereby providing a high-level of data insight which can also be readily utilised, thus providing a greater contribution within an organisation than is otherwise possible.

Section 3:

The final section of this study will conclude the research, thereby analysing the findings and performance of not only the framework, yet, in addition providing a conclusion for this study and the approach adopted to develop the framework. The section will also propose future recommendations which may be observed to further this study.

Chapter 7:

Conclusion and Further Research

This final chapter concludes the research conducted within this study. This chapter encompasses a summary of the research findings, based upon the performance of the proposed framework within the case studies. In addition to an analysis of the reviewed literature, techniques and methodology, which were explored to formulate the framework: KDDS-BI, which has been proposed by this study to address the identified gap in the current body of knowledge. Furthermore, an evaluation of the research is made, with key conclusions arising from the thesis subsequently reported, as is, the original contribution to research in this field. Finally, noting that all research is fallible and subject to a finite timescale, the limitations that bounded the study are discussed along with proposals to further the study.

7.1 Conclusion

Effective BI integration often requires a combinatorial approach, thereby amalgamating a number of techniques through various tools, which must be seamlessly integrated in organisational activity. CRISP-DM[7] has demonstrated within data mining, how a meta-level framework which can be applied to structure projects, can greatly increase the effectiveness of investigations, Consequently, it was this necessity for a framework tailored to BI which motivated this research. Guided by this motivation, it was proposed in the opening chapter of this study that this research would address the following hypothesis:

"A meta-level framework for Business Intelligence will facilitate knowledge discovery and decision support."

Hence, for this hypothesis to be realised it was necessary to undertake a number of objectives:

- Conduct an in depth critique and review of the literature and the theories which underpin the domain of BI and related strategies. This will subsequently, provide a background to this study. Whilst permitting the identification of a suitable research methodology, which is able to capture and generate data to test the focal theory, via a specific research design.

In order to fully understand the area in which the study was to take place it was necessary to examine this focal domain from a number of perspectives. Chapter 2, reviewed the theory and logic underpinning BI in addition to the various techniques and strategies that constitute BI, such as data mining; unsupervised learning; supervised learning; association rule mining; and intelligent agents, were reviewed and analysed, in addition to distinguishing BI from knowledge management, and reviewing decision making mechanisms and decision support systems. This provided an in-depth understanding of the theoretical background of the research area. With an analysis of the theories underpinning the research completed, it was possible to select a suitable research model which would provide a valid, methodological approach for proposing a framework tailored to BI.

Chapter 3 examined the various approaches which could be considered for this study. A 'hard systems approach' was selected as a methodology to underpin the formulation of a framework. This approach provided the design through which, to discover a scientifically valid framework based upon implicit tangible knowledge. Since many organisations lack the technical expertise and additionally employ ad-hoc techniques, a soft systems methodology would not have provided the essential criteria upon

[7] http://www.crisp-dm.org/: Accessed December, 2008.

which to formulate a valid framework. This lack of significant BI knowledge is a limitation which has also been identified by Forsling (2007, p. 1) who has recommended that organisations:

"Establish a BI Competence Centre. A steering committee has been a proven approach in many successful companies."

Consequently, in light of insufficient support for a soft systems approach and as discussed within Chapter 3, a framework, especially one which is applied at an abstract level provides a system, through which design and development can take place, for which a hard systems methodology provided the required investigative modelling criteria. A key issue for the formulation of an effective framework, however, was to investigate and identify the strengths and weakness of conventional approaches. It was discovered that conventional approaches although capable of executing effective projects using an explicit strategy, lacked the dynamic requirements required by BI projects.

- Undertake exploratory studies within the research methodology to hypothesis a meta-level framework which can facilitate the selection of suitable BI techniques to develop solutions and address organisational requirements for providing decision support to executives.

Since currently no framework designed explicitly for BI exists, BI projects are explored through ad-hoc techniques, exploring frameworks developed for a particular technology. This approach however, does not permit the full exploration of the potential of BI. Consequently, much as CRISP-DM has achieved for data mining, BI too, requires a tailored and focused framework. This is a view which is supported by the Gartner Group, who has specified that if BI is to reach it full potential, organisations must recognise the explicit requirements of BI projects, and understand that BI is not limited to advanced analytics, or reporting capabilities (Forsling, 2007). In order to discover the requirements that a tailored approach to BI must address, it was imperative to discover the advantages of conventional approaches to BI, as well as the short-comings of these approaches. This in turn provided the criteria which the framework proposed and developed through this study must address. This documentation was realised in Chapter 4, which provided the opportunity to review and analyse the technical aspects of conventional approaches as pertaining to this study.

- Develop a meta-level framework to structure BI investigations and select suitable techniques.

Through the use of a systems engineering approach to structure this research and the motivation for analysing conventional techniques, specifically when applied to BI established which has been documented in Chapter 4. This approach permitted the critical issues and key features required for a BI framework that are not apparent in conventional approaches to be initially identified. It is noteworthy, that although many of the selected features were apparent in a variety of conventional

approaches. No single framework encapsulated all necessary requirements. This enabled the framework to be proposed in Chapter 5. KDDS-BI, which is tailored to BI at a meta-level, thereby providing a tool, technique and domain neutral approach to structuring BI investigations.

- Evaluate the framework through three case studies which will assess the proposed framework and its capability to support the process of selecting suitable techniques for BI integration.

Documented in detail in Appendices B, C & D. KDDS-BI was initially investigated for the purpose of direct marketing (Appendix B: Case Study 1). Direct marketing proposed an interesting challenge since it is an aspect of marketing that could previously been achieved through mass distribution. However, with the ever-increasing scale upon which organisations must compete, mass marketing is no longer a viable or sustainable marketing strategy. In contrast it is imperative that organisations be able to predict trends and classify consumers through the identification of key characteristics. The identification of these characteristics will enable organisations to execute more targeted and focused marketing strategies, which will consequently yield a greater return on investment. In order to discover these characteristics, that can guide decision makers in the most appropriate direction. KDDS-BI was explored to structure a BI investigation. Studying historical organisational data which is routinely accumulated, the data relating to the records of the 'The Insurance Company' regarding the sale of caravan policies was examined and interrogated through novel and innovative BI strategies. The analysis of an existing data set through BI techniques permits an organisation to reap further benefits from a data set which would otherwise be stored and conceal useful, yet hidden, patterns and trends. Upon studying the data set, it was determined that based upon the attributes and data characteristics classification could be explored to discover these trends. Consequently, examining the data through a variety of classification techniques not only facilitated the discovery of these significant results, yet further provided an opportunity to analyse and evaluate the performance of classification techniques. The analysing of various techniques was an essential objective since the application domain of direct marketing is dynamic. Consumer preference are constantly changing and evolving. Hence, KDDS-BI is a cyclical, iterative framework, however, the analysis of techniques will ensure that the most suitable technique is selected; consequently, the results will provide the most accurate basis upon which to provide decision makers with support. As a result the DSS discovered through KDDS-BI will display emergent intelligence properties, since with each iteration of the framework, even if only the 'Decision Support' and 'Business Decision' phases, the performance can be expected to increase.

The second case study (Appendix C: Case Study 2), analysed the performance of KDDS-BI within 'sales promotion'. Sales promotion provides organisations with an opportunity to target consumer at

the point-of-sale. Furthermore, sales promotion if effectively executed will not only entice consumer to purchase in greater quantity at the point-of-sale, but additionally enable managers to attract a greater amount of consumers into a particular store. However, if sales promotions are to be effective then it is essential that decision makers amass as much knowledge as possible on their target audience. This knowledge will facilitate the decision makers to implement promotions and strategies which most aptly serve the needs and requirements of the explicit consumers' particular to their environment. To realise this objective KDDS-BI was explored to structure the analysis of 'Tesco Clubcard' data pertaining to fresh meat products. The Tesco Clubcard has proven to be one of the most effective loyalty schemes. Moreover, the success of the scheme has resulted in rival organisations not only implementing similar schemes, but additionally, permitting the use of Clubcard rewards in their own stores to 'poach' customers away from Tesco's. It is therefore imperative that an organisation, such as Tesco, reinforce and supplement schemes as successful as the Clubcard, with strategies such as sales promotion. In order to increase the effectiveness of decisions made with regard to sales promotion strategies, data relating to Clubcard data regarding the sale of fresh meat products was analysed and explored through KDDS-BI. This structured analysis unveiled that a combinatorial approach of BI techniques executed to uncover as much knowledge as possible of consumer behaviour can significantly increase an organisations competitive capability. By analysing the performance of the lowest performing region enabled the discovery of significant information. Due to the objectives of the investigation requiring systematic analysis, the cyclical nature of KDDS-BI, resulted in the iterative analysis of output, extraction of knowledge, and exploration of this discovered knowledge to realise further objectives. Furthermore, even in the event that initially identified techniques, such as clustering, which despite being theoretically suitable, resulted in less then optimal performance when applied. KDDS-BI could be dynamically manipulated to discover techniques which yielded greater accuracy.

Whilst, the initial two case studies explored KDDS-BI for various approaches within marketing (direct marketing, customer profiling, the Marketing Mix and sales promotion). The third case study (Appendix D: Case Study 3), analysed the performance of KDDS-BI in the other major activity of business organisations identified by Stevenson (2006) 'operations'. A major facet of operations is the allocation of resources. This allocation has gained a greater amount of significance since the quantity of resources managed by an organisation has not only increased, yet additionally in many instances, become geographically disperse. In effective resource management can result in diseconomies of scale, significantly reducing organisational efficiency. Hence, this essential aspect of organisational behaviour was scrutinised through KDDS-BI. As a domain within which to conduct this analysis, the 'London Underground Ltd. (LUL) network' was selected. Forming a major component of Transport for London, the LUL network spans nearly 400 route miles with 275 stops. The busiest railway line within the network carries over 180 million passengers annually. The number of journeys undertaken

within the network consistently exceeds one billion, with 150,000 people an hour entering the LUL system[8]. In addition to this logistic pressure LUL employs approximately 16,000 people. Consequently, the task of allocating resources throughout the network proposes a formidable challenge. However, the dynamic, unpredictable and flexible nature of LUL results in this challenge being well suited to a BI investigation. Encountering this formidable challenge through a BI technique such as intelligent agents provided the opportunity to analyse the environment of scheduling resources through an innovative approach. Whilst innovative, intelligent agents provide more then just a unique approach to a challenging quandary, rather they purpose a technique which can greatly enhance the capabilities of decision makers, though the provision of 'what-if' opportunities. Through the functionality of KDDS-BI two Multi-Agent based DSS were discovered. BIDS permitted agents representing drivers to negotiate with agents that could assign suitable shifts. While, BITS provided a high granularity Multi-Agent system that was able to negotiate the optimum allocation of resources, explore 'what-if' opportunities, and simulate real-world predicaments, to provide decision makers with real-time, accurate information in addition to possible schedules. The process illustrated through this case study could effortlessly be extrapolated to identify opportunities within organisation supply-chains.

The three case studies analysed the performance of KDDS-BI to structure a BI investigation in a variety of independent domains, each posing it own individual challenges. To realise these objectives, the full dynamic nature of KDDS-BI had to be explored, this is also why the analysis of sub-steps of each phase was not explicitly defined by headings, since each phase required a high level of dynamic analysis to be fully comprehended. Moreover, to efficiently resolve these challenges required KDDS-BI to permit the analysis and management of commercial and open-source platforms in a unique and novel method. The platforms were innovatively explored using advanced analytics, intelligent agents and reporting through the interrogation of flat files and relational databases to provide decision makers with support for both managing resources and executing strategies through a combination of techniques which provide opportunities to gain a completive advantage in the global market.

- Asses the significance of the framework within each case study and the implications that occur through the development life cycle.

The selected case studies have permitted the evaluation and testing of KDDS-BI in a number of scenarios. These scenarios have simulated real-world situations, with live real-world data. Albeit, each case study, as documented in the Appendix has culminated with an implicit conclusion specific to the respective chapter, these findings can be further evaluated to asses the performance and functionality

[8] http://www.tfl.gov.uk/: Accessed December, 2008.

of KDDS-BI at an abstract level. This was realised in Chapter 6, which provided the opportunity to not only asses the performance of KDDS-BI, within each case study, but further cross-analyse this performance to discover the similarities and overlap between each instance.

Since, BI investigations are inherently complex, often drawing upon a number of techniques for a single purpose, it was discovered that the lack of flexibility and support provided by conventional techniques was a major hindrance of conventional approaches for BI, thus requiring an ad-hoc approach to be adopted. Since it is fundamental to a BI investigation that the correct approach be utilised for effective use, a gap in the current body of knowledge was identified. It is this gap which has been addressed by this study, for which a framework has been proposed, developed and tested. Through the realisation of each of the objectives, the framework; KDDS-BI, can be validated as a scientifically valid, technique and tool independent means of structuring BI investigations at an abstract level.

7.2 Contribution to Knowledge

The contributions of this research have been alluded to throughout this study. However, the primary contribution of this research stems from an engineered framework for exploring data through BI techniques to provide decision support. BI is a novel technology, which has received significant interest from both the commercial and academic interest. Nevertheless, much of this interest has been limited to the application of BI technology. Hence, despite providing large-scale appeal, much of the potential of BI remains unexplored, since it is largely applied in an ad-hoc manner, consequently a vast majority of the research remains unstructured and difficult to truly analyse.

The gap which was identified within the current body of knowledge revealed that currently no suitable framework for BI has been formulated. This is especially true for one which can be applied at a meta-level and is therefore tool, technology and domain independent. This study has endeavoured to fill this gap through the proposition of a framework which can be applied at an abstract level and therefore structure a BI investigation, irrespective of the end user or their explicit requirements. To this end, the framework: KDDS-BI, which has been proposed and developed through this research, has been formulated with a view to providing a tailored and structured approach to BI investigation. This need for a tailored, abstract model can greatly increase the effectiveness of solutions, as witnessed in the field of data mining through the introduction of CRISP-DM[9]. The framework itself provides a number of contributions; based upon an analysis of conventional techniques, KDDS-BI provides an innovative combination of stage which can be explored to structure the integration of BI within organisational activity, furthermore, KDDS-BI has demonstrated how existing data within an organisation that is

[9] http://www.crisp-dm.org/: Accessed December, 2008.

routinely acquired, can be further analysed and explored to uncover deep and hidden data structures. These data structures can ascertain future trends, consumer behaviour, and allocation of resources in addition to various other patterns which can provide significant competitive advantages to an organisation. However, this is not the only contribution, in addition to a tailored BI framework that supports knowledge discovery. The process of evaluating the framework has provided this research with a number of novel, unique and innovative approaches to conventional organisational problem predicaments, while also uncovering potential opportunities. Within direct marketing and customer profiling it has been revealed how a novel use of open-source software can result in a customised platform thorough which to not only recognise potential customers, and knowledge regarding the behaviour of consumer segments, through a unique combinatorial approach using a number of supervised techniques. Yet the performance of these techniques has been compared to asses the accuracy of the said techniques within the context of identifying consumer segments. Not only was the combination of techniques explored, but the performance of the techniques using both numeric and nominal data was analysed. The second case study, explored the framework within sales promotion, through a combination of software packages, Tesco Clubcard data pertaining to fresh meat sales was intricately analysed to discover not only the performance of a region; but the impact of promotions; the consumer segments that respond to particular promotions, the areas within which to target these customers and the products that can be marketed together. This valuable novel information was discovered through a unique combination of supervised, unsupervised techniques, in addition to association rule mining, thereby combining a broad range of advanced analytics. The final case study analysed the performance of KDDS-BI to allocate resources in addition to develop advanced schedules and plans, in addition to reports detailing these schedules. These contributions were realised in a unique combination, through a novel technology which has limited exposure to the domain of resource scheduling, let alone train scheduling. The resources of the selected domain for the case study; LUL, provided a formidable challenge, which was resolved in a unique investigation of two intelligent DSS based upon agent technology.

Whilst each of the case studies has provided a contribution to knowledge through the novel use of BI technology, it is noteworthy the approach which provides the core contribution to knowledge. Thus, this study requires that each case study not be viewed in isolation. In contrast the case studies conducted as part of this study should be cross-analysed, as is the case in Chapter 6. Chapter 6, provides an analysis of how, despite significantly different in each instance, the process of technique selection, development, knowledge management and decision support has been consistent. The case studies thus illustrate that the application domain or method bears no significance upon the performance of KDDS-BI. However, in each case study KDDS-BI, has provided a structured approach which has guided the process of discovering objectives, selecting techniques, modelling

solutions, determining an appropriate means for developing these models, and ultimately providing decision support to an organisation, through BI and an increased level of data insight.

7.3 Limitations of Research

Due to constraints of time and limitations in the available resources, this research has not been without its limitations. The research discussed within this thesis, has not fully examined data collection in marketing. In contrast emphasis has been placed upon the utilisation of organisational data. This is due to organisations routinely collecting significant quantities of data which is seldom interrogated. KDDS-BI endeavours to provide support and structure to the process of analysing these data sets. This represents an extremely appealing prospect to organisations since it provides the means through which to interrogate existing data sets in a novel and unique fashion, with little or no data collection expense. Although, data was explicitly collected with regard to resource management, this can be considered more as an evaluation of existing data (resources) rather then specific data collection for the purpose of BI interrogation.

Furthermore, KDDS-BI provides a greater emphasis upon extracting objectives directly from the data, as opposed to users. This is due to the limitation imposed by prospective users. The end user or decision maker, need not be a BI expert, therefore should be capable of operating a proposed solution with limited theoretical knowledge of the techniques employed. In addition and key to BI investigations, data is interrogated to uncover hidden and innovative information. As a result, the full objectives or potential of the data may not be apparent to individuals who do not posses intimate theoretical knowledge of BI techniques. As supported by a hard systems engineering approach, it is thus, the task of the designer/s and implementer/s to conduct this analysis. This provided the designer has knowledge of the application domain being explored; it is the task of this individual to determine the objectives In order to provide support for this process, requirements are assessed upon the analysis of techniques, and KDDS-BI phases may be repeated if required. Chapter 6 and Chapter 7, illustrated the necessity to asses requirements, in addition to the reality, that techniques theoretically suited to a particular objective, do not always yield the greatest results. Consequently, requiring techniques and at times objectives to be re-assessed as was discovered.

For effective investigation, KDDS-BI requires a basic level of prerequisite knowledge regarding BI techniques. However, KDDS-BI does provide and as demonstrated through the case studies, successfully provides a framework through which an organisation can investigate existing datasets at an abstract level, irrespective of the domain, tool or technique. This investigation will aid organisations to discover deep and hidden information and furthermore exploit this information to ensure the organisations decision makers and managers can apply novel and innovative strategies for a

competitive edge. Nevertheless, the requirement for an expert within an organisation is supported by the recommendations of the Gartner Group of establishing a BI excellence centre (Forsling, 2007) which can oversee, and execute a BI integration project.

Once the investigation of organisational data has enabled preliminary objectives to be determined, possible techniques to realise these objectives can be explored. Since this requires explicit knowledge of BI, KDDS-BI is not proposed as a replacement for an expert. Rather, the aim of KDDS-BI is to augment the capabilities of an expert by providing support and direction to the process of selecting a suitable technique and developing a solution. Furthermore, the decision support provided is not intended to replace a decision maker. In contrast, the proposed solution, even if autonomous, as with an intelligent agent-based solution, should provide a greater level of support for decisions. This support, will serve to increase efficiency through more significant information upon which to formulate decisions. Moreover, the quality of decisions and the performance of solutions will display emergent intelligence, therefore, increase in quality, with each iteration of the framework. However, time constraints have only permitted the KDDS-BI cycle to be executed once within this study, yet the development of the framework and nature of BI techniques has permitted the integration of 'emergent intelligence' capabilities. Furthermore, KDDS-BI ensures a degree of consistency within the integration of BI throughout all departments.

7.4 Future Recommendations

To further this research, it is recommended that the emergent properties of proposed solutions be further investigated. This can be realised through repeated iterations of not only the 'Decision Support' phase and 'Business Decision' phase, but also through the further exploration of the iterative and flexible nature of KDDS-BI on a deeper level. Furthermore, this will enable the discovery and analysis of solutions explored through KDDS-BI to predict and evolve with trends and changes in consumer behaviour. However, such an investigation would require longer then the duration of this research, which was subject to time constraints. This would provide the opportunity to not only analyse the capability of KDDS-B to provide not only structured support for technique and solution discovery, but further assess the impact of the discovered solutions and the support provided to decision makers.

Furthermore, it would be of significant interest to analyse the performance of KDDS-BI within a live setting. Thereby, providing the opportunity to analyse and evaluate KDDS-BI within an organisation for real-world deployment. Thereby, further demonstrating the benefits and advantages of a tailored framework within BI, this will serve to reinforce not only the capabilities and options, but moreover the effectiveness of executive decisions. This would also provide further opportunities to analyse an

even greater number of datasets, then was possible through this study alone. As discussed in Chapter 2, BI has been applied in various industries and for a variety of applications; however, it would be of significant interest to explore the performance of KDDS-BI in further domains, enabling more diverse datasets to be evaluated through a structured approach as has been conducted in this study.

References

de Abajo N., Diez A. B., Lobato V. & Cuesta S. R. (2004) 'Artificial Neural Networks: Quality Diagnostic Models for Packaging Manufacturing: An Industrial Data Mining Case Study', *KDD '04-Conference Proceedings*, 10th ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, WA, USA.

Abi-Antoun M. (2007) 'Making Frameworks Work: A Project Retrospective', *ACM- OOPSLA '07*: Companion to the 22nd ACM SIGPLAN Conference on Object Oriented Programming Systems and Applications. Montreal, Canada.

Abidi, S.S.R. & Ong, J. (2000) 'A Data Mining Strategy for Inductive Data Clustering: A Synergy Between Self-Organising Neural Networks and K-Means Clustering Techniques',
TENCON, vol 2, September, pp. 568 – 573.

Abraham, A., Jain, L. C. & an der Zwaag, B. J. (2004) *Innovations in Intelligent Systems; Design, Management and Applications*: New York: Springer.

Adderley R. and Bond J. W. (2007) 'Police Forensic Science Performance Indicators – A New Approach to Data Validation', *Applications and Innovations in Intelligent Systems XV*: London: Springer.

Addriaans, P. & Zanteg, D. (1996) *Data Mining*: London: Addison-Wesley.

Agarwal R., Prasad J., Tanniru M. & Lynch J. (2000) 'Risks of Rapid Application Development', *Communications of the ACM*, vol. 43, no. 11, November, ACM.

Agrawal R. & Srikant R. (1994) 'Fast Algorithms for Mining Association Rules in Large Databases', *VLDB*, September, pp. 487-499.

Alonso, F., Caraça-Valentea, J. P., Gonzálezb, A. L. & Montes, C. (2002) 'Combining Expert Knowledge and Data Mining in a Medical Diagnosis Domain', *Expert Systems with Applications*, vol. 23, no. 4, November, pp. 367-375.

Amershi S. & Conati C. (2007) 'Unsupervised and Supervised Machine Learning in User Modelling for Intelligent Learning Environments', *IUI '07- Conference Proceedings*, 12th International Conference on Intelligent User Interfaces, Island of Madeira, Portugal, ACM.

Anand S. S. & Buchner A. G. (1998) *Decision Support Using Data Mining:* Pennsylvania, USA: Trans-Atlantic Publications

Anderson, L. & Rönnbom, A. (1999) *Intelligent Agents- A New Technology for Future Distributed Sensory Systems***:** Göteborg University, Department of Informatics.

Adriaans, P. & Zantinge, D., 1996. *Data Mining*. Harlow, England: Addison-Wesley.

Alfred, R. & Kazakov, D. (2007) 'Handling Datasets in a Multi-Relational Environment: Cluster Dispersion vs Cluster Purity', *Intelligent Data Acquisition and Advanced Computing Systems*, Technology and Applications, September, pp. 196 – 201.

Barry, M. J. & Linoff, G. S. (2004) *Data Mining Techniques. 2nd Edition*: San Francisco: Addison-Wesley.

Arnott, D. & Pervan, G. (2008) 'Eight Key Issues for the Decision Support Systems Discipline', *Decision Support Systems*. Vol. 44, No. 3, February, pp. 657-672.

Badiru, A. B. & Cheung, J. Y. (2002) *Fuzzy engineering Expert Systems with Neural Network Applications*: New York: John Wiley and Sons.

Baker, M. J. (2000) *Marketing Strategy and Management, 3rd Edition*: Hampshire: Macmillan Press Ltd.

Badjonski, M., Ivanovic, M. & Budimac, Z. (2006) 'Adaptable Java Agents– A Tool for the Programming of Multi-Agent Systems' in *Artificial Intelligence Applications and Innovations*: Boston: Springer.

Beck, K. & Andres, C. (2005) *Extreme Programming Explained, 2nd Edition*: London: Addison-Wesley.

Bellifemine, F. L., Giovanni C. & Greenwood, D. (2007) *Developing Multi-agent Systems with JADE*: New York: John Wiley and Sons.

Belch, G.E. & Belch, M. A. (2007) *Advertising and Promotion; An Integrated Marketing Perspective*: New York: McGraw-Hill.

Berry, M. J. A. & Linoff, G. S. (2004) *Data Mining Techniques for Marketing, Sales and Customer Relationship Management, 2nd Edition*: New York: John Wiley & Sons Ltd.

Beynon-Davies, P., Mackay, H. & Tudhope, D. (2000) 'It's Lots of Bits of Paper and Ticks and Post-It Notes and Things . . .: A Case Study of a Rapid Application Development Project', *Information Systems Journal*, vol. 10, no. 3, pp 215-216, Blackwell Science Ltd.

Bigus, J.P. & Bigus, J. (2001) *Constructing Intelligent Agents Using JAVA, 2nd Edition:* Chichester: John Wiley & Sons Ltd.

Binder, R. V. (2000) *Testing Object-Orientated Systems: Models, Patterns & Tools:* Reading, Massachusetts: Addison-Wesley Longman Inc.

Bond, A.H. & Gasser, L. (1988) *Readings in Distributed Artificial Intelligence:* San Mateo, CA., Morgan Kaufmann.

Bourne, A. (2000) 'Managing Human Factors in London Underground', *The Human Challenge- IEEE Seminar on Control of Railways*, Ref. no. 2000/049, pp. 1-3, London, UK.

Bowling, M. & Veloso, M. (2000) *An Analysis of Stochastic Game Theory for Multi-Agent Reinforcement Learning:* Pittsburgh : Carnegie Mellon University.

Brachman, R. J. & Anand, T. (1996) 'The Process of Knowledge Discovery in Databases', *Advances in Knowledge Discovery and Data Mining*, pp. 37-57.

Bradshaw, J. M. (1997) 'Introduction to Software Agents' in Bradshaw, J. M. (ed.) Software *Agents*, Menlo Park, Cambridge, MA, AAAI Press/MIT Press.

Bretier, P. & Sadak, D. (1997) *A Rational Agent as a Kernel of a Cooperative Spoken Dialogue System; Implementing A Logical Theory Of Interaction*: Intelligent Agents III; Berlin; Springer.

Broardman, J. & Sauser, B. (2008) *Systems Thinking: Coping with 21st Century Problems*: New York: CRC Press.

Burstein, F. & Holsapple, C. W. (Eds.). (2008). Handbook on Decision Support Systems 1: Basic Themes: Berlin, Germany: Springer-Verlag.

Butler Group (2006) *Business Intelligence: A Strategic Approach to Extending and Standardising the Use of BI*: London: FT Intelligence/Europe Intelligence Wire.

Campbell, H.M. (2006) 'The Role of Organizational Knowledge Management Strategies in the Quest for Business Intelligence', *IEEE International Engineering Management Conference*, September, September, pp. 231 – 236.

Campbell, P.R.J., Adamson, K. & Ford-Hutchinson, R. (2006) 'Towards DSS: Enhancing Domain Knowledge through Knowledge Discovery', *Innovations in Information Technology*, November, pp. 1-5.

Caruana, R. & Niculescu-Mizil, A. (2006) 'An Empirical Comparison of Supervised Learning Algorithms', *Proceedings of the 23rd international Conference on Machine Learning*, ICML '06, June, ACM.

Carey, M. & Carville, S. (2003) *Scheduling and Platform Trains at Busy Complex Stations*: Transportation Research: Elsevier Science Ltd.

Cawsey, A. (1998) *The Essence of Artificial Intelligence*: Prentice Hall: London.

Ceglar, A. & Roddick , J. F. (2006) 'Association Mining', *Computing Surveys (CSUR)* , vol. 38, no. 2 , July, ACM.

Chalmers, A.F. (1999) *What Is This Thing Called Science (3rd Edition)*: Buckingham: Open University Press.

Charlton, P, Cattoni, R. Potrich, A. & Mamdani, E (2000) *Evaluating The FIPA Standards and Its Role In Achieving Cooperation in Multi-Agent Systems*: Proceedings of the 33rd Hawaii International Conference on Systems Sciences.

Champandard, A. J. (2008) Path-*Planning From Start to Finish*: Available on-line: http://ai-depot.com/: Accessed November, 2007.

Checkland, P. (1999) *Soft Systems Methodology in Action*: Chichester: John Wiley & Sons Ltd.

Cheng, H., Lun Y. & Sheu, C. (2008) 'An Ontology-Based Business Intelligence Application In a Financial Knowledge Management System', *Expert Systems with Applications*, In Press, Corrected Proof, Available online 4 March 2008.

Chun, S. & Park, Y. (2006) 'A New Hybrid Data Mining Technique Using a Regression Case Based Reasoning: Application to Financial Forecasting', *Expert Systems with Applications*, vol. 31, no. 2, August, pp. 329-336.

Chung, W., Chen, H. & J. Nunamaker, J. (2003) 'Business Intelligence Explorer: A Knowledge Map Framework for Discovering Business Intelligence on the Web', *36th Annual Hawaii International Conference on System Sciences*, January.

Dehning, B. & Richardson, V.J. (2002) Returns on Investments in Information Technology: A Research Synthesis', *Journal of Information Systems*, November, vol. 16, no. 1, pp.7–30.

Cody, W. F., Kreulen, J. T., Krishna, V. & Spangler, W. S. (2002) 'The Integration of Business Intelligence and Knowledge Management', *IBM Systems Journal*, vol 41, no 4.

Combes,C., Meskens, N., Rivat,C. & Vandamme, J. (2008) 'Using a KDD Process To Forecast the Duration of Surgery', *International Journal of Production Economics*, vol. 112, no. 1, March, pp. 279-293.

Cohan, P.R. & Levesque, H.J. (1990) 'Intention is choice with commitment', *IEEE Artificial Intelligence*, vol, 42, pp. 2-3.

Cox, J. F. & Blackstone, J. H. (2008) *APICS Dictionary 12th Edition*: Chicago: APICS Educational Society for Resource Management.

CRISP-DM Process Guide and User Manual: Available online: http://www.crisp-dm.org/: Accessed November, 2008.

Curotto, C.L. & Ebecken, N.F.F. (2005) Implementing Data Mining Algorithms in Microsoft SQL Server: Southampton, Boston: WIT press.

Curtis, G., & Cobham, D., (2008) *Business Information Systems: Analysis, Design & Practice*: Prentice Hall: London.

Dandawate, Y.H., Joshi, M.A. & Gawande, P.G. (2008) Image Compression Using Enhanced Vector Quantizer Designed with Selective Training of Unsupervised Neural Network: *Wireless, Mobile and Multimedia Networks*, January, pp. 263 – 266.

Davenport, T. H. & Harris, J. G. (2007) *Competing on Analytics: The New Science of Winning:* Boston: Harvard Business School Press.

Davidson, P. & Wernstedt, F. (2002) *A Framework for Evaluation of Multi-Agent System Approaches to Logistic Network Management:* Department of Software Engineering and Computer Science, Blekinge Institute of Technology, Sweden.

Dasgupta, S. (1991) *Design Theory and Computer Science:* Cambridge: Cambridge University Press

Debowski, S. (2006) *Knowledge Management*: Australia: John Wiley & Sons Ltd.

Delahunty A. & McDonald, F. (Eds.) (2007) *Oxford Dictionary*: Oxford: Oxford University Press, p. 699.

Demestichas, K. D, Artemis A. Koutsorodi, A. A., Adamopoulou, E. F. & Theologou, M. E. (2008) 'Modelling User Preferences and Configuring Services in B3G Devices', *Wireless Networks*, vol. 14, no. 5, October, Kluwer Academic Publishers.

Dennis, A. & Wixcom, B. H. (2000) *Systems Analysis and Design: An Applied Approach*: New York: John Wiley & Sons Ltd.

Dhar, V. & Stein, R. (1997) *Seven Methods for Transforming Corporate Data into BI*: London: Prentice-Hall.

Dibb, S., Simkin, L., Pride, W. M. & Ferrell, O. C. (2005) *Marketing: Concepts and Strategies: 5th Edition*: Boston: Houghton Mifflin.

Dunham, M. H. (2003) Data mining Introductory and Advanced Topics: New Jersey: Pearson Education Inc.

Eck, M.V. (2004) *Advanced Planning and Scheduling: Is Logistics Everything?;* Vrije Universiteit Amsterdam: BWI paper.

Eckerson, W. (2008) *Ad Hoc Business Intelligence is Killing Us!:* Available on line http://www.businessintelligence.com/: Accessed August, 2008.

Elbashir, M. Z., Collier, P. A. & Davern, M. J. (2008) 'Measuring the Effects of Business Intelligence Systems: The Relationship between Business Process and Organizational Performance', *International Journal of Accounting Information Systems*, In Press, Corrected Proof, Available online 20 August 2008.

Esterby-Smith, M., Thorpe, R. & Lowe, A. (1999) *Management Research: An Introduction*: Sage.

Evelson, B. (2008) *The Forrester Wave: Enterprise Business Intelligence Platforms, Q3 2008*; July, Available online: http://www.sas.com/news/analysts/: Accessed June, 2008.

Elkan, C. (1997) 'Naive Bayesian Learning', *Technical Report No. CS97-557,* Dept. of Computer Science and Engineering, California: University of California, San Diego.

Farber, J. (1999) *Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence*: Reading, Massachusetts: Addison-Wesley Longman Inc.

Fawcett, T. & Provost, F. (1997) Adaptive Fraud Detection, *Data Mining and Knowledge Discovery*, vol. 1 no .3, pp. 291-316.

Fawcett, T. (2003) 'ROC Graphs: Notes and Practical Considerations for Data Mining Researchers', *Technical Report HPL-2003-4*, California: HP Laboratories.

Feng, Q. & Guoqiang Lu, G. (2003) **'**FIPA-ACL Based Agent Communications in Plant Automation: Emerging Technologies and Factory Automation', *Proceedings ETFA '03: IEEE Conference*, September, vol. 2, pp 74 – 78.

Finnie, G. & Barker, J. (2005) 'Real-Time Business Intelligence in Multi-Agent Adaptive Supply Networks', *The IEEE International Conference on e-Technology, e-Commerce and e-Service*, March, pp. 218 – 221.

FIPA (2008) FIPA Homepage: available online: http://www.fipa.org: Accessed June, 2008.

Fong, A.C.M., Hui, S.C. & Jha, G. (2002) 'Data mining for Decision Support', *IT Professional,* vol. 4, no. 2, April, pp. 9 – 17, IEEE Xplore.

Foulds L, R. (2000) 'Direction manager: A DSS for one-way street system design', *Proceedings for Urban Transport 2000 Conference*, Cambridge University.

Foulds, L. R. & Johnson, D.G. (2000) 'Slotmanager: A DSS for Timetabling'; *Decision Support Systems*, vol. 27, no. 4, April, pp. 1044-1061.

Fornora, N. & Colombetti, M. (2002) 'Operational Specification of a Commitment Based Agent Communication Language', *International Joint Conference on Autonomous Agents and Multi -Agent Systems*, July.

Forsling, C. (2007) Gartner Press Releases: Available online: http://www.gartner.com/it/: Accessed September, 2008.

Fowler, A. (1998) 'Innovative Design of Management Systems through Simulation', *International Conference on Simulation*, October, pp. 215 – 222.

Furnkranz, J. & Flach, P. (2005) 'ROC'n'Rule Learning – Toward a Better Understanding of Rule Covering Algorithms', *Machine Learning*, vol. 58, no. 1, pp. 39-77.

Gachet, A. (2004) *Building Model-Driven Decision Support Systems with Dicodess*. Zurich, VDF.

Gadomski A.M. (1998) *Integrated Parallel Bottom-up and Top-down Approach to the Development of Agent-based Intelligent DSSs for Emergency Management*, TIEMS98, Washington.

Gayialis, S. P. & Tatsiopoulos, I. P. (2002) 'Design of IT Driven Decision Support Systems for Vehicle Routing and Scheduling', *European Journal of Operational Research*, vol. 15, no 2, Elsevier Science Ltd.

Geist, I. (2002) 'A framework for data mining and KDD', *Proceedings of the ACM Symposium on Applied Computing*, SAC '02, March, ACM.

Gervais, E., Liu, H., Nussbaum, D., Roh, Y-S., Sack, J-R. & Yi, J. (2007) 'Intelligent Map Agents -An Ubiquitous Personalised GIS', *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 62, no. 5, October, pp. 347-365.

Gilbert, D. C. & Jackaria, N. (2002) 'The Efficiency of Sales Promotion in UK Supermarkets: A Consumer View', Int*ernational Journal of Retail and Distribution Management*, vol. 30 no. 6, pp 315-322.

Golfarelli, M., Rizzi, S. & Cella, I. (2004) 'Beyond Data Warehousing: What's Next in Business Intelligence?', *Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP***:** November, ACM Press.

Goa, S., Wang, H., Xu, D. & Wang, Y. (2007) 'An Intelligent Agent-Assisted Decision Support Systems for Financial Planning', *Decision Support Systems*, vol. 44, no. 1, May, pp. 60-78.

Goncalves, M. (2008) Global Management Strategies: Sales, Design, Manufacturing and Operations; New York : ASME.

Guerlain, S., Brown, D.E. & Mastrangelo, C. (2000) 'Intelligent Decision Support Systems', *IEEE International Conference on Systems, Man, and Cybernetics*, vol. 3, October, pp. 1934 **-** 1938.

Haag, S., Cummings, M., McCubbrey, D., Pinsonneault, M. & Donovan, J. (2003). *Management Information Systems: For the Information Age: 4th Edition*: New York: McGraw-Hill Ryerson Limited.

Hadaya, P. & Pellerin, R. (2008) 'Determinants of Advance Planning and Scheduling Systems Adoption', *The Third International Conference on Software Engineering Advances*, 2008, October, pp. 494 – 499.

Han J & Kamber M. (2006) *Data Mining: Concepts and Techniques*: Los Altos, CA: Morgan Kaufman.

Hand, D., Mannila, H. & Smyth, P. (2001). *Principles of Data Mining*: Cambridge: MIT Press.

Hanfling, O. (1981) *Logical Positivism*: Oxford: Basil Blackwell.

Hanson, E.A. & Zilberstein, S. (2001) 'Monitoring and Control of Anytime Algorithms: A Dynamic Programming Approach', *Artificial Intelligence, Special Issue on Computational Tradeoffs under Bounded Resources*, vol. 126, no. 1, February, pp. 139-157, Elsevier Science Ltd.

Hastie, T., Tibshirani, R. & Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*: New York: Springer.

Hay G. J. & Castilla, G. (2006) 'Object based Image analysis: Strengths, Weakness, Opportunities and Threats (SWOT)', International Conference on Object-based Image Analysis (OBIA 2006), July, Salzburg, Austria.

He, K., Shao, J., Liu, Y. & Dong, S. (2008) 'Conceptual Design of Rail Transit Based Urban Logistics Delivery System', *IEEE International Conference on Industrial Informatics*, July, pp. 221 – 226.

He, X.J. & Wu, W. (2006) 'Factors Affecting Adoption of ERP in China', *International Conference on Computational Intelligence for Modelling, Control and Automation, 2006 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce,* November, pp. 156 – 156.

Heckerman, D. (1995) A Tutorial on Learning Bayesian Networks. *Technical Report MSRTR-95-06*, Washington: Microsoft Research.

Helin, H. (2003) *Software Agent Technology: Agent Communication*: Department of Computer Science: University of Helsinki: Available online: www.cs.helsinki.fi/u/hhelin/opetus/oat/communication-240203.pdf: Accessed November 2007.

Hettich, S. & Bay, S. D. (2007) The UCI KDD Archive. Irvine, CA: University of California, Department of Information and Computer Science. Available Online: http://kdd.ics.uci.edu: Accessed July, 2007.

Herschel, R. T. & Jones, N. E. (2005) 'Knowledge Management and Business Intelligence: The Importance of Integration', *Journal of Knowledge Management*, vol. 9, no. 4, pp. 45 – 55.

Hitchens, D.K. (2007) *Systems Engineering: A 21st Century Systems Methodology*: New York: John Wiley & Sons Ltd.

Hitt M. A. & Collins, J. D. (2007) Business ethics, strategic decision making and firm performance; *Business Horizons*, vol. 50, no. 5, pp 353-357.

Holsapple, C.W. & Whinston, A. B. (1996) *Decision Support Systems: A Knowledge-Based Approach*: New York: St. Paul, West Publishing.

Howson, C. (2008) *Successful Business Intelligence: Secrets to Making a Killer BI App*: New York: McGraw-Hill.

Huang, J. & Ling, C. (2005) 'Using AUC and Accuracy in Evaluating Learning Algorithms', *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299 - 310.

Hung, S., Yen, D. C. & Wang, H. (2006) 'Applying Data Mining to Telecom Churn Management', *Expert Systems with Applications*, vol. 31, no. 3, October, pp. 515-524.

JADE (2008) JADE online documentation: Available online: http://jade.tilab.com/doc/index.html: Accessed December, 2008.

James, M. (1991) *Rapid Application Development*: Oxford: Macmillan Publishing Co.

Jars, I., Kabachi, N. & Lamure, M. (2004) 'Social Learning in Multi Agents System: An Additional Step toward Intelligence', *IEEE International Conference on Systems, Man and Cybernetics*, vol 6, October, pp. 5542 – 5547.

Jennings, N.R., Norman, T.J. & Faratin, P. (1998) 'Adept: An Agent-Based Approach to Business Process Management', *SIGMOD Record*, vol. 27, no. 4.

Johansson, A. (2004) 'Professional Work and Its Impact on Development of Information and Communication Technology', *Special Interest Group on Computer Personnel Research Annual Conference Archive*, pp. 65 – 69.

Joyce, K. M. (2005) 'Riding the Tide', *PROMO Industry Trends Report*, April, pp. 3 - 6.

Juan, T., Pearce, A. & Sterling, L. (2002) 'ROADMAP: Extending the Gaia Methodology for Complex Open Systems', *AAMAS*, July, Bologna, Italy.

Kalim, K., Carson, E. & Cramp, D. (2004) 'The Role of Soft Systems Methodology in Healthcare Policy Provision and Decision Support', *IEEE International Conference on Systems, Man and Cybernetics*, vol. 6, October, pp. 5025 – 5030.

Kalos, A. & Rey, T. (2005) 'Data Mining in the Chemical Industry', *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, August, ACM.
Publisher: ACM

Kantardzic, M. & Zurada, J. (2005) *Next Generation of Data-Mining Applications*: New Jersey: John Wiley & Sons-IEEE Press.

Keep, C, McLaughlin, T. & Parmar, R. (2000) *Defining Postmodernism*: Available online: http://www.iath.virginia.edu/elab/hfl0242.html: Accessed September, 2008.

Kelle, P. & Akbulut, A. (2004) 'The Role of ERP Tools in Supply Chain Information Sharing, Cooperation and Cost Optimisation', *International Journal of Production Economics*, vol 12, no 4, Elsevier Science Ltd;

Kendall, E. A., Krishnan, P. V. M., Suresh, C. B. & Pathak, C. V. (2000) *An Application Framework for Intelligent and Mobile Agents*: ACM Computing Surveys.

Khazanchi, D. & Munkvold, B. E. (2003) 'On the Rhetoric and Relevance of IS Research Paradigms: A Conceptual Framework and Some Propositions', *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*, January, pp. 10.

Klösgen, W. & Zytkow, J. M. (2005) *Data Mining and Knowledge Discovery Handbook*: New York: Oxford University Press, Inc.

Knights, M. (2008) *BI Growth to Buck Economic Trends*: Available online: http://www.itpro.co.uk/183480/bi-growth-to-buck-economic-trends: Accessed December, 2008.

Knoors, F. (2002) *'Establish the door-to-door management system', European Commission: DG TREN 5th FP*: Sequoyah International Restructuring.

Kobti, Z. & Sharma, S. (2007) 'A Multi-Agent Architecture for Game Playing', *IEEE Symposium on Computational Intelligence and Games*, April, pp. 276-281.

Kock, N. (2006) *Systems Analysis and Design Fundamentals: A Business Process Redesign Approach*: London: Sage Publications.

Koenig, S. (2004) 'A Comparison of Fast Search Methods for Real-Time Situated Agents', *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS, pp. 864-871.

Kotler, P., Wong, V., Saunders, J. & Armstrong, G. (2005) *Principles of Marketing: European , 4th Edition*: London: Pearsons Education Limited.

Krümpelmann, P., Thimm, M., Ritterskamp, M. & Kern-Isberner, G. (2008) 'Belief Operations for Motivated BDI Agents', *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems*, May, vol. 1, International Foundation for Autonomous Agents and Multiagent Systems.

Kudyba, S. (2004) *Managing Data Mining: Advice from Experts*: USA: Cybertech Publishing.

Kuo, Y-T., Lonie, A., Sonenberg, L. & Paizis, K. (2007) 'Domain Ontology Driven Data Mining: A Medical Case Study', *Proceedings of the 2007 International Workshop on Domain Driven Data Mining*, August, DDDM '07, ACM.

Kwon, O. B. & Lee, J. J. (2001) 'A Multi-Agent System for Efficient ERP Maintenance', Expert *Systems with Applications*, vol. 21, no. 4, November, pp. 191-202, Elsevier Science Ltd.

Landqvist, F. & Pessi, K. (2004) 'Agent Action: Business Cases with Individualised Information Services in a Business Intelligence Context', *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, January, pp. 9.

Langley, P., Iba, W. & Thompson, K. (1992) 'An Analysis of Bayesian Classifiers', *Proceedings of the 10th National Conference on Artificial Intelligence*, California, pp. 223-228, AAAI Press.

Lawrence, K. D., Kudyba, S. & Klimberg, R. K. (2008) *Data Mining Methods and Applications*: New York: Auerbach Publications.

Lazcorreta, E., Botella, F. & Fernández-Caballero, A. (2008) 'Towards Personalised Recommendation by Two-Step Modified Apriori Data Mining Algorithm', *Expert Systems with Applications*, vol. 35, no. 3, October, pp. 1422-1429.

Lavrac, N., Todorovski, L. & Jantke, K. P. (Editors) (2006) 'Discovery Science', *Lecture Notes in Computer Science*, Vol. 4265, Springer-link.

Lawton, G. (2006) 'Making Business Intelligence More Useful', *Computer*, vol. 39, no. 9, September, pp. 14 – 16.

Lee, Y. H., Jeong, S. K. & Moon, C. (2002) 'Advanced Planning and Scheduling with Outsourcing in Manufacturing Supply Chain', *Computers & Industrial Engineering*, vol. 43, no. 1/2, pp. 251-274.

Lee H. & Park, S. C. (2005) 'Intelligent Profitable Customer's Segmentation System Based on Business in Tier Science', *Expert Systems with Applications*, vol. 29, no. 1, July, pp. 145-152. Elsevier Science.

Leist, S. & Zellner, G. (2006) 'Evaluation of Current Architecture Frameworks', *Proceedings of the 2006 ACM Symposium on Applied Computing*, SAC '06, April, ACM.

Li, H., Yang, S. X. & Biletskiy, Y. (2008) 'Neural Network Based Path Planning for a Multi-Robot System with Moving Obstacles', IEEE International Conference on Automation Science and Engineering, 2008. August, pp. 163 – 168.

Li, J. & Ruhe, G. (2007) 'Decision Support Analysis for Software Effort Estimation by Analogy', *PROMISE '07: Proceedings of the Third International Workshop on Predictor Models in Software Engineering*, May, IEEE Computer Society.

Li, S-T., Shue, L-Y. & Lee, S-F. (2006) 'Enabling Customer Relationship Management in ISP Services through Mining Usage Patterns', *Expert Systems with Applications*, vol. 30, no. 4, May, pp. 621-632.

Lin, H. (2007) *Architectural Design of Multi-Agent Systems: Technologies and Techniques*: USA: Information Science Reference.

Lomuscio, A. & Raimondi, F. (2006) 'Model Checking Knowledge, Strategies and Games in Multi-Agent Systems', Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS '06, May, ACM.

Luhn, H. (1958) 'A Business Intelligence System', *IBM Journal*, vol. 2, no. 4, pp. 314, IBM. Available online: http://www.research.ibm.com/journal/: Accessed August, 2008.

Luo; Q, (2008) 'Advancing Knowledge Discovery and Data Mining', *International Workshop on Knowledge Discovery and Data Mining*, WKDD 2008, January, pp. 3 – 5.

Ma, K-L. (2007) 'Machine Learning to Boost the Next Generation of Visualization Technology', *Computer Graphics and Applications*, vol. 27, no. 5, September, pp. 6 – 9, IEEE.

Magee, B. (1985) *Popper*: London: Fontana Press.

McConnell, S. (1996) *Raid Development*: Washington: Microsoft Press.

McCorkell, G. (1999) *Direct and Database Marketing: London*: The Institute of Direct Marketing / Kogan Page.

Mclaughlin, S. (2007) 'Knowledge Management - Achieving Success with KM', *Engineering Management Journal*, vol. 17, no. 5, October, pp. 42- 45.

Merril D. M. & Tennyson, R. D. (1977) *Teaching Concepts: An Instructional Design Guide*: Englewood Cliffs, New Jersey: Educational Technology Publications.

Miao, Q., Li, Y., Wang, F-Y. & Tang, S. (2006) 'Modelling and Analysis of Artificial Transportation System Based on Multi-Agent Technology', *Intelligent Transportation Systems Conference*, ITSC '06, pp. 1120 – 1124, IEEE.

Michalewicz, Z., Schmidt, M., Michalewicz, M. & Chiriac, C. (2005) 'Case Study: An Intelligent Decision Support System', *Intelligent Systems*, vol. 20, no. 4, July, pp. 44 – 49, IEEE.

Mirkin, B. (2005) *Clustering for Data mining: A data Recovery Approach*: Oxford: Chapman & Hall.

Mladenic´ D, Lavrac˘ N, Bohanec M. & Moyle S. (2003) *Data Mining and Decision Support: Integration and Collaboration*: Dordrecht: Kluwer Academic Publishers.

Monahan, G. E. (2000) Management Decision Making: Spreadsheets, Modelling Analysis and Applications: UK: Cambridge University Press.

Murata, T. & Nakamura, T. (2006) 'Multi-Start Node Genetic Network Programming for Controlling Multiple Agents', *IEEE International Conference on Systems, Man and Cybernetics*, SMC '06, vol. 3, October, pp. 1927 – 1932.

Nawana, H. S. (1996) 'Software Agents: An Overview', *Knowledge Engineering Review Journal*, vol. 11, no. 3, September, pp. 1-40.

Nemati, H. R. (2002) *Enhancing Enterprise Decisions through Organisational Data Mining*: Available online: http://www.allbusiness.com/technology/915422-1.html: Accessed November, 2008.

Nemati, H. R., Steiger, D. M., Iyer, S. L. & Herschel, R. T. (2002) 'Knowledge Warehouse: An Architectural Integration of Knowledge Management, Decision Support, Artificial Intelligence and Data Warehousing', *In Decision Support Systems*, vol. 33, no. 2, June, pp. 143-161. Elsevier Science.

Nguyen, T. M., Schiefer, J. & Tjoa, A. M. (2005) 'Sense & Response Service Architecture (SARESA): An Approach towards a Real-Time Business Intelligence Solution and Its Use for a Fraud Detection Application', *Proceedings of the 8th ACM International Workshop on Data Warehousing and OLAP*, DOLAP '05, November, ACM.

Ortiz, S. (2002) 'Is Business Intelligence a Smart Move?', *In Computer*, vol. 35, no. 7, July, pp. 11-14, IEEE Xplore.

Osborne, M.J. (2002) *An Introduction to Game Theory*: Oxford: Oxford University Press.

Ossowski S. (1999) *Co-Ordination in Artificial Agent Societies: Social Structures and Its Implications for Autonomous Problem-Solving Agents*: Berlin: Springer-Verlag.

Padgham, L. & Winkoff, M. (2004) *Developing Intelligent Agent Systems: A Practical Guide*: New York: John Wiley & Sons.

Padgham, L., Thangarajah, J. & Winikoff, M. (2007a) 'AUML Protocols and Code Generation in the Prometheus Design Tool', *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '07, May, ACM.
Publisher: ACM.

Padgham, L., Thangarajah, J. & Winikoff, M. (2007b) 'The Prometheus Design Tool - A Conference Management System Case Study', *In Agent-Oriented Software Engineering VIII: 8th International Workshop*, AOSE 2007, Honolulu, May, pp. 197-211, Springer.

Padgham, L., Thangarajah, J. & Winikoff, M. (2008) **'**Prometheus Design Tool', *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence* (AAAI-2008), Chicago, July, AAAI.

Papadakis, V. M. & Barwise, P. (2002) 'How Much Do CEOs and Top Managers Matter in Strategic Decision-Making?', *British Journal of Management*, vol. 13, no. 1, pp. 83-95, British Academy of Management.

Pedersen, T.B. & Jensen, C.S. (2001) 'Multidimensional Database Technology', *In Computer*, vol. 34, no. 12, November, pp. 40 – 46, IEEE Xplore.

Pedrycz, W. (2005) *Knowledge-Based Clustering: From Data to Information Granules*: New Jersey: John Wiley & Sons.

Peng, W., Sun, T., Rose, P. & Li, T. (2007) 'A Semi-Automatic System with an Iterative Learning Method for Discovering the Leading Indicators in Business Processes', *In Proceedings of the 2007 International Workshop on Domain Driven Data Mining*, DDDM '07, August, ACM.

Peng, Y., Kou, G., Shi, Y. & Chen, Z. (2006) 'A Systemic Framework for the Field of Data Mining and Knowledge Discovery', Proceedings of the Sixth IEEE International Conference on Data Mining Workshops, ICDM Workshops 2006, December, pp. 395 – 399.

Pearl, J. (1985) 'Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning', *In Proceedings of the 7th Conference of the Cognitive Science Society*, August, University of California, Irvine, CA, pp. 329-334.

Parsons, S. (2005) 'Decision Theory and Game Theory in Agent Design', *In Special Issue: Decision Support Systems*, vol. 39, no. 2, April.

Petal, N. V. (2005) *Critical Systems Analysis and Design: A Personal Framework Approach*: Oxford: Routledge Publishing.

Piatetsky-Shapiro, G. (1991) 'Knowledge Discovery in Real Databases', AI Magazine vol. 11, no. 5, pp. 68-70.

Piatetsky-Shapiro, G., Khabaza, T. & Ramaswamy, S. (2003) 'Capturing Best Practice for Microarray Gene Expression Data Analysis', *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, August, ACM.

Phillips-Wren, G., Sharkey, P. & Dy, S. M. (2008). 'Mining Lung Cancer Patient Data to Assess Healthcare Resource Utilisation', *Expert Systems with Applications*, vol. 35, no. 4, November, pp. 1911 – 1619.

Helen Pinto, H., Han, J., Pei, J., Wang, K., Chen, Q. & Dayal, U. (2001) 'Multi-Dimensional Sequential Pattern Mining', *Proceedings of the 10th International Conference on Information and Knowledge Management*, CIKM '01, October.

Pollack, M. E. (1990) *Plans as Complex Mental Attitudes: Intentions in Communication*: Cambridge: MIT Press

Pooley, R. & Wilcox, P. (2004) *Applying UML: Advanced Applications*: Oxford: Elsevier Butterworth-Hienemann.

Power, D. J. (2002) *Decision Support Systems: Concepts and Resources for Managers*: Westport, Conn.: Quorum Books.

Power, D. J. (2007) '*A Brief History of Decision Support Systems, Version 4.0*': Available online: http://www.DSSResources.com: Accessed July, 2008.

van der Putten P. & van Someren M. (eds)(2000) *CoIL Challenge 2000: The Insurance Company Case*: Leiden Institute of Advanced Computer Science Technical Report 2000: Amsterdam: Sentient Machine Research.

Quinlan, J.R. (1990) 'Decision Trees and Decision Making', *IEEE Transactions on Systems, Man and Cybernetics*, vol. 20, no. 2, pp. 339-346.

Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning*: California: Morgan Kaufmann.

Raivio, K. (2006) 'Analysis of Soft Handover Measurements in 3G Network', *Proceedings of the 9th ACM International Symposium on Modelling Analysis and Simulation of Wireless and Mobile Systems*, MSWiM '06, October, ACM.

Reichel, K., Hochgeschwender, N. & Voos, H. (2008) 'OpCog: An Industrial Development Approach for Cognitive Agent Systems in Military UAV Applications', *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems*, *Industrial Track*, AAMAS '08, May, International Foundation for Autonomous Agents and Multiagent Systems.

Roiger, R. J. & Geatz, M. W. (2003) *Data Mining a Tutorial Based Primer*: San Francisco: Addison-Wesley.

Ruggeri, F., Kenett, R. & Faltin, F. (2007) *Encyclopaedia of Statistics in Quality and Reliability*: New Jersey: John Wiley & Sons.

Rupnik, R., Kukar, M., Bajec, M. & Krisper, M., (2006) 'DMDSS: Data Mining Based Decision Support System to Integrate Data Mining and Decision Support',In *International Conference on Information Technology Interfaces*, pp. 225 – 230.

Russell, S. & Norwig, P. (1995) *Artificial Intelligence: A Modern Approach*: London: Prentice Hall.

Rzevski, G. A. (2002) *Multi-Agent Systems and Distributed Intelligence (Paper 021, Version 1.0):* Available online: www.brunel.ac.uk/research/madira: Accessed February, 2007.

Simmers, C. (2004) 'A Stakeholder Model of Business Intelligence', *Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04*), January.

Schroeder, M., Gilbert, D., van Helden, J. & Noy, P. (2001) 'Approaches to Visualisation in Bioinformatics: From Dendrograms to Space Explorer', *Information Sciences*, vol. 139, no. 1-2, November, pp. 19-57.

Schwarz, A., Mehta, M., Johnson, N. & Chin, W. W. (2007) 'Understanding Frameworks and Reviews: A Commentary to Assist Us in Moving Our Field Forward By Analysing Our Past', *ACM SIGMIS Database*, vol. 38, no. 3, July, ACM.

Sethuraman, A., Yalla, K. K., Sarin, A. & Gorthi, R. P. (2008) 'Agents Assisted Software Project Management', *Proceedings of the 1st Bangalore Annual Compute Conference*, Compute '08, January, ACM.

Sharma, S. & Osei-Bryson, K.-M. (2008) 'Organisation-Ontology Based Framework for Implementing the Business Understanding Phase of Data Mining Projects', Proceedings of the 41st Annual Hawaii International Conference on System Sciences, January, pp. 77 – 79.

Shoham, Y. (2008) 'Computer Science and Game Theory', *Communications of the ACM*, vol. 51, no. 8, August, ACM.

Shoham, Y. & Leyton-Brown, K (2008) *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*: Cambridge: Cambridge University Press.

Shvertner, K. (2003) 'Rapid Application Development (RAD) and Deployment Using Oracle', *Proceedings of the 4th International Conference on Computer Systems and Technologies: E-Learning*, CompSysTech '03, June, ACM.

Sickle, J. V. (1997) 'Using Mean Similarity Dendrograms to Evaluate Classifications', *Journal of Agricultural, Biological and Environmental Statistics*, vol. 37, no 4, December.

Singh, M. P & Huhns, M. N. (2005) *Service-Oriented Computing: Semantics, Processes, Agents*: New Jersey: John Wiley & Sons.

Smith K. & Gupta, J. (2002) Neural Networks in Business; Techniques and Applications: London: Idea Group Publishing.

Sprague, R. H., Watson, H. J. & Sprague (Jr.), R. H. (1996) *Decision Support for Managers*: London: Pearson Higher Education.

Stadler, H. (2005) 'Supply Chain Management and Advanced Planning - Basics, Overview And Challenges', *European Journal of Operational Research*, vol. 163, pp. 575–588.

Stamelos, J. G. & Sfetsos, P. (2007) *Agile Software Development Quality Assurance*: USA: Information Science Reference.

Stanhope, P. (2002) *Get in the Groove: Building Tools and Peer-To-Peer Solutions with the Groove Platform*: New York: Hungry Minds.

Stevenson, W. J. (2006) *Operations Management: International Student Edition with Global Readings*: New York: McGraw-Hill-International.

Subramaniam, V. & Hunt, A. (2006) *Practices of an Agile Developer*: Texas: Pragmatic Bookshelf.

SUN MICRO-SYSTEMS (2008): *CUSTOMER SUCCESS STORIES*: Available online: http://www.sun.com/smi/Success/IndustrySpecific/Transportation/London.Under.html: Accessed December, 2008.

Tan, P.-N., Steinbach, M. & Kumar, V. (2005) *Introduction to Data Mining*: USA: Addison-Wesley.

Tapp, A. (2005) Principles of Direct Marketing and Database Marketing: London: Prentice Hall.

Taylor, D. A. (2004) *Supply Chains: A Manager's Guide*: Boston: Pearsons Education Inc.

Thompson, E. (2002) *OLAP Solutions: 2nd Edition*: Building Multidimensional Information Systems: New Jersey: John Wiley & Sons.

Tozicka, J., Rovatsos, M. & Pechoucek, M. (2007) 'A Framework for Agent-Based Distributed Machine Learning and Data Mining', *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '07, May.

Trajkovski, I., Zelezný, F., Tolar, J. & Lavrac, N. (2006) 'Relational Subgroup Discovery for Descriptive Analysis of Microarray Data', *CompLife*, vol. 4126, pp. 86-96, Springer-Link.

Tran, N., Weyerman, W., Giraud-Carrier, C., Seppi, K., Warnick, S. & Johnson, R. (2005) 'Studies in the Dynamics of Economic Systems', *Proceedings of 2005 IEEE Conference on Control Applications*, CCA 2005, August, pp. 861 – 866.

Trestian, I., Ranjan, S., Kuzmanovi, A. & Nucci, A. (2008) 'Unconstrained Endpoint Profiling (Googling the Internet)', *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 4, October, ACM.

Turban, E. & Aronson, J. E. (2001) Decision Support Systems and Intelligent Systems: New Jersey: Pearsons Education.

Tveit, A. (2001) *A survey of Agent-Oriented Software:* Engineering Department of Computer and Information Science**,** Norwegian University of Science and Technology: Available online: http://www.abiody.com/jfipa/publications/AgentOrientedSoftwareEngineering/: Accessed: July, 2008.

Venter, J., de Waal, A. & Willers, C. (2007) *Specialising CRISP-DM for Evidence Mining*
IFIP International Federation for Information Processing: Advances in Digital Forensics III: Boston: Springer.

Vidal, J. M. (2007) *Fundamentals of Multiagent Systems with NetLogo Examples*: Available online: http://multiagent.com/fmas/: Accessed: November, 2007.

Vidgen, R. & Braa, K. (1997) 'Balancing Interpretation and Intervention in Information Systems Research: The "Action Case" Approach', *Proceedings of the IFIP TC8 WG 8.2 International Conference on Information Systems and Qualitative Research,* Philadelphia, Pennsylvania, January pp. 524-541, 1997.

Watson, R. (2001) 'The Effect of Privatisation on Train Planning: A case study of the UK', *Transport Reviews*, vol. 21, no. 2, April, pp. 181-193(13), Routledge.

Wasserman, T. J., Martin, P., Skillicorn, D. B. & Rizvi, H. (2004) 'Developing a Characterisation of Business Intelligence Workloads for Sizing New Database Systems', Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP, DOLAP '04, November, ACM.

Weiss, G. (1999) *Multi-Agent Systems: A Modern Approach to Distributed Artificial Intelligence*: Cambridge: MIT Press.

Weiss, N. (2007) *Introductory Statistics: 8th Edition*: New York: Pearsons International.

Wang, X.F., Zhang, S.C., Khosla, P.K. & Kilicote, H. (2001) 'Anytime Algorithm for Agent-Mediated Merchant Information Gathering', *Proceedings of the 4th International Conference on Autonomous Agents*, Agents 2000, June, pp. 333-340, Barcelona, Spain, ACM.

Watson H. & Wixom, B. 'The Current State of Business Intelligence', *Computer*, vol. 40, no. 9, September, pp. 96 – 99.

Watson, R. T. 2002; Data management; Databases and Organisations: Chichester: John Wiley & Sons.

Weka (2008): Available online: http://www.cs.waikato.ac.nz/ml/weka/: Accessed: September, 2008.

Wei, X., Xiaofei, X., Lei, S., Quanlong, L. & Hao, L. (2001) 'Business Intelligence Based Group Decision Support System', *Proceedings of the International Conferences on Info-Tech and Info-Net*, ICII 2001, October, vol. 5, Beijing, pp. 295 – 300.

Winston, W. L. (2004) *Operations Research: Applications and Algorithms: 3rd Edition*: California: Duxbury Press.

William, C., Shanmuganathan, S. & Ghotbi, N. (2007) 'Text Mining in Radiological Data Records: An Unsupervised Neural Network Approach', *Proceedings of First Asia International Conference on Modelling & Simulation*, AMS '07, March, pp. 329 – 333.

Williams, S. & Williams, N. (2003) 'The Business Value of Business Intelligence', *Business Intelligence Journal*, TDWI, vol 3, no. 4.

Witten, I. & Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*: San Francisco Morgan Kaufmann.

Wooldridge, M. (2002) *An Introduction to Multi-Agent Systems*: Chichester: John Wiley & Sons ltd.

Wooldridge, M., Jennings. N. R. & Kinny, D. (2000) The Gaia Methodology for Agent-Oriented Analysis and Design: *Autonomous Agents and Multi-Agent Systems*, vol. 3, pp 285-312, Kluwer Academic Publishers.

Wren, A. & Wren, D. O. (1995) 'A Genetic Algorithm for Public Transport Driver Scheduling', *Journal of the Operational Research Society*, vol. 22, no. 1, January, pp. 101 - 110.

Wu, L., Barash, G. & Bartolini, C. (2007) 'A Service-oriented Architecture for Business Intelligence', *IEEE International Conference on Service-Oriented Computing and Applications*, SOCA '07, June, pp. 279 - 285.

Wu, X. (2004) 'Data Mining: Artificial Intelligence in Data Analysis', *Proceedings of IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, pp. 7.

Xu, L., Zeng, L., Shi, Z., He, Q. & Wang, M. (2008) 'The Research and Design of Grid-Based Business Intelligence Network-BGBIN', *International Symposium on Electronic Commerce and Security*, August, pp. 129 – 132.

Xu, L., Zeng, L., Shi, Z., He, Q. & Wang, M. (2007) 'Research on Business Intelligence in Enterprise Computing Environment', *IEEE International Conference on Systems, Man and Cybernetics*, ISIC, October, pp. 3270 – 3275.

Yuefeng, L. (2007) 'Interpretations of Discovered Knowledge in Multidimensional Databases', *IEEE International Conference on Granular Computing*, GRC 2007, November, pp. 307 – 309.

Zan, M., Shan, Z., Li, L. & Ai-jun, L. (2007) 'A Predictive Model of Churn in Telecommunications Based on Data Mining', *IEEE International Conference on Control and Automation*, ICCA 2007, May, pp. 809 – 813.

Zhang, D. & Zhou, L. (2004) 'Discovering Golden Nuggets: Data Mining in Financial Application', *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 34, no. 4, November, pp. 513 – 522.

Zhang, F., Yang, B., Song, W. & Li, L. (2007) 'Intelligent Decision Support System Based on Data Mining: Foreign Trading Case Study Control and Automation', *In IEEE International Conference*, ICCA 2007, May, pp. 1487 – 1491.

Zhao, M., Chen, Y., Liu D. & Li, J. (2006) 'The Application of Optimized Fuzzy Decision Trees in Business Intelligence', *Proceedings of the* International *Conference on Machine Learning and Cybernetics*, Dalian, China, August, pp. 2212 – 2217.

Zhengwen, D., Zhang, A., Anosike, I., Lim, M-K. & Akanle, O. M. (2006) 'An Agent-Based Approach for E-Manufacturing and Supply Chain Integration', *Computers & Industrial Engineering*, vol. 51, no. 2, October, pp. 343-360.

Zhon, Y. (2007) 'Soft Systems Methodology Based on Decision Making Knowledge Integration', *Proceeding of the International Conference on Wireless Communications, Networking and Mobile Computing*, WiCom 2007, September, pp. 5733 – 5736.

Zhou, F., Yang, B., Li, L. & Chen, Z. (2008) 'Overview of the New Types of Intelligent Decision Support System', 3rd International Conference on Innovative Computing Information and Control, ICICIC '08, June, pp. 267 – 270.

Appendix A:

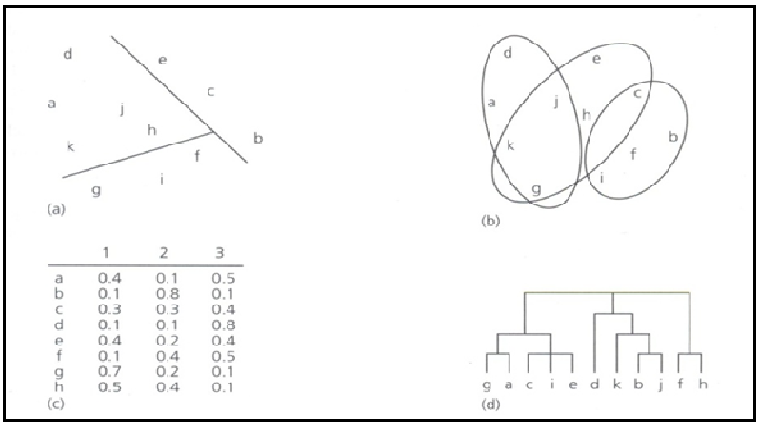
Technical Documentation Relating to Case Studies

This appendix provides an extension of Chapter 2 and more importantly the background details for tools and techniques which were explored and used for the case studies contained in appendices: B, C and D. The techniques documented within this section demonstrate the functionality of KDDS-BI to require background analysis to only be conducted once and the information shared between projects irrespective of the end-user or domain.

A.1 Unsupervised Clustering

Unspervised Clustering (otherwise known as simply: 'Clustering') can be defined classification or grouping of physical or abstract objects into different groups, or more accurately, the partitioning of a data set into subsets (clusters). The data items in each subset share some common trait. Clustering is generally regarded as an unsupervised learning technique. Within clustering, there is no dependent variable that can guide the learning process. In contrast, learning takes place through knowledge structures that are developed by utilising a common trait or some measure of cluster quality; this may be a characteristic or even proximity according to some defined distance measure, the data/ instances are then grouped into two or more classes (Alfred & Kazakov, 2007; Roiger & Geatz, 2003).

In Clustering, data is often placed upon a two dimension plane and partitioned, thereby, identifying the various clusters and constituent values. Figure A.1 (a) depicts one possible option for portioning data using this method. It should be noted that an instance does not have to be exclusive, various clustering algorithms allow (whilst some do not) for an instance to belong to multiple clusters, as depicted in figure A.1 (b). Figure A.1 (c) represents instances that have been portioned probabilistically rather than categorically, thus in this example each cluster sum to 1. Figure A.1 (d) the clusters have been divided by hierarchy, resulting in few clusters at the highest level, with each cluster dividing into a number of sub-clusters (Witten & Frank, 2005).



	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1

Figure A.1: Methods for representing clusters.

Since there are various options for identifying clusters, and as a result there are a number of applications for Clustering. Clustering methods such as the one depicted in figure A.1 (d) are known as 'dendrograms', this method is frequently investigated to cluster species in biology and agriculture (Sickle, 1997; Schroeder, 2001). Clustering is also often investigated to identify outliers. Outliers are anomalies or values that are 'irregular' and as result, do not conform to the clusters within the data set. Outliers can often be of greater interest then the clusters themselves, since an outlier can be investigated to identify unexpected behaviour. Outlier detection is of great interest within fraud detection as they can identify behaviour or traits that lie outside the normal expected behaviour, an example of this can be a value inconsistent with an individual's credit card spending pattern may be due to the card having been stolen (Mirkin, 2005). Other common uses of unsupervised clustering are to:

- Determine whether meaningful relationships, correlations or trends can be identified within the dataset.
- Evaluate the probable performance of a supervised learner model.
- Determine which attributes of a dataset will be most suited as input attributes for a supervised learner.

Clustering is a challenging field of research and one within which its potential applications pose their own requirements (Han & Kamber, 2006). As a result, Clustering is a technique suited to datasets where the target variable or a test set is unavailable.

A.1.1 K-means Algorithm

The K-means algorithm for Clustering is one of the simplest unsupervised learning algorithms that can be investigated to resolve statistical Clustering problems. The procedure for K-means Clustering follows a simple method for classifying a given data set through a certain ('K') number of disjoint clusters. The algorithm can be defined by the following 5 stages (Roiger & Geatz, 2003):

1. Select a value for 'K'. K = the total number of clusters to be determined.
2. The algorithm will procede to randomly select 'K' data points as initial centre points for the clusters. The centre point of the cluster is known as the 'centroid'.
3. Each instance in then placed within a cluster to which it is most 'similar'. Although the criteria explored to determine similarity can be altered, the most frequent criteria for similarity is 'Euclidean distance'.
4. Once all instances have been placed within their designated cluster, the centroids for the clusters are updated, through the computation of a new mean for each cluster.
5. If the recalculated mean value for the centroid results in no movement of the centroid the process terminates, thus it is considered that a 'convergence criterion' has been met. If however the centroid is replaced, the process will reiterate stages 3-5 until the ideal mean value for the cluster has been discovered.

The main advantages of this algorithm are its simplicity and speed with which can be investigated upon large datasets. However, this method is not without its drawbacks since the algorithm does not yield the same result with each application. The cause of this variance between applications is that the initial assignment of the centroid can bear a distinct effect upon the outcome. The objective of the algorithm is to minimise the total squared distance or squared error function between the centroid and the points within the cluster. This error or distance can be calculated using formula (A.1).

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2 \qquad \text{(A.1)}$$

Where $||x_i^{(j)} - c_j||^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the n data points from their respective cluster centres.

The minimum squared distance does not always represent the global maximum, since the initial assignment of the centroid is randomly selected. In order to obtain a global minimum for the squared distance, the algorithm

will often be applied several times, prior to selecting the result that returns the smallest total squared distance and accepting this instance to be the global minimum (Witten & Frank, 2005).

'Cluster analysis' is widely used in market research when working with multivariate data from surveys and test panels. Market researchers use cluster analysis to partition the general population of consumers into market segments and to better understand the relationships between different groups of consumers/potential customers (Mirkin, 2005). Therefore, K-means Clustering can be investigated to determine the segmentation within customer groups in addition to the products that would benefit from joint sales promotions.

A.1.2 EM (Expectation-Maximisation) Algorithm

Like the K-means Clustering algorithm and furthermore, considered to extend the algorithm, the EM (Expectation-Maximisation) algorithm is another statistical Clustering technique which can be explored. The EM algorithm employs the finite Gaussian mixtures model. Gaussian mixture models are among the most statistically mature methods for Clustering. A mixture is a set on 'n' probability distributions, where each distribution represents a cluster. The mixtures model, assigns each data instance a probability, thereby assuming all attributes to be independent random variables. The EM algorithm can be defined as an iterative technique for estimating the value of some unknown quantity, given the values of some correlated, known quantity. EM is frequently used for data clustering in machine learning and computer vision, since the EM algorithm is an efficient iterative procedure to compute the 'maximum likelihood' estimate in the presence of missing or hidden data. In maximum likelihood estimation, the motivation is to estimate the model parameter(s) for which the observed data are the most likely (Mirkin, 2005).

The algorithm initially assumes that the quantity is represented as a value in some parameterised probability distribution (e.g. the Gaussian mixture model). The EM algorithm can then be defined through the following steps (Han & Kamber, 2006):

1. Initialize the distribution parameters.
2. E-Step: estimate the 'expected' value of the unknown variables, given the current parameter estimate.
3. M-Step: re-estimate the distribution parameters to 'maximise' the likelihood of the data, provided the expected estimates of the unknown variables.
4. Repeat the second and third stages until convergence.

Hence, the iterated steps of the EM algorithm consist of two processes: The E-step (step 2) and the M-step. In the 'expectation', or 'E-step', the missing data are estimated given the observed data and current estimate of the model parameters. This is achieved using the conditional expectation. In the 'M-step', the likelihood function is maximised under the assumption that the missing data is known. The estimate of the missing data from the 'E-step' is used in lieu of the actual missing data. The EM algorithm, terminates once a local maximum is achieved, known as convergence. Convergence is assured since the algorithm is guaranteed to increase the likelihood with each consecutive iteration. However, like the K-means algorithm, a local maximum, may not be the global maximum, hence the algorithm should be repeated several times, with varying initial parameters. The optimal convergence will be the highest likelihood score (Pedrycz, 2005).

A.2 Supervised Learning

In contrast to unsupervised Clustering, supervised learning techniques, construct models through input attributes that can be explored by the learning model to predict output attribute variables. In general (although not exclusively), supervised learning techniques only permit the prediction of a single output attribute. The output attributes are known as dependent variables, since their outcome is dependent upon one or more of the input attributes, known as independent variables. Due to unsupervised Clustering models not requiring an output attribute, all attribute in unsupervised learning are independent (Amershi & Conati, 2007). Learning algorithms are now used in many domains (Caruana & Niculescu-Mizil, 2006). Supervised learning algorithms are regarded as supervised since they employ a 'Training set'. A training set provides the output values for each of the training examples. The training set encompasses the data instances that are used to create supervised learning models. The models that are built are then tested via a set is data instances known as the test set. In the event that several models have been built using a particular training set a validation set is data instances that are applied to optimise parameter settings for supervised learning models, or aid in the selection of a model (Watson, 2002).

A number of supervised learning techniques have been introduced in the last decade these supervised learning techniques' can be further categorised according to the data type of the output attribute. Classification outputs are categorical, while estimation models output numeric values/attributes (Caruana & Niculescu-Mizil, 2006). 'Classification' is a procedure in which individual items are placed into groups based on quantitative information on one or more characteristics inherent in the items (referred to as traits, variables, characters or attributes). Classification techniques have been comprehensively researched perhaps more so then any other supervised learning technique. Classification techniques are generally employed to classify, current rather then future behaviour. Examples of situations to which classifiers are applicable (Roiger & Geatz, 2003):

- Determine the characteristics that differentiate patients that have suffered a stroke to those who have not.
- Determine whether an individual is likely to default on loan repayments, or whether an individual should even be considered for a loan.
- Develop a profile of a dangerous driver.
- Identify false insurance claims.

The Classification models are developed using a training set of previously labelled items the classifier builds models to learn how best to classify future instances. As discussed, when investigating with supervised classification the output attribute, its range of values and the training data are provided, these attributes are then analysed in newly presented examples when determining the most suitable class to assign the new instance. These models, however, are usually optimised with test and validation sets prior to using real data. Classification often employs 'Decision Trees' to model results. Decision Trees create a map from observations about an item to obtain conclusions about its target value to model classes (Witten & Frank, 2005). Similar to Classification, 'Estimation' examines pre-determined datasets (training sets) to develop models and analyse new instances, which are assigned to labels. However, unlike Classification, which analyses discrete or categorical attributes, Estimation is the technique utilised for analysing datasets where the outcome variable is continuous or numeric. Hence, provided with training data to develop a model, estimation provides a value for some unknown variable

such as income, height or credit balance. An advantage of Estimation over Classification is that the outcome variables can more easily be ranked. Regression models and Neural Networks are examples of techniques that are often investigated for Estimation.

Majority of supervised learning algorithms can be applied for Classification or Estimation; however these algorithms are usually not able to examine both discrete and continuous datasets attributes within a single dataset. A common approach for such an instance is to transform the continuous variables to discrete values (Li & Rhue, 2007). Furthermore, whilst Classification and Estimation are investigated to determine a current condition, *Prediction* can be investigated to determine future outcomes. Prediction models are investigated when examining datasets to analyse a future variable or outcome. Although similar to classification or estimation, through prediction models, records are classified according to predicted future behaviour or estimated future value. Most Classification or Estimation techniques can be adapted to analyse numeric or discrete outcome variables for Predication models provided the outcomes has yet to occur. The primary reason for prediction being considered distinct from Classification or Prediction is the additional issues regarding the temporal relationships between the input variables (predictors) to the output variable (Barry & Linoff, 2004). Sequential pattern and time-series mining is often investigated for prediction. Supervised learning can be investigated to discover patterns or trends where one event (or value) results in a particular later event (or value). One example of such an instance is that generally an increase in the rate of inflation, tends to result in a fall within the stock market (Zhang & Zhou, 2004). Bayes Theorem; Naïve Bayes; Bayesian Networks; Decision Trees; Production Rules; and Artificial Neural Networks are supervised learning techniques which have been used within this study.

A.2.1 Bayes' Theorem

Bayesian Classifiers are a fundamental statistic approach in pattern recognition, which assumes that the prior probabilities of classes and the conditional probabilities of patterns given each class are all known. There have been a number of Bayesian classification algorithms, however, they are all based upon 'Bayes' theorem. Named after Thomas Bayes an 18th century philosopher, Bayes' theorem relates the conditional and marginal probabilities of class 'ω_i'and pattern 'x', where 'x' has a non-vanishing probability. Bayes' theorem thus can be expressed as:

$$Posterior = \frac{Likelihood \cdot Prior}{Evidence} \qquad (A.2)$$

Consequently, in a multiple class pattern recognition case, where class label 'ω_i' comes from $\{\omega_i \ldots \omega_c\}$:

- '$P(x)$' is the prior or marginal probability or evidence of a pattern 'x'. Furthermore this function acts as a normalising constant.
- '$P(\omega_i)$' is the prior probability or marginal probability of class 'ω_i', where '$i = 1 \ldots L$'. It is considered 'prior' since 'ω_i' does not take into account any information about pattern 'x'.
- '$P(x|\omega_i)$' is the conditional probability of pattern 'x' given class 'ω_i'.
- '$P(\omega_i|x)$' is the conditional probability of class 'ω_i', given pattern 'x'. It is also called the 'posterior' probability because it is derived from or depends upon the specified value of 'x'.

Thus, from Bayes' theorem, posterior probability is calculated as (Frank & Witten, 2005):

$$P(\omega_i|x) = \frac{P(x|\omega_i) \cdot P(\omega_i)}{P(x)} \quad \text{(A.3)}$$

Where the evidence '$P(x)$' can be calculated as

$$P(x) = \sum_{j=1}^{c} P(x|\omega_j) \cdot P(\omega_j) \quad \text{(A.4)}$$

Bayes' theorem has been the basis for a number of classification techniques. 'Naïve Bayes Classifier' is the simplest approach in Bayes' classifier family, and sometimes the most effective approach although it has a strong constraint on the independence between features of patterns.

A.2.2 Naïve Bayes

The naïve Bayes algorithm assumes attribute values are independent with no dependence relationships, and as a result is considered naïve. Although the constraint of independence between features is not always satisfied in reality, this method is still attractive to many classification problems. Due to the simplicity of the model and its inexpensive cost in computation, the model is especially suited to investigations involving machine learning (Langley et al., 1992). Han & Kamber (2001) and Witten & Frank (2005) amongst others have expressed how empirical studies illustrate that the performance and accuracy of naïve Bayes is comparable to that of more sophisticated classifiers, such as Decision Trees, even though the class conditional independence assumption is made. Formally expressed, naïve Bayes also classifies a pattern 'x' into a class 'ω_i', if 'ω_i' has the largest posterior probability '$P(\omega_i|x)$'.

In order to calculate the posterior probability of each class, the conditional probabilities of a pattern 'x' given each class 'ω_i', which are '$P(x|\omega_i)$' must first be discovered. To discover these probabilities they must be estimated from training data. Therefore given that a pattern '$\boldsymbol{x}$' is represented by 'm' features, '$\boldsymbol{x} = [x_1, x_2 \ldots x_m]$', the conditional probability '$P(x|\omega_i)$' is computed as:

$$\begin{aligned} P(x|\omega_i) &= P(x_1, x_2 \ldots x_m|\omega_i) \\ &= P(x_1|\omega_i)\, P(x_2 \ldots x_m|\omega_i, x_1) \\ &= P(x_1|\omega_i)\, P(x_2|\omega_i, x_1)\, P(x_3 \ldots x_m|\omega_i, x_1, x_2) \\ &= P(x_1|\omega_i)\, P(x_2|\omega_i, x_1) P(x_3|\omega_i, x_1, x_2) \cdot \ldots \cdot P(x_m|\omega_i, x_1, x_2 \ldots x_{m-1}) \end{aligned} \quad \text{(A.5)}$$

Confirming to the discoveries that have been observed in statistics, 'x_i' and 'x_j' are conditional independent to each other given 'C'. Thus, if '$P(x_i|C, x_j) = P(x_i|C)$', it can be naïvely assumed that all the features in a pattern $\boldsymbol{x}$ are conditionally independent given class 'ω_i'. The computation of conditional probability '$P(x|\omega_i)$' can consequently be simplified as (Han & Kamber, 2001):

$$P(x_1, x_2 \ldots x_m|\omega_i) = P(x_1|\omega_i)\, P(x_2|\omega_i) \cdot \ldots \cdot P(x_m|\omega_i) \quad \text{(A.6)}$$

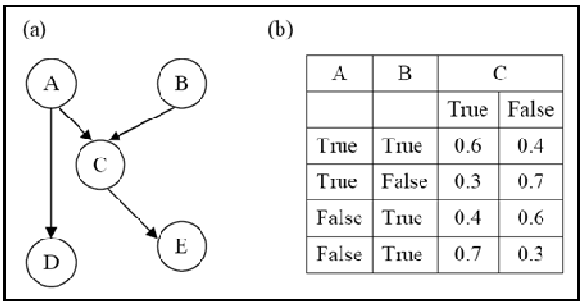
As discussed due to its simplicity naïve Bayes has proven to be a popular approach to computational classification problems. Furthermore, the performance of naïve Bayes has proven to be comparable to that of far more sophisticated algorithms. However, naïve Bayes is not the only classifier that is grounded in Bayes' theorem. Bayesian networks have also proven to be a popular classification approach.

A.2.3 Bayesian Networks

Originally popularised by Pearl (1985), 'Bayesian networks' also known as belief networks, is a probabilistic graphical modelling approach. Thus, Bayesian networks are utilised to represent a set of variables and their probabilistic independencies. In contrast to naïve Bayes which can be defined as a probability estimator, Bayesian networks facilitate causal relationships between variables to be modelled as a directed acyclic graph (Heckerman, 1995). Naïve Bayes is able to provide an estimate of the probability that a particular instance belongs to a certain class, these estimates can then be ranked to plan predictions. However, naïve Bayes can only represent simple distributions, whereas Bayesian networks can be investigated to concisely and comprehensibly represent more complex probability distributions. Furthermore, Bayesian networks permit this distribution to be represented graphically (Witten & Frank, 2005). Bayesian networks are explored to represent knowledge pertaining to an uncertain domain; consequently each node in the graph represents a random variable, while the edges between the nodes represent probabilistic dependencies among the corresponding random variables. These conditional dependencies in the graph are often estimated by using known statistical and computational methods that combine principles from graph theory, probability theory, computer science, and statistics (Ruggeri et al, 2007). A Bayesian network is a represented by an acyclic directed graph, together with an associated set of probability tables. Acyclic graphs are a popular graphical modelling structure, often utilised within statistics, machine learning and artificial intelligence. The structure of an acyclic graph is defined by two features:

- A set of nodes (vertices) that represent random variables and are drawn as circles labelled by the variable names.
- A set of directed edges which represent direct dependence among the variables and are drawn by arrows between nodes.

Figure A.2 illustrates a Bayesian network. Figure A.2 (a) depicts an acyclic graph, the directed edges indicate probability dependency and are connected in such a way that no cycles are present, hence the graph is considered acyclic. The nodes of the graph represent attributes. Each of the attributes is assigned a conditional probability distribution. Figure A.2 (b) illustrates the associated probability table. Consequently, for this example the probability table depicts the combination of probabilities for the True and False values of attribute C and its parent attributes A and B. Assuming there are no missing values; a given sample has attribute values which correspond to a conditional probability in the table at its node. For each class label these probabilities are multiplied together and then normalized to sum to 1, resulting in the joint probability for each class label for the given sample (Witten & Frank, 2005).



A	B	C	
		True	False
True	True	0.6	0.4
True	False	0.3	0.7
False	True	0.4	0.6
False	True	0.7	0.3

Figure A.2: (a) A simple Bayesian network. (b) Conditional probability table for attribute C.

The probability that a certain sample belongs to a particular class label can be predicted by applying Bayes theorem, using naïve Bayes and Bayesian networks as classifiers. Both methods of classification have been highly regarded in the field of machine learning and have been proven to for yield impressive results (Elkan, 1997). Bayesian networks provide a means through which statistics can be graphically represented; however, Bayesian networks are not the only graphical modelling approach that is available for classification investigations. 'Decision Trees' are another popular statistical modelling approach which can be explored for classification.

A.2.4 Decision Trees

Decision Trees are a predictive model, thus explored for mapping observations about an item to analyse or ascertain relevant conclusions regarding an explicit target value. More descriptive names for such tree models are 'Classification Tree' (discrete outcome) or 'Regression Tree' (continuous outcome). In these tree structures, leaves represent classifications and branches represent the relationships between features that lead to the classifications. Decision trees offer not only a graphical representation, but offer several other advantages to the user since Decision Trees are (Roiger & Geatz, 2002):

- Simple to understand and interpret. A Decision Tree models can be easily understood with a very little guidance or prior knowledge.
- Valuable information and knowledge can be discovered from Decision Trees that have been generated with limited data.
- Decision Trees can be transformed into rules.

In addition to the above advantages, decision trees have proven to be successful for a variety of applications. Decision Trees have proven to be successful, since just like naïve Bayes classifiers, decision trees provide a simple means through which data can be analysed. Decision Trees provide a 'divide-and-conquer' approach to classification. Decision Trees utilise internal nodes, branches and leaves to learn from a set of independent instances. Figure A.3 illustrates a Decision Tree, samples initiate at the root node and branch down to other nodes until a leaf is reached. Each branch represents a particular outcome of an attribute test at the node and the leaf represents a class label. Once an attribute have been selected as the root node, a branch and node is created for each attribute value. An attribute is selected at each new node according to training samples contained there and splits further. This continues recursively until the node contains training samples with the same class which then becomes a leaf, or until another specified stopping condition is met (Roiger & Geatz, 2002).

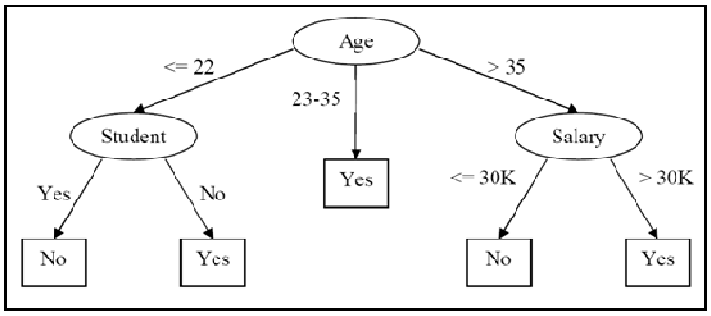


Figure A.3: Decision Tree for deciding whether to 'mail' a customer.

This 'divide-and-conquer' approach of Decision Tree models was coined by Quinlan (1990) who developed a popular Decision Tree induction algorithm 'Id3'. The basic premise of Id3 was that in the Decision Tree each node should be associated the non-categorical attribute which is most informative among the attributes not yet considered in the path from the root. Id3 has been further refined into the 'C4.5 algorithm'. C4.5 is an extension of Id3 that accounts for unavailable values, continuous attribute value ranges, pruning of Decision Trees, and rule derivation (Quinlan, 1993). Decision Trees constructed in this manner can perform well on a training set, however in the absence of pruning, they tend not to generalise well to independent test sets.

This problem is known as 'overfitting'. There are various pre- and post-pruning methods to avoid overfitting. A fundamental post-pruning method is 'subtree replacement'. Subtree replacement constructs a complete tree prior to pruning the tree backwards and replacing sub-trees with single leaves. Subtree replacement is illustrated in Figure A.4. Figure A.4 (a) depicts the complete tree, the sub-tree 'C' is then pruned and replaced by leaf 'a' in Figure A.4 (b). An alternative approach 'subtree raising' is illustrated in Figure A.4 (c). Subtree raising can be utilised when nodes can be raised to encompass other nodes, consequently figure A.4 (c) illustrates an event where node 'C' subsumes node 'B' (Witten & Frank, 2005).

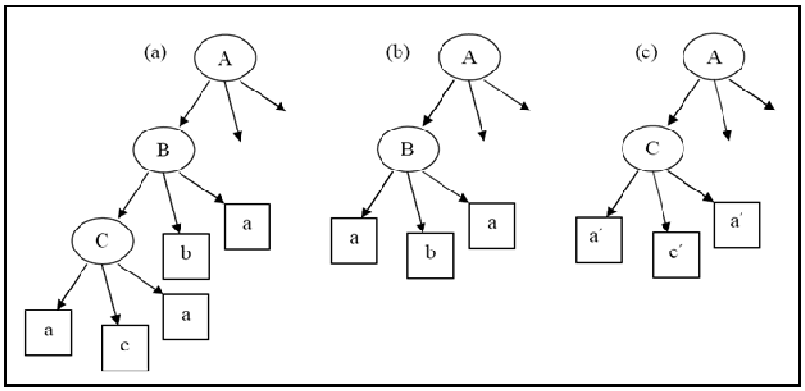


Figure A.4: (a) Unpruned tree. (b) Sub-tree 'C' replaced by leaf 'a' (c) Sub-tree 'C' is raised.

Both these pruning approaches are used in conjunction with one of two methods to compute the error rate 'reduced-error pruning' or 'error-based pruning'. Reduced-error pruning employs validation sets to assess whether the internal node would lead to a greater expected error rate if pruned, if so then the subtree is kept (Han & Kamber, 2001). Error-based pruning implemented in C4.5 uses the same training data to assign the majority

class of samples at a particular node to represent it. Subsequently the error estimates for each leaf in each subtree is combined into the ratio of the number of samples they cover. If the parent node has a lower error estimate when compared, then the children (leaves) are pruned away (Witten & Frank, 2005). All these methods are available in the influential C4.5 algorithm (Quinlan, 1993), allowing pruning parameters to be adjusted to empirically determine which yields the best Decision Tree for a given dataset.

The aforementioned 'divide-and-conquer' approach to constructing Decision Trees, in addition to the ease through which decision trees can be followed and understood facilitates the extraction of rules from the Decision Tree. Consequently, Classification Rules provide another means through which classification problems such as those underpinning this research can be addressed.

A.2.5 Production Rules

'Rule-based classification' is a popular alternative to Decision Trees. Although providing similar functionality to that which is provided by the previously investigated classification techniques. Rule-based classification, unlike Bayesian classifiers and Decision Trees, does not provide graphical representation, rather a set of rules that can be investigated to provide conditions which when met provide the correct classification of an instance. Rules, such as these that provide classification these rules are known as 'Production Rules' or 'Classification Rules' (Roiger & Geatz, 2003).Production Rules consist of two distinct elements, and are a result illustrated in the form of:

IF (antecedent conditions), ***THEN*** (consequent conditions).

The 'IF' element, also referred to as the 'left-hand side of a rule', describes the rule *antecedent*, these are the pre-conditions that must be observed. In general, in the event of multiple antecedent conditions these are intersected through a logical 'AND', and must, therefore, all be satisfied prior to the rule coming into effect. The 'THEN' element, or 'right-hand side of a rule', is the rule *consequent*. Accordingly, the antecedent provides the details for the value or value ranges of one or more attribute variables. Whilst the consequent provides the value or value ranges for the output variables (Han & Kamber, 2006). Production Rules provide an effective method for detailing rules that can be applied to classify new instances; furthermore, these rules can be derived (mapped) directly from Decision Trees. This novel feature has led to the popularity of Production Rules, since not only can they be mapped from Decision Trees, but unlike Decision Trees, provide 'nugget' of knowledge. Consequently, new rules can be added to an existing set of rules, without disrupting the order of the rule set. In contrast if a new antecedent was to be added to Decision Tree, the entire tree may require reshaping to ensure that the integrity of the tree is maintained.

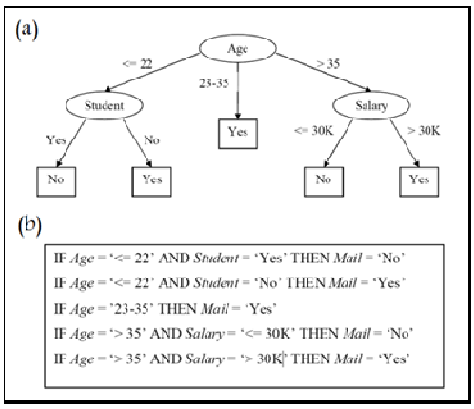


Figure A.5: (a) Decision Tree (b) Production Rules.

If a rule set is to be mapped from a Decision Tree, the rule initiates at the root node of the tree. Each node and branch (attribute-value pair) in the path correlates to form the antecedent, whilst the leaf (the class label), provides the consequent values. Figure A.5 (a) illustrates a Decision Tree. Mapping rules from the corresponding branches-leafs of a Decision Tree will provide a set of Production Rules (figure A.5 (b)). The individual Production Rules that are discovered are considered to be intersected through a logical 'OR', thus, if any rule applies to an instance, the consequent is applied. Conflicts may arise in the event that more than one of the Production Rules applies to an instance. In such an incident, there are several solutions, amongst these are to not provide a classification in the event multiple classifications apply, alternatively once all instances within the sample have been classified, the most popular rule is applied to classify the conflicting instances. Although, Production Rules mapped from a Decision Tree are usually unambiguous, therefore, the order in which they execute is irrelevant (Witten & Frank, 2005). In the event that the rule set is produced from a means other then mapping the order of the rules bears relevance to the accuracy of classification. The rules should however be viewed as a decision list, and applied in order, since altering the order may result in the rules being read out of context, thus 'misclassifying' instances.

Furthermore, production rules that are mapped directly from a Decision Tree are often found to be far more complex than necessary. For this reason, it is often advisable to employ pruning measure to ensure the efficiency of the Production Rules, thereby removing redundant rules. Alternatively covering algorithms can be employed (Furnkranz & Flack, 2005). Covering algorithms employ a separate-and-conquer approach. This separate-and-conquer approach implies that one rule is learnt at a time which 'covers' a certain amount of training samples. These samples are in turn separated from the training samples and subsequent rules are learned which cover the remaining samples. The benefit of this approach is that the algorithm generates rules for one class at a time, with a view to maximising the number of positive samples covered and disregarding the negative examples. This differs to divide-and-conquer Decision Trees, which aim for overall purity concerning all classes. Consequently, datasets with a disproportionately low number of positive samples will be classed less accurately with divide-and-conquer Decision Trees than with separate-and-conquer rules.

A.2.6 Artificial Neural Network

An Artificial Neural Network, often just referred to simply as a 'Neural Network', is a mathematical model or computational model, based on biological Neural Networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an artificial Neural Network is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. In more practical terms, Neural Networks are non-linear statistical data modelling tools. They can be used to model complex relationships between inputs and outputs or to find patterns and/or structures in data. Neural networks are traditionally considered a means through which to examine data and develop models that enable them to uncover and identify these patterns. Since Neural Networks examine data to develop models, Neural Networks are considered to be a supervised learning technique. However, not all Neural Networks are supervised (Dandawate et al, 2008; William et al, 2007). Neural Networks can be considered both a supervised or unsupervised technique, given the means that are employed to develop models.

The simplest and most widely used form of a Neural Network is a forward feed Neural Network (Berry & Linoff, 2004). As illustrated by figure A.6, a forward feed Neural Network generally, Neural Networks consist of at least three layers; an 'input layer', 'hidden layer' and 'output layer'. These layers contain a number of interconnected nodes that mimic the neurons of a human brain. Each node, much like a human neuron has a limited 'simplistic' function, however it is the emergent sum of all activity within the network that enables complex calculations.

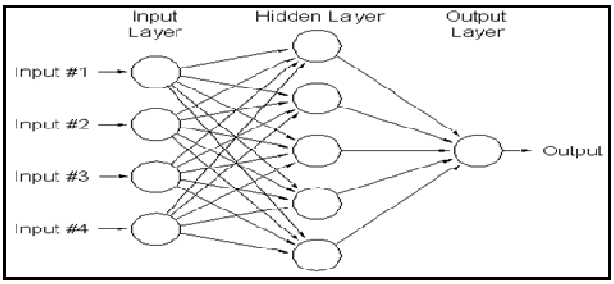


Figure A.6: A forward-feed Neural Network.

A forward feed Neural Network calculates output values from input values. The nodes of the input layer represent the raw information that is fed into the network. Each node is connected to exactly one input source and maps the value to a range between -1 to 1. The hidden layer (of which there can be many) is connected to neither the raw inputs nor output values of the network hence, is considered hidden. The hidden layer can be considered a 'black box', which performs complex calculations to provide an output, these calculations are not evident to the user of the network. Each node within the hidden layer is connected to each node of the input layer. The hidden layer will then multiply the mapped values of the input by its associated 'weighting'. As a result, the activity of each hidden unit is determined by the activities of the input units and the weights on the connections between the input and the hidden units. The final layer, the 'output layer' is fully connected to the hidden layer, i.e. each node in the hidden layer is connected to each layer in the output layer. The number of nodes in the output layer depends upon the number of values required as an output. Each node of the output

layer will therefore, calculate a single value from the values of the hidden layer, by mapping them back into the original range of the raw input values (Badiru & Cheung, 2002). Such a Neural Network topology is typical of one that is used for classification and prediction. Neural networks have been successfully investigated for a variety of purposes (Smith & Gupta, 2002):

- Predict future outcomes/ events, through models developed with training data.
- Classify unseen data into pre-defined groups based upon structures that have been observed whilst developing models.
- Cluster data into natural groups based upon similarity in the characteristics of the data.

Furthermore, Neural Networks have been successfully applied in a number of industries for multiple applications, such as predicting financial data, diagnosing medical conditions, identifying fraudulent transactions, identifying failure rates in engines, amongst others (Berry & Linoff, 2004). Evidently, Neural Networks have become extremely popular since their use has proven successful in variety of applications. Neural Networks, purpose a suitable approach for tackling BI problems, such as this investigation. Neural networks can be explored to interrogate the data and identify hidden patterns that can allude to the realisation of the objectives such as identifying the impact of sales promotions.

A.3 Association Rule Mining (ARM)

Frequent pattern are constantly appearing within datasets. Items that frequently appear together within data, such as bread and milk transactions are referred to as 'itemsets'. A frequent sequential pattern in behaviour or dependency of items is known as a subsequence, for example, the purchase of a digital camera, will precede the purchase of a memory card. (Han & Kamber, 2006; Mirkin, 1996). Association rule mining (ARM) otherwise known as affinity grouping, is a popular and well researched method for discovering interesting correlations such as these between attributes within a dataset (Pinto et al, 2001).

The increasing amount of data being collected and stored, and the escalating frequency with which companies are introducing loyalty cards combined with the greater number of online transactions, it is becoming of immense interest to organisations to be able to effectively utilise this data for a competitive advantage. ARM is a technique often utilised by retail organisations to gain such an advantage trough the investigation of customer behaviour, consequently ARM is often referred to as 'market basket analysis' when applied within the context of a transaction database. ARM enables all possible combinations of product groupings to be explored. This is of interest as it can uncover potentially interesting relationships between items that may not always be obvious. However a few input item sets can result in a large number of rules, of which some may be obvious or of no real interest. An example of this is that transactions consisting of Beer, may also be highly likely to contain Crisps, this may not be a novel discovery within itself, however the rule set may also uncover that customers that purchase Beer may also purchase nappies, since those with young families are likely to spend majority of their time at home. ARM can therefore aid in many business decision processes through the discovery of relationships (Hand et al, 2001). These relationships can be investigated when considering what items to place of offer, especially with regard to multi-buy offers, be this to increase sales of two items (itemsets), or increase the sales of an item that depends upon another for greater interest (subsequence), in addition to the purchasing habits that

must be considered when contemplating the most effective store/isle layout, or even the layout of printed media, be it advertising or a catalogue (Witten & Frank, 2005).

Association rules are implications of the form 'If X then Y' (X →Y) where 'X' and 'Y' are two disjoint subsets of all available items, therefore have no common items. Association rules are computed from the data and, unlike 'if-then' rules of logic, association rules are probabilistic in nature. 'X' is called the antecedent or LHS (left hand side) and 'Y' is called the consequent or RHS (right hand side). Association rules are able to find relationships within datasets without the restriction of a single dependent variable. Association rules can also have one or more output attributes unlike classification rules, which generally limit the consequent of a rule to a single output attribute. Furthermore, attributes that are antecedent or precondition for one rule may appear as a consequent of another rule, hence output attributes can serve as input attributes for additional rules (Roiger & Geatz, 2003).

Due to the large number of rules that can be generated, even by a single input attribute care must be taken when interpreting association rules. To ensure that only meaningful rules of interest are mined, many ARM algorithms employ 'confidence' and 'support' measures. Confidence and support measures ensure that the rules that are mined are able to satisfy constraints or measures of significance or interest. Every association rule has a level of confidence (or accuracy) and support (or coverage) attached to it. The confidence is the conditional probability attached to consequent of a rule being true in the event that the antecedent is known to be true, hence the number of correctly predicted instance as a proportion of the number of instance to which the rule is applicable. However, the confidence does not measure the percentage of instances within the dataset that contain all items that are listed in a particular association rule; this statistic is known as the support (Han & Kamber, 2006). Support is simply the number of instances within a dataset that include all items in the antecedent and consequent parts of the rule, usually expressed as a percentage. The confidence can therefore be thought of as the ratio of the number of instances that include all items in the consequent as well as the antecedent (the support) to the number of transactions that include all items in the antecedent. One more statistic of an association rule that is of interest is the 'lift'. Lift is the ratio of confidence to the expected confidence. Expected confidence is the number of instances that include the consequent divided by the total number of instances. Hence, lift is a parameter that provides information about the increase in probability of the consequent, given the antecedent (Berry & Linoff, 2004).

A.3.1 Apriori Algorithm

First proposed by Agrawal& Srikant (1994) the Apriori algorithm is designed to operate on databases containing transactions, thus providing a suitable ARM algorithm for this study. The algorithm generates an 'item set'. An item set is a combination of attributes and values that meet a pre-defined coverage requirement. Since only the attribute-value combinations that satisfy the coverage requirement are generated, the algorithm can analyse datasets in a reasonable amount of time. Apriori rule generation is a two-step process. Initially an item set is generated; the pseudo code for this process is illustrated in code-table A.1 (Lazcorreta et al, 2008).

1	$_k = 0$
2	**repeat**

3	$_k$++
4	**if** $_k > 1$ **then**
5	$\{C_k = \text{generate\_candidates}(V_{k-1})\}$
6	**else**
7	$\{C_1 = E\}$
8	**endif**
9	**for all** *0* **do**
10	$C_o = \{ c \mid c \sqsubseteq o \wedge c \in C_k \wedge o \in O\}$
11	**for all**
12	$C_k^i.count^{++} \mid C_o \in C_k$
13	**end for**
14	**end for**
15	$L_k = \{C_k \mid C_k^i.count \geq \text{minsup}\}$
16	**until** $V_k = \emptyset$
17	return $L = \bigcup_{k=o}^{n} L_k$

Code-table A.1: Apriori itemset generation.

The Aprioiri algorithm employs a 'level-wise' search, where 'k-item sets' are used to explore '(k-1) – item sets'. Initially '1-item sets' is discovered by scanning the dataset to accumulate the count for each item whilst selecting those items that satisfy the level of minimum support. The resulting set is denoted 'L_1'. This set is then used to discover 'L_2', which is the set of '2-item sets', this process will iterate until k-item sets have been discovered. The discovery of 'L_k' therefore, requires one full scan of the data set. Once an item set has been generated, the second step of the Apriori algorithm is to investigate the item set to create a set of association rules. It is this process of generating the association rules from the knowledge obtained from the item set that denotes the algorithms name (Ceglar & Roddick, 2006). The generated rules will provide information of interesting relationships between the attributes, where the relationship satisfies both a minimum support and minimum confidence. These rules can be explored to find the relationship between attributes such as types of meat within this study. These relationships can then be further investigated be it for decisions regarding which items should be cross-promoted due to a relationship, or decisions regarding store layout, thereby placing items that sell together in close proximity.

A.4 Intelligent Agent Decision Mechanism

Since a Multi-Agent system consists of a number on intelligent agents that communicate and perform limited tasks. It is the sum of these tasks that determines the effectiveness of the system. It is, therefore, imperative that the constituent agent be able to effectively communicate and make decisions. The FIPA architecture was discussed in Chapter 2, as an architecture upon which to base communication. However, the decision mechanism must also be investigated. Tasks can be specified to agent by writing a program that the agent can execute; this nevertheless, has the drawback of reducing the flexibility of the agent. Since the agent will, much like in an object-oriented program follow a predetermined algorithm to achieve its objectives. This reduces the agent's ability to deal with unforeseen circumstances. Consequently, this approach is not desirable if all features

that make a Multi-Agent systems novel are to be retained. It is preferential if an agent can be specified its function through indirect means, allowing the agent to decide upon what course of action would best satisfy the objective it has been created to achieve (Vidal, 2007). Decision Theory and Game Theory can be applied to understand the action of agents and provide models upon which to base the decision mechanism of agents, allowing agents to be indirectly provided with means upon which to act (Shoham & Leyton-Brown, 2008).

Decision Theory provides a general paradigm for designing agents that can operate in complex, uncertain environments and act rationally to maximise their preferences. Decision-theoretic models use precise mathematical formulism to define the properties of the agent's environment, agent's sensory capabilities, the extent to which the agent's actions alter the environment, in addition to the agent's goals and preferences (Parsons, 2005). In contrast, Game Theory consists of a number of models to aid the understanding situations in which decision-makers interact, especially environments such as Multi-Agent systems, where there are a large number of interacting entities (Osbourne, 2002; Bowling & Veloso, 2002). Hence, Game-theoretic adds to the decision-theoretic framework, the idea of multiple agents interacting within a common environment. Based upon the assumption that agents are rational and self-interested, game theory can be used to specify how agents, separately or jointly, can alter the environment they are in and the effect that, these alterations, bear upon their own preferences (Shoham, 2008).

An extension of 'Game Theory' and 'Decision Theory' can be investigated, to specify tasks to agents and equate the state the environment is in, through the introduction of a performance measure. By specifying a numeric value to represent the state that an environment is in, known as a utility function, an agent can be informed of how good the environment is, enabling it to autonomously function until the desired level of utility is achieved, i.e. the specified goal and task is represented by the function (Wooldridge, 2002):

$$u : E \rightarrow \mathbb{R} \tag{A.7}$$

This function (A.7) enables a real value to be assigned to every state that the environment can be in. A drawback to this approach is that values are only assigned to local states, as a result, it is difficult to specify to a long-term view, thus if an agent is required to operate independently over long periods of time then it is more suitable to assign a utility not to individual states but the run itself:

$$u : \mathcal{R} \rightarrow \mathbb{R} \tag{A.8}$$

An eminent example of this utility function can be found in 'The Tileworld' (Pollack, 1990); Wooldridge, 2002; Murata & Nakamura, 2006). The Tileworld provides an example of a dynamic environment within which agents are able to act, and alter. Based upon parameters that are set by the user, a domain is randomly generated. This randomly generated domain consistently varies over a period of time, due to the random appearance and disappearance of holes. It is the task of he agent to fill these holes, thereby directly interacting with a dynamic environment. If the utility function 'u', has some upper bound to the utility that can be assigned to state, then:

$$k \in \mathbb{R} \text{ such that for all } r \in \mathbb{R} \Rightarrow u(r) \leq k \tag{A.9}$$

By assigning upper bounds to the utility that can be assigned enables agents to reach an optimal state, thus maximise expected utility.

A.4.1 Anytime Algorithm

The concept of the utility function can be extended to operate in conjunction with an 'Anytime Algorithm'. An Anytime Algorithm is an algorithm which can return a result at anytime, thus not requiring it to be operated for a predetermined duration of time, or execute a complete algorithm, prior to a result being returned. However the greater the period of time that an Anytime Algorithm is permitted to calculate a result, the more optimal the result will be (Hansen & Zilberstein, 2001). In the event that the algorithm is terminated prematurely it will return the best result found, it should be noted that this may not be the optimum result but an intelligent Anytime Algorithm can be programmed to re-commence from the best previously found result, unlike a conventional Anytime Algorithm that will commence from a random point (Wang et al, 2001). A 'Co-ordinate Ascent Anytime Algorithm' (CAAA) initiates at a random choice for a combined action, and then loops over all agents, with each agent optimising their own actions while the actions of all other agents stay the same. The loop will continue until no improvement can be made, at which point a new random starting point is selected and the process repeated. The CAAA can return the highest value found at any time (Hansen & Zilberstein, 2001):

1	**define:** scope(*p*) is the set of all agents that define an action in *p;*
2	**define:** a_i = the action of an agent i;
3	**define:** A_i = Dom (a_i)
4	**define:** $a_{\_i}$ = the action of all agents but agent i;
5	**define:** a* = the best combined action found so far;
6	**define:** *p*(a) = the payoff rule *p* gives, in the event of combined action a;
7	
8	P ← all value rules
9	**loop**
10	**for** *i* ∈ agents **do**
11	$a_i \leftarrow$ random (A_i)
12	**end for**
13	**do**
14	*i* ← next (agent)
15	$p \leftarrow \{p_j \in P \mid i \in \text{scope}(p_j)\}$
16	$a^*_i \leftarrow \arg\max_{ai \in Ai} (\sum p_j (a^*_{\_i} \cup a_i))$
17	**while** no agent has changed action since last time *i* was checked
18	**end loop**
19	

Code-table A.2: Pseudo-code for a Co-ordinate Ascent Anytime Algorithm.

The CAAA in code-table 8.1 is looking for the 'Nash-Equilibria' (NE). In this instance all value rules are shared, however, the NE technically requires all agents to have their own payoff rules. Payoff rules enable the system to differentiate between two agents when deciding whom to give control (Osbourne , 2002). Due to their being no guarantee that the NE found is optimal, a CAAA will infinitely calculate. This CAAA could be modified to commence intelligently from a predetermined point (the last best value) rather than randomly, the agents could also be ordered to ensure a uniform pattern for the search rather than randomly queuing the agents for the next operator. The 'argmax' function also returns the same result in the event that that two actions result in the same

payoff. However, a preset preference toward an action in the event of a conflict would suffice to resolve any conflicts between agents.

A.5 Evaluation of Advanced Analytical Platforms

There are various advanced analytical analysis platforms, which are available from a number of vendors. These are both commercial and open-source. In order to discover suitable BI platforms for the objectives of this study, a number of these packages were investigated as illustrated in table A-1.

Vendor	Solution name	Commercial or open-source	Key features
Actuate	Actuate iServer Enterprise	Commercial	Actuate is a market-leader in Enterprise reporting applications. Actuate iServer Enterprise is a scalable, server for generating, managing and securely delivering reporting and analytic content.
Actuate	BIRT	Open-source	Eclipse-based reporting system that integrates with Java/J2EE applications to produce reports.
Business Objects	Business Objects XI Release 2	Commercial	A combination of Crystal enterprise with Business Objects' semantic layer technology. A suite of BI products is provided. These suites deliver cross-enterprise, data access and management, in addition to information delivery services.
IBM	DB2 Data Warehouse Edition (DWE) v.9	Commercial	Can be used to build complete data warehousing solutions that include scalable relational databases, data access capabilities, business intelligence analytics, and front-end analysis tools. Furthermore, DWE integrates core components for warehouse administration, data mining, OLAP and inline analytics and reporting.
IBM Cognos	IBM Cognos 8 BI	Commercial	Delivers a complete range of BI capabilities on a single, service-oriented architecture (SOA). Facilitates the authoring, sharing, and use of reports that can draw upon data from all organisational sources.
InfoAcumen Corporation	iData Analyzer	Commercial	A Microsoft Excel add-on iData Analyser utilises artificial intelligence technology, such as, neural network building tool coupled with proprietary data mining technology to unveil hidden knowledge from any spreadsheet of source data. iData Analyzer can not only classify in domains containing both categorical and numerical data but also help to determine the best attributes for inclusion within a data warehouse. Using a data structure that allows domain instances to be retained once they are processed. Due to this, valuable knowledge contained within the instances is not lost once data mining is complete.
Information Builders	WebFOCUS 7.1	Commercial	Deployable as a platform for BI applications or as a BI technology stack WebFOCUS facilitates data handling through an extensive array of data integration technologies including data, application and technology adapters.
JasperSoft	Business Intelligence Suite	Open-Source	A modular suite of BI tools that provide support for data integration, reporting and analysis. Consisting of: Jasper Analysis: provides a powerful web-based online analytical processing (OLAP) capability to securely analyse data. JasperAnalysis leverages standard RDBMS technologies and can be deployed against existing application databases, operational data stores, as well as data marts and data warehouses. JasperETL: developed through a technology partnership with Talend. It is a complete and ready-to-run data integration platform for data integration, transformation, movement, cleansing and enrichment. JasperReports: an embedded Java reporting library. It provides accelerated report development, high-performance and scalability. JasperReports handles all manners of complexity, including multiple data sources, sub-reports, and crosstab reports. Unlike most reporting solutions, the built-in virtualisation capability enables output of arbitrarily large reports, limited only by available disk storage resources. JasperReports
KXEN	IOLAP	Commercial	Provides support for executive scorecards, planning and budgeting, and operational performance reporting and analysis. This is achieved through the discovery of Key Performance Indicators (KPIs) that are predictive and quantifiable, to produce reliable forecasts. Furthermore, IOLAP facilitates with the organisation of information to focus upon the reporting of the metrics and KPIs that have the most direct impact upon the organisation.
The MathWorks	MATLAB	Commercial	A high-level language and interactive environment that enables you to perform computationally intensive tasks faster than with traditional programming languages such as C, C++, and Fortran. MATLAB supports the entire data analysis process, from acquiring data from external devices and databases, through pre-processing, visualization, and numerical analysis. All the graphics features that are required to visualise engineering and scientific data are available in MATLAB. Furthermore

			MATLAB can be integrated with other languages and applications or in contrast MATLAB algorithms and applications can be deployed as stand-alone programs or software modules.
Microsoft	Business Intelligence Development Studio	Commercial	Microsoft Visual Studio 2008 with additional project types that is specific to SQL Server. Business Intelligence Development Studio is the primary environment for the development of business solutions that include Analysis Services, Integration Services, and Reporting Services projects. Each project type supplies templates for creating the objects required for business intelligence solutions, and provides a variety of designers, tools, and wizards to work with the objects.
Microsoft	Excel & Microsoft SQL Server Data Mining Add-Ins	Commercial	Microsoft Excel forms a part of the Microsoft Office Package providing spreadsheet based features along with analytical functionality to users. Microsoft Excel can be explored in conjunction with Microsoft SQL Server Data Mining Add-Ins for realising BI objectives. Microsoft SQL Server Data Mining Add-Ins for Office 2007 is a set of easy to use data mining capabilities that enable predictive analysis at every desktop. Being able to harness the highly sophisticated data mining algorithms of Microsoft SQL Server 2005 Analysis Services within the Microsoft Office, business users can easily gain valuable insight into complex sets of data with just a few mouse clicks. The Data Mining Add-Ins for Office 2007 facilitates end users to perform advanced analysis directly from within Microsoft Excel.
Microsoft	Excel Pivot Tables	Commercial	One of the most powerful features in Microsoft Excel and provided free. Pivot tables are a way to extract data from a long list of information, and present it in a readable form and facilitate the analysis of data. Enabling users to create multidimensional data views by dragging and dropping column headings to move data around, Pivot tables are especially well-suited to the task of taking enormous amounts of data and summarising that data into useful reports.
Microsoft	SQL Server Enterprise Edition	Commercial	Provides a complete end-to-end BI platform. The additional features over previous incarnations of SQL Server form the core of the BI solution. Contained within the database enhancements these are bundled free with the database licence. The Reporting Services (in its second release) is provided in two forms Developer and Builder. Developer provides tools to create reports that are developed via technical developers. Whilst Builder is provides tools for business analysts to write reports. Furthermore new management tools simplify administration, whilst ETL tools have been improved and re-architected as Integration Services. New algorithms have been added to the data mining functionality. Furthermore, embedded within the database is a .NET Common Runtime Library to expand the options available to developers building BI applications. However, as a BI solution, SQL Server 2005 EE is lacking in end-user accessibility applications and limited business analytics. Microsoft has addressed these issues with Business Scorecard manager, which provides a score carding and dashboard framework. In addition to Office 2007, this has a number of BI elements that have been focused toward end-user accessibility.
Microsoft	SQL Server Management Studio	Commercial	An integrated environment for accessing, configuring, managing, administering, and developing all components of SQL Server. SQL Server Management Studio combines a broad group of graphical tools with a number of rich script editors to provide access to SQL Server to developers and administrators of all skill levels. SQL Server Management Studio combines the features of Enterprise Manager, Query Analyzer, and Analysis Manager, included in previous releases of SQL Server, into a single environment. In addition, SQL Server Management Studio works with all components of SQL Server such as Reporting Services, Integration Services, and SQL Server Compact 3.5 SP1. Developers are consequently provided with a familiar structure, whilst database administrators get a single comprehensive utility that combines easy-to-use graphical tools with rich scripting capabilities.
MicroStrategy	MicroStrategy platform	Commercial	The MicroStrategy platform architecture is capable of delivering all of the functionality requirements for BI, such as scorecards and dashboards, enterprise reporting, OLAP analysis, advanced and predictive analysis, alerts and proactive notification within a single architecture.
Miner3D Inc.	Miner3D	Commercial	Data visualization software for multidimensional exploratory data analysis. Integrated model builders automatically create charts for current data loaded in the system. Miner3D includes support for creating several basic 3D and 2D charts that can be customised and converted to a totally different graph type.
Oracle	Hyperion System 9 BI +	Commercial	Provides all types of reporting and analysis, ad-hoc query and data warehouse reporting functionalities. Furthermore it has a variety of output options such as PDF, HTML and Microsoft Office integrating a sophisticated calculation engine.
Oracle	Oracle Data	Commercial	An addition for Oracle Database 11g Enterprise Edition. It facilitates the rapid

	Mining (ODM)		development and deployment of applications that deliver predictive analytics and new insights using ODM's SQL and Java API's that automatically mine Oracle data. Since the data, models and results remain in the Oracle Database, data movement is eliminated, security is maximised and information latency is minimised. Oracle Data Mining models can be included in SQL queries and embedded in applications to offer improved business intelligence.
Oracle	Oracle Data Miner	Commercial	The graphical user interface for Oracle Data Mining (Release 10.1 (10g) and above) that helps data analysts to mine their Oracle data in order to find valuable hidden information, patterns, and new insights. Easy-to-use wizards that guide them through the data preparation, data mining, model evaluation, and model scoring process. As the data analyst transforms the data, builds models, and interprets results, Oracle Data Miner can automatically generate code needed to transform the data mining steps into an integrated data mining/BI application.
Pentaho	BI Suite Version 2	Commercial	A complete business intelligence platform that includes reporting, analysis (OLAP), dashboards, data mining and data integration (ETL). Can be used as a full suite or as individual components that are accessible via web services.
Pentaho	Mondrian	Open source	An OLAP server written in Java. It enables users to interactively analyze very large datasets stored in SQL databases without writing SQL. The main functions provided by Mondrian are high performance, interactive analysis of large or small volumes of information from "Dimensional" views. Parsing of Multi-Dimensional eXpression (MDX) language into SQL in order to retrieve answers to dimensional queries. High-speed queries through the use of aggregate tables in the RDBMS and advanced calculations using the calculation expressions of the MDX language.
Salford Systems	CART	Commercial	Classification And Regression Trees (CART) is a robust, easy-to-use decision tree that automatically sifts large, complex databases, searching for and isolating significant patterns and relationships. Designed for both non-technical and technical business users, CART can quickly reveal important data relationships The most recent 2008 release, CART 6.0, includes modelling automation technology that dramatically accelerates the process of generating accurate and robust models for deployment in core business functions.
Sage	Sage 200 Business Intelligence	Commercial	Provides a unified and integrated view of all your data, with complete management dashboards and analysis. Furthermore, no technical knowledge or pre-requisites are required for Sage 200 Business Intelligence Sage 200 Business Intelligence provides an alternative to complex spreadsheet reporting with reports. In addition the features of Microsoft Excel, such as ease of formatting, calculations and macros can all be used to enhance these reports which are not only easy to create and automatically refresh when the data is updated. Sage BI can also be integrated with Sage Line 500 and Sage 1000,
Sage	Intelligent Reporter (IR)	Commercial	Gathers data from Sage Line 50 into the Microsoft Excel environment. It facilitates users to analyse data easily, quickly and efficiently, from many different angles. Providing an insight into how the business is performing and enabling them to make informed and reliable business decisions.
SAP	NetWeaver BI	Commercial	Is an integrated BI solution that has been developed upon the NetWeaver integration platform with SAP's Data Warehouse and Business Warehouse (BW) at its core. Designed for use in conjunction with other SAP applications or as a stand-alone application, NetWeaver BI is available with vertical business domain templates (called Business Content). These models and templates provide a means through which the implementation can be templates
SAS	Enterprise Miner	Commercial	Supports the entire data mining process with a broad set of tools. Consequently, regardless of data mining preference or skill level, Enterprise Miner addresses complex problems facilitating users to transform raw data to accurate, business-driven data mining models.
Spago	SpagoBI 2.0	Open-source	Is a professional Business Intelligence suite entirely developed and released according to the best Free Open Source Software community's practices. SpagoBI allows the end-user to compose the most suitable BI platform, also mixing open source and proprietary products providing results quickly with a smooth insertion within pre-existing environments.
SPSS Inc.	Clementine	Commercial	Facilitate the access, preparation and integration of structured data in addition to text, Web, and survey data. This can be used to build and validate models, using advanced statistical and machine-learning techniques. This enables the deployment of analysis and predictive models on a scheduled basis or in real time.
Stanford University	Cluster 3.0	Open-source	Implements the most commonly used clustering methods for gene expression data analysis. The clustering methods can be used in several ways. Cluster 3.0 provides a Graphical User Interface to access to the clustering routines. It is available for Windows, Mac OS X, and Linux/Unix. Python users can access the clustering routines

			by using Pycluster, which is an extension module to Python. Users that want to make use of the clustering algorithms in their own C, C++, or Fortran programs have the option to download the source code of the C Clustering Library.
University of Waikato	Weka	Open-source	A popular suite of machine learning software written in Java, developed at the University of Waikato. WEKA is free software available under the GNU General Public License. The workbench contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality. Originally developed in C, Since 1997 Weka has been developed in Java, thereby providing a portable collection of data preprocessing and modelling techniques capable on running on almost any platform Weka supports several standard data mining tasks, more specifically, data pre-processing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Furthermore, Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka. Another important area that is currently not covered by the algorithms included in the Weka distribution is sequence modelling. However projects such as that by Patro[10] have addressed this issue.
Rapid-I / University of Dortmund	RapidMiner	Open-source	Previously known as YALE (Yet Another Learning Environment) is an environment for machine learning and data mining experiments. Developed completely in Java, to ensure portability and completely integrating the Weka workbench. Yale facilitates experiments to be made up of a large number of arbitrarily nestable operators, described in XML files which can easily be created with RapidMiner's graphical user interface. The modular operator concept of RapidMiner allows the design of complex nested operator chains for a huge number of learning problems in a very fast and efficient way (rapid prototyping).

Table A-1: Platforms for advanced analytical analysis.

A.6 Evaluation of Intelligent Agent Platforms

There are various Intelligent Agent platforms, which are available from a number of vendors. These are both commercial and open-source. In order to discover suitable BI platforms for the objectives of this study, a number of these packages were investigated as illustrated in table A-2.

Vendor	Solution name	Commercial or open-source	Key features
AOS	JACK (Agent Development Environment)	Commercial	JACK is a mature, cross-platform environment for building, running and integrating commercial-grade multi-agent systems. Consists of a collection of autonomous agents that take input from the environment and communicate with other agents. This provides system builders with a very powerful form of encapsulation. It is built on a sound, logical foundation: BDI (Beliefs/Desires/Intentions) Hence, each agent is defined in terms of its goals, knowledge and social capability, and is then left to perform its function autonomously within the environment it is embedded in.
BT	ZEUS	Open-source	An agent building toolkit, which provides frameworks for development of collaborative agent systems. Zeus systems are constructed using large, course-grained agents, whom accentuate autonomy or co-operation. Zeus provides visualisation, scripting and monitoring tools as well as a simple agent building environment. Zeus is an open source agent development tool kit that was created as part of the Midas and Agentcities research projects at BT in the late 1990's and early 2000's. Zeus has won a BCS Gold medal for its technical innovation, and has been successfully used in many experimental projects within BT and in the rest of the world. Zeus is written in Java and features facilities to implement BDI style (not quite BDI) agents, agents with reactive rule bases, agents with intelligent message handling functionality and DAML-S service descriptions. The reasoners in Zeus include a Rete style rule engine which uses CLIPS type rule definitions that can be extended and plugged in Java, a simple distributed planner and graph style agent rationality (you can build graphs to describe the problem solving behaviour of the agent that you want). A simple ontology is used to integrate the concepts in the agent between the different sub systems.
DARPA	Cougaar	Open-Source	A Java-based architecture for the construction of large-scale distributed agent-

[10] http://davis.wpi.edu/~xmdv/weka/

research project			based applications. It is a product of two consecutive, multi-year DARPA research programs into large-scale agent systems spanning eight years of effort. The first program conclusively demonstrated the feasibility of using advanced agent-based technology to conduct rapid, large scale, distributed logistics planning and re-planning. The second program is developing information technologies to enhance the survivability of these distributed agent-based systems operating in extremely chaotic environments. The resultant architecture, Cougaar, provides developers with a framework to implement large-scale distributed agent applications with minimal consideration for the underlying architecture and infrastructure.
GMD FOKUS and IKV++	Grasshopper	Open-source	Grasshopper has been designed in conformance with the first mobile agent industry standard, namely the Object Management Group's Mobile Agent System Interoperability Facility.In addition, the latest Grasshopper version is also compliant with the specifications of the Foundation for Intelligent Physical Agents. Grasshopper realises a Distributed Agent Environment (DAE). The DAE is composed of regions, places, agencies and different types of agents. A place provides a logical grouping of functionality inside of an agency. The region concept facilitates the management of the distributed components (agencies, places, and agents) in the Grasshopper environment. Agencies as well as their places can be associated with a specific region by registering them within the accompanying region registry. All agents that are currently hosted by those agencies will also be automatically registered by the region registry. If an agent moves to another location, the corresponding registry information is automatically updated.
IBM / Alpha Works Services	AGENT BUILDING AND LEARNING ENVIRONMENT (ABLE)	Commercial	A Java framework, component library, and productivity tool kit for building intelligent agents using machine learning and reasoning. The ABLE framework provides a set of Java interfaces and base classes used to build a library of JavaBeans called AbleBeans. The library includes AbleBeans for reading and writing text and database data, for data transformation and scaling, for rule-based inference using Boolean and fuzzy logic, and for machine learning techniques such as neural networks, Bayesian classifiers, and decision trees. Developers can extend the provided AbleBeans or implement their own custom algorithms. Rule sets created using the ABLE Rule Language can be used by any of the provided inference engines, which range from simple if-then scripting to light-weight inference to heavy-weight AI algorithms using pattern matching and unification. Java objects can be created and manipulated using ABLE rules. User-defined functions can be invoked from rules to enable external data to be read and actions to be invoked.
IBM	Aglets Software Development Kit (ASDK)	Open-source	Aglets is a Java mobile agent platform and library that eases the development of agent based applications. An aglet is a Java agent able to autonomously and spontaneously move from one host to another. Aglets allow the freedom to meander servers running host software and travel along a predefined pathway, which can be suspended, re-routed and resumed.
Magent-A	Intelligent Enterprise (I-Enterprise) suite	Commercial	i-Enterprise is MagentAs' fully integrated system designed to compete directly with products offering comprehensive support for all enterprise-wide activities, such as ERP. All systems agents within i-Enterprise are interconnected on the 'peer-to-peer' principle. Hence, e-Commerce Systems communicate with the Real-Time Logistics System to resolve delivery problems and the Logistics Systems agree with the e-Commerce System to issue tender to employees on the internal corporate market, or to assemble a team for a new project. All these capabilities result in synergetic increases in the intellectual capacity, flexibility and efficiency of i-Enterprise and the enterprise itself. All constituent components of i-Enterprise System can be delivered separately and integrated with existing software
Nortel Networks	FIPA-OS	Open-source	Developed at Nortel Networks, an open-agent platform that supports communication using the FIPA (Foundation for Intelligent Physical Agents) communication language (Posland et al, 1999). Distributed under an open-source licence and designed as a set of coupled parts to allow the open-source development community to make advances, and implement FIPA specifications using Java.
ObjectSpace, Inc.	VOYAGER SYSTEM	Open-source	An agent-enhanced Object Request Broker (ORB) written in Java by ObjectSpace, Inc. Voyager agents are autonomous and mobile but not intelligent.
Perdue University	Bond	Open-source	Bond is a Java based, FIPA compliant agent framework. Providing, Multi-plane state machine agent model, component-based architecture (strategies and planes) and a Python based agent description language (Blueprint). Bond places a strong emphasis upon introspection, visual modelling and software verification. In addition to dynamic agent behaviour (agent assembly, mobility, surgery, trimming, and lazy loading of strategies).
Reticular Systems Inc	AGENTBUILDER		Developed by Reticular Systems Inc. using a high-level, agent-orientated programming language, providing a suite of graphical programming tools used to configure agents and specify behaviour. This toolkit allows intelligent agents to

			be constructed in a Java environment.
Sandia National Laboratories.	JAVA EXPERT SYSTEM SHELL (JESS)	Open-source	Java implementation of CLIPS (C Language Integration Production Language System) expert system rule-base environment, developed at Sandia National Laboratories. While not strictly an agent environment, provides a rule-based inference support in Java.
Stanford University	JAVA AGENT TEMPLATE LITE (JATLite)	Open-source	Developed at Stanford University, a set of lightweight Java packages, providing a layered architecture for building Multi-agent systems.
Swarm Development Group	SWARM	Open-source	Swarm is a collection of software libraries which provide support for simulation programming. Swarm Provides Various Tools: The Swarm libraries provide a number of convenient pieces of code that will facilitate the design of an agent-based model. These tools facilitate the management of memory, the maintenance of lists, scheduling of actions, and many other chores. Users build simulations by incorporating Swarm objects in their own programs. Users are encouraged to study a number of tutorial examples in order to make full use of the Swarm libraries and the strategy of modeling that inspires them.
Telecom Italia Lab	The Jade Framework	Open-source	Jade (Java Agent DEvelopment) Framework is a software framework fully implemented in Java language. It simplifies the implementation of multi-agent systems through a middle-ware that claims to comply with the FIPA specifications and through a set of tools that supports the debugging and deployment phase. The full FIPA communication model has been implemented within Jade and its components have been fully integrated: interaction protocols, envelope, ACL, content languages, encoding schemes, ontologies and, finally, transport protocols. The transport mechanism, in particular, is like a chameleon because it adapts to each situation, by transparently choosing the best available protocol. Java RMI, event-notification, HTTP and IIOP are currently used, but more protocols can be easily added. Most of the interaction protocols defined by FIPA are already available and can be instantiated after defining the application-dependent behaviour of each state of the protocol. SL and agent management ontology have been implemented already, as well as the support for user-defined content languages and ontologies that can be implemented, registered with agents, and automatically used by the framework. Jade is being used by a number of companies and academic groups, both members and non-members of FIPA, such as BT, CNET, NHK, Imperial College, IRST, KPN, University of Helsinky, INRIA, ATOS and many others.
Toshiba	Bee-gent	Open-source	Bee-gent (Bonding and Encapsulation Enhancement aGENT) is a new type of development framework in that it is a 100% pure agent system. As opposed to other systems which make only some use of agents, Bee-gent completely "Agentifies" the communication that takes place between software applications. The applications become agents, and all messages are carried by agents. Thus, Bee-gent allows developers to build flexible open distributed systems that make optimal use of existing applications. In simple terms, create wrappers for those legacy applications.
Tryllian Systems Inc.	GOSSIP		Gossip implements learning technology to profile user preferences, which shall perform, automated actions upon the users behalf.
TuCSoN	TuCSoN	Open-source	TuCSoN exploits a notion of local tuple-based interaction space, called tuple centre, which is a tuple space enhanced with the notion of behaviour specification. By programming its behaviour in response to communication events, a tuple centre can embody coordination laws. Several issues critical to Internet applications, such as heterogeneity and dynamicity of the Internet nodes, can then be charged upon tuple centres, and transparently to agents, which can then be designed independently of the different node architectures, according to a straightforward interaction protocol.
Whitestein Technologies	The Living Systems® Technology Platform	Commercial	The Living Systems Technology Platform is based on Java and Eclipse technology. Consisting of three major technology suites (products): LS/TS (Living Systems Technology Suite), an industry-grade, Java-based development and run-time environment for autonomic, self-managing systems and applications. LS/ABPM (Living Systems Autonomic Business Process Management), an extension of LS/TS towards direct support for autonomic business process management. LS/ASCO (Living Systems Autonomic Service Composition and Orchestration), a further LS/TS extension, currently in development, to cover autonomic service composition and orchestration in the context of Service Oriented Architectures (SOA).

Table A-2: Platforms for Intelligent Agent analysis.

Appendix B: Case Study 1

Direct Marketing: The Insurance Company

In order to investigate the advantages and disadvantages of the proposed framework KDDS-BI, it will be explored with the aid of case studies. This case study will examine the performance of KDDS-BI when investigated for the purpose of 'direct marketing'. Direct marketing has been selected since it represents a challenging area of marketing. In this age of global telecommunications, which enables even small scale retailer to compete globally, ensuring that consumers are aware of an organisations goods and services is of paramount importance.

B.1 Marketing

The evolution of telecommunications and global markets has facilitated organisations to access previously untapped markets. However for these markets to be effectively exploited, it is imperative that an organisation be capable of ensuring that these potential customers are made aware of the products and services that are available. As a result, organisations invest vast amounts of resources for marketing objectives. Marketing is a term that is often ill-defined and frequently used as an 'umbrella term' to group together a number of concepts and theories. The Chartered Institute of Marketing defines marketing as:

"Marketing is the management process responsible for identifying, anticipating and satisfying customer requirements profitably."

The term marketing is often applied synonymously with selling or advertising. Marketing extends further then merely the advertising or selling of goods and services, rather a complete marketing strategy which further considers the assessment of customer needs and requirements through market research, in conjunction with the development of a product, its pricing promotion and distribution (Dibb et al, 2005). If an organisation is to compete and be profitable, it is imperative that its goods and services are marketed effectively. However, effective marketing does not function in isolation. In contrast, marketing is an integrated process which must be observed throughout all operational activity within a business. During the early period of industrialisation, output could not only maintain a steady pace with demand, yet production exceeded demand. Consequently, the organisations were considered to operate within a 'sellers market'. Thus, firms were 'production oriented', permitting them to produce goods that they deemed necessary. Since the Second World War, the advancements in technology and increased competition have resulted in organisations being required to adopt a 'marketing orientated' approach to business processes. Consequently, organisations must currently compete in a 'buyers market'. Consumers are therefore, in a position from which they are able to select the goods and services they wish to acquire, from a variety of vendors. This in turn, compels organisations to ensure they can not only meet the needs of their customers, yet in addition, penetrate the market more effectively than their competitors. This has become further crucial given the progressive and integrated global marketplace and increased online transactions. This global marketplace permits consumers to purchase goods and services from not only a number of vendors, but vendors that may be located throughout the world. Accordingly, organisations must be able to identify and assess the needs and requirements of consumers in precise detail, not only accurately but also more promptly than their competitors. Consequently, paramount to an effective marketing strategy, is the identification and analysis of the constantly changing requirements of customers, whilst identifying new potential customers by anticipating the future needs of customers. An effective marketing strategy must further encompass an intimate understanding of general market trends and developments that may influence the views of customers and the activities of organisations operating within a particular market area. However, this marketing must be targeted or directed at the correct consumer to ensure that maximum return can be reaped form the marketing efforts of an organisation.

B.1.1 Direct Marketing

In the current digital global marketplace, an organisation's marketing strategy must be aimed at a variety of consumers from various backgrounds, often located in a variety of countries. Consequently, given an organisations ability to collect a variety of data not only on past and current customers, but also those market segments that can contain potential customers, conventional 'mass' marketing is becoming increasingly inefficient. Where mass marketing targets broad markets with homogeneous messages and offers distributed through intermediaries. Direct marketing in contrast, describes a growing trend that refers to a more targeted and narrowly defined marketing strategy aimed at specific customers or market segments.

Direct marketing is conducted via direct communication with specifically selected individuals for an immediate response. Furthermore, direct marketing facilitates an organisations ability to enforce stronger more robust customer relationships and augment customer loyalty. Direct marketing can be selected as a primary marketing strategy or even as a supplementary marketing strategy implemented in unison with more conventional techniques permitting deeper market saturation and understanding (Kotler et al, 2005). Direct marketing is based upon theories of income distribution initially proposed by Vilfreado Pareto in 1897. Figure B.1 depicts Pareto's principle that can be investigated as a model for the theory of direct marketing. Despite not always adhering strictly to the 80/20 percentages, Pareto's principle can be a very effective marketing tool, and underpins the theory of direct marketing (McCorkell, 1999).

Figure B.1: Pareto's principle.

Pareto's principle underpins direct marketing theory, however, the origins of direct marketing in practice can be found in early mail order companies. Prior to the widespread integration of computer science, information systems and business practices, mail order companies were collecting information and categorising these customers into current customers, potential customers and lapsed customers. Further information could then be recorded on these customers, permitting a company to discover the value and frequency of past order in addition to other behaviour indicators. Thus, rather then classifying these customers with a broad group such as age or region, the information could be analysed to provide a more accurate profile of the potential future behaviour of a particular individual (McCorkell, 1999). The expansion of direct marketing from a mail order strategy to a mainstream marketing strategy took place during the 1970s; American Express was amongst the initial companies to recognise the potential of direct marketing and pioneered this expansion outside of the mail order industry. In Europe, direct marketing did not gain widespread recognition until the 1980s. As in the USA,

according to research untaken by the Henley Centre[11], it was the financial services companies that were the first to exploit account records for the purpose of marketing. By the early 1990s, the charity sector and companies such as BT had emerged as highly accomplished practitioners. Furthermore, the late 1990s and early 2000s have witnessed the advent of 'database-driven marketing'. The integration of database technology within the business process model has increased the efficiency of developing customer profiles. The increased use of database technology along with more flexible technology, increasingly well-educated managers has resulted in the growth of direct marketing and its proliferation into a number of industries (Tapp, 2005). However, if direct marketing is to reach its full potential then it is imperative that managers be able to as accurately as possible identify potential and repeat customers, in addition to the most critical customer base within marketing, the customers that will only repeat custom if marketed correctly.

Thus, to identify customers that may otherwise go undetected, new and novel technologies must be investigated. Since direct marketing is a strategy that is entwined with database technology; it is one that can be enforced and improved through BI strategies. BI strategies will permit the discovery of trends and patterns that cannot be discovered through conventional database techniques, thus may not be evident without investigation through BI. If organisations wish to identify existing, repeat and potential customers, in addition to those that may otherwise be prone to being 'poached' by competitors, the customer groups must be marketed in a way that identifies there individual needs and requirements.

B.2 KDDS-BI Case Study: The Insurance Company

Within Direct marketing, often a data set is not collected for a specific reason. Instead, organisation will record various details and preferences of customers routinely, as a part of their daily organisational activities. This has become more viable given the falling cost of data storage and processing power. In such circumstances KDDS-BI can be investigated as a framework to provide structure to the process of interrogating the data set. In order to ensure the practice of direct marketing can be more accurately focused through BI techniques. BI techniques must be applied in a systematic manner that enables decision makers within an organisation to discover, identify and exploit opportunities. Consequently, KDDS-BI can be investigated to address the requirements of direct marketing investigations with the application of BI techniques in a systematic manner.

B.2.1 Data Investigation

In order to demonstrate the structure that can be provided to the interrogation of a data set for decision support through BI, KDDS-BI can be applied to structure the investigation. A dataset will be analysed with a view to discovering those individuals that are likely to purchase insurance policies for caravans. If individuals that are likely to purchase these policies can be correctly identified then it can greatly focus the capabilities of direct marketing strategies, such as 'direct mailing'. Direct mailing refers to the marketing strategy of distributing advertising material to the postal address of a (usually large) number of individuals. Direct mail is often pejoratively referred to as 'junk mail' and discarded by some of recipients, as it is often of no interest to them.

[11] http://www.hchlv.com: Accessed September, 2008.

Advancements within digital communication have mirrored this, through direct e-mails, are pejoratively referred to as, 'spam mail'. However, if direct mailing campaigns could be restricted to the individuals who have a high likelihood of purchasing the goods or services advertised the effectiveness of direct mailing will provide a significantly higher return on investment. Furthermore, as a result the cost to organisations can be reduced as an effective targeted direct mailing will result in significantly less individuals targeted that will discard the advertising campaign.

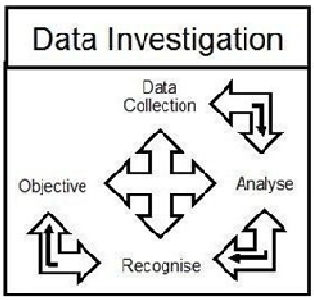


Figure B.2: Data Investigation stage of KDDS-BI.

The initial 'Data Investigation' phase of KDDS-BI involves exploration of the dataset. Once the data set has been obtained it can be further investigated. The primary objective of this phase is to obtain the data. Once obtained, the data set can be analysed to discover the opportunities that are available through the application of BI strategies. This will facilitate the recognition of objectives that can be further investigated. Since, data sets are often collected for purposes other then the investigation of BI or may be historical data that has been compiled by an organisation, it can be more robust to discover objectives direct from the data, as a primary objective of BI is to discover opportunities that may not be instantly evident. In order to study the performance of KDDS-BI for enhancing marketing decisions a data set detailing information about customers consisting of 86 variables and includes product usage data and socio-demographic data derived from area codes can be investigated. This dataset is owned and supplied by the Dutch data mining company Sentient Machine Research[12] (Hettich & Bay, 1999; Putten & Somerson. 2000), and is based on real world business data, which this research will investigate the data within a specific framework for a variety of objectives. Furthermore, the data set has been selected provides a benchmark through which the performance of KDDS-BI as a framework for BI investigations can be investigated. In addition, the data set contains noisy, correlated, redundant and high dimensional data, in addition to a weak relationship between the input and target variables. Hence, the data is representative of the type of data that is stored by many organisations.

The data set consists of 9822 records in a tab delimited format. Each record consists of 86 attributes, containing socio-demographic data and product ownership. The socio-demographic data is derived from area codes. All customers living in areas with the same area code have the same socio-demographic attributes. Table B.1 describes the attributes, labels and description. The data set provides information for a number of distinct customers, through details ranging from socio-demographic attributes, to other insurance policies the individual

[12] http://www.smr.nl/

may already possess. Although, the attributes in addition to the values for a number of the attributes are represented numerically, they are in fact nominal, and consequently serve as labels. Hence, within the data set, information detailing individuals is provided; these details can be further investigated to profile customers and those likely to buy a policy whilst describing the reasons/attributes which result in these customer buying policies. In addition to profiling, those customers that are potentially interested in purchasing a caravan insurance policy can be predicted. These objectives will facilitate the insurance company to directly target customers who are likely to buy policies, thereby ensuring that the provided profiles enable marketing campaigns to be directed toward high value customer profiles.

1	Customer subtype	1-41	1 = High income, expensive child … 41 = Mixed rurals	
2	Number of houses	1-10	N/A	
3	Average size of household	1-6	N/A	
4	Average age	1-6	1 = 20-30 years 2 = 30-40 years 3 = 40-50 years	4 = 50-60 years 5 = 60-70 years 6 = 70-80 years
5	Customer main type	1-10	1 = Successful hedonists 2 = Driven growers 3 = Average family 4 = Career loners 5 = Living well	6 = Cruising seniors 7 = Retired and religious 8 = Family with grown ups 9 = Conservative families 10 = Farmers
6-43	Other socio-demographic attributes	0-9	0 = 0% 1 = 1-10% 2 = 11-23% 3 = 24-36% 4 = 37-49% (% = proportion of area code covered by the attribute)	5 = 50-62% 6 = 63-75% 7 = 76-88% 8 = 89-99% 9 = 100%
44-64	Contribution level for policy	0-9	0 = *f* 0 1 = *f* 1-49 2 = *f* 50-99 3 = *f* 100-199 4 = *f* 200-499 (*f* is the symbol of the old Dutch currency, now it's euro)	5 = *f* 500-999 6 = *f* 1000-4999 7 = *f* 5000-9999 8 = *f* 10,000-19,999 9 = *f* 20,000+
65-85	Number of policies	1-12	N/A	
86	Caravan policy	0-1	0 = No	1 = Yes

Table B.1 Attributes and value descriptions.

Therefore, the objective for the application of KDDS-BI is to facilitate decision makers in identifying which customers have caravan insurance and the attributes that differentiate these customers from others, furthermore, the customers that are likely to buy policies in the future must also be identified. This will ensure that future marketing campaigns not only target the correct individuals, but the nature and details within the campaign are tailored to the specific attributes of potential customers. As a result, the objectives can be identified as:

- Identify which customers are likely to purchase a caravan policy.
- Identify the attributes that distinguish these customers, so that they can be effectively targeted.

- Identify any attributes that identify consumers unlikely to purchase a policy.
- Ensure that the results provided are accurate and can, therefore, be proposed as scientifically valid results.

B.2.2 Data Modelling

Once the possible objectives have been established, the investigation can enter the second phase 'Data Modelling' (figure B.3). The primary deliverable of this phase will be the modelling of the solution with a view to discovering the technical requirement for a successful investigation thus the requirements of the objectives must be examined to discover the BI strategies that are available and those that will ensure the objectives can be successfully achieved.

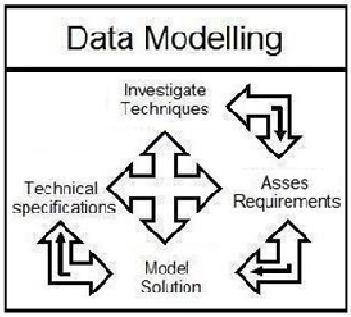


Figure B.3: Data Modelling stage of KDDS-BI.

The objectives identified during the 'Data Investigation' phase require key attributes to be identified. These key attribute will enable decision makers to determine the attributes that can be explored to identify which customers should be targeted with direct marketing campaigns for caravan insurance, to ensure that the campaign can be aimed at consumers likely to purchase a policy. Having assessed the objectives, it was determined that this is an exploration of data that will reap maximum benefit via analysis through advanced analytics, especially those that can employ intelligent analysis techniques. As a result, the initial stage of the 'Data Modelling' phase is to explore analytic techniques that will enable the investigation objectives to be fulfilled. Consequently, 'Classification models' can be investigated to predict those customer groups that are likely to purchase caravan insurance policies and are therefore a key demographic group that the company should target. Classification models demonstrate a 'supervised learning' approach to advanced analytics However, in order to further this research three types of classification techniques can be further investigated these are 'Bayesian Classifiers', 'Decision Trees' and 'Production (Classification) Rules'. The theory that underpins these 'supervised learning' techniques has been discussed in the Literature Review (Chapter 2) of this study. However, for each of these techniques there are a number of algorithms that can be investigated. Furthermore, in contrast to conventional methods of investigating a particular classifier, a combination of these classifiers can be investigated. Investigating a combination of multiple classification models will provide more robust and detailed information, upon which business decisions can be concluded. With careful deliberation of the objectives of this research, it was determined that, Bayesian classifiers, Decision Trees and Rule-based classification provide the most suitable means through which the investigation of direct marketing upon the caravan insurance dataset can most effectively be explored (further details of these techniques can be found in Appendix A). However, prior to this,

the objectives of the investigation must be assessed to ensure that the identified techniques will enable the realisation of the objectives. Fundamentally, it is imperative to ensure the dataset is considered when deciding which technique will yield the most accurate results. Furthermore, it is the aim of this study to apply a variety of techniques, not to compare which will yield the most accurate results, but rather provide a variety of results. The results can subsequently be individually investigated to discover information that can aid with business decisions. This will permit the development of a conceptual model must be investigated, this will facilitate with the discovery of which techniques can be applied and the pre-requisites that must be observed when investigating a technical BI solution.

The information encapsulated within the dataset represents various information detailing customers, where each instance within the dataset provides socio-demographic and insurance product information on an individual customer, including whether or not they posses a caravan insurance policy. Classification will permit the identification of the customers that are likely to purchase a policy in addition to applying class label to these customers that will enable the discovery of the attributes that identify these individuals. Since, it is a supervised learning approach that proposes the most suitable means through which to realise the objectives of this investigation, and a number of suitable supervised techniques have been identified. To ensure the validity of the results, a combination of these techniques will be explored, providing the opportunity to cross-analyse the techniques and discover which the most reliable is, given specific circumstances. Upon having re-assessed the objectives, these can be redefined as:

- Interrogate the data set using a variety of supervised learning techniques.
- Analyse the performance of these supervised learning techniques.
- Identify which customers are likely to purchase a caravan policy.
- Identify the attributes that distinguish these customers, so that they can be effectively targeted.
- Identify any attributes that identify consumers unlikely to purchase a policy.
- Ensure that the results provided are accurate and can therefore be proposed as scientifically valid results.

Since, classification (through a number of possible techniques) has been discovered to be the most appropriate means through which to interrogate and analyse the dataset. As discussed in the literature review, classification is a supervised learning technique; correspondingly, a key target attribute (variable) must be identified. In this instance, the target attribute is whether a particular customer possesses a caravan policy. This attribute is the dependent variable or target variable, hence is the attribute that must be predicted for future instances. Since it has been observed that a supervised learning technique will provide the most effective means through which to make the discoveries necessary to satisfy the objective that were identified in the 'Data Investigation' phase of KDDS-BI the data must be appropriately modelled. As a result, during the model solution sub-step of this phase the dataset must be modified to ensure there is a training set, which will provide the means through which models can be trained, and a test set, with the target variable omitted. For this reason, the data set must be partitioned to provide sufficient data for the learning process. The learning process is induction-based, thereby permitting the classifier to build classification models, which will facilitate with the classification of new instances. Therefore, the data set has been provided in two distinct sets, a training set and test set. The training

set consists of 5822 customer description records each instance is detailed via 86 attributes. These have been compiled from 43 socio-demographic attributes which themselves have been derived from the customers area code. However, as a result, customers who have the same area code, will also posses the corresponding socio-demographic attributes values. A further 42 attributes relate to data regarding insurance products. 21 of these indicate the number of different insurance policies the customer posses, whilst the remaining 21 attributes indicate the premium level (contribution) of each policy. Since, the training set will be investigated to train and validate prediction models the target variable, thus specifying weather the customer possess a caravan insurance policy.

The test set contains 4000 customer records with the same attributes; however, the customers detailed are independent from those contained within the training set. The test set will be investigated to evaluate the capabilities of the classification techniques. The training set contains only 348 positive instances which accounts for 6% of the data set. Thus, the classifier will have limited information from which to build accurate classification models from. Consequently, of paramount importance is the ability to validate and evaluate the discoveries of the algorithms that are applied to the data set. Various validation techniques exist which can be employed to assess the accuracy of what classifiers have learned from the training data. Once validated, the classification model can be evaluated on an independent test set where performance is measured in relation to the error rates. The error rate is a measure of the classifier's performance. The classifier predicts the class of each instance, if correctly classified then it is counted as a success; if incorrect the instance is counted as an error. The error rate is therefore the proportion of errors made over the whole set of instances, accordingly, it measures the overall performance of the classifier (Witten & Frank, 2005).

A common validation technique which is employed is to draw upon two-thirds of the training set to enable the classifier to 'learn', whilst the remaining third is employed as a 'holdout' test set, to validate the data. However, this holdout method, risks using training and testing samples which do not fully represent the distribution of the class. To ensure that this ineffective representation does not affect the validity of the model learned, 'stratification' can be employed (Roiger & Geatz, 2003). Stratification is defined as the process of partitioning data into distinct or non-overlapping groups, thereby, approximately allowing for each class to be represented in a suitable proportion. Thus, stratification will ensure that an adequate number of customers are sampled. In order, to ensure the efficiency of stratification in representing the class distribution, stratification will be performed in conjunction with cross validation. Since only 6% of the instances within the data set are positive, the application of cross validation can help ensure that there is an equal distribution of classes within the data set. Cross validation partitions the data randomly into '*n*' fixed-sized mutually exclusive subsets known as 'folds', '*n-1*' folds are used as training data, whilst the remaining 'n^{th}' is employed as the test set. The process is repeated until all folds have been utilised as the test nest. Consequently, the overall error estimate yielded by averaging the '*n*' error estimates provides the average accuracy. Experimental results have revealed 10 as the value of '*n*' to be the most efficient, thus in the event $n = 10$, the method is known as '10-fold cross validation. Furthermore, stratified 10-fold cross-validation has become the practical standard due to its relatively low bias and variance (Han & Kamber, 2001).

As discussed the most effective method for calculating classification correctness is to present new unseen data to a classification model in the form of a test data set. The accuracy of the model upon the test data set can be summarised visually within a table known as a 'confusion matrix'. A confusion matrix is employed to compensate for the shortcomings of using error rate alone to judge the performance of the classifier. Since the error rate is the proportion of errors made over the whole set of instances. If there is a large variance amongst the class distribution, the error rate can be misleading. In multiclass prediction, the results of a test set are often displayed as a two-dimensional matrix with a row and column for each class (table B.2).

		Predicted Class	
		YES	**NO**
Actual Class	**YES**	True positive (TP)	False negative (FN)
	NO	False positive (FP)	True negative (TN)

Table B.2: Confusion matrix.

A confusion matrix illustrates the number of classified samples which are true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) for each class label. The TP & TN figures represent correctly classified instances. Whereas, the FP figure represents instances that have been falsely classified as positive e.g. yes is classified, when actual outcome is no. In contrast the FN figure represents instances that which have been falsely classified as negative. As a result the higher the TP & TN figures, the more accurate a classifiers performance can be deemed.

A confusion matrix can further more aid with the discovery of the true positive rate (TPR) and the false positive rate (FPR). The TPR represents the percentage of the total amount of positive samples in the dataset:

$$TPR = \frac{TP}{TP + FN} \quad \text{(B.1)}$$

In contrast, the FPR is the proportion of negative instances that were incorrectly reported as being positive:

$$FPR = \frac{FP}{FP + TN} \quad \text{(B.1)}$$

The relationship between the TPR and FPR can be visually plotted using a ROC (Receiver Operating Characteristic) curve. A ROC curve depicts the performance of a classifier without regard to the class distribution or error costs. Closely related to 'Lift charts', which are widely employed within marketing, ROC curves are be utilised within machine learning to compare the performance of classifier algorithms, since they provide a richer measure of classifier performance, then that of recall-precision graphs and lift curves (Fawcett, 2003). Figure B.4 depicts a ROC curve. The jagged line represents the sample of test data plotted upon the curve. The dotted curved live, reflects the average of the sample taken from the results of cross validation, thereby producing a smooth ROC curve. The closer the ROC curve is to the straight diagonal line, which

represents the likelihood that for every 'TP' there will be corresponding 'FP', the less accurate the model can be considered.

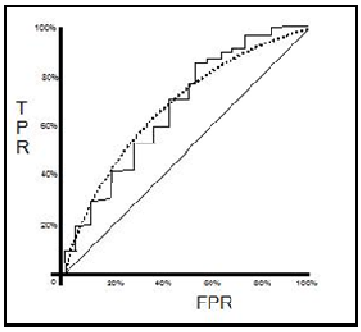


Figure B.4: ROC curve.

The area under the ROC curve (AUC) is a measure which can additionally be used to assess the predictive ability of a learning model. Also known as *discrimination*; it is expressed as a value between 0 and 1, and accounts for classifications that are random. A perfect test would be 1, whilst 0.5 or below is considered inadequate, since it falls on or below the diagonal line therefore representing no model (Berry & Linoff, 2004). Furthermore, it has been proposed that the AUC is a better measure overall and should therefore replace accuracy as they primary means for measuring and comparing classifiers (Huang & Lin, 2005).

The training data set and test data set, in addition to the considered techniques and evaluation methods will provide sufficient criterion through which the objectives of this research can be realised. However this has also unveiled the requirements of the project that must be considered from a technical perspective. Consequently, a BI application that provides functionality and capabilities that would facilitate the investigation of multiple classification models, which can not only be compared but individually applied to interrogate the data and extract interesting information, is required. Since approaches such as Bayesian classification, Classification Rules and Decisions Rules have advantages as well as disadvantages, by performing multiple calculations on the data with multiple approaches and a combination of algorithms, the results should be more reliable and the various approaches will compensate for the shortcomings of individual approaches. A primary requirement is that the BI software utilised to explore the data, facilitate exploration of the various approaches and techniques that have been proposed thus far. Furthermore, given the limited nature of the domain within which the solution must operate, namely a desktop PC of a decision maker, the software whilst capable of handling large data sets, must not require significantly large processing power. For this reason, a mid-specification Desktop PC has been selected as a platform:

- Intel Celeron 440 Processor (2 GHz. 800 MHz FSB, 512 KB Cache),
- Windows Vista Home Edition,
- 1 GB RAM,
- 80 GB HDD.

The specification of the PC has been selected to ensure that results can be obtained without necessitating a

complex hardware infrastructure. Furthermore, the relatively low cost of the PC will reflect the resources available to the majority of decision-makers. Thereby, providing a suitable solution that can be integrated in many (if not all) working environments at a relatively low cost.

B.2.3 Development

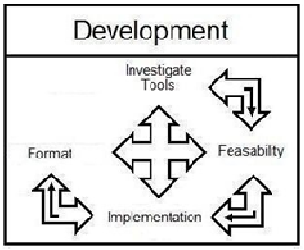


Figure B.5: Development stage of KDDS-BI.

In addition to investigating a conceptual model that will provide a greater insight into the nature of the data, the requirements, techniques and specifications which must be observed if the objectives of this investigation are to be attained have been explored and determined. It is now possible to examine various applications that can facilitate in the realisation of the objectives. Thus enabling the raw data to be analysed and informative discoveries uncovered that will permit decision makers to direct business objectives with a greater competitive edge. There are various software solutions that are available from a number of vendors; these are both commercial and open-source. In order to find a suitable BI solution for this investigation, a number of these packages were investigated.

University of Waikato	Weka	Open-source	A popular suite of machine learning software written in Java, developed at the University of Waikato. WEKA is free software available under the GNU General Public License. The workbench contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality. Originally developed in C, Since 1997 Weka has been developed in Java, thereby providing a portable collection of data preprocessing and modelling techniques capable on running on almost any platform Weka supports several standard data mining tasks, more specifically, data pre-processing, clustering, classification, regression, visualisation, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Furthermore, Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka. Another important

			area that is currently not covered by the algorithms included in the Weka distribution is sequence modelling. However projects such as that by Patro[13] have addressed this issue.

Table B.3: Subsection of Appendix A: Section A.5-Table A-1.

After consideration of the various available tools (for which the review is contained in Appendix A: Section A.5-Table A-1) and despite the availability of a number of commercial BI solutions, for this study an open-source solution was considered more suitable for the purposes of this investigation. Thus, the motivation for choosing open-source software is due to the ease with which open source platform can not only be customised, but further integrated within existing software and hardware infrastructures. Thereby, providing the option of implementing only desired functionality, in addition many of the open-source solutions are developed in Java. Thus the platform upon which the solution is to run can be given less consideration. Of the open-source solutions available Weka chosen as the most suitable since it provides algorithms that meet the requirements of investigation and will facilitate the realisation of the objectives of this investigation, table B.3provides a subsection of Appendix A: Section A.5-Table A-1 and illustrates the key features of the Weka workbench.

During the Data Modelling phase, it was identified that to successfully investigate the data for meaningful discoveries that can provide decision support for direct marketing, Bayesian Classification; Decision Trees; and Production Rules would provide effective means through which to study the data set. Weka provides following algorithms that can be investigated and support the findings of the data modelling phase of this investigation:

- For Bayesian networks a selection of algorithms are available to estimate conditional probability tables in combination with different learning algorithms, allowing for increased levels of experimentation:
    - *NaïveBayes*: Standard probabilistic naïve Bayes classifier.
        - NaïveBayesSimple: Numeric attributes are modelled by a normal distribution.
    - *BayesNet*: Base class for a Bayes Network classifier. Provides data structures (network structure, conditional probability distributions, etc.) and facilities common to Bayes Network learning algorithms. The algorithm learns Bayesian networks using the 'SimpleEstimator' to find the conditional probability tables in conjunction with these search/learning algorithms:
        - HillClimber: Uses a hill-climbing algorithm which, inserts, deletes and reverses directed edges until a better solution is found.
        - *RepeatedHillClimber*: Searches for Bayesian network structures by repeatedly generating a random network and applying the hillclimber algorithm and return the best structure of the various iterations.
        - *TAN (Tree Augmented Naïve Bayes)*: uses the naïve Bayes classifier and attaches directed edges to it.
        - *K2*: This Bayes Network learning algorithm uses a hill climbing algorithm restricted by an order on the attributes.
- For Decision Tree classifiers, Weka provides a selection of algorithms. The parameters for many of these can be adjusted through the error-pruning settings, thereby allowing for detailed investigation:

[13] http://davis.wpi.edu/~xmdv/weka/: Accessed July, 2008.

- *J48*: Generates an unpruned or pruned Decision Tree based up Quinlan's C4.5 algorithm. Allows for a combination of error-pruning and reduced-error-pruning (REP) with subtree raising and subtree replacement pruning.
- *ADTree*: Builds an alternating Decision Tree and only applies to two-class problems. Each node is assigned either a positive or negative value which each represents a class. Filtering down the tree the value at each node is summed. Whether the total is negative or positive determines the class outcome.
- *Id3*: Constructs an unpruned Decision Tree based on an earlier version of Quinlan's C4.5 algorithm. However, this classifier cannot handle numeric attributes.
- *RandomTree*: Constructs a tree using randomly chosen attributes at each node without performing any pruning.
- *REPTree***:** Fast Decision Tree learner. Builds a Decision/Regression Tree using information gain/variance reduction and prunes it using REP (with backfitting). Only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5).

- Weka also provides a selection of algorithms for Production Rules. Similar to the algorithms provided for Decision Trees, the parameters for many of the Production Rule algorithms can also be adjusted through the error-pruning settings, thereby allowing for detailed investigation:
  - *DecisionTable*: Builds a simple decision table majority classifier.
  - *JRip*: Implements a propositional rule learner – 'Repeated Incremental Pruning to Produce Error Reduction' (RIPPER) can be used with or without incremental reduced-error rate pruning. By incremental it differs to standard reduced-error pruning by pruning rules immediately after they are generated.
  - *PART*: Builds a partial C4.5 Decision Tree with each iteration and makes the 'best' leaf into a rule. Uses separate-and-conquer and can be applied with or without reduced-error rate pruning.

The above represent a selection of the algorithms that are provided by Weka. Consequently, Weka provides many of the functions and capabilities that underpin this investigation. As established since Weka is an open-source platform it can be further customised to meet the explicit requirements of this investigation. Initially, the Weka source files were downloaded via CVS, through the Eclipse tool (Figure B.6-B.8). The Eclipse tool is an open source development platform comprised of extensible frameworks, tools and runtimes for building, deploying and managing software across a java project lifecycle.

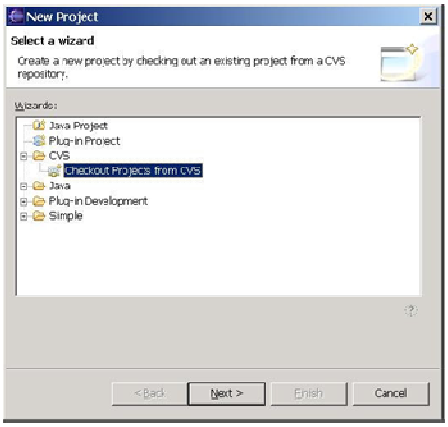


Figure B.6: Eclipse tool used to checkout a project from CVS.

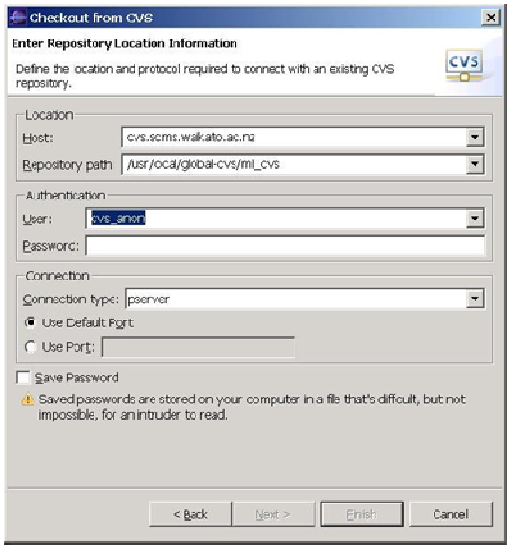


Figure B.7: Enter repository location information.

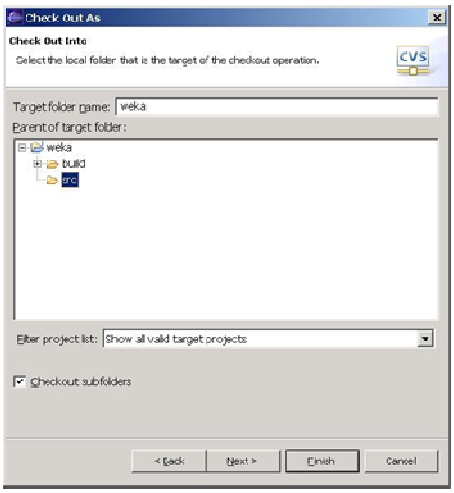


Figure B.8: Select source code to checkout from CVS.

Once the source files were downloaded, the Eclipse tool was used as an IDE (Integrated Development Environment), thereby facilitating the manipulation of the Weka source code to permit the required customisation (Figure B.9).

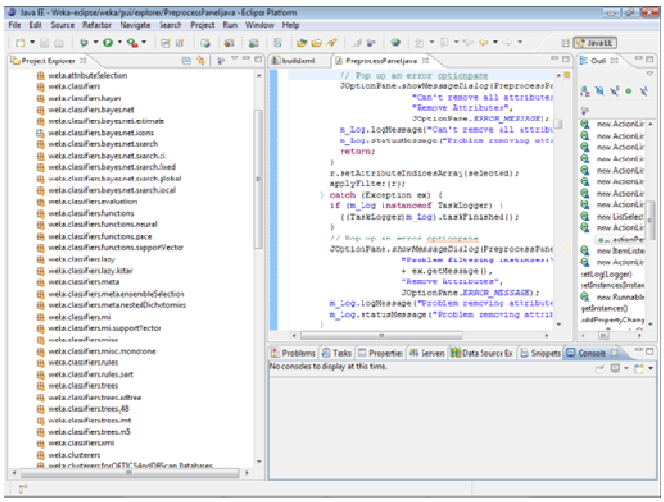


Figure B.9: Eclipse employed as a Java IDE to investigate Weka source code.

The customisation of the Weka source code allowed for the exact requirements of this investigation to be integrated with the functionality of Weka, providing a solution that will explicitly meet these requirements. As part of the customisation obsolete algorithms and functionality were removed through manipulation of the original code, this enabled the memory management and efficiency of the tool to be increased. Further modifications were implemented in the form of a modified Graphical User Interface (GUI) to ensure that the internal modifications are extended to the user, in addition to a more unambiguous selection of algorithms. These modifications will provide faster executing, more focused tool that will be able to meet the requirements of this project.

Once the modifications to the source code had been completed, the Eclipse tool was used as a means through which an external executable '.jar' file could be created. These stages are illustrated via figure B.10 through to figure B.13.

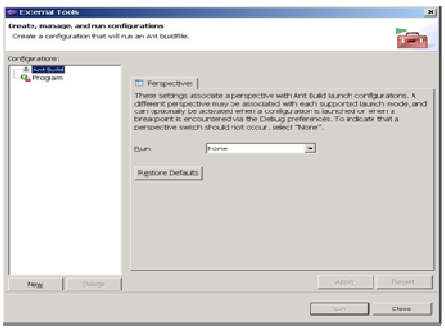


Figure B.10: Exporting source code to an external (from Eclipse) file.

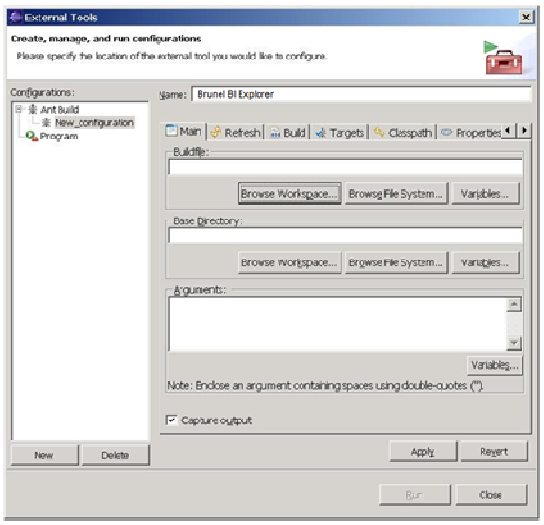


Figure B.11: Set name for external tool.

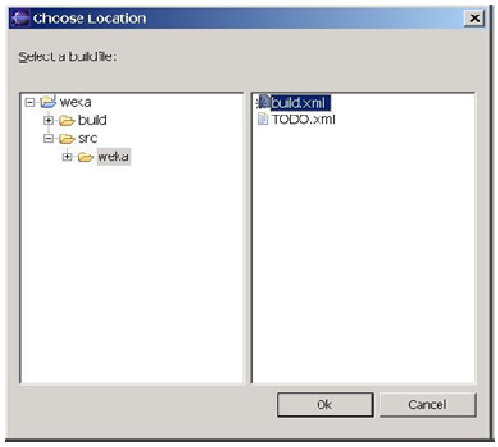


Figure B.12: Selection of a 'build file' to create an executable file.

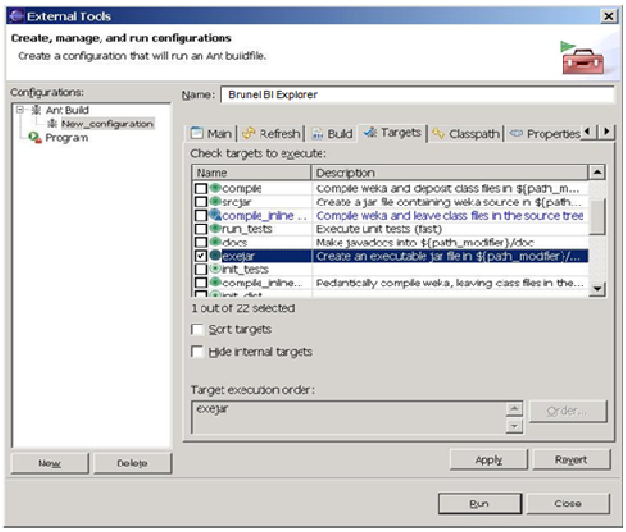


Figure B.13: Select target file.

The executable file that has been created for the solution, which will be referred to as Brunel BI Explorer (figure B.14), can be implemented independently from the Eclipse tool on any system that has a Java Runtime Environment configured.

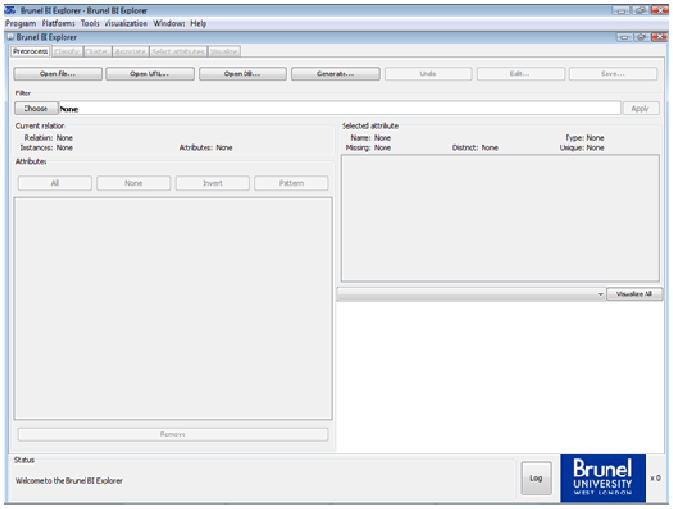


Figure B.14: Brunel BI Explorer.

In addition to the desktop executable interface, a web based interface which provides access to the algorithms was developed (figure B.15).The web-based GUI has been designed as a Java applet, provides the user with the option of investigating the provided classification algorithms to analyse datasets from any location (figures B.16 & B.17). The two interfaces provide a complete BI solution. However, the web-GUI does not provide the full functionality of the desktop solution and early sages of the investigation unveiled that the web-GUI was subject to limited performance in the event that the dataset to be analysed was significantly large. The interface is,

therefore, one that is better suited to small datasets as the time taken for calculating the results is far greater then the desktop executable; or for the purpose of presenting, demonstrating functionality.

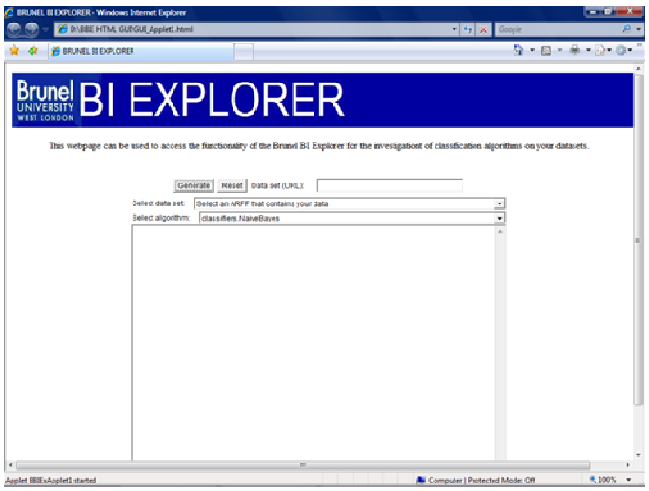


Figure B.15: Web based interface for Brunel BI Explorer.

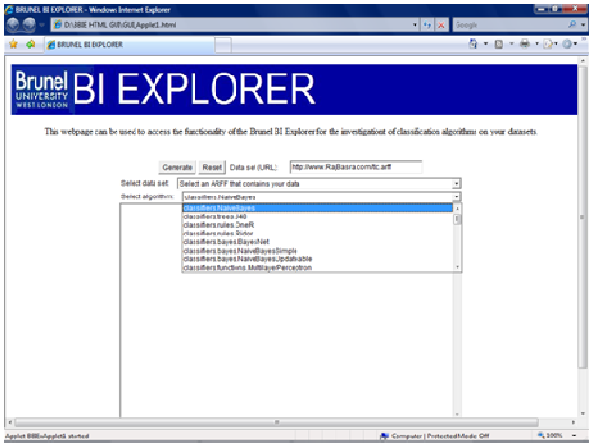


Figure B.16: Web-based classification options.

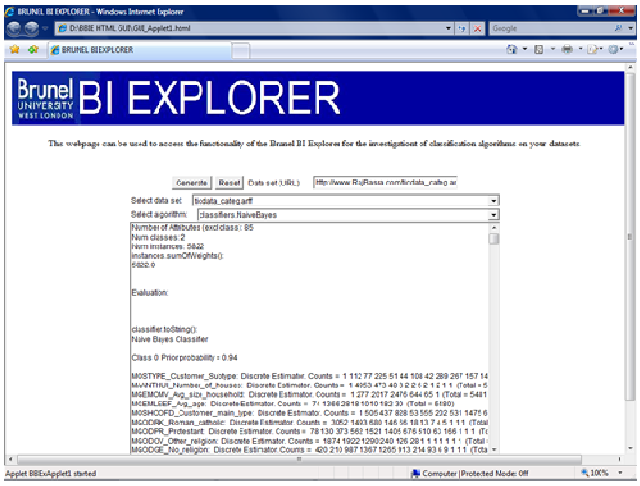


Figure B.17: Results generated through the analysis of the Insurance dataset with the 'NaïveBayes' algorithm.

Due to the limitations of the web-GUI, for the purpose of this investigation the analysis will be conducted via the desktop executable. Since the desktop executable not only provides greater functionality but also greater efficiency when dealing with a large dataset. In addition to the implemented solution, prior to analysis the dataset must be transformed into a suitable format. Consequently, the next stage was to investigate the data set and realise any formatting that would be required for the data set to be applied within the Brunel BI Explorer. The dataset was initially in an unlabeled, tab-delimited format. Although the Brunel BI Explorer can interpret this format by assigning attribute names using the first row of values, this approach can be inefficient, since it can prohibit clarity when viewing output from data mining. In order to preserve the integrity of the output, the data must be transformed into a file format such as Attribute-Relation File Format (ARFF). ARFF are an ASCII text file that describes a list of instances sharing a set of attributes. The ARFF has been developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato. The ARFF provides a more effective means since rather than assign the initial row of value as the descriptor label. ARFF permits descriptive labels for attributes to be set in a header section, with corresponding values contained in the data section. ARFF consists of two distinct parts 'header' and 'data'. The Header of the ARFF contains the name of the relation, a list of the attributes (the columns in the data), and their data types. The data section of the ARFF contains the values that correspond to the attributes. A typical structure for an ARFF is illustrated in the code-table B.1.

1	@relation NameOfFile
2	@attribute attribute1Name {StringValue1, StringValue2, StringVlaue3}
3	@attribute attribute2Name numeric
4	@attribute attribute3Name { true, false}
5	@attribute attribute4Name date <dd-mm-yyyy>
6	@data
7	StringValue3, NumericValue, true/false, 00-00-0000
8	StringValue1, ?, true/false, 00-00-0000
9	StringValue2, NumericValue, true/false, 00-00-0000
10	…

Code-table B.1: Typical structure for an ARFF file.

Lines 1-9 represent the 'header'. Line 3 represents the name of the file, preceded with the '@relation' declaration, if the name is a string that includes spaces, it must be quoted, i.e. Line 3 could have been written as:

@relation <*Name of File*>

Attribute declarations take the form of an ordered sequence of '@attribute' statements. Each attribute in the dataset has its own '@attribute' statement which uniquely defines the name of the attribute and data type. The order the attributes are declared indicates the column position in the data section of the file. Thus, if an attribute is the third one declared, then all values that correspond to that attribute will be found in the third comma delimited column. The format for the '@attribute' statement is:

@attribute <*attribute-name*> <*datatype*>

The 'attribute-name' must start with an alphabetic character, as with the file name, if spaces are to be included in the name then the entire name must be quoted. The data type can consist of four formats:

- *String*: denotes arbitrary textual values (code-table B.1, line 5).
- *Numeric*: can be real or integer numbers (code-table B.1, line 6).
- *Nominal*: defined by providing an attribute name and listing all possible values: {name1, name2, name3, ...} (code-table B.1, line 7).
- *Date*: the name of the attribute is followed by the keyword date. The keyword date is usually followed with an optional string '<*date-format*>' specifying how date values should be parsed and printed (code-table B.1, line 8).

The keywords 'numeric', 'string' and 'date', in addition to the '@relation', '@attribute' and '@data' declarations are case insensitive. The '@data' (line 10) signals the start of the data instances. Each line of data corresponds to a particular instance, with each value distinguished via a comma. Missing values (line 15, value 2) must be substituted with a single question mark.

The ARFF file format therefore provides a robust data file for representing data sets that are to be interrogated (Witten & Frank, 2005). As discovered in the 'Data Modelling' phase of KDDS-BI, the dataset consists of two files, the training set and test set. The training set contains 5822 customer records consisting of 43 socio-demographic attributes in addition to 43 attributes that detail the contribution level and insurance policies that the customer possesses. The test set contains 4000 customer records with the same attributes; however, all customers are unique (as was illustrated in table B.1)

Of the dataset attributes 2, 3, 4, 6-43, 44-64 and 65-85 are considered ordinal attributes as the values can be compared i.e. relative. The remaining attributes are not relative to each other and simply correspond to a label contained in the data dictionary. Since classifier algorithms handle attributes with numeric and nominal data types differently, both training and test datasets were transformed into a numeric and nominal ARFF. Furthermore, the transformation of the dataset provided the opportunity for more explanatory labelling. More explanatory labels will provide a greater level of clarity especially when viewing antecedents and consequents of Production Rules. This data transformation was conducted via the pre-process tab within the Brunel BI Explorer. Initially the dataset is loaded, basic dataset information such as the number of instances (samples) and attributes are displayed, along with attribute-specific data such as name and type (nominal or numeric), with value counts and class distribution visualised in a histogram (figure B.18).

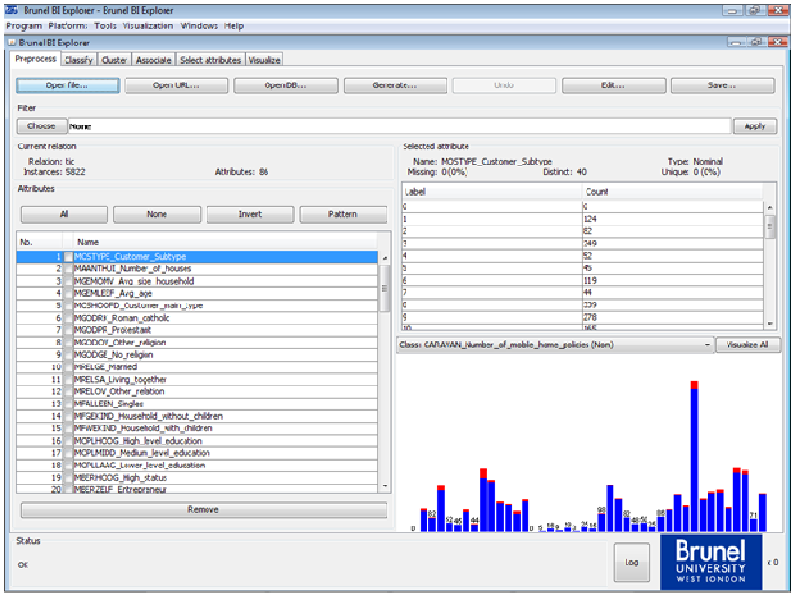


Figure B.18: Brunel BI Explorer pre-process.

The supplied tab-delimited training set containing unlabeled attributes once loaded, can have the descriptive labels set. This is achieved through the dataset viewer, via the 'edit' button (figure B.19).

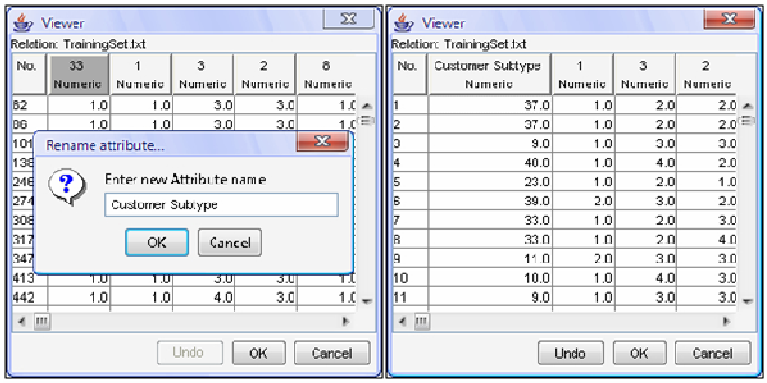


Figure B.19: Assigning descriptive labels to attributes.

The target attribute 'Caravan pol' had its class values '0' and '1' replaced by 'No' and 'Yes' respectively to aid clarity when interrogating the data. In order to replace the values, the 'Replace values…' function was used (figure B.20).

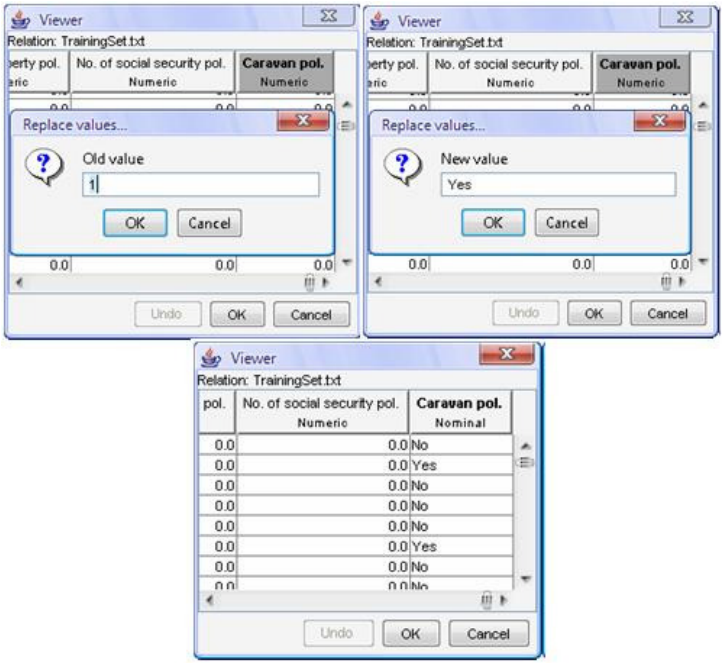


Figure B.20: Attribute value '1' replaced with 'yes'.

The modified training dataset was then saved as an ARFF. Code-table B.2 illustrates a sub-section of the dataset in ARRF. As previously described, the @relation declaration (line 1) indicates the name of the dataset, whilst each attribute in the dataset has its own @attribute statement which uniquely defines the name of that attribute and data type, which in this case is numeric (lines 2 – 87). The order in which the attributes are declared indicates the column position in the data section of the file. Thus, for the first sample (customer) in the dataset (line 89), the 'Customer Subtype' is '37', 'No. of houses' is '1' this pattern continues, with each value corresponding to an ordered attribute until the attribute 'Caravan pol.' which is displayed as 'No'.

1	@relation TrainingSet-Numeric.arff
2	@attribute 'Customer Subtype' numeric
3	@attribute 'No. of houses' numeric
4	@attribute 'Avg. size household' numeric
…	…
45	@attribute 'Contr. car pol.' numeric
46	@attribute 'Contr. delivery van pol.' numeric
47	@attribute 'Contr. motorcycle/scooter pol.' numeric
…	…
87	@attribute 'Caravan pol.' {Yes, No}
88	@data
89	37,1,2,2,8,1,4,1,4,6,2,2,0,4,5,0,5,4,0,0,0,5,0,4,0,2,3,5,0,2,7,7,1,2,6,3,2,0,5,2,0,5,4,2,0,0,0, 0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,2,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,No
90	37,1,2,2,8,0,4,2,4,3,2,4,4,4,2,0,5,4,0,0,0,7,0,2,0,5,0,4,0,7,2,7,0,2,9,0,4,5,0,0,0,3,4,2,0,0,6, 0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,No
…	…

Code-table B.2: Sub-section of numeric ARFF for the insurance dataset.

Although, the attribute values were converted to numeric values, they are in reality nominal. Consequently an additional ARFF was created. However, in contrast to the numeric ARFF, which was formatted via the Brunel BI Explorer, the conversion from numeric to nominal must be manually conducted. Therefore, a copy of TrainingSet-Numeric.arff was transferred to the text editor 'Notepad'. Via Notepad it was possible to modify each @attribute statement as illustrated in Code-table B.3. For each attribute the numeric data type has been replaced by all possible values that can correspond to the attribute lines (2 – 87).

1	@relation TrainingSet-Nominal.arff
2	@attribute 'Customer Subtype'
	{0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,151,6,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31}
3	@attribute 'No. of houses' {1,2,3,4,5,6,7,8,9,10}
4	@attribute 'Avg. size household' {1,2,3,4,5,6}
…	…
45	@attribute 'Contr. car pol.' {0,1,2,3,4,5,6,7,8,9}
46	@attribute 'Contr. delivery van pol.' {0,1,2,3,4,5,6,7,8,9}
47	@attribute 'Contr. motorcycle/scooter pol.' {0,1,2,3,4,5,6,7,8,9}
…	…
87	@attribute 'Caravan pol.' {Yes, No}
88	@data
89	37,1,2,2,8,1,4,1,4,6,2,2,0,4,5,0,5,4,0,0,0,5,0,4,0,2,3,5,0,2,7,7,1,2,6,3,2,0,5,2,0,5,4,2,0,0,0,
	0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,0,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,No
90	37,1,2,2,8,0,4,2,4,3,2,4,4,4,2,0,5,4,0,0,0,7,0,2,0,5,0,4,0,7,2,7,0,2,9,0,4,5,0,0,0,3,4,2,0,0,6,
	0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,No
…	…

Code-table B.3: Sub-section of nominal ARFF for the insurance dataset.

This procedure was repeated for the test dataset, which resulted in 'TestSet-Numeric.arff' and 'Test-Set-Nominal.arff' datasets. Since there are no missing values no further steps were necessary prior to interrogating the dataset.

B.2.3 Decision Support

Once the data had been transformed into the correct format to that which is required by the Brunel BI Explorer, the data could then be interrogated to discover hidden information that can be exploited for decision support. The initial stage of the 'Decision Support' phase of KDDS-BI is to gather the output (figure B.26). Thus, the dataset will be investigated using a variety of classification models. The result of these models can then be further analysed to extract novel information that can provide decision support.

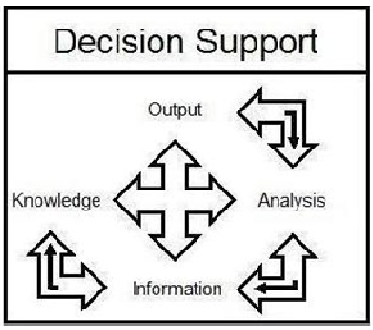


Figure B.21: Decision Support stage of KDDS-BI.

The initial stage of the Decision Support stage required the ARFF containing the numeric training dataset to be loaded into the Brunel BI Explorer via the 'Preprocess' tab. Once loaded the dataset can be applied as a means through which the classifier can be trained. The 'Classify' tab was selected and the 'NaïveBayes' classifier was selected from the list of algorithms, in addition to the 'Test options' set. As discovered when modelling the data, the most suitable test is to cross-validate the data with 10 folds (figure B.22).

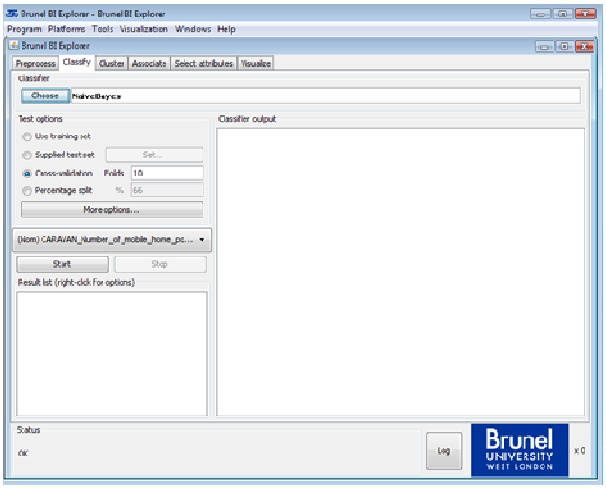


Figure B.22: Classifier and test options settings.

The parameters for the classifier can be further refined by clicking on the name of the classifier. For the Decision Tree classifiers and rule learners the pruning settings were adjusted with all other settings kept as default (figure B.23). Once all necessary parameter settings had been configured the classifier was initiated. Figure B.24 depicts a sub-section of the initial output. This output provides the results summary for the 'NaïveBayes' classification upon the training set. Subsequently, the results are employed at this stage as a means thorough which the classifier can be trained. The output itself therefore does not require further analysis, until they have been evaluated with the test data.

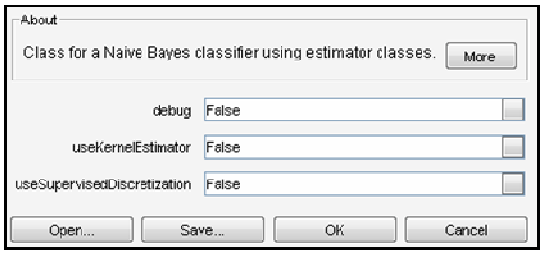


Figure B.23: Classifier settings.

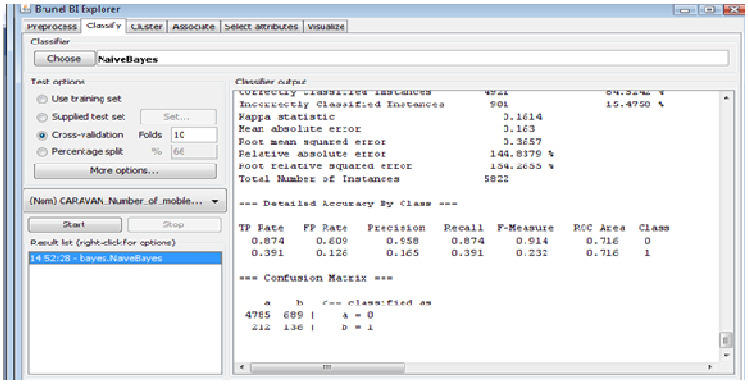


Figure B.24: 'NaïveBayes' classification output for training data.

The described procedure for training the classifier was then repeated for the various other BayesNet, Decision Tree learners and Production Rule classification algorithms that were identified within the feasibility sub-step of the 'Development' phase of KDDS-BI. The investigated models were then stored within the Brunel BI Explorer, prior to loading the ARFF containing the nominal training dataset. Once the nominal training data was loaded the procedure for developing models was repeated, thus training models developed for both the numeric and nominal datasets, to ensure a broad coverage. This process was investigated for all classification algorithms aside from Id3, since this algorithm is unable to analyse numeric attributes. Table B.4 depicts the classification models that were investigated and the various pruning settings were applicable.

NaiveBayes	N/A
BayesNet – HillClimber	N/A
BayesNet - RepeatedHillClimber	N/A
BayesNet - TAN	N/A
BayesNet - K2	N/A
J48	Error-based pruning & Subtree replacement
J48	Error-based pruning & Subtree raising
J48	Reduced-error pruning & Subtree replacement
J48	Reduced-error pruning & Subtree raising
J48	No pruning
Id3 (nominal only)	N/A
RandomTree	N/A
REPTree	Reduced-error pruning

REPTree	No pruning
ADTree	N/A
DecisionTable	N/A
JRip	Incremental reduced-error pruning
JRip	No pruning
PART	Error-based pruning
PART	Reduced-error pruning
PART	No pruning

Table B.4: Classifiers and pruning settings that were applied to data set.

Once the classification models had been trained, the trained models were then evaluated upon the independent test data. Remaining on the 'Classify' tab; 'Supplied test set' was selected under the test options and enabling the numeric ARFF to be loaded (figure B.25). The models were then evaluated using the test data (figure B.26).

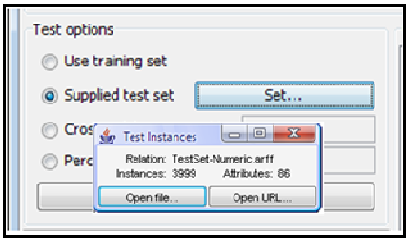


Figure B.25: Loading Numeric ARFF.

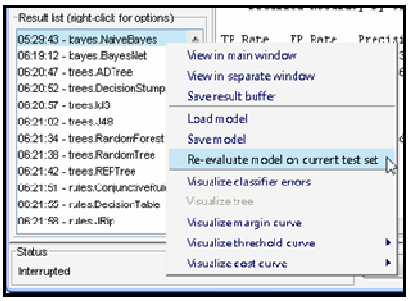


Figure B.26: Models used to investigate test data.

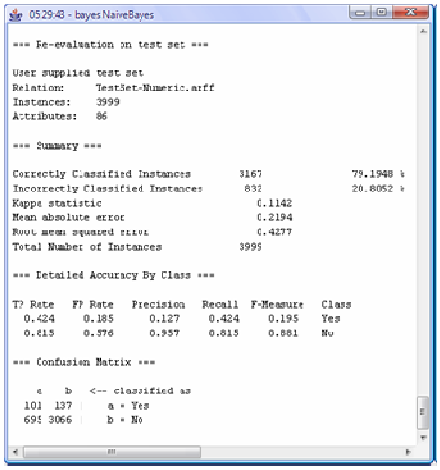


Figure B.27: Results of the numeric test data re-evaluated with trained 'NaïveBayes' algorithm.

The evaluation results are displayed under the existing classifier output (figure B.27). As identified with in the Data Modelling phase of KDDS-BI, the 'TP Rate', 'TN Rate' (same as TP Rate for 'No'), 'Correctly Classified Instances/Samples' (CCS), 'Root Mean Squared Error' (RMSE) and 'Mean Absolute Error' (MAE), can be investigated to evaluate the performance of the classifier. The number of CCS represent the percentage of samples that have been correctly classified, thus the CCS value represents the classification success rate. The RMSE is a quadratic scoring rule which measures the average magnitude of error. This is obtained by measuring the differences between values predicted by a model and the values actually observed. The individual variation between the values is known as the 'residual' value, these are aggregated to provide the RMSE. The RMSE is considered a good measure of accuracy and predictive power, a low RMSE signifies a high level of accuracy. The MAE is the same as RMSE, however, the statistic is calculated by the absolute differences instead of the squared difference. The MAE is also a statistic used to calculate how close forecasts or predictions are to the eventual outcomes.

The performance and accuracy of the classifiers is of paramount importance since it provides an indicator of the functionality of the classifier, thereby providing a basis upon which to ascertain the reliability of the discovered results. The reliability must be analysed as, more reliable the results the greater the level of fundamental support provided to decision-makers. Further to this, as discussed in the 'Data Modelling' phase the 'TP Rate' and 'TN Rate' provide the percentage of the total amount of positive samples in the dataset which can be used to formulate a ROC Curve. The ROC Curve, whilst providing a graphical representation of classifier accuracy also functions as a means through which to assess the predictive ability of a learning model. Thus as discussed in the 'Data Modelling' phase the 'Area Under the Curve' (AUC) or level of discrimination, accounts for classifications which are random. The ROC Curve can be visualised by accessing the model from the list and visualising the threshold curve for the target class value 'Yes' (figure B.28).

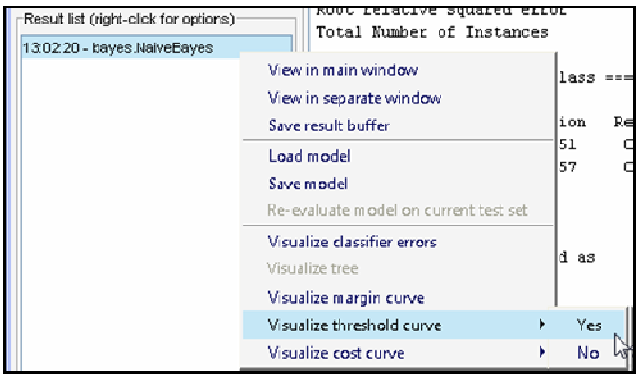


Figure B.28: Generating a ROC curve for the 'NaïveBayes' model.

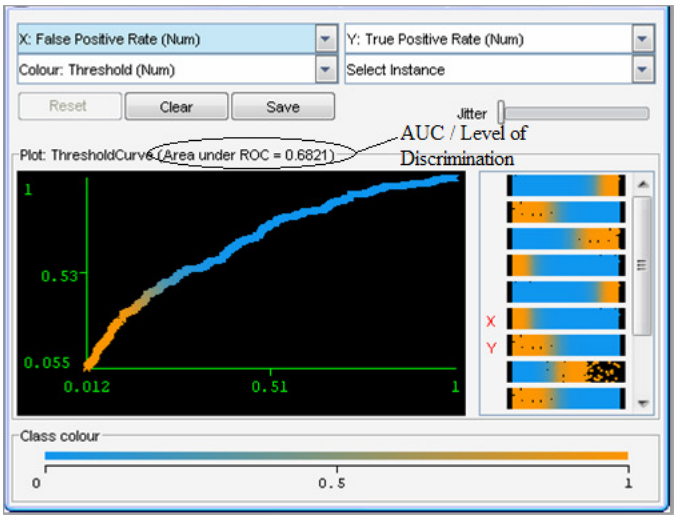


Figure B.29: ROC curve and level of discrimination for the 'NaïveBayes' model.

The Brunel BI Explorer will then generate the curve and provide the statistic for the AUC (figure B.29). The level of discrimination is expressed as a value between 0 and 1, since the further the curve is from a perfectly diagonal line through the axis the better the classification, correspondingly, a perfect statistic would be 1, the following values however can be used as a guide:

- .90-1 = very accurate,
- .80-.90 = above average,
- .70-.80 = average,
- .60-.70 = below average,
- .50-.60 = inaccurate.

The process of evaluating the numeric test data against the trained model was repeated for all Bayesian classifiers. In addition the process was also repeated to evaluate models trained with nominal training set against the nominal test data. These results can be viewed in table B.6. For each measure the highest scoring result has been underlined.

NaïveBayes (Nominal)	37.4	86.1	83.2	0.380	0.177	0.706
NaïveBayes (Numeric)	**42.4**	81.5	79.2	0.428	0.219	0.682
BayesNet - HillClimber (Nominal)	31.9	89.7	86.3	0.329	0.160	**0.718**
BayesNet - HillClimber (Numeric)	37	88	84.9	0.348	0.171	0.702
BayesNet - RepeatedHillClimber (Nominal)	31.9	89.7	86.3	0.329	0.160	**0.718**
BayesNet - RepeatedHillClimber	3.7	88	84.9	0.348	0.171	0.702

(Numeric)						
BayesNet – TAN (Nominal)	9.7	97.8	92.6	0.260	**0.079**	0.679
BayesNet – TAN (Numeric)	6.3	**99**	**93.4**	**0.24**	0.112	0.717
BayesNet – K2 (Nominal)	40.3	84.6	82.0	0.395	0.191	0.709
BayesNet – K2 (Numeric)	37	88	84.9	0.348	0.171	0.702

Table B.5: Evaluation results for Bayesian classifiers.

Table B.5 illustrates that the TP rate for 'NaïveBayes (Numeric)' correctly predicted 42.4% of the customers who have a caravan insurance policy, and outperformed the other Bayesian classifiers. Furthermore, it is significant that the same classification model has the lowest TN rate of 81.5% i.e. those customers that do not posses a caravan insurance policy. This implies that the accuracy was gained for predicting the 'Yes' value for the caravan policy class at the expense of classifying negative samples ('No' value samples) resulting in the lowest CCS rate of 79.2%. BayesNet - TAN (Numeric) achieved the highest TN rate of 99% as well as correctly classifying the highest number of instance (93.4%). However, the classifier only predicted 6.3% of positive samples; this clearly reflects the proportion of 'Yes' and 'No' classes in the test set, which is 238 and 3762 respectively. Furthermore, the same classifier possesses the best (implied through the lowest figure) RMSE rate and MAE rate with 0.24 and 0.079 respectively, indicating that the 6.6% of incorrectly predicted samples were of low magnitude. The results for BayesNet - HillClimber and BayesNet - RepeatedHillClimber display no disparity between their predictive ability as their results were identical. Consequently, they both possess an identical AUC measure of 0.718 which is the highest, indicating that the level of random classifications is slightly below the median within the desired range (0.5 - 1).

Prior to further investigating these results, it is imperative to obtain all of the output collected through the classification algorithms that were applied to the dataset. ROC curves are not the only visual method through which results can be observed. Certain classification models such as Bayesian networks and Decision Trees can be visualised graphically. By selecting the model in the 'Result list' and choosing 'Visualise tree' (figure B.30), the Brunel BI explorer can generate a graphical representation of the model.

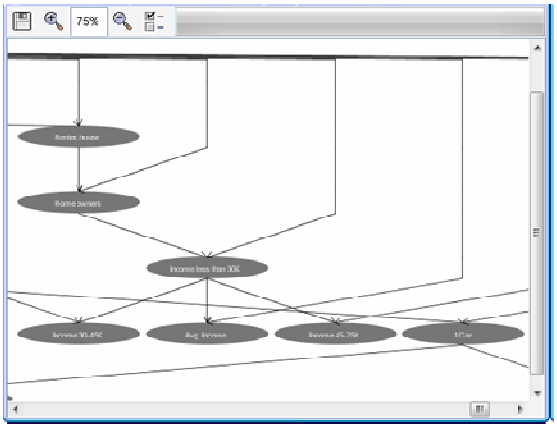


Figure B.30: Visualisation of a BayesNet acyclic graph (partial view).

The Bayesian network nodes in figure B.35 represent an attribute and the unidirectional arrows indicate probability dependency. The conditional probability table for each attribute can be viewed by clicking on the node. Figure B.31 illustrates this table for the 'Income 45-75K' attribute whose parent nodes are 'Income less than 30K' (as visible in figure B.30) and the target variable: 'Caravan pol', (this is not visible since figure B.30 is a partial view of a Bayesian network).

Probability Distribution Table For Income 45-75K

Caravan pol.	Income less than 30K	0	1	2	3	4	5	6	7	8	9
Yes	0	0.16	0.024	0.063	0.17	0.248	0.189	0.044	0.034	0.024	0.044
Yes	1	0.035	0.008	0.076	0.178	0.432	0.127	0.076	0.042	0.026	0.008
Yes	2	0.026	0.026	0.109	0.307	0.307	0.151	0.036	0.026	0.005	0.005
Yes	3	0.118	0.045	0.245	0.391	0.155	0.009	0.009	0.009	0.009	0.009
Yes	4	0.135	0.058	0.442	0.173	0.058	0.058	0.019	0.019	0.019	0.019
Yes	5	0.159	0.432	0.159	0.114	0.023	0.023	0.023	0.023	0.023	0.023
Yes	6	0.25	0.321	0.107	0.107	0.036	0.036	0.036	0.036	0.036	0.036
Yes	7	0.375	0.292	0.042	0.042	0.042	0.042	0.042	0.042	0.042	0.042
Yes	8	0.083	0.25	0.083	0.083	0.083	0.083	0.083	0.083	0.083	0.083
Yes	9	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
No	0	0.131	0.013	0.069	0.156	0.259	0.194	0.031	0.058	0.017	0.072
No	1	0.035	0.042	0.142	0.121	0.345	0.171	0.076	0.015	0.052	0.001
No	2	0.043	0.087	0.226	0.293	0.243	0.079	0.02	0.008	0	0
No	3	0.086	0.108	0.312	0.343	0.116	0.022	0.013	0	0	0
No	4	0.174	0.147	0.344	0.215	0.08	0.035	0.003	0.001	0.001	0.001
No	5	0.384	0.226	0.235	0.103	0.044	0.004	0.001	0.001	0.001	0.001
No	6	0.251	0.375	0.216	0.147	0.002	0.002	0.002	0.002	0.002	0.002
No	7	0.607	0.295	0.075	0.003	0.003	0.003	0.003	0.003	0.003	0.003
No	8	0.279	0.644	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
No	9	0.92	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009

Figure B.31: Conditional probability table for 'Income 45-75K' node.

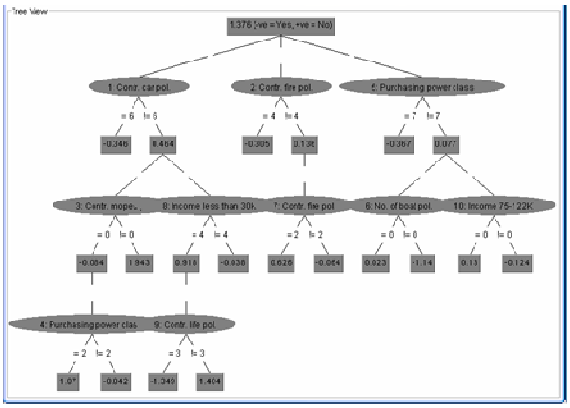


Figure B.32 Visualisation of ADTree.

The tree visualised in figure B.32, is an alternating Decision Tree that has been discovered through the application of the ADTree algorithm. As defined in the 'Development' phase, an alternating Decision Tree assigns each node either a positive or negative value. Filtering down the tree the value at each node is summed. Whether the total is negative or positive determines the class outcome. This instance a positive value is classified as 'No' whereas in the event that the total is negative, the class is valued as 'Yes'. Thus, once a tree has been visualised in this way it enables a person to manually classify a customer. Consequently, a customer who has a contribution level of 2 for a car policy, lives in an area where 37-49% of the population have an income less than 30K and has a contribution level of 3 for a life policy would be 1.376 + 0.464 + 0.918 – 1.349 = 1.409; therefore, classified as not having a caravan insurance policy.

J48 (Error-based pruning & subtree replacement) (Nominal)	0	**100**	94.0	0.237	0.112	0.5
J48 (Error-based pruning & subtree replacement) (Numeric)	0.4	99.9	**94.9**	0.238	0.111	0.498
J48 (Error-based pruning & subtree raising) (Nominal)	0	**100**	94.0	0.237	0.112	0.5
J48 (Error-based pruning & subtree raising) (Numeric)	0.4	99.9	**94.9**	0.238	0.111	0.498
J48 (Reduced-error pruning & subtree replacement) (Nominal)	2.5	99.4	93.6	0.242	0.104	0.639
J48 (Reducued-error pruning & subtree replacement) (Numeric)	3.8	99	93.3	0.421	0.105	0.615
J48 (Reduced-error pruning & subtree raising) (Nominal)	3	99.8	93.8	0.237	0.108	0.645
J48 (Reducued-error pruning & subtree raising) (Numeric)	3.8	99	93.3	0.241	0.105	0.615
J48 (No pruning) (Nominal)	12.6	95.1	90.2	0.294	0.108	0.557
J48 (No pruning) (Numeric)	**17.2**	95.5	90.8	0.282	**0.099**	0.58
Id3 (Nominal)	16.5	93.6	86.6	0.316	0.103	0.552
Id3 (Numeric)	N/A	N/A	N/A	N/A	N/A	N/A
RandomTree (Nominal)	14.3	94.8	90.0	0.298	0.102	0.549
RandomTree	13.9	93.9	89.1	0.317	0.105	0.541

(Numeric)						
REPTree (Nominal)	3.4	99.6	93.9	0.236	0.113	0.669
REPTree (Numeric)	2.5	99.6	93.8	**0.234**	0.106	0.675
REPTree (No pruning) (Nominal)	15.1	94.9	89.7	0.294	0.113	0.669
REPTree (No pruning) (Numeric)	11.8	97.3	92.2	0.266	0.106	0.675
ADTree (Nominal)	0.4	99.9	94.0	0.262	0.200	**0.712**
ADTree (Numeric)	0.4	**100**	94.0	0.262	0.202	0.699

Table B.6: Evaluation results for Decision Tree classification models.

Table B.6 represents the results for all Decision Tree classification models that were applied to the dataset, as previously the most efficient statistic has been underlined. 'Pruned Decision Trees', discovered through the 'J48' classifier, performed poorly at predicting customers with a caravan policy. This was not improved through trials with various pruning settings, as illustrated by table B.7 the 'J48' models obtained a very low TP rate of between 0 and 3.8%. Furthermore, 'J48 error-based pruned trees' were found to be the least accurate at classifying positive samples, yet correctly classified 99.9% of the negative samples. This resulted in the highest amount of correctly classified samples overall with 94.9%, which is similar to the results of the 'BayesNet-TAN (Numeric)' model. There was, however, no variation in performance between subtree replacement or subtree raising in conjunction with error-based pruning as both models, since these models yielded identical results. In contrast the 'unpruned J48 (Nominal) Decision Tree' had the highest TP rate amongst the Decision Tree classifiers with 17.2%, along with the lowest mean absolute error rate of 0.099. Unlike with the 'BayesNet-Tan' classifier; a different classifier has the lowest root mean-squared error (RMSE) rate – 'REPTree (Numeric)' with 0.234. This suggests that the sizes of the prediction errors of the 'REPTree' model are larger than that of the 'unpruned J48 (Numeric)' model. The 'Id3' classifier had a slightly lower TP rate than 'J48 (No pruning)' with 16.5%. As stated previously the 'J48' an implementation of the 'C4.5' classifier which itself is an enhanced version of 'Id3'. Since the 'Id3' model only accepts nominal attributes, there are was no numeric results for this algorithm. 'ADTree (Numeric)' only predicted 0.4% of the customers who have a caravan policy however predicted all 100% of the customers who do not possess a policy. It also had the second-highest AUC statistic of 0.699. However the nominal 'AD Tree' model performed marginally better with an AUC statistic of 0.712. The final models that were discovered were those using Classification Rule learners (Production Rules). The Classification Rule learner models outputted IF-THEN rules were outputted amongst the results (figure B.38).

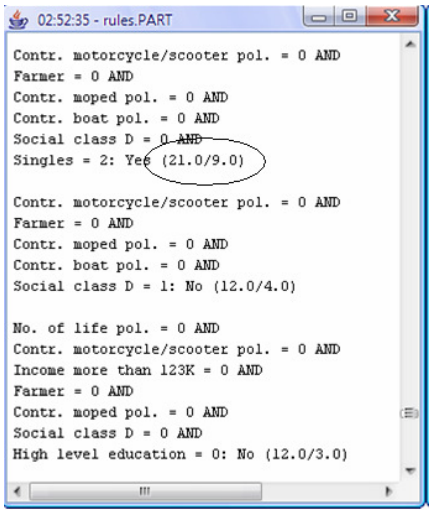


Figure B.33: Rules generated via the PART classification model.

The initial rule indicates that if the customers pays no contribution to a motorcycle/scooter policy, moped policy or a boat policy and presides in an area where the population of farmers and social class D is 0%, and 11-23% are single, then these customers will posses a caravan policy. The following the rule such as '(21.0/9.0)' (circled in figure B.33) indicates that 21 instances are classified by the rule, of which 9 are incorrectly classified.

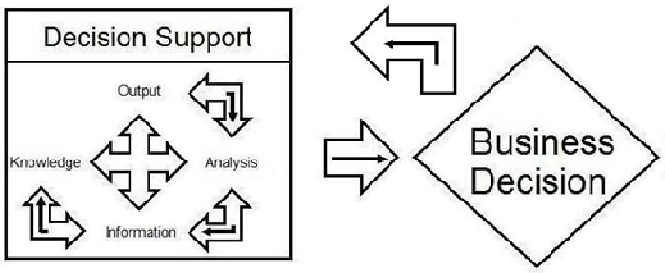


Figure B.34: Decision Support stage of KDDS-BI to provide business decisions.

As illustrated by figure B.34, the output must be further analysed, to extract information which can then be applied in context to provide decision support. A complete summary of the Classification Rule learner results is illustrated in table B.7. The 'unpruned PART classifier (Nominal)' has the highest TP rate of 19.3%. Whilst, 'JRip' and 'Decision Table' classifiers performed significantly worse than the 'PART' models, as illustrated by very low TP rates which range between 0% and 2.9%, despite the 'JRIP' Classification Rule learner performing well, when identifying those that do not posses a policy.

DecisionTable (Nominal)	0.4	99.7	93.7	0.240	0.111	0.528
DecisionTable (Numeric)	0.4	99.9	93.9	0.236	0.108	0.604
JRip (Reduced-error pruning) (Nominal)	0.8	99.9	94.0	0.237	0.111	0.504

JRip (Reduced-error pruning) (Numeric)	0	**100**	94.0	0.237	0.112	0.5
JRip (No pruning) (Nominal)	2.9	99.9	**94.1**	0.236	0.108	0.514
JRip (No pruning) (Numeric)	2.5	99.8	94.0	0.239	0.111	0.511
PART (Error-based pruning) (Nominal)	4.6	98.4	92.8	0.246	0.105	0.662
PART (Error-based pruning) (Numeric)	01.5	95.9	90.8	0.291	0.108	0.666
PART (Reduced-error pruning) (Nominal)	0	99.9	94.0	**0.235**	0.106	**0.670**
PART (Reduced-error pruning) (Numeric)	3.8	99.1	93.4	0.250	0.106	0.627
PART (No pruning) (Nominal)	**19.3**	92.8	88.4	0.317	0.112	0.571
PART (No pruning) (Numeric)	16.4	95.1	90.4	0.302	**0.099**	0.565

Table B.7: Evaluation results for Classification Rule learners.

Thus, once this preliminary analysis had been completed, the results and discoveries could be further investigated to extract valuable information that can be used within the context of providing decision support for direct marketing.

1	NaiveBayes (Numeric)	42.4	79.2	81.5	0.428	0.219	0.682
2	BayesNet - K2 (Nominal)	40.3	84.6	82.0	0.395	0.191	0.709
3	NaiveBayes (Nominal)	37.4	86.1	83.2	0.340	0.177	0.706
4	BayesNet - K2 (Numeric)	37	88	84.9	0.348	0.171	0.702
5	BayesNet -HillClimber (Nominal)	31.9	89.7	86.3	0.329	0.160	0.718
6	PART - No pruning (Nominal)	19.3	92.8	88.4	0.317	0.112	0.571
7	J48 - No pruning (Numeric)	17.2	95.5	90.8	0.282	0.099	0.588
8	Id3 (Nominal)	16.5	93.6	86.6	0.316	0.103	0.552
9	PART - No pruning (Numeric)	16.4	95.1	90.4	0.302	0.099	0.565
10	REPTree - No pruning (Nominal)	15.1	94.4	89.7	0.294	0.112	0.568

Table B.8: Top ten classification models based upon TP Rate.

Table B.8 illustrates the ranking of the top ten classification models, irrespective of classification technique. This ranking has been based upon the TP Rate, thus the number of customers that were correctly predicted to posses a caravan insurance policy. The ranking is dominated by Bayesian classifiers (1-5). The Bayesian classifiers obtained a success rate of at least 30% with regard to correctly classifying positive instances. However, variations amongst numeric and nominal models may be small for the same classifier, it still bears

relevance since the ranking appear interlaced. Thereby, confirming the notion of classification algorithms treating attributes dependent upon the data types despite the values being represented remaining the same. When descending down the rank from 1 to 10 the values of TN rate, CCS, RMSE and MAE generally decreases. Whereas, with regard to the AUC; the statistic steadily increases with the Bayesian classifiers. However, becomes much lower with the 'PART rule' learner and hovers above 0.5 thereafter. Figure B.35 illustrates multiple ROC curves for various models upon the same axis.

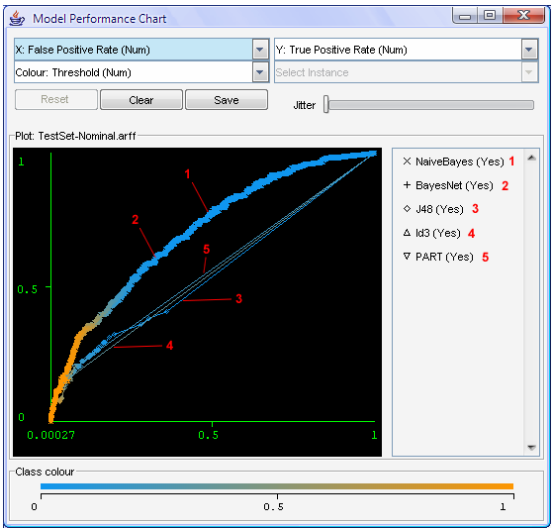


Figure B.35: Multiple ROC curves represented on the same axis.

Due to the class colours remaining the same resulting in closely plotted curves for 'NaïveBayes (Numeric)' and 'BayesNet - K2 (Nominal)' indistinguishable nevertheless, figure B.40 is able to illustrate the variation between the classification models and their associated ROC curves. 'NaïveBayes (Numeric)' and 'BayesNet - K2 (Nominal)' performed significantly better reflected by the arches toward the ideal position of the top-left corner, point (0, 1). Since Bayesian classifiers use probabilistic methods, thus each sample is assigned a probability of it being positive. This results in each instance becoming a data point and thereby subsequently becomes plotted on the ROC curve, consequently resulting in the vast visual difference to the curves of the other two types of classifier. In contrast to Bayesian classifiers, the 'J48 (Tree)', 'Id3 (Tree)', and 'PART (Rule)' algorithms, posses fewer data points since only the positive samples they correctly classified are used. The results show these to be minimal. This characteristic results in the 'curve', resembling a straight line. Thus the use of ROC curves in this scenario is far less useful at ascertaining which classifier performed better; when compared to using statistical measures. Consequently, the evaluation statistic provide a more accurate assessment of classifier performance

The AUC statistic will indicate the level of randomness among classifications for the positive samples. Results show that all classifiers had below 0.75 for AUC and majority between 0.5-0.6 suggesting that positive classifications were subject to a degree of randomness. Furthermore, 'NaïveBayes' has the best TP rate with 42.4%, yet also yielded the worst AUC result amongst the top five classifiers. In contrast, the 'BayesNet - HillClimber (Nominal)' model is ranked 5th highest on TP rate with 31.9% and in addition yielded the most

significant AUC statistic of 0.718 overall. Furthermore, the findings of table B.9 and the related measurements of TP rate and AUC which are similar for both the 'NaïveBayes' and 'BayesNet' classifiers and in turn higher than the other three classifier models support these results. Consequently, indicating that in comparison to other types of classifiers; investigating Bayesian classifiers to determine which customers to market a caravan insurance policy too will produce the best response rate.

Thus the highest ranked classifier was 'NaïveBayes' which correctly predicted 42.4% of customers who have a caravan insurance policy; the target attribute for the analysis, whilst 'BayesNet - HillClimber (Nominal)' model is ranked 5th highest on TP rate with 31.9% yielded the most significant AUC statistic of 0.718 overall, hence, most accurately classifying the customers who posses a caravan insurance policy. As a result, the evaluation of the statistical, measures not only provides information with regard to direct marketing, but enables the reliability and accuracy of the results to be assessed. Furthermore, by implementing a variety of classification models, a greater quantity of information can be discovered from the data set. However, it was the 'NaïveBayes' classification model which yielded the greatest TP rate and 'BayesNet – Hill Climber' which yielded the greatest AUC. These two classification models, can be further investigated to identify characteristics that can be explored to identify those customers that are likely to purchase a caravan insurance policy. Since, 'NaïveBayes' and 'BayesNet-Hill climber' have been found to yield the highest and most accurate results respectively, the models can be visualised (as graphs) from the results list to provide independent statistics that with regard to the attributes that were of the most significant interest (figure B.36).

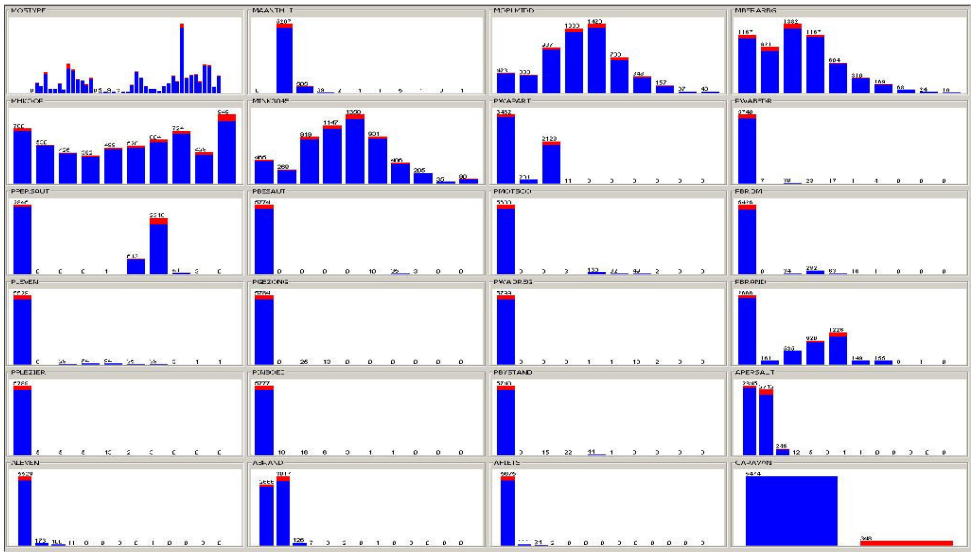


Figure B.36: Attributes individually visualised once the 'NaïveBayes' model had been investigated.

Once the models had been re-evaluated the following 7 attributes were amongst the most significant when assessing the criteria for those who purchase caravan insurance policies.

- PPERSAUT = Contribution to car policies.
- PBRAND = Contribution to fire policies.
- APERSAUT = Number of car policies.
- MOSTYPE = Customer Subtype.

- PPLEZIER = Contribution to boat policies.
- MINKGEM = Average income.
- PWAPART = Contribution to private third party insurance.

The above 7 attributes were found to be the most significant when both the 'NaïveBayes' and 'BayesNet- Hill Climber' models were re-evaluated and analysed. Although, there may be many reasons for the significance of these attribute as being key to defining the type of customer who is most likely to purchase a caravan insurance policy, the aim of this investigation is to discover these attributes so that they can be exploited for targeted marketing. The information extracted from the analysis provided the following knowledge that can be used for decision support. Customers who were found to contribute the most to car policies are also be likely to purchase a caravan policy. Furthermore, those customers that possessed the largest number of car policies were also a key demographic. However, in contrast, those that contributed a large amount to fire policies, boat policies and/or private third party insurance were unlikely to purchase a caravan policy. In addition, to the policy purchasing habits of key customers, the customer subtype also bared significance upon the likelihood of a customer purchasing a caravan insurance policy, it was found that 'double income no kids', 'Senior cosmopolitans', 'Middle class families' were amongst those most likely to purchase a caravan policy. Furthermore, 'Average income' was also found to be an attribute that was of significance, however, those who had very high or very low average income were found to not be likely to purchase a policy, in contrast to those customers that were found to earn around the median of average income. Consequently, these findings can be placed into formal reports such as this document and presented to stakeholders and/or decision makers, with regard to marketing policies. The application of KDDS-BI has provided a scientific method for finding key customer types and socio-economic demographics that should be considered when instigating a marketing campaign, and in turn it is these customer groups that should be most aggressively marketed.

B.3 Conclusion

This case study has endeavoured to investigate the applicability of KDDS-BI to discover valuable information to support the decision making process within the field of 'direct marketing'. Direct marketing is an interesting domain since it is one which represents a challenging area of marketing. In this age of global telecommunications, which enables even small scale retailer to compete globally, ensuring that consumers are aware of an organisations goods and services is of paramount importance. However, in contrast to traditional marketing strategies, which are conducted on a large scale. Due to the increase in the size of the potential market it is essential that organisations be able to identify targeting customers. Hence, if these key consumer segments can be identified, through predictive analysis, and targeted with explicit focused campaigns, an organisation can justifiably expect a significant return on investment (in marketing strategies), as proposed by Pareto's principle. This case study has demonstrated the applicability of KDDS-BI for direct marketing with regard to postal advertising campaigns, but can be just as effective with regard to telemarketing campaigns, or those conducted via the internet. The explicit domain, despite being explored with live data, serves an experimental simulation which can be extrapolated to any form direct marketing campaign. Live data consisting of socio-demographic and purchasing attributes of consumer was explored to identify target consumers. Through analysis of the data,

objectives were extracted based upon identification of the key properties of the attributes. Furthermore, this is data collected routinely by organisations, especially those which provide insurance policies. Accordingly, the cost of data collection to an organisation is negative, since it stands to benefit from stored data. Once the objectives had been identified the data could be further analysed using the structure provided by KDDS-BI to investigate techniques, requirements, and explore a conceptual model for the study.

Prior to analysing the result a mean through which the investigation of supervised techniques could be investigated was required. This tool was discovered in the form of the Brunel BI Explorer. Of more significance than the tool itself, is the techniques that were explored. Since the objectives of the investigation required a supervised approach. The performance of various classifiers could be analysed and compared to ensue that the reliability of the results could be as high as possible. This provided interesting results, since the variation between the accuracy and performance of the various classifiers could be compared thereby providing insight as to the circumstances in which these classifiers perform optimally. Therefore, the investigation provides not only a means through which to investigate this data set, but corresponding with the iterative nature of KDDS-BI. The investigation further provides a template that can be explored to assess future investigations. It has been the objective of this investigation to not identify individuals but rather a sub-set of customers who represent the segment of the total population most likely to purchase a policy. Given the details that were provided, it is only possible to identify the characteristics that identify these potential customers. However, if further study was to be conducted to discover exactly why these characteristics are key factors, then these identified customers, who are likely to buy a policy could be directly targeted with a marketing strategies such as a survey. Since these customers are likely to purchase a policy, they will consequently also be the most likely to respond to such a campaign, that could be based upon these findings. As a result, an investigation such as this not only provides the functionality to explore and integrate BI techniques into organisational practices. Yet, additionally displays emergent accuracy, with each iteration of the ‘Decision Support’ and ‘Business Decision’ phase of KDDS–BI the accuracy of the results can be expected to increase. Moreover, as illustrated by Figure B.34, the process of assessing the classification models against new data should be periodically repeated to ensure that the decision support can reflect market changes and trends, whilst providing an organisation with a competitive edge. The results of the investigation, however even with a single cycle of the framework, provided significant results which can provide substantial insight into consumer behaviour thereby substantially increasing the astuteness of judgement and decisions elected by decision makers. Consequently, these findings can be placed into formal reports and presented to stakeholders and/or decision makers, with regard to marketing policies. The application of KDDS-BI has provided a scientific method for finding key customer types and socio-economic demographics that should be considered when instigating a marketing campaign, and in turn it these customer groups that should be most aggressively marketed.

Appendix C: Case Study 2

Sales Promotion: Tesco

In order to investigate the advantages and disadvantages of the proposed framework; KDDS-BI, it will be explored with the aid of case studies. This case study will examine the performance of KDDS-BI, when investigated to increase the effectiveness of 'sales promotion'. The term 'Marketing Mix', defines the marketing strategy encompassing all elements of the planning and operation of marketing. Sales promotion, provides an interesting marketing strategy to investigate, since it can be combined with various other marketing strategies through BI, such as those intended to increase consumer loyalty.

C.1 Marketing Mix

Case Study 1 (Appendix B) defined marketing as the management process responsible for identifying, anticipating and satisfying customer requirements profitably. For this purpose, direct marketing was examined through the investigation of BI techniques. However, direct marketing provides one facet of the 'Marketing Mix'. The Marketing Mix defines the marketing strategy encompassing all elements of the planning and operation of marketing. The Marketing Mix, therefore, consists of activities such as branding, pricing, packaging, sales, distribution, advertising, sales promotions and marketing research. The Marketing Mix is frequently simplified so that it can be concisely defined as the 'Four P's' (Kotler et al, 2005):

- *Product*: The totality of 'goods and services' that are offered by an organisation to the target market. Thus, the product can be defined as anything, which an organisation can offer to a market for attention, acquisition, use or consumption, to satisfy a consumer want or need. Thus, a product can be a physical object, service person, place, organisation or idea.
- *Price*: The amount of money charged for a product or service. However further to an exchange of monetary value for the goods or services, the price can be defined by the sum of the values that a consumer exchanges for the benefit of having or using the product or service.
- *Place*: Corresponds to all organisational activities that facilitate the availability of the product to the target market.
- *Promotion*: Activities that communicate the product or service and its merits to the target market. Furthermore, the objective of promotion is not only to identify the product or service to the target market, rather simultaneously encourage the target audience to buy/use the product.

The 'Four P's' therefore represent the set of controllable, tactical marketing tools that an organisation can amalgamate to produce a desired response from its target market. Figure C.1 illustrates the 'Four P's' along with some of the marketing tools that can be investigated to achieve the desired outcome: influencing the target market.

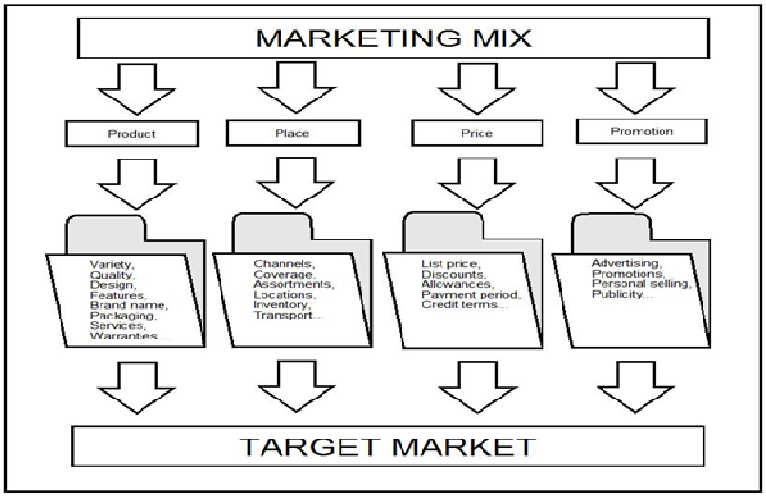


Figure C.1: The four P's of the Marketing Mix and associated marketing tools.

Figure C.1 illustrates how one or more of the 'Four P's' can be investigated for an effective marketing strategy to influence the behaviour of the target market. 'Price' is one of the Marketing Mix tools which an organisation can examine to increase sales and/or profit. However pricing decisions must be co-ordained with product design, distribution and promotion decisions to provide an effective and sustainable marketing strategy. Of the 'Four P's', price is a factor that can be considered for a faster response than product; place; or promotion, since these may require a greater amount of time for the effective implementation of a marketing strategy (Belch & Belch, 2007). Furthermore, prices can be set to retain the loyalty of customers and support of re-sellers; in addition prices may be manipulated to avoid intervention from governing bodies. As discussed pricing is a strategy often co-ordinated with various factors and as a result often infused with promotion, which like price, embodies a factor that can be more easily examined and varied than product or place. Consequently, prices can be reduced temporarily to enhance the sales or enthusiasm that surrounds a particular product or line of products in addition to drawing a greater level of customer into a store. Price and promotion is also frequently infused in order to increase the sales of two or more products through combined/cross-product promotion. This can be a very effective marketing strategy for organisations, this strategy of combining price and promotion to increase interest form the target market is known as 'Sales Promotion'.

C.1.1 Sales Promotion

Belch & Belch (2007) has defined sales promotion as:

"…a direct inducement that offers an extra value or incentive for the product to the sales force, distributors or the ultimate consumer with the primary objective of creating an immediate sale."

Sales promotion involves an inducement; as a result, this provides an additional incentive to the consumer to purchase a particular product. In contrast to 'above the line' promotion in which the advertiser pays an advertising agency to place the promotion. The inducement or incentive in sales promotion refers to short-term 'below-the-line' activities usually (although not exclusively) taking place at the point-of-sale to launch a product or maintain or increase sales. The terms 'below the line', refers to forms of non-media communication, or non-media advertising. Below-the-line activities include premium offers, flash packs, charity linked schemes, branded packs, on-pack offers, mail-ins, sampling, in-store promotions, prize contests, multi-buy offers amongst others activities. Many of these activities are intended to be subtle enough for the consumer to be unaware that the promotion is taking place, thereby seamlessly integrating into purchasing habits. These activities can be divided based upon the intended target market. Sales promotions may be intended toward consumers or trade, examples of such activities are illustrated in table C.1. Consumer-orientated sales promotion represents a 'pull strategy', hence, intended to operate in unison with advertising to encourage consumers to purchase a particular product, thereby creating a demand for that product. Furthermore, consumer promotions are employed by retailers to entice consumers to a particular store, or as defined by the Marketing Mix; 'place'.

Samples	Contests and dealer incentives
Coupons	Trade allowances

Premiums	Point-of-purchase displays
Contests	Training programs
Refunds/rebates	Trade shows
Bonus packs	Co-operative advertising
Price-off / Limited time offers	Push money (additional commission)
Multi-buy	
Loyalty programs	
Event marketing	

Table C.1: Examples of trade and consumer-oriented sales promotion activities.

Sales promotion is an established component of the marketing process. Yet, its role within the marketing process has dramatically increased over the past decade. The World Advertising Research Centre[14] and Joyce (2005) concur, that sales promotion spending has increased from $56 billion in 1991 to approximately $343 billion in 2005; by 2007 this figure had increased to $498 billon worldwide. In addition, marketers spend an estimated $150 billion annually on promotions targeted at retailers. Furthermore, as identified by the Institute of Sales Promotion[15], in the UK alone, 2007 witnessed a net increase of 8.3% in marketing budgets despite the economic down turn. It is estimated that sales promotion accounts for 60-75% of the expenditure of promotional budgets across retail industries (Belch & Belch, 2007). During 2007 companies trading consumer packaged goods devoted 74% of marketing expenditure on sales promotion. The rapid growth of sales promotion can be accredited to several factors (Kotler et al, 2005; Gilbert & Jackaria, 2002):

- Greater pressure upon managers to increase sales, particularly in consumer markets.
- Manufacturers are striving to increase market share.
- Sales promotion can provide an effective means through which a brand can be differentiated from competitors.
- The efficiency of above-the-line advertising has declined, due to rising costs, media clutter and legal restraints.
- Consumers have become more deal-oriented, in conjunction with a increase in large-scale retailers that are demanding increased incentives from manufacturers.

Sales promotion is considered an effective mean through which to obtain these objectives. Furthermore, sales promotion is considered an acceleration tool, hence, a strategy implemented with a view to increasing the rate of the selling process, in addition, to maximising sales volume, often at a particular location. However, it is imperative that organisations are able to efficiently discover the promotions that are likely to reap the maximum return for a product in addition to the locations (place) at which certain promotions are likely to attract consumers. Thus, to identify which ‘prices’ and ‘promotions’ are likely to work in unison with which ‘products’ in which ‘place’, new and novel technologies must be investigated. Such an investigation will facilitate the full potential of the Marketing Mix to not only be realised in the short-term, but furthermore, for sustainable Marketing Mix activities. Since, Marketing Mix activities are marketing strategies that rely upon the identification of consumer habits and purchasing trends, it is one that can be enforced and improved through BI

[14] http://www.warc.com: Accessed October, 2008.
[15] http://www.isp.org.uk: Accessed October, 2008.

strategies. Yet, BI can be explored not only to improve the efficiency and impact of sales promotion but also to asses the impact. Marketing strategies such as sales promotion and other 'below-the-line' activities are intended to be subtle enough for the consumer to be unaware of the promotion. This is due to effective 'below-the-line' strategies seamlessly integrating into purchasing habits. As a result, it can be difficult for an organisation to recognise the impact of sales promotion. Additionally the growing use of sales promotion has resulted in 'Promotion Clutter'. Promotion Clutter, similar to 'Media Clutter' and 'Advertising Clutter', implies that the consumer has been overwhelmed by the quantity of promotion/media/advertising, thereby becoming desensitised and requiring more novel targeted strategies.

BI techniques will permit the discovery of these trends, patterns and consumer requirements that cannot be discovered through conventional techniques, which may not be evident without investigation through BI. If organisations wish to identify the most effective Marketing Mix activity that will provide the greatest return in a particular location then it is imperative that the trends and behaviour of consumers be analysed to provide decision-makers with the detailed information required to execute the most effective marketing campaign whilst maximising the organisations competitive-edge.

C.2 KDDS-BI Case Study: Tesco

In the current digital age, multi-national organisations must compete with an ever-increasing number of competitors. Consequently, these organisations must discover novel and innovative methods to gain a competitive edge. Advertising alone is no longer sufficient, thus in an attempt to ensure repeat business from consumers many organisations have invested in loyalty schemes. Loyalty schemes endeavour to reward customers for repeat business by providing consumers with the opportunity to regain 'points' against purchases that can be redeemed, thereby providing further incentive for repeat custom. This is considered a 'below-the-line' marketing strategy. However, loyalty schemes whilst providing consumers with incentive to return for future purchases, simultaneously provides organisations with an additional benefit, since these whilst providing a consumer with loyalty points, information of the consumer, their purchasing habits and hopping patterns, thereby facilitating organisations to develop profiles of their consumers through socio-demographic information and purchasing traits.

Combined with reduced storage costs, this information can be collected and retained in vast quantities. Furthermore, this information rather than collected for any specific reason is collected for general marketing purposes. This information can be interrogated with a view to uncovering hidden trends within the purchasing habits of the consumer. Consequently, BI is a technology that can be investigated to uncover, examine and exploit the hidden information within this data. Thereby providing information to an organisation with regard to the promotions that are providing the desired return in addition to any details of which market groups should be targeted with which particular promotions. For this purpose KDDS-BI can be applied to an existing dataset, in order to provide a structured and scientific method through which to interrogate a dataset to provide decision makers with more accurate information for decision support within the retail sector.

C.2.1 Data Investigation

In the past decade, loyalty schemes have gained large scale recognition within the retail sector, especially as a strategy through which return custom can be ensured and an organisations brand differentiated from competitors, in addition to, established as a unique preferred brand. American Airlines was amongst the first major companies that implemented a loyalty scheme with its frequent flyer program in 1981. However, since then, the concept of the loyalty scheme has gained widespread recognition. Supermarkets have been at the forefront of this trend with almost all large supermarkets employing such schemes. The UK supermarket sector is dominated by Tesco, Asda, Sainsbury's and Morrisons which are the only chains which operate full-scale superstores of 40,000 square feet or more. Of the major supermarkets in the UK, only Tesco and Sainsbury's offer a loyalty card-scheme to customers. Tesco's Clubcard scheme has been operating since 1995 and has now become the largest loyalty card in the UK, with around 13 million active Clubcard holders[16]. The Tesco Clubcard scheme reward consumers with points for purchases that are then redeemable in the form of coupons against future purchases. The Clubcard scheme has proven to be an overwhelming success for Tesco's, the success of the scheme has resulted in rival supermarkets such as Asda accepting Tesco Clubcard vouchers to be redeemed within Asda stores.[17] Consequently, it is imperative that Tesco's ensures that its consumers needs are directly met, providing consumers with customised offers and coupons that will ensure that Tesco remain at the forefront. In addition this must be supplemented with in-store offers that meet the needs and requirements of the consumers within explicit regions thereby ensuring that the coupons will not be redeemed in a rival store. Since BI is a technology that can be investigated for these requirements a dataset will be investigated through the application of KDDS-BI, to examine the information that can be discovered through the application of BI technology. Furthermore, this information will be additionally investigated to provide more focused decision support providing valuable information upon which to guide the future actions of an organisation with a view to gaining a competitive-edge over rivals.

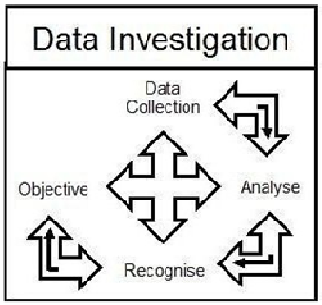


Figure C.2: Data Investigation stage of KDDS-BI.

The initial phase of KDDS-BI is 'Data Investigation' (figure C.2). In accordance, the data will be collected and analysed to discover the opportunities that are available for further investigation. Consequently, a dataset has been collected in conjunction with Dunnhumby[18]. Dunnhumby not only co-developed the Tesco Clubcard

[16] http://www.Tesco.com: Accessed October, 2008.
[17] http://wwwAsda.com: Accessed October, 2008.
[18] http://www.dunnhumby.com: Accessed October, 2008.

scheme, but has further established the Dunnhumby Academy of Consumer Research. The Academy established in April 2005, has been motivated by the opportunity to provide a mechanism for farmers and farmer controlled businesses to access the Tesco Clubcard database. The database is managed by Dunnhumby, on behalf of Tesco and contains unique customer insight from the purchasing behaviour of over one million households in the UK. To further this research a dataset extracted from this database will be interrogated and analysed. The dataset details the sales of 43 fresh meat products; 20 Beef; 13 Lamb: and 12 Pork. The dataset details the sales of these products for a 104 week period, dating from the 29/05/06 till 14/01/08. During this period Tesco's are running two promotions; temporary price reduction and a multi-buy promotion. Containing 4300 instances, each instance describes the sales figures individually for the 43 fresh meat products on for the week according to socio-demographic details of the consumer. Consequently, each instance is illustrated via 81 attributes. Table C.2 provides an overview of the attributes contained within the dataset and the value descriptors associated with these attributes; a full version of this table can be found in Appendix D. The dataset provides information regarding the sales of fresh meat products during a particular week, in addition to information regarding any offer that may have been associated with the given product. The data set also provides information detailing socio-demographic information associated with the consumers, thus specifying the life stage, and region of the consumer. The sales figures are numeric, however for all other attributes the numeric attribute corresponds to a given label, thereby resulting in nominal values. The data set, however, contains no missing data.

Given the information that is provided by the dataset, the dataset can be interrogated for a variety of purposes. Consequently, the objective for the application of KDDS-BI is to facilitate decision makers in identifying the opportunities that are available to Tesco's with regard to effective marketing and ensuring that consumer segments are targeted appropriately given their life stage and region. Furthermore, it would be of significant interest to analyse whether there is any differentiation between the effectiveness of sales promotions in particular regions or consumer segments. This will ensure that future marketing campaigns not only target the correct individuals, but the nature and details within the campaign are tailored to the specific attributes of potential customers. Correspondingly, the objectives can be identified as:

- Identify region with lowest sales.
- Identify the level of impact of promotions on fresh meat sales.
- Explore opportunities for sales promotions to enhance sales.
- Identify which products are purchased in together and therefore benefit from joint sales promotions. This information can further be investigated when determining the most effective store/shelf layout, thereby placing products with strong relationships in close proximity.
- Ensure that the results provided are accurate and can, therefore, be proposed as scientifically valid results.

1	Week beginning (date)	1-86	1=29-May-06 2=05-Jun-06 3=12-Jun-06 4=19-Jun-06 5=26-Jun-06 6=03-Jul-06 7=10-Jul-06 8=17-Jul-06 9=24-Jul-06 10=31-Jul-06 11=07-Aug-06 12=14-Aug-06 13=21-Aug-06 14=28-Aug-06 15=04-Sep-06 16=11-Sep-06 17=18-Sep-06 18=25-Sep-06 19=02-Oct-06 20=09-Oct-06 21=16-Oct-06 22=23-Oct-06 23=30-Oct-06 24=06-Nov-06 25=13-Nov-06 26=20-Nov-06 27=27-Nov-06 28=04-Dec-06 29=11-Dec-06	30=18-Dec-06 31=25-Dec-06 32=01-Jan-07 33=08-Jan-07 34=15-Jan-07 35=22-Jan-07 36=29-Jan-07 37=05-Feb-07 38=12-Feb-07 39=19-Feb-07 40=26-Feb-07 41=05-Mar-07 42=12-Mar-07 43=19-Mar-07 44=26-Mar-07 45=02-Apr-07 46=09-Apr-07 47=16-Apr-07 48=23-Apr-07 49=30-Apr-07 50=07-May-07 51=14-May-07 52=21-May-07 53=28-May-07 54=04-Jun-07 55=11-Jun-07 56=18-Jun-07 57=25-Jun-07 58=02-Jul-07	59=09-Jul-07 60=16-Jul-07 61=23-Jul-07 62=30-Jul-07 63=06-Aug-07 64=13-Aug-07 65=20-Aug-07 66=27-Aug-07 67=03-Sep-07 68=10-Sep-07 69=17-Sep-07 70=24-Sep-07 71=01-Oct-07 72=08-Oct-07 73=15-Oct-07 74=22-Oct-07 75=29-Oct-07 76=05-Nov-07 77=12-Nov-07 78=19-Nov-07 79=26-Nov-07 80=03-Dec-07 81=10-Dec-07 82=17-Dec-07 83=24-Dec-07 84=31-Dec-07 85=07-Jan-08 86=14-Jan-08
2	ID	1-50	1= Older families & Scotland 2= Older families & East England 3= Older families & London 4= Older families & Midlands 5= Older families & North East 6= Older families & North West 7= Older families & South West 8= Older families & South & South East 9= Older families & Wales & West Country 10= Older families & Yorkshire 11= Older Adults & Scotland 12= Older Adults & East England	28= Pensioners & South & South East 29= Pensioners & Wales & West Country 30= Pensioners & Yorkshire 31= Young Adults & Scotland 32= Young Adults & East England 33= Young Adults & London 34= Young Adults & Midlands 35= Young Adults & North East 36= Young Adults & North West 37= Young Adults & South West 38= Young Adults & South & South East	

			13= Older Adults & London 14= Older Adults & Midlands 15= Older Adults & North East 16= Older Adults & North West 17= Older Adults & South West 18= Older Adults & South & South East 19= Older Adults & Wales & West Country 20= Older Adults & Yorkshire 21= Pensioners & Scotland 22= Pensioners & East England 23= Pensioners & London 24= Pensioners & Midlands 25= Pensioners & North East 26= Pensioners & North West 27= Pensioners & South West	39= Young Adults & Wales & West Country 40= Young Adults & Yorkshire 41= Young Families & Scotland 42= Young Families & East England 43= Young Families & London 44= Young Families & Midlands 45= Young Families & North East 46= Young Families & North West 47= Young Families & South West 48= Young Families & South & South East 49= Young Families & Wales & West Country 50= Young Families & Yorkshire
3	Life stage of Customer	1-5	1=Older families 2=Older adults 3=Pensioners 4=Young adults 5=Young families	
4	Region	1-10	1= Scotland 2=East England 3=London 4=Midlands 5=North East	6=North West 7=South West 8=South & South East 9=Wales & West Country 10=Yorkshire
5-42	Weekly sales per Beef product / corresponding offer	Numeric / 0-1	0 = No Promotion 1 = Promotion	
43-60	Weekly sales per Lamb product / corresponding offer	Numeric / 0-1	0 = No Promotion 1 = Promotion	
61-85	Weekly sales per Pork product / corresponding offer	Numeric / 0-1	0 = No Promotion 1 = Promotion	

Table C.2: Attributes and values description.

C.2.2 Data Modelling

Once the possible objectives have been established, the investigation can enter the second phase 'Data Modelling'. The primary deliverable of this phase will be the modelling of the solution with a view to discovering the technical requirement for a successful investigation thus the requirements of the objectives must be examined to discover the BI strategies that are available and those that will ensure the objectives can be successfully achieved.

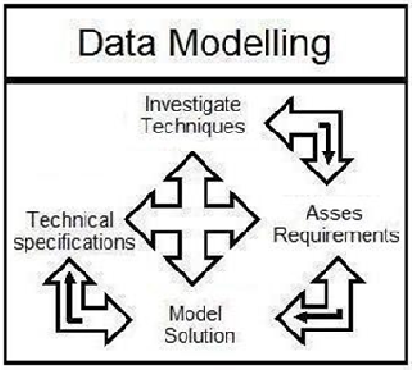


Figure C.3: Data Modelling stage of KDDS-BI.

The objectives defined during the Data Investigation phase must be explored to discover which techniques can most efficiently be investigated to realise these objectives successfully, so that the results can be studied to provide significant decision support. As discovered during the 'Data Investigation' phase of KDDS-BI, the information encapsulated within the dataset represents various details regarding customers. Each instance of the dataset provides socio-demographic information in addition to information pertaining to various types of fresh meat products and promotions associated with these products. This is therefore an exploration of data that will reap maximum benefit through deep investigation of the data via advanced analytics, especially those that can employ intelligent analysis techniques. This has been determined as the most suitable approach, since it is the correlation between various attributes and the strength of these relationships that is of paramount interest.

Clustering models and Association Rule Mining (ARM), in addition to Classification, are various advanced analytic techniques which can be further investigated to facilitate in the realisation of the objectives of this investigation. Hence, in order to achieve the specified objectives, a combination of techniques must be explored. The theory, underpinning these techniques has been studied in the literature review. Furthermore, as discussed in the Literature Review (Chapter 2), Clustering is an unsupervised learning technique; consequently there are no dependent variables. In contrast to supervised learning, the objective of Clustering techniques is to discover concept structures and relationships hidden within the dataset. There are a number of explicit algorithms and techniques that can be further investigated to realise the investigation objectives. Classification techniques such as Decision Trees (reviewed in Appendix A) and Neural Networks can be examined since these techniques can be studied to find relationships between attributes. Since there is no specific target variable i.e. a particular variable to be discovered, the focus is to find patterns and trends that may be hidden within the data. For this reason a combination of Clustering and ARM will be explored (See Appendix A for technical details relating to these techniques). Clustering techniques such as K-Means clustering and the EM algorithm, which have proven to be a proficient technique for data clustering in machine learning applications, furthermore, the unsupervised clustering is a technique that can aid the discovery of deep, hidden data as required by this study. Another technique that will be applied is Association Rule Mining (ARM). Unlike K-means Clustering and the EM algorithm, the Apriori algorithm is one which can be investigated for ARM. As discussed in the literature review, ARM can be investigated to discover interesting relationships between attributes within a dataset, however, unlike 'Production Rules', 'Association Rules' can return one or several output attributes.

To successfully accomplish the objectives of this investigation, it will be required to employ both supervised and unsupervised analysis techniques; such as Artificial Neural Networks, Clustering, ARM and Decision Trees. Thus, the investigation will require a novel and innovative combination of these techniques. After having explored a selection of these techniques, a conceptual model for the realisation of the objectives can be investigated. Since the analysis will require a combination of techniques to cross-analyse the data, the data will have to be extended to include training sets and test sets. Moreover, dependent upon the objective it will be necessary to separate particular attribute variables, such as regions, meat varieties etc. Since this can be a time consuming endeavour that given no explicit output variables will be subject to change as there is no specific output variable. Consequently, support for the efficient realisation of select datasets is a feature which should be observed when investigating a suitable tool. The requirement to cross examine the data lends to the modelling of a relational database. A relational database can be created through explicit tables for the meat varieties, and demographic attributes, these tables can then be further analysed using SQL, furthermore, this will provide a greater level of functionality when maintaining the integrity of the data. Consequently the data can be modelled through an entity-relationship diagram (figure C.4).

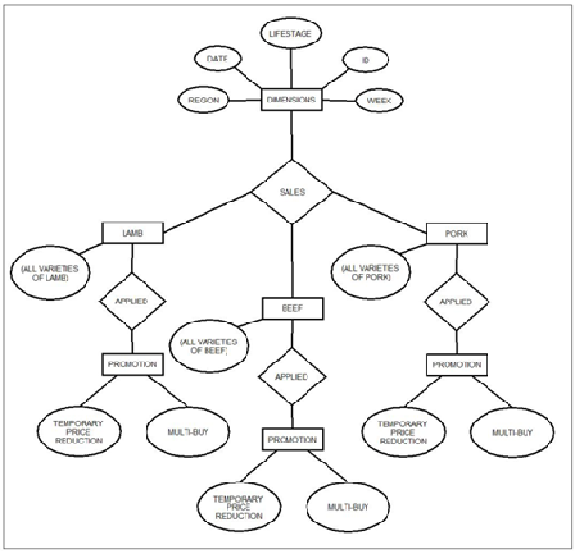


Figure C.4: Entity-Relationship diagram for data set.

Through examination of these relationships it was determined, that due to the absence of dependent variables in conjunction with the size of the dataset the objectives will have to be fulfilled through the combination of techniques. This combined approach to problem solving will allow for the results to be cross-analysed, thereby ensuring they are valid with a high level of accuracy. To ensure a focused investigation, the objectives identified during the data investigation stage have been focused toward the lowest performing region. The motivation for selecting a single region to focus upon is that this will permit the investigation and objectives to gain greater significance since the results can be focused and therefore more intimately scrutinised. In addition, since the dataset contains an identical number of instances for each region, the region selected bears no real significance rather, ensures that the theory and procedure underpinning the process can be demonstrated without out

compiling substantial results that do not interrogate the data at a deep level. The objectives can henceforth be redefined as:

- What is the worst and best performing area in terms of sales?
- In the both the best and worst performing region, which is the largest and smallest consumer and what are the products they purchase the most?
- Is the most product purchased the most / least by the highest / lowest consumer group the same in each area or does this vary for each area, i.e. is it a case of a particular group just being a high consumer or is there a particular trend in an area which can be replicated in others.
- Is this product consistent with the overall, nationwide highest selling product?
- What products can be cross promoted with the highest selling product?
- What is the impact of all sales promotion techniques on this product and is one sales promotion technique more successful?
- What is the overall impact of sales promotion (nationwide) and who is the consumer who sales promotion has the smallest and largest impact on (All products)?
- For both types of sales promotions which consumer responds best and worst to which promotion (lowest performing region)?

The objectives have been defined with a view to increasing sales through sales promotions and analysing these implications to promote profit maximisation. Although, the objectives can be extended for a nation wide study, this is not necessary to demonstrate the suitability of the KDDS-BI framework and is therefore, beyond the scope of this research. To ensure that the decision support provided via the results of this investigation can be justified a means of validation is required. This verification will be achieved through the analysis of each objective through multiple techniques, thereby permitting the results to be validated through a unified conclusion. In addition ROC curves and if appropriate 10 fold cross-validation (both of these techniques explored in Appendix B) can be employed and analysed to discover the overall accuracy of the models to ensure that the decision support provided can be scientifically validated.

Given the limited nature of the domain within which the solution must operate, namely a desktop PC of a decision maker the technical requirements must be assessed. The software whilst capable of handling large data sets must not require significantly large processing power. For this reason, a mid-specification desktop PC has been selected as a platform:

- Intel Celeron 440 Processor (2 GHz. 800 MHz FSB, 512 KB Cache),
- Windows Vista Home Edition,
- 1 GB RAM,
- 80 GB HDD.

The specification of the PC has been selected to ensure that results can be obtained without necessitating a

complex hardware infrastructure. Furthermore, the relatively low cost of the PC will reflect the resources available to the majority of decision-makers. Thereby, providing a suitable solution that can be integrated in many (if not all) working environments at a relatively low cost.

C.2.3 Development

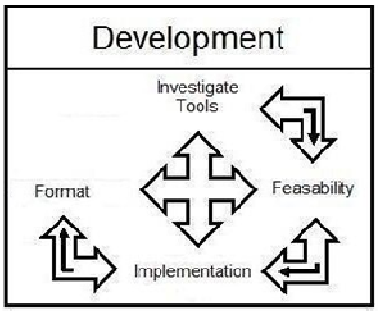


Figure C.5: Development stage of KDDS-BI.

The 'Data Investigation' phase has explored the requirements, techniques and specifications that must be observed if the objectives of this investigation are to be attained, in addition to a conceptual model that will provide greater insights into the nature of the data. It is now possible to examine various applications that can facilitate in the realisation of the objectives. Thus, enabling the raw data to be analysed and informative discoveries uncovered that will permit decision makers to direct business objectives with a greater competitive edge. There are various software solutions which are available from a number of vendors reviewed in Appendix A: Section A.5-Table A-. In order to find a suitable BI solution for this investigation, a number of these packages were investigated. Once the various software packages that have been identified, carefully considered and tested it was discovered that a combination of commercial platforms would provide the functionality and algorithms, which could be further investigated to realise the objective of this investigation. Table C.3 illustrates a subsection of Appendix A: Section A.5-Table A-1, and provides insight into the key features of the Microsoft applications. The combination of applications provides a BI solution that can be investigated for the realisation of the objectives of this investigation.

Microsoft	Business Intelligence Development Studio	Commercial	Microsoft Visual Studio 2008 with additional project types that is specific to SQL Server. Business Intelligence Development Studio is the primary environment for the development of business solutions that include Analysis Services, Integration Services, and Reporting Services projects. Each project type supplies templates for creating the objects required for business intelligence solutions, and provides a variety of designers, tools, and wizards to work with the objects.
Microsoft	Excel & Microsoft SQL Server Data Mining Add-Ins	Commercial	Microsoft Excel forms a part of the Microsoft Office Package providing spreadsheet based features along with analytical functionality to users. Microsoft Excel can be explored in conjunction with Microsoft SQL Server Data Mining Add-Ins for realising BI objectives. Microsoft SQL Server Data Mining Add-Ins for Office 2007 is a set of easy to use data

			mining capabilities that enable predictive analysis at every desktop. Being able to harness the highly sophisticated data mining algorithms of Microsoft SQL Server 2005 Analysis Services within the Microsoft Office, business users can easily gain valuable insight into complex sets of data with just a few mouse clicks. The Data Mining Add-Ins for Office 2007 facilitates end users to perform advanced analysis directly from within Microsoft Excel.
Microsoft	Excel Pivot Tables	Commercial	One of the most powerful features in Microsoft Excel and provided free. Pivot tables are a way to extract data from a long list of information, and present it in a readable form and facilitate the analysis of data. Enabling users to create multidimensional data views by dragging and dropping column headings to move data around, Pivot tables are especially well-suited to the task of taking enormous amounts of data and summarising that data into useful reports.
Microsoft	SQL Server Enterprise Edition	Commercial	Provides a complete end-to-end BI platform. The additional features over previous incarnations of SQL Server form the core of the BI solution. Contained within the database enhancements these are bundled free with the database licence. The Reporting Services (in its second release) is provided in two forms 'Developer' and 'Builder'. Developer provides tools to create reports that are developed via technical developers. Whilst Builder is provides tools for business analysts to write reports. Furthermore new management tools simplify administration, whilst ETL tools have been improved and re-architected as Integration Services. New algorithms have been added to the data mining functionality. Furthermore, embedded within the database is a .NET Common Runtime Library to expand the options available to developers building BI applications. However, as a BI solution, SQL Server 2005 EE is lacking in end-user accessibility applications and limited business analytics. Microsoft has addressed these issues with Business Scorecard manager, which provides a score carding and dashboard framework. In addition to Office 2007, this has a number of BI elements that have been focused toward end-user accessibility.
Microsoft	SQL Server Management Studio	Commercial	An integrated environment for accessing, configuring, managing, administering, and developing all components of SQL Server. SQL Server Management Studio combines a broad group of graphical tools with a number of rich script editors to provide access to SQL Server to developers and administrators of all skill levels. SQL Server Management Studio combines the features of Enterprise Manager, Query Analyzer, and Analysis Manager, included in previous releases of SQL Server, into a single environment. In addition, SQL Server Management Studio works with all components of SQL Server such as Reporting Services, Integration Services, and SQL Server Compact 3.5 SP1. Developers are consequently provided with a familiar structure, whilst database administrators get a single comprehensive utility that combines easy-to-use graphical tools with rich scripting capabilities.

Table C.3: Subsection of Appendix A: Section A.5-Table A-1.

Although, the platforms identified in table C.3 are commercial, hence, requiring a greater cost then that which is associated with an open-source platform. It was determined that the functionality provided by this combination of platforms, resulted in the additional cost being justifiable. Since the objectives of this investigation must be satisfied using a multitude of BI techniques it was no single stand alone software application provided the full

functionality, hence the number of applications selected that must be configured and implemented unanimously. Thus the proposed BI solution will amalgamate the following platforms:

- .Net Framework,
- Business Intelligence Development Studio,
- Microsoft Excel,
- SQL Server 2008,
- SQL Server Data Mining Add-Ins for Excel,
- SQL Server Management Studio,
- SQL Server Visual Studio.

The limited nature of the technical requirements, in addition to the number of techniques required and the size of the dataset resulted in the combination of these applications providing the most suitable means through which to further this investigation. Furthermore, as discussed in the 'Model Solution' sub-step, this combinatorial platform will permit the analysis of a relational database at a deep level. Thereby, facilitating a high degree of data manipulation to ensure an in-depth analysis of the data to uncover hidden, meaningful information and knowledge resulting in more focused business decisions for future activities. Once the software had been investigated and selected the various applications had to be downloaded and installed. Available as downloads from the Microsoft Developer Network website[19] (figure C.6).

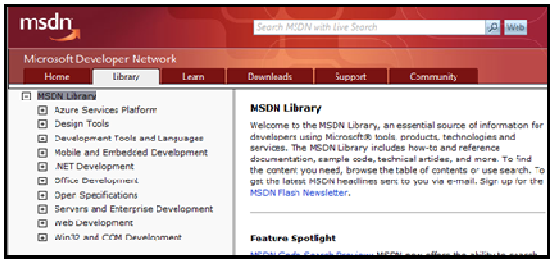


Figure C.6: MSDN Homepage

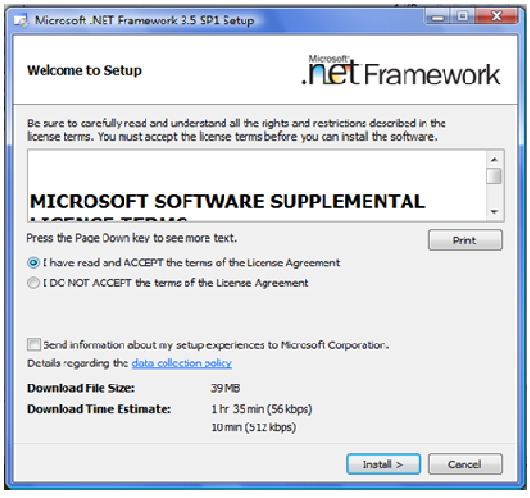


Figure C.7: Software download.

[19] http://msdn.microsoft.com/en-gb/default.aspx: Accessed January, 2008.

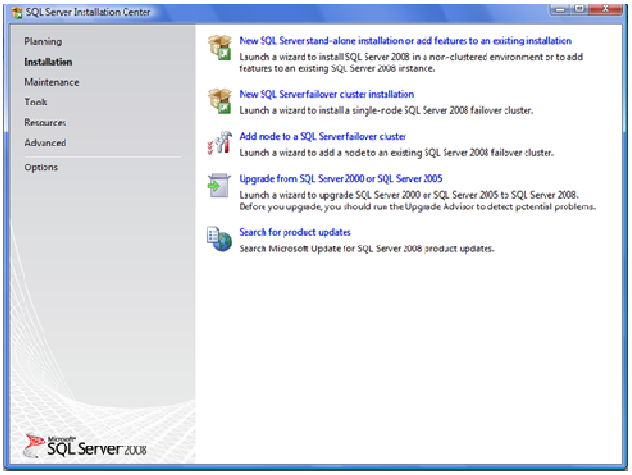


Figure C.8: Software installation.

Once the appropriate platforms had been downloaded (figure C.6), they required to be installed (figure C.8) and configured. The process of configuration required for all applications to be configured with Microsoft SQL Server and the data mining add-ins were required to be configured with Microsoft Excel. Once configured the dataset could be inspected to investigate any formatting requirements. Fundamental to the formatting of the dataset was the explicit definition of a primary key (figure C.9) a pre-requisite for analysis through advanced analytics. Consequently, due to the dataset not containing a unique identifier, all instances were not being recognised. Initially it was endeavoured to overcome this through the insertion of a composite primary key, thereby retaining the original structure of the dataset. However, due to Microsoft Business Intelligence Development Studio not providing support for this function, the dataset was modified through SQL Server Management Studio, with the addition of a column named 'MeatID' which provided a primary key for each instance of the dataset (figure C.10). Since the combination of software provides the functionalities for only particular attributes to be analysed from within the dataset, the dataset had to be defined into specific tables. This was realised through the creation of a relational database, which enables the attributes to be optimally distributed for analysis (figure C.11). Hence, once configured the dataset could then be investigated and interrogated to obtain output that can be analysed.

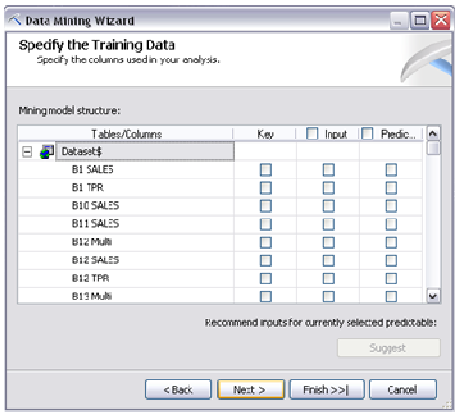


Figure C.9: Unique identifier required.

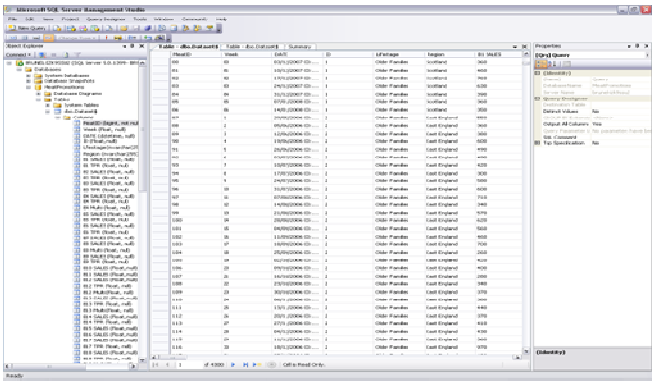


Figure C.10: Primary key defined for the dataset.

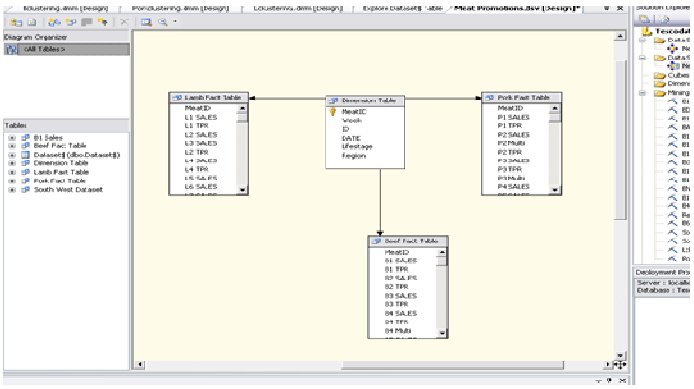


Figure C.11: Subset of tables that compile the relational database.

Once the initial formatting had been completed, the data set had to be further pre-processed, since the dataset must be loaded and selected within SQL Server Business Intelligence Studio. As illustrated by figure C.12 the data source must first of all be declared, once declared a suitable data mining structure can be explored. The data mining structure will permit the investigation of the dataset, whilst ensuring the appropriate technique (figure C.13) is applied to the correct relational database (figure C.14). Having selected the database, the tables that contain the specified key attributes must be declared in addition (figure C.15) to the attributes that are to be analysed (figure C.16). It should be noted, that the described process of selecting the technique, table and attributes is dependent upon the analysis that is to be completed, and as a result must be repeated in the event that further analysis is to be conducted using varying techniques or data selection.

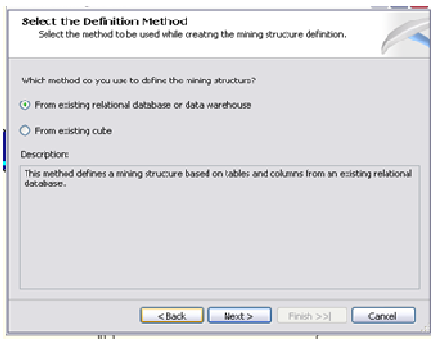


Figure C.12: Define Data source.

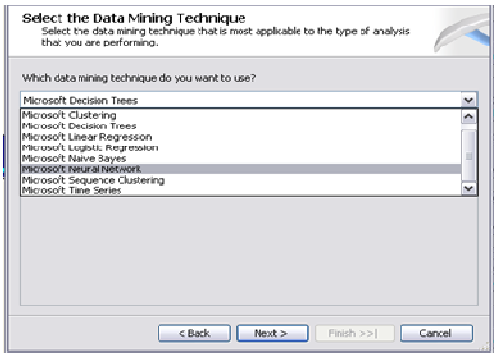


Figure C.13: Select data mining technique.

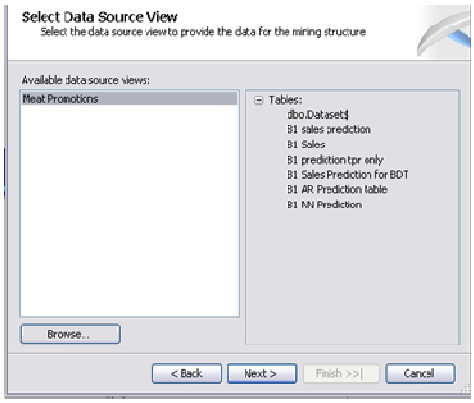


Figure C.14: Select the relational database that contains the require data.

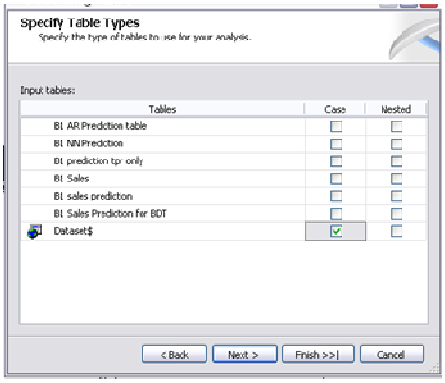


Figure C.15: Select the table that is to be analysed from within the relational database.

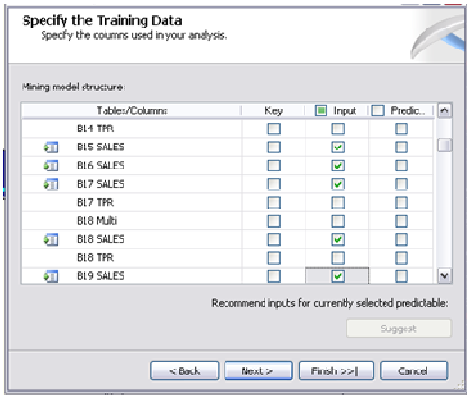


Figure C.16: Select which attributes to analyse and predict.

C.2.4 Decision Support

With the data had been transformed into the correct format to that which is required by the combination of software applications. The data could then be interrogated to discover hidden information that can be exploited for decision support. The initial stage of the 'Decision Support' phase of KDDS-BI is to gather the output by applying the BI techniques to the dataset (figure C.17). Thus, the dataset will be investigated using a variety of BI techniques. The result of these models can then be analysed to extract valuable knowledge.

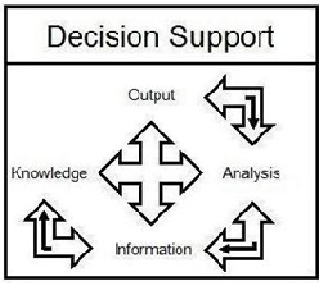


Figure C.17: Decision Support stage of KDDS-BI to provide business decisions.

The initial analysis that was conducted endeavoured to investigate how sales could be increased in the least optimally performing region. Hence, once the appropriate attributes had been selected (figure C.18), a Decision Tree was explored which provided details as to the performance of the regions through sales figure. The analysis was initially conducted using only Beef products, since it is this meat type that contains the highest sales figures and most variety within the dataset.

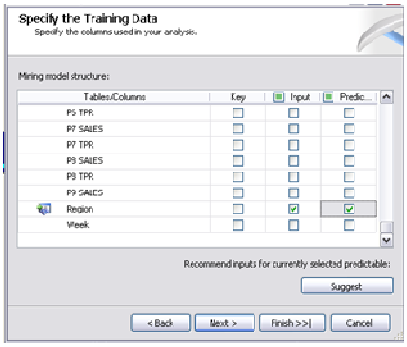


Figure C.18: Attribute selection.

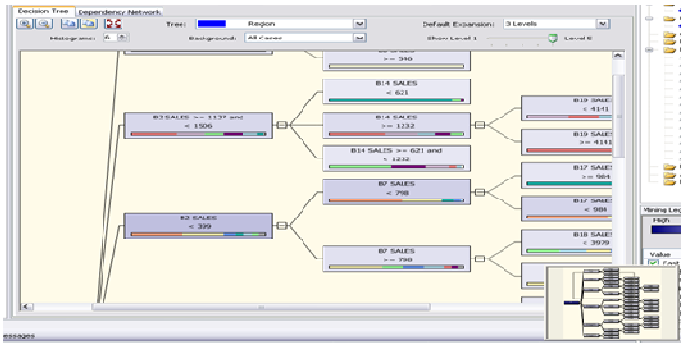


Figure C.19: Decision Tree detailing sales figures for Beef by region.

Analysis of the Decision Tree led to the discovery that it is the South West region that is most poorly performing. Consequently, in order to examine whether sales promotions could be employed to increase sales, a Decision Tree was created that would facilitate the analysis of the region individually (figure C.20). In order to further analyse this hypothesis, in addition to verifying the findings, the data was examined through Clustering techniques (figure C.21) and Neural Networks (figure C.22). Verifying the findings of the Decision Tree, it was discovered that the South West region has the lowest sales figures for Beef products, accounting for 70% of the sales in the lowest boundary: 0-168 (figure C.23). Consequently, to analyse the opportunities that could be explored for increasing these figures the nationwide impact of sales promotions was investigated using statistical analysis tools with Excel (figure C.23).

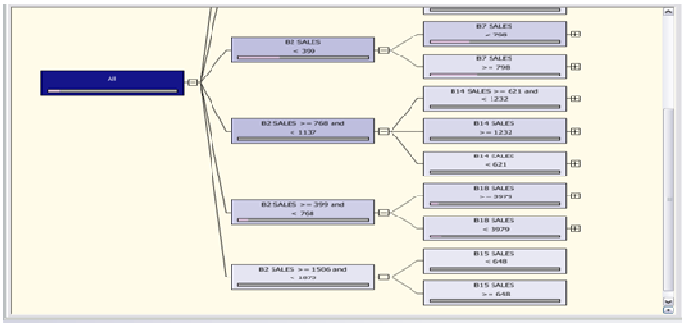


Figure C.20: Decision Tree illustrating sales figures for South West region only.

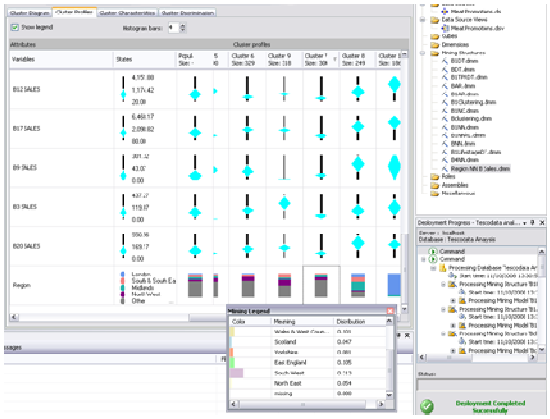


Figure C.21: Cluster analysis of the sales figures of each region.

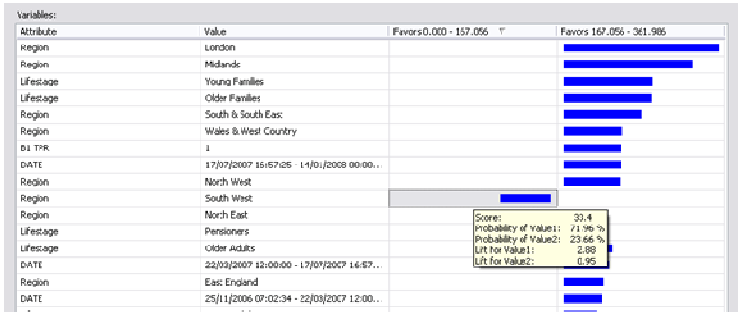


Figure C.22: Neural Network analysis of lowest sales region.

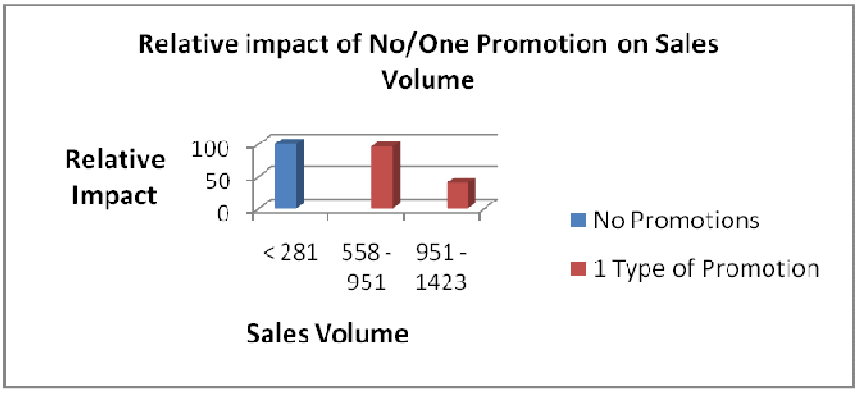


Figure C.23: Statistical analysis for the impact of sales promotions.

The statistical analysis of the impact of sales promotions illustrated that there is a significant increase in sales figures. Consequently, to investigate the impact of sales promotions as a means through which sales in the South West could be increased further analysis upon the Decision Tree, was conducted within SQL Server Business Intelligence Studio to uncover the type of Beef that sells the least within the South West. The analysis unveiled that it is 'Specially Reared/Organic Fry/Grill Beef', labelled B6 within the data dictionary. Once the lowest performing type of Beef in the region had been found, the effect of sales promotions within the region could be analysed. This investigation was conducted to discover if past sales figures have been improved through the utilisation of sales promotions (figure C.24). Figure C.25 illustrates the mining structure that was created to asses the impact of sales promotion within each region.

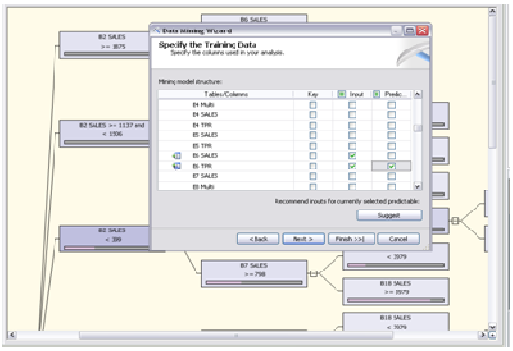


Figure C.24: Impact of sales promotions upon the least optimal Beef variety.

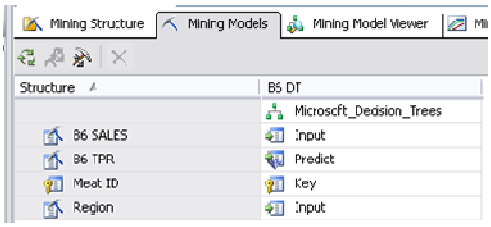


Figure C.25: Mining structure for the analysis of the impact of sales promotions by region.

Figure C.26 illustrates the Decision Tree that was created using the mining structure (figure C.25). This Decision Tree facilitated the examination of the impact of sales promotions within the South West, studying the figures for the average sales and cross-analysing these findings with Neural Network analysis, it was discovered that there is an 77.78% increase in sales of 'Specially Reared/Organic Fry/Grill Beef' in the event that a sales promotion is activated (figure C.26), as illustrated in figure C.27b it is the statistic for value 2 that is most significant since this implies that a temporary price reduction promotion is active. Figure C.28 illustrates a scatter plot lift chart that can be analysed to asses the accuracy of the model in a similar fashion to a ROC curve. The closer the plotted points are to the diagonal line the higher the more accurate the model can be considered.

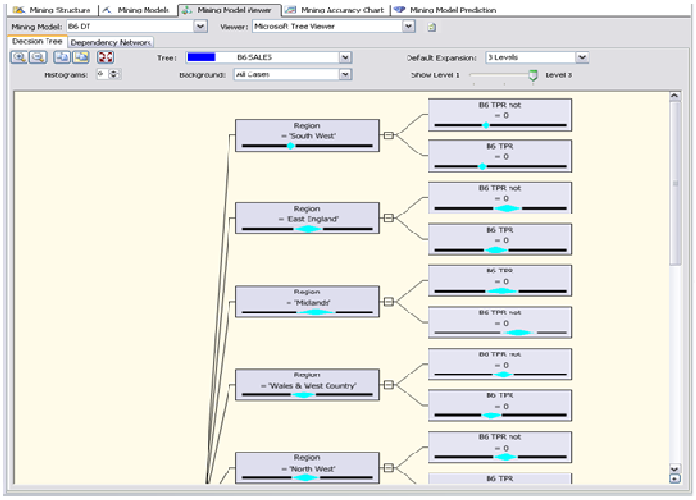


Figure C.26: Impact of sales promotions by region.

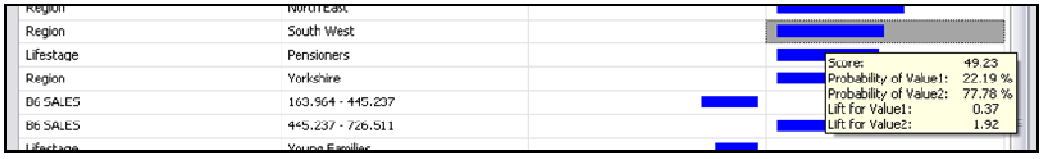


Figure C.27(a): Impact of temporary price reduction using Neural Network analysis.

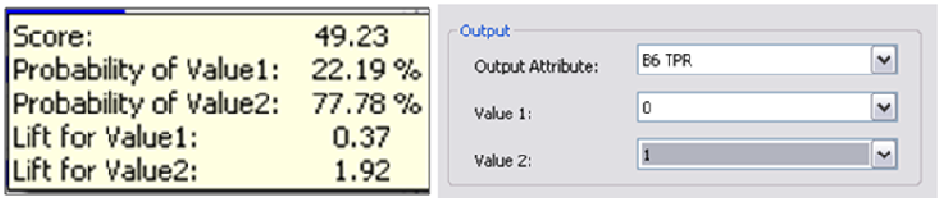


Figure C.27(b): Impact of temporary price reduction using Neural Network analysis.

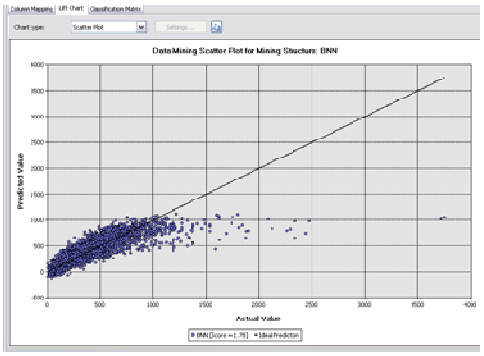


Figure C.28: Accuracy for Neural Network analysis of temporary price reduction.

To assess the significance of a 77.78% increase, using the previously defined processes, an alternate Decision Tree was created that provided the means to examine the impact of sales promotions for all varieties of Beef in London (figure C.29). London was chosen as the alternate region as it had been discovered that this is the region with the highest sales figures.

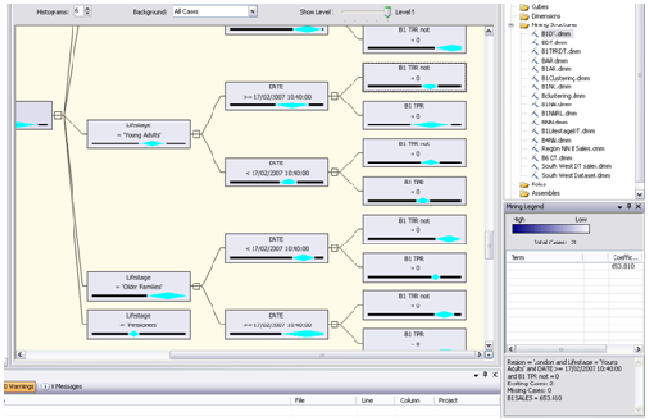


Figure C.29: Impact of sales promotion in London.

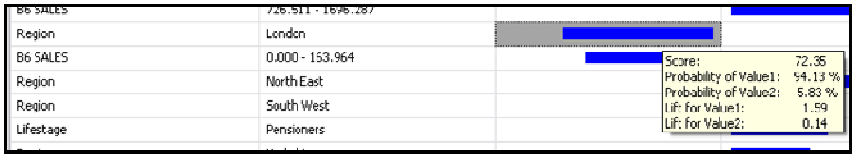


Figure C.30: Neural network analysis of impact of sales promotion in London.

Cross-analysing the results of the Decision Tree analysis (figure C.29) on the impact of sales promotions in London with that of Neural Network analysis. It was discovered and validated that there is no significant increase (5.83%) in sales figures through sales promotions (figure C.30). Hence, sales figures remained consistent, often showing no difference in sales whether there was a sales promotion active or not. Consequently, not only does this determine that stores in London are over discounting by applying sales promotions. Thereby reducing their profit margin, but also that 77.78% increase in the South West region, provides a significant increase and incentive for sales promotions. As a result, it is a suitable marketing strategy to apply sales promotions to increase sales figures within the South West. However, for sales promotion to be

truly effective it must be discovered which consumer segments should be targeted. Hence, those that constitute for the largest purchasers and those that buy the least amount. Thus to realise this objective, a SQL statement was defined that would select all attributes from the complete dataset, however, return only instances where the region is 'South West':

'SELECT * FROM Meat_Promotions WHERE Region LIKE 'South West''

The SQL statement provided a table detailing activity only in the South West region. Due to the format through which the database had been saved SQL server did not require individual tables to be listed rather just the table containing all foreign IDs. The SQL declaration, once entered with SQL Server Visual Management Studio was converted to a native elongated format a sub-section of this is illustrated in figure C.31.

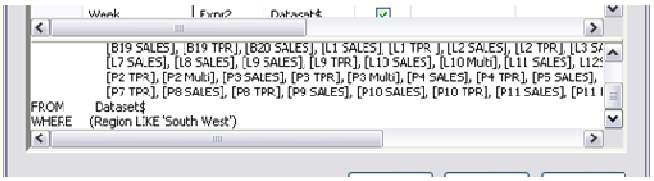


Figure C.31: Sub-section of the SQL statement declaration to discover all activity in South West region.

Using this new dataset defined by the SQL statement, a Decision Tree was developed detailing sales figures within the South West region by life stage (figure C.32). Analysis of this tree uncovered that the largest consumers in the South West were 'Pensioners' followed by 'Older Adults', in contrast, 'Young Families' was the consumer group that purchased the least. This informs us that the demographic in the South West is an older population which can be explored when proposing promotional schemes within the area, thereby eliminating the need to target younger people with a great volume.

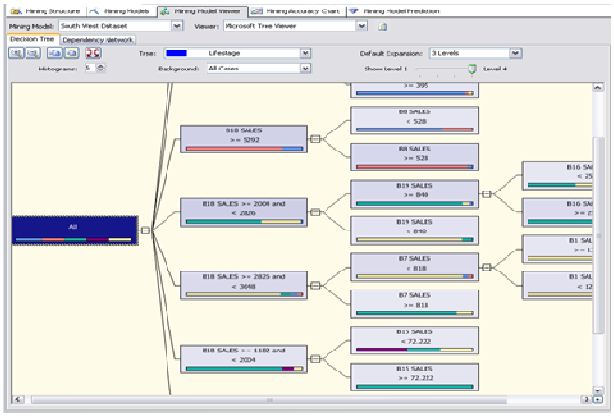


Figure C.32: Sales figures in the South West distributed by life stage.

However, cross-analysing these findings with other regions it was discovered that pensioners are not a large consumer group when inspected on a nationwide level. Pensioners as a consumer group were then analysed with ARM to find strong relationships between the group and sales figures (for Beef, as this meat type has the most variety). The results of the association rules were configured into a 'dependency network' to illustrate the

strongest relationships as illustrated by figure C.33. A dependency network, provides a graphical representation of the rules discovered through the exploration of the data set using ARM. As illustrated by the dependency network in figure C.33, it is possible to discover that Pensioners when observed in all regions have a strong correlation with low sales figure boundaries. Thereby reasoning that a nationwide marketing scheme aimed at older consumers would be inefficient however could provide increased sales if confined to the South West. Thus, Sales figures can be increased in the South West using sales promotion. In addition the customer segments that constitute the biggest and smallest consumer groups have been established. As a result, the investigation can now be directed toward exploring which sales promotion technique; 'temporary price reduction' or 'multi-buy savings', bears the greatest impact upon sales figures, thereby proving decision support for deciding which sales promotion strategy to execute.

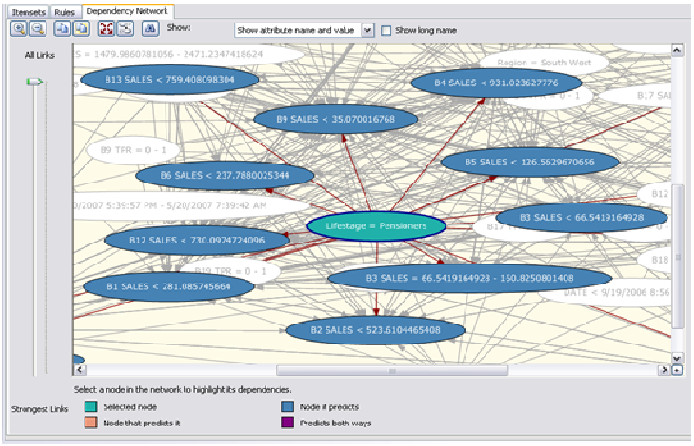


Figure C.33: Dependency network created through ARM, to uncover the strongest relationships between sales figures and pensioners nationwide.

By creating clusters for each of the meat types, the data set was analysed for statistical averages. These averages have been compiled in to table C.4. Table C.4 details the averages for the values, thus, the mean and standard deviation have been included in addition to the average value for the coverage of each sales promotion technique. Reviewing the data set and table C.4, it was determined that 'Premium Mince Beef' (B13) can be investigated to discover and compare the impact that either 'temporary price reduction' or 'multi-buy savings' produces upon sales figures, permitting the analysis of which sales promotion technique is more effective. The motivation for using Premium Mince Beef as the selected meat variety was that this meat variety had not only been promoted with both forms of sales promotion, but also had significant sales figures and combined sales promotion coverage, which permitted a more conclusive analysis.

Label	Meat type	Mean	Std. Dev	Max. Value	TPR Coverage (%)	Multi-buy Coverage (%)
B1 SALES	Premium Roasting Beef	361.99	289.03	1229.1	44	0
B2 SALES	Premium Fry/Grilling Beef	849.68	571	2562.69	77	0
B3 SALES	Premium Diced Beef	119.87	105.8	437.27	06	0
B4 SALES	Std Sirloin Steak Beef	1251.62	839.91	3771.36	27	01
B5 SALES	SR/Organic Roasting Beef	210.48	195.99	798.46	29	0
B6 SALES	Specially Reared/Organic Fry/Grill Beef	445.24	417.06	1696.43	43	0

B7 SALES	Healthy Diced Beef	1600.9	1043.95	4732.77	0	0
B8 SALES	Healthy Beef Mince	1460.29	1330.38	5451.44	0	22
B9 SALES	Healthy Fry/Grill Beef	43.07	86.08	301.32	01	0
B10 SALES	Organic Beef Mince	429.97	452.69	1788.02	0	0
B11 SALES	Specially reared/ Organic Diced Beef	185.61	204.53	799.19	0	0
B12 SALES	Std Other Fry/Grill Beef	1174.42	992.49	4151.88	33	06
B13 SALES	Premium Mince Beef	1237.01	933.75	4038.25	21	1
B14 SALES	Std Rump Steak Beef	1239.88	1012.22	4276.54	24	0
B15 SALES	Std Fillet Steak Beef	420.76	272.66	1238.74	0	0
B16 SALES	Value Fry/Grill Beef	702.82	457.07	2074.04	0	0
B17 SALES	Std Diced Beef	2098.82	1456.45	6468.17	07	0
B18 SALES	Std Beef Mince	7818.21	6184.94	26373.02	03	08
B19 SALES	Std Roasting Beef	2224.54	1374.48	6348	33	0
B20 SALES	Value Roasting Beef	169.17	127.13	550.56	0	0
L1 SALES	SR/Organic Roasting Lamb	125.29	125.06	500.47	21	0
L2 SALES	SR/Organic Fry/Grilling Lamb	324.62	241.11	1047.95	34	0
L3 SALES	Healthy Fry/Grilling Lamb	322.98	267.49	1125.45	0	0
L4 SALES	Organic Mince Lamb	158.17	198.09	752.44	02	0
L5 SALES	Healthy Mince Lamb	85.42	104.98	400.36	0	0
L6 SALES	Organic Diced Lamb	28.31	82.73	276.5	0	0
L7 SALES	Healthy Diced Lamb	1626.24	1363.13	5715.62	0	0
L8 SALES	Premium Roasting Lamb	137.3	125.7	514.38	0	0
L9 SALES	Std Roasting Lamb	561.85	1101.73	3867.04	74	0
L10 SALES	Std Mince Lamb	712.05	655.7	2679.16	0	21
L11 SALES	Premium Mince Lamb	9.53	27.86	93.1	0	0
L12 SALES	Std Fry/Grilling Lamb	1809.54	1232.72	5507.68	0	0
L13 SALES	Premium Fry/Grilling Lamb	14.74	25.14	90.15	0	0
P1 SALES	Premium Roasting Pork	879.97	800.17	3280.47	26	0
P2 SALES	Premium Fry/Grill Pork	74.84	141.8	500.24	33	0
P3 SALES	Std Roasting Pork	732.65	961.14	3616.07	57	01
P4 SALES	Specially Reared/ Organic Roasting Pork	276.52	221.29	940.39	43	0
P5 SALES	Value Roasting Pork	113.22	128.02	497.3	0	0
P6 SALES	Std Fry/Grill Pork	1883.72	3862.85	13472.27	42	26
P7 SALES	Specially Reared/Organic Fry/Grill Pork	144.8	341.33	1168.8	19	0
P8 SALES	Value Fry/Grill Pork	229.17	208.91	855.91	0	0
P9 SALES	Healthy Fry/Grilling Pork	201.23	230.98	894.17	0	0
P10 SALES	Std Mince Pork	671.61	586.84	2432.13	08	0
P11 SALES	SR/Organic Mince Pork	18.27	44.2	150.85	0	0
P12 SALES	Healthy Diced Pork	1449.75	1212.05	5085.91	0	0

Table: C.4: Average values for data set.

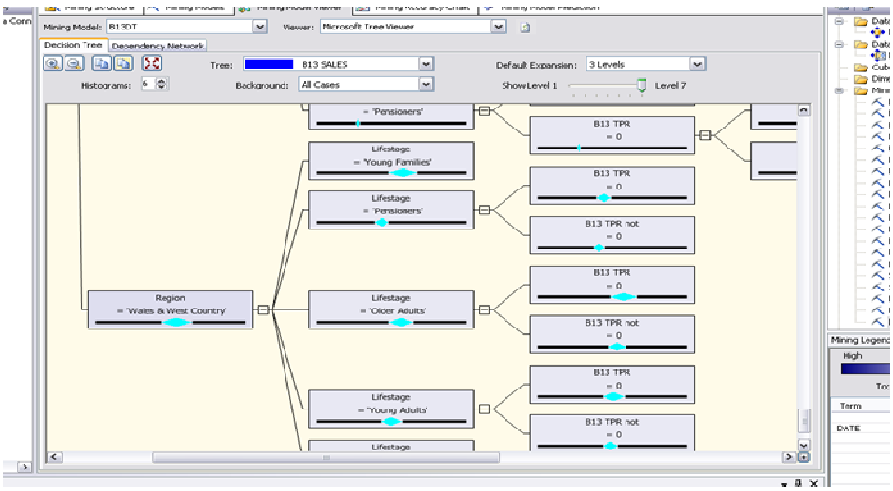


Figure: C.34: Decision Tree to analyse the impact of sales promotion strategies; Temporary price reduction or Multi-buy promotion.

Exploring; MeatID; DATE; Lifestage; Region; B13 SALES (Predictable Value); B13 TPR; and B13 MULTI, as the mining structure, Decision Tree analysis was inspected the results for which are illustrated in figure C.34. It was found that, temporary price reduction did not bear a significant effect upon the sales figures for Premium Mince Beef. This result was further analysed to discover if the time of year had any bearing upon the impact of temporary price reduction. Consequently, in contrast to the overall findings, it was discovered that temporary price reduction actually increases sales significantly during the summer period. This can be ascribed to the fact that Beef is often replaced as the meat of choice during the winter months, due to Beef not being a traditional festive meat variety, where as Beef popular for meals, is also a popular BBQ meat, which may constitute to the greater demand in the summer months.

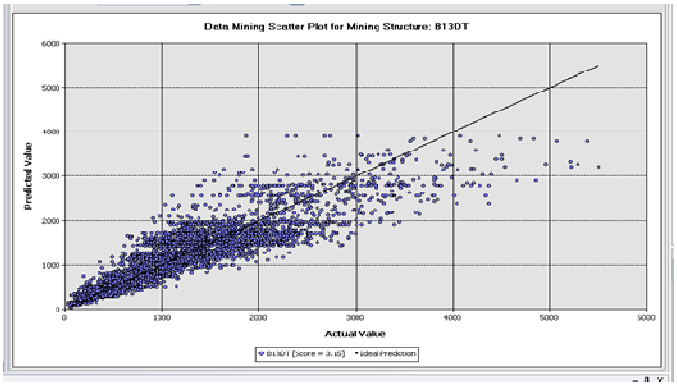


Figure: C.35: Scatter plot lift chart to examine model accuracy.

The results were further scrutinised as a scatter plot lift chart (figure C.35) to ensure the accuracy of the findings. Since majority of the values are close to the diagonal line (figure C.35), verifying a significant level of accuracy, the results can be validated, permitting further inspection. Thus, more in-depth analysis to probe the effect of 'multi-buy promotions' as the chosen sales promotion technique was conducted, which resulted in similar findings. Thereby emphasising the discovered trend; of a greater impact of the promotion in the summer months.

However, when the impact of the two sales promotion techniques was compared it was uncovered that temporary price reduction bears a more significant impact then that of multi-buy promotions.

In order to cross-analyse the findings to discover the effect of a particular type of sales promotion upon the other types of meat. The same process for analysis was conducted using Pork and Lamb meats. Consequently, the analysis process was repeated upon 'Standard Fry/Grill Pork' since it was this variety that contained the largest combined mean coverage for both temporary price reduction and multi-buy promotion (table C.4). It was discovered that again temporary price reduction increased sales more significantly then multi-buy promotions, in contrast to Premium Mince Beef; the sales remained consistent throughout the year. However, for Lamb, since no explicit Lamb variety had historical data for both sales promotion techniques, two varieties were selected; 'Organic Mince Lamb' and 'Standard Mince Lamb'. These two varieties were selected since they are both 'Mince Lamb' and between them encompass the highest sales promotion coverage (table C.4). Again, temporary price reduction was found to have a greater impact. As with Beef, the impact of the sales promotion was most significant during the summer period. It can therefore be deemed that temporary price reduction can bear a greater significance upon increasing sales. Thus, if the promotion is motivated by the incentive of increasing sales, temporary price reduction is a more ideal form of sales promotion. This should not result in multi-buy promotions being disregarded. Multi-buy promotions can be applied to increase the sales of a combination of related products. In order to uncover relationships between products initially Neural Networks were explored to explore the effect that products bear upon one another (figure C.36).

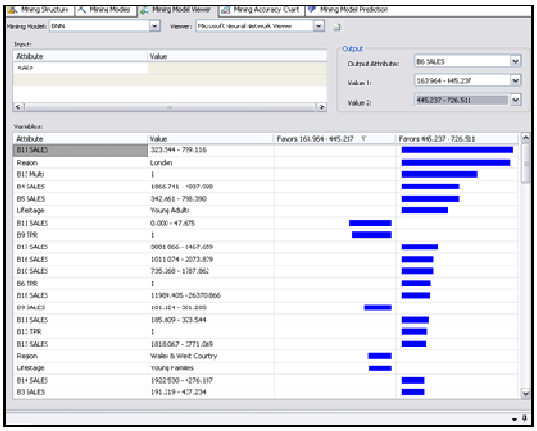


Figure C.36: Neural network analysis to examine relationship between Beef varieties.

The Neural Network analysis illustrated in figure C.36, unveils the relationship that is evident between 'Specially Reared/Organic Fry/Grill Beef' (B6), which was the lowest selling Beef variety in the least efficient region (South West), and the other attributes in the dataset. The strongest correlation of Specially Reared/Organic Fry/Grill Beef sales was with 'Specially Reared/ Organic Diced Beef' (B11), thus the two varieties could be prompted in conjunction. Furthermore, in contrast to the South West the sales of Specially Reared/Organic Fry/Grill Beef are high in London. Of most significance in this instance is that, Specially Reared/Organic Fry/Grill Beef sales are high in conjunction with 'Premium Mince Beef multi-buy offers'. Thus the two varieties are ideal for cross-promotion. Using ARM to develop dependency networks, Premium Mince Beef multi-buy offers was investigated to discover the correlation it bears with various other Beef varieties.

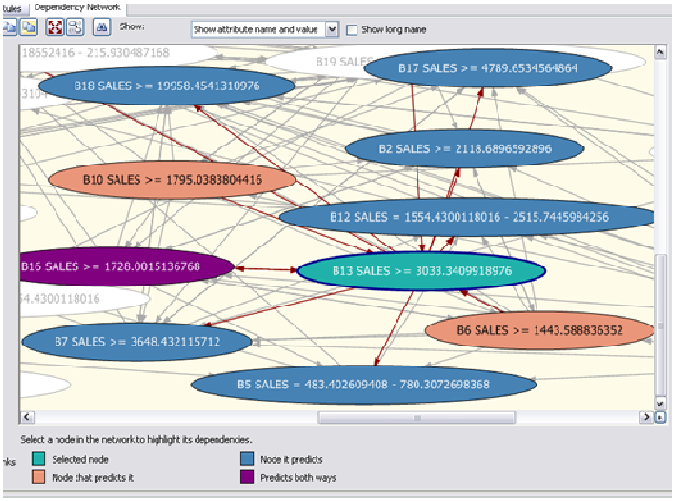


Figure C.37: Dependency network developed using ARM to examine impact of Premium Mince Beef upon other Beef varieties.

The dependency network (figure C.37) unveils that high Premium Mince Beef sales can predict, and therefore bear a strong correlation to high sales of:

- Premium Fry/Grilling Beef (B2),
- Specially Reared Organic Roasting Beef (B5),
- Healthy Diced Beef (B7),
- Standard Other Fry/Grill Beef (B12),
- Value Fry/Grill Beef (B16),
- Standard Diced Beef (B17),
- Standard Beef Mince (B18).

High Premium Mince Beef sales are also high in conjunction with high sales of 'Organic Beef Mince', since this variety, like, Specially Reared/Organic Fry/Grill Beef predicts high sales of Premium Mince Beef. Interestingly, all varieties of Mince Beef sell well together. Value Fry/Grill Beef, is also another significant product as all other varieties are of higher quality. This process was repeated with Lamb (figure C.38) and Pork (figure C.39). In each instance, the highest selling variety was selected; Standard Mince Lamb (L10) and Standard Roasting Pork, to identify strong relationships with other varieties. Consequently, Standard Mince Lamb was found to sell highly in conjunction with:

- Specially Reared Roasting Lamb (L1),
- Specially Reared Organic Fry/Grilling Lamb (L2),
- Healthy Fry/Grilling Lamb (L3),
- Organic Mince Lamb (L4),
- Healthy Mince Lamb (L5),
- Organic Diced Lamb (L6),
- Healthy Diced Lamb (L7),

- Standard Roasting Lamb (L9)
- Premiums Mince Lamb (L11),
- Standard Fry/Grilling Lamb (L12).

Standard Mince Lamb therefore sells highly in conjunction with a variety of other Lambs, permitting multi-buy offers to be explored amongst all these varieties. If a more focused multi-buy promotion is required, then Healthy Fry/Grilling Lamb is the most suitable cross-promotional product. These two varieties are suggested since they predict each other in high volume. Standard roasting Lamb can also be explored since this predicts high sales of Standard Mince Lamb. Standard Fry/Grilling Lamb is another option since this is predicted in high volume by Standard Mince Lamb.

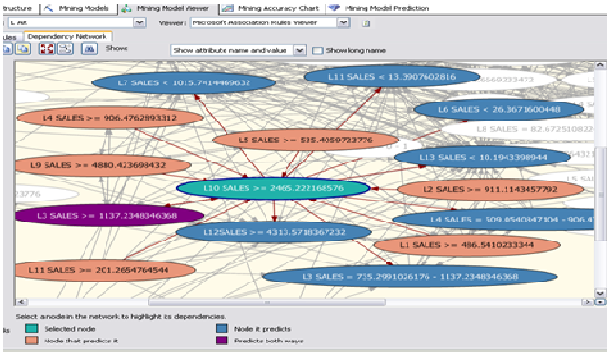


Figure C.38: Dependency network developed using ARM to examine impact of Standard Mince Lamb upon other Lamb varieties.

As was the case with Standard Mince Lamb, Standard Roasting Pork can also be cross-promoted with a variety of meat types, both; Healthy Fry/Grilling Pork (P9); and Specially Reared Organic Mince Pork (P11) were predicted in high volume. Standard Roasting Pork furthermore bared a positive correlation upon:

- Premium Fry/Grill Pork (P2),
- Value Roasting Pork (P5),
- Standard Fry/Grill Pork (P6),
- Specially Reared/Organic Fry/Grill Pork (P7),
- Value Fry/Grill Pork (P8),
- Standard Mince Pork (P10),
- Healthy Diced Pork (P12).

Of these Pork varieties Standard Mince Pork, is the Pork variety that had the most significant sales volume in conjunction with Standard Roasting Pork.

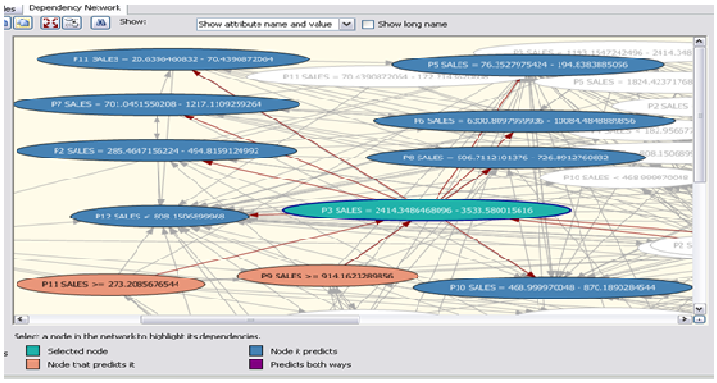


Figure C.39: Dependency network developed using ARM to examine impact of Standard Roasting Pork upon other Pork varieties.

KDDS-BI facilitates a high-degree of flexibility, within the 'Decision Support' phase as the output is collected, analysed, information extracted and transformed in to knowledge for decision support. This investigation has illustrated why such a high level of flexibility is required with BI analysis. Throughout section C.2.3, output has been analysed, information extracted, converted to knowledge and decision support provided, prior to further analysis, based upon the results, thereby re-iterating the stage. However, this knowledge can be reviewed and enhanced, ensuring that decisions based upon the analysis can significantly enhance the opportunities available to an organisation.

Using a combination of techniques applied to gather and verify results, it has been discovered that the South West is the lowest performing region. Sales promotions however, bear an extremely positive effect within in this region, in contrast to high performing regions such as London. In London, sales promotions do not impact sales figures, therefore, managers in this region, unless motivated by the objective to increase the quantity of products sold, e.g. in the event that they are nearing a sell-by-date. Else, managers should avoid sales promotions as they reduce profit margins, whilst bearing no real impact upon sales. In contrast, in low performing regions such as the South West; it is advisable for managers to include sales promotions in their marketing strategy. Sales promotions have proven to significantly increase sales and thereby increase the performance of the region. In addition, Pensioners, who nationwide are a low impact consumer segment, are in fact the highest consuming group in the region; consequently it is this consumer group that can be targeted as a positive response can be expected. When targeting consumer segments, it is not only the consumer group that constitutes the highest consumers, but also the lowest that should be targeted. In this region the lowest consuming group is young adults, this knowledge provides incentive to target these consumers, to increase the interest and response within this consumer segment.

In terms of potential strategies, temporary price reduction has proven to be more successful as a sales promotion technique. In addition these promotions have proven to be even more significant in the summer months. Multi-buy promotions should not be discarded. This technique is not only viable but can facilitate in the easy promotion of multiple product. A number of relationships between products have been identified that can be explored through promoting a popular product with an unpopular product, to increase consumption of the lower

performing product. Alternatively, two high performing products can be promoted together, thereby increasing sales figures within the region. The relationship between products has a duel benefit. Not only can it be explored for sales promotions, yet it can be studied when determining the most effective layout of shelves. Unpopular products can be placed amongst popular items; to increase the exposure they receive In addition unpopular products should be placed in high visibility areas, to ensure that they can be marketed with a view to increasing impulse buys. Or if motivated by the aim of increasing sales figures, popular products can be placed in high visibility areas to ensure that they are consumed in even greater quantities.

The process of obtaining information to provide decision support, in addition to findings of this study should not be considered absolute. The process of obtaining output, analysing this output and exploring knowledge and information gained for positive decision making, is a highly iterative process. Correspondingly, the process should be repeated frequently to ensure that managers can remain at the forefront of trends. This will assure that promotional strategies reflect the dynamic nature of consumers. Remaining at the forefront of trends and meeting the needs of target audiences will ensure that not only low performing regions can increase sales figure, but high performing regions, which are more prone to rival firms entering the market can remain dominant in any region.

C.3 Conclusion

This case study has endeavoured to investigate the applicability of KDDS-BI to discover valuable information to support the decision making process within the field of 'sales promotion'. Sales promotion has been selected as an area in which to investigate BI, since this provides a significant technique which blends all four aspects of the Marketing Mix. The Marketing Mix defines the marketing strategy encompassing all elements of the planning and operation of marketing. Sales promotion provides an interesting area to investigate, since it can be combined with various other marketing strategies such as those intended to increase customer loyalty. With the increasing use of loyalty cards/schemes, which are now operated by not only many of the large multinational organisations, but also smaller independent retailers, such as those conducting operations via the internet, provides an efficient manner through which to access customer data. Tesco was therefore explored as a source for data. The Tesco Clubcard scheme has proven itself to be one of the most successful loyalty schemes. However, marketing strategies such as loyalty schemes have not been immune to the competitive nature of organisations. Competitors of Tesco, such as Asda, are willing to accept vouchers which have been issued as part of the Clubcard scheme. As a result, it is essential that loyalty scheme be combined with sales promotion to ensure a competitive-edge.

BI provides an apt technology through which to investigate such a combinatorial marketing strategy. Thus, exploring Tesco Clubcard data, KDDS-BI was investigated to provide a structured means through which the data could be explored and analysed. Initially various objectives were extrapolated from the data, permitting the investigation of suitable techniques. Given the nature of the data, it was discovered that a single technique would not be sufficient. In contrast, a combination of BI strategies could be explored in order to fully analyse the data. Conducting the study through a combination of techniques not only provided an innovative approach, but furthermore a means through which to verify the results. However, upon assessing the potential scope and scale

of this study, it was determined that to illustrate the key influences of KDDS-BI, the study would be focused upon increasing the impact of sales promotion within the lowest performing region. Thus, the objectives were redefined to asses the extent to which a combinatorial BI approach could increase the impact of sales promotions within low performing regions, by exploring consumer segments, techniques and relationships between products thereby providing focused and targeted sales promotion knowledge explicit to the region. This knowledge could then be scrutinised as a basis for business decisions and future direction.

Upon exploring and analysing the data through a combination of advanced analytical software, it was found that Clustering, although initially appearing as an ideal technique, given the dataset, had limited results. The limitation of Clustering mainly stemmed from the ambiguous nature of the output. Clustering was found to be an effective technique, consequently was explored for aspects of the analysis, however in many instances although returning correct results, the results required substantial deciphering, in comparison to other techniques. As a result, facilitated by the iterative and flexible nature of KDDS-BI, further techniques were explored providing more effective output. However, Clustering was nevertheless scrutinised as a means through which to corroborate the results. Since the investigation consisted of multiple objectives, that had to be methodically analysed, the results of each objective provided information and knowledge which could be further assessed for subsequent objectives. Again, KDDS-BI provided a structured framework within which to conduct this iterative process. Hence, as illustrate by the ‘decision support phase’, the iterative nature of the four sub-steps facilitated the high level of flexibility required by this investigation. As a result, output was analysed, information extracted, converted to knowledge, this knowledge could then provide the basis to acquire further output until the objectives had been satisfied, permitting the results to be presented to provide decision support.

The analysis unveiled substantial results which can be examined to gain a significant and invaluable competitive edge over competitors thereby increasing the gains which can be reaped through current marketing strategies. Although, the objectives of this study focused upon the lowest performing region, the process can be extended to any region with regard to fresh meat sales, and/or product. The process provided noteworthy results, which demonstrated the potential of BI strategies, and the advantages of structuring such a project within the framework provided by KDDS-BI.

Appendix D: Case Study 3

Managing Organisational Resources: The London Underground Ltd.

In order to investigate the advantages and disadvantages of the proposed framework KDDS-BI, it will be explored with the aid of case studies. This case study will examine the performance of KDDS-BI when investigate to increase the effectiveness of 'managing organisational resources'. Due to organisations competing on a global level, in previous; untapped markets, these organisations are required to amass an ever-increasing amount of resources. However, increasing capacity in this manner can result in an organisation falling prone to diseconomies of scale. It is therefore, imperative that to remain competitive organisations manage and distribute these resources, be they physical, financial or human, effectively and optimally. Optimally allocating resources throughout an organisation can exponentially increase the level of efficiency. Accordingly, BI provides a strategy which can be explored to increase the efficiency with which an organisation is able to allocate resources.

D.1 Operations Management

Organisations are expanding at an ever-increasing rate. The advent of digital technologies has provided organisations with the opportunity to target previously untapped consumer segments. Consequently, many organisations have increased the scale of their operations in addition to increasing the quantity and distribution of resources. Having investigated the strategies through which BI can be explored to identify, target and meet the needs of these consumers. It is essential to look at the various aspects that have effected the internal operations of an organisation. To ensure that the needs of an increasing consumer base can be effectively satisfied, organisations must increase their capacity. However, if organisations are to ensure that they are not subject to 'diseconomies of scale', it is imperative that they are able to efficiently manage and allocate resources. 'Operations management' is the management of the aspect of an organisation that is responsible for the creation of goods and/or services (Goncalves, 2008). The creation of goods and/or services entails the transformation from inputs such as capital, labour, and information to outputs, namely the goods or services provided by the organisation. Furthermore, the transformation process is considered to add value, thereby ensuring that the value of the output exceeds the cost of the input.

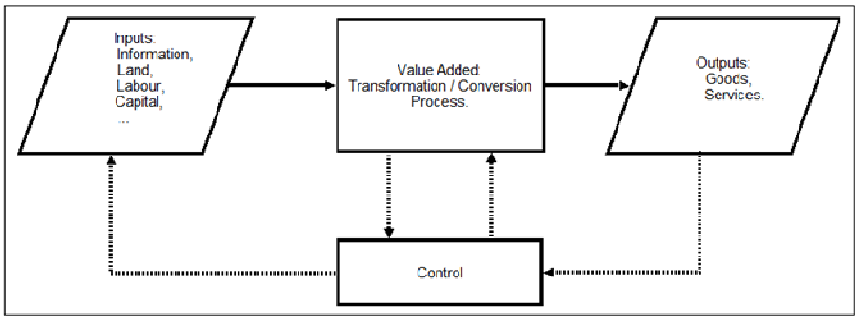


Figure D.1: Operations process.

Figure D.1, illustrates the operations process of transforming/converting inputs to outputs, whilst adding value. The greater the value added via the operations process the greater the profits for an organisation. As a result, the greater the profits, the more an organisation can invest in research and development, new facilities and equipment, in addition to covering any over-heads. However, the process of adding value not only refers to the production of a high quality product. In addition to a high equality product an organisation must effectively manage and allocate resources, both human and physical. Effective allocation of resources will permit an organisation to provide a high quality product at the lowest possible cost, hence, maximising profits. An organisation relies upon the process of adding value to ensure that it is able to most efficiently meet the needs and requirements of consumers which are obtained through various marketing strategies). As a result to ensure that it is the organisations goods and services that are selected by consumers over that of rival organisations, in addition to maintaining a competitive edge, quality control must be observed. To ensure that a desired level of control and quality is maintained, an organisation will obtain measurements at various points of the operations process via feedback. This procedure is illustrated by the dotted arrows in figure D.1. The procedure of gathering feedback to ensure and maintain the required certain standard requires a control system that will

ensure whether any corrective actions are necessary. These control systems provides information to the operations manager. The operations manager can as a result be considered a planner or decision maker. The decisions and actions taken result can directly affect the process of an entire organisation. Thus the key decisions made by an operations manager will determine (Stevenson, 2006):

- *What* resources are required and the quantity in which they are required.
- *When* will the resources be required? *When* should he work be scheduled? *When* should materials or other supplies be ordered? *When* is the influence of an operations manger required to refine the operations process?
- *Where* is the work to take place?
- *How* will the process be conducted and resources allocated?

These circumstances that require the observation and influence of an operations managers input will seldom have a clear solution. In contrast, the operations manager will be required to determine the option that provides the greatest return from a number of alternatives. Consequently, it is imperative that the operations manager be provided the highest quality of information upon which to base decisions. Since the quality of the decisions and course of action can only be as effective as the information upon which it is based. Due to the complex nature and significance of the operations managers decisions there have been a number of systems developed that can support the decision making capabilities of the operations manager, with a view to providing the highest quality of information. Amongst these, the key developments have been in the form of ERP (Enterprise Resource Planning) systems and APS (Advanced Planning and Scheduling) systems.

D.1.1 Enterprise Resource Planning (ERP) & Advanced Planning and Scheduling (APS)

In an attempt to successfully manage organisational activities and support operations managers, companies have implemented a variety of Enterprise Resource Planning (ERP) systems. ERP systems are enterprise-wide integrated software packages designed to uphold the highest standards of quality within business processes (Kelle & Akbulut, 2004). Although, ERP systems are generally focused upon the internal activities within manufacturing processes, many ERP applications can be extended to encompass the entire supply chain.

An ERP system consists of a collection of planning modules that translate the anticipated demand into plans that can be implemented to coordinate the various aspects of managing supply, production and distribution. Other modules in the ERP software aid the organisation to implement the plans and integrate them into daily operations in addition to providing computerised support for purchasing goods/services, receiving goods/services, sales and various other operations (Taylor, 2004). ERP systems are nevertheless, far more difficult to maintain then general software applications due to the size, complexity and dynamic nature of the environments within which they are expected to function. Consequently, despite many companies having over the last two decades, implemented an ERP system designed to integrate all internal business processes. ERP systems however, still exhibit a number of limitations for production management. The flexibility and performance of ERP systems has frequently been less than desirable. As a result many organisations have

augmented ERP systems with Advanced Planning and Scheduling (APS) systems (Hadaya& Pellerin, 2008). The APICS Dictionary (Cox & Blackstone, 2008) defines an APS system as:

"… any computer program that uses advanced mathematical algorithms or logic to perform optimization or simulation on finite capacity scheduling … These techniques simultaneously consider a range of constraints and business rules to provide real-time planning and scheduling, decision support, available-to-promise, and capable-to-promise capabilities."

APS systems provide functionality to enable operations managers to control the logistic flow in a supply-chain in addition to various other domains (Davidson & Wernstedt, 2002). APS systems include a range of capabilities, facilitating finite-capacity scheduling at the shop floor, vehicle routing and constraint-based planning amongst various other uses (Lee et al, 2002). Vehicle routing and scheduling performed within the context of supply-chain management is a complex process and difficult to solve with the use of empirical methods. Early papers, dating back to the 1960's were published which claimed to have solved the problem of generating distribution schedules. Furthermore, many of these proposed methodologies were seen as programmable. Despite the significant research which has taken place to transform the proposed methodologies into executable algorithms. It is only now, that technological evolution has permitted these algorithms to be implemented through APS systems and be integrated into daily operations as decision support tools (Gayialis & Tatsiopoulos, 2002). There are various advantages and disadvantages that must be observed when integrating ERP or APS systems. ERP systems are able to efficiently process transactions and execute standard repetitive tasks; however tend to perform sub-optimally when implemented in the capacity of a decision support system. This shortcoming can largely be attributed to limited capability, thereby failing to deliver optimum results or meet all expected requirements. A number or explanations for the shortcomings of ERP systems have been put forth, the principle of which is that the level of detail specified within the systems can be too vague for any substantial decisions to be conclude. Moreover, the technology frequently used for ERP systems is unable to cope with a great deal of real-time analysis and simulation in addition to many of the tools within the systems being incomprehensible, therefore, used infrequently by senior management. These are some of the criticisms which have been directed towards ERP systems (He & Wu, 2006). In contrast, APS systems are not prone to many of these failings and are also capable of tasks such as multi-site planning (Eck, 2003). However, APS systems are only able to find near optimal solutions therefore further investigation is required for systems that can provide optimal solutions in real-time. As a result BI is a technology that can be investigated to explore the opportunities to improve the quality of information provided to an operations manager. Since, BI technology will permit the investigation of techniques that can augment the functionality of ERP/APS systems. The process of effectively allocating resources in addition to the planning and management of these resources is an integral component of organisation activities. BI will however, permit the discovery of trends and hidden patterns which can be explored for optimum allocation of resources and increase the functionality of ERP/APS systems.

D.2 KDDS-BI Case Study: London Underground Ltd.

The optimal allocation of resources, in addition to the planning of these resources, is an integral component of organisation activities. Consequently, BI can be investigated a means through which the allocation of resources can be improved and optimised. BI provides the capabilities to allocate resources through the identification of trends and patterns. These trends and patterns can be explored to develop models. These models can be exploited to provide decision support within an organisation, thereby refining the operational process. To successfully investigate BI within resources allocation, KDDS-BI can be applied to structure the study. This will ensure a valid investigation that will permit the interrogation and application of data detailing the resources of an organisation. In order to explore the applicability of KDDS-BI to investigate BI techniques for resource planning within a complex and dynamic organisation that adds value to its services through the effective allocation of resources, London Underground Ltd. (LUL) can be investigated.

D.2.1 Data Investigation

LUL, popularly known as The Tube, represents one of the greatest scheduling challenges. LUL is the oldest underground rail system in the world and spanning nearly 400 route miles with 275 stops. The busiest railway line within the network carries over 180 million passengers annually. LUL passengers consistently make over 930 million trips - the same as passengers on all other trains throughout the UK put together. More recently, with the introduction of schemes such as the Oyster Card and Congestion Charging, the number of journeys consistently exceeds one billion, with 150,000 people an hour entering the LUL system[20]. In addition to this traffic and the physical resources, such as, the network control centre, line service control centres, functional control centres, station control rooms and ticket offices, route track, trains 'Rolling Stock Cabs', stations and signalling infrastructure. As identified by Transport for London[1] and Andrew Bourne (2000), LUL currently employs approximately 16,000 people. Around 9,000 can be considered to be operators. In the course of their work they 'operate' station facilities, trains, control systems in order to safely and effectively deliver London's tube service. This places considerable pressure upon LUL logistics systems, which must ensure that all trains are running on time, with crews in place and all required resources accurately allocated. In addition, LUL logistics must cope with the occurrence of unpredictable events such as track or train failures, unaccounted absence of crew members, accidents and many other unforeseen occurrences.

Currently all routine maintenance is scheduled for night-work, when the LUL network is closed. However, any maintenance that requires a longer period of time can result in considerable disruption of operation, resulting in large sections of track requiring closure. The majority of notices for these disruptions must be manually displayed and overseen by operators. The 2 month closure of the Circle Line in 2003 for repairs due to extensive work being required upon the track and its associated tunnels in the wake of a derailment affected 400,000 commuters. In addition, the terrorist attacks of 7th July 2005 required extensive repairs to the network. The repairs required included replacement of cabling carrying signalling information, communications and power. In addition, the stations also required a deep clean and, in some areas, redecoration. This required over 200

[20] http://www.tfl.gov.uk/: Accessed November, 2008.

engineers, working round-the-clock shifts, to repair the damage[21]. Despite the best efforts of LUL staff, there was still mass disruption throughout the network. In addition to such extreme circumstances, commuters are often encountered with closures regularly throughout the year. These frequent closures highlight the scope of disruptions that can be caused by an accident and amplified by the resources not optimally allocated. This can result in LUL logistics being unable to effectively provide the most effective service possible. The capabilities of LUL operators can be even more arduous in the event that they are required to reallocate resources in emergencies. Experienced LUL schedulers manually prepare all scheduling at busy stations. These schedulers have over the years developed acceptable methods for resolving these combinatorial problems[22]. Academic researchers and engineers have investigated alternative approaches for train scheduling as reviewed by Carey & Carville (2003). The planning process (especially for busy passenger trains) is segregated over several stages. Initially, train operators such as Metronet or Tubelines create outline draft timetables for the services. These estimates are based upon approximations of resource requirements. These timetables are then adjusted and revised to eliminate all conflicts. Nevertheless, this is generally a heuristic approach. Consequently, this slow ad hoc process does not allow operators to investigate 'what-if' options due to time constraints (Watson, 2001). Due to the conventional approach to scheduling being extremely time consuming, there are several advantages of computer based algorithms over manual methods. The principle of which is that results can be prepared more rapidly then possible through manual schedulers, thus allowing for a greater number of options to be examined.

CMC Limited is one organisation that in conjunction with SunMicro Systems and Oracle has been investigating the potential to computerise many of LUL's operations. Since being awarded a contract by LUL in 1989, CMC has designed and developed of a database for its timetabling and operation plan generation software system; CART (Computer Aided Railway Timetabling). In 1998 CART was made available to the scheduling department and was implemented through the departments Windows PCs through Hummingbirds 'Exceed emulator'. The CART software was later migrated to SUN machines enabling access to not only more processing power, but also enabling operators to access to the variety of Window based applications, such as MS Word. In 2002 LUL decided that it would re-engineer CART and use Oracle as the DBMS. CMC recommended that Developer 6i be used with Oracle and provided training to LUL personnel. CMC has since developed a highly customised Workflow Management Application (WMA) through Developer 6i and Oracle 8i for use by the scheduling department. WMA provides automated Task Management, Resource Management and Allocation Function for the Scheduling Department. The Scheduling Department has been using WMA to generate schedules since May 2003 (Winston, 2004; Sun, 2008). In order to guarantee compliance with regulations and passenger service efficiency, every six weeks LUL must, submit to the Department of Transportation a statistical report indicating its record in meeting the preset schedules. Current approaches to scheduling such as those provided by CMC, SunMicro Systems and Oracle have enabled LUL to include a level of automation.

[21] http://www.metronetrail.com/: Accessed November, 2008.
[22] *http://www.sun.com/smi/Success/IndustrySpecific/Transportation/London.Under.html*: Accessed November, 2008.

LUL is a highly dynamic complex and unpredictable environment, and although CMC has managed to provide LUL with an automated system the system is still based around an Oracle database that requires the expertise of manual schedulers and train operators such as Metronet or Tubelines to create outline draft timetables for the services (Watson, 2004). Despite reducing the burden placed upon the operators the capabilities for exploring 'what-if' options is still limited. As a result, LUL performance can be improved through the exploration of BI techniques. BI techniques provide the opportunity to explore an intelligent system that would provide the functionality of exploring 'what-if' options, in addition to increasing the speed with which the system can recover in the event of an accident. Table D.1 illustrates a contrast between intelligent and automated systems. Studying the table, it is clear that the logistic problems faced by LUL are an ideal environment for a BI-based intelligent solution.

Key Features	Predictability, Reliability, Economy of scale.	Agility, Responsiveness, Self-organisation.
Mechanisms For Achieving Key Features	Determination of algorithms, Memory, Integration.	Informed guessing, Knowledge, Learning, Networking.
Key Limitations	Rigidity.	Risk of mistake.
Mechanism For Dealing With Key Limitation	Modularisation.	Distribution of intelligence, Full use of local knowledge, Learning from experience.
Area For Application	Stable environment, Long production runs, Mass production.	unpredictable environment, Frequent changes in production runs, Customised products, Short lead times.

Table D.1: Contrast of automated and intelligent systems.

Consequently, for this application of KDDS-BI, in contrast to a conventional dataset, it is the resources that need be allocated that provide the opportunities for a BI investigation. Hence, the data set in this event is compiled of the resources that an organisation possesses and must manage.

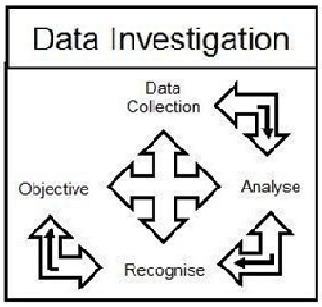


Figure D.2: Data Investigation stage of KDDS-BI.

Having reviewed the problems faced by London underground which proposes the data set for this investigation. As illustrated by the 'Data Investigation' phase of KDDS-BI, the data set must be analysed. Since the LUL provides a public service, in contrast to profits. The value added, can be measured through the optimal achievement of the highest level of service provided at the lowest possible cost. Hence, from the analysis of the data in addition to table D.1, it can be determined that LUL would benefit from the exploration of an intelligent advanced scheduling and planning system, that can provide support for operation managers.

Since a key limitation of an intelligent system is the risk of mistake (table D.I) which in the case of LUL, is a high risk factor. The intelligent system will not be required to autonomously execute actions, rather provide the details of the optimal allocation of resources, in addition to alternatives, to provide decision support. Consequently, the objectives for this investigation can be defined as:

- Investigate an intelligent system for optimal allocation of LUL resources.
- Investigate the capability of an intelligent system to provide decision support.
- Ensure that the proposed system is effective and robust, thereby capable of providing reliable decision support.

D.2.2 Data Modelling

Once the possible objectives have been established, the investigation can enter the second phase 'Data Modelling'. The primary deliverable of this phase will be the modelling of the solution with a view to discovering the technical requirement for a successful investigation thus the requirements of the objectives must be examined to discover the BI strategies that are available and those that will ensure the objectives can be successfully achieved.

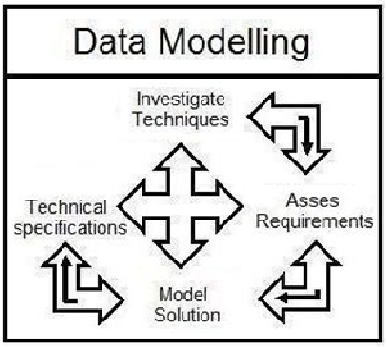


Figure D.3: Data Modelling stage of KDDS-BI.

The objectives defined during the data investigation phase must be explored to discover what techniques can most efficiently be investigated to realise these objectives successfully, so that the results can be studied to provide significant decision support. As discovered during the 'Data Investigation' phase of KDDS-BI, the information encapsulated within the dataset details the requirement for an intelligent solution that can provide decision support, to LUL on the optimal allocation of resources. The sheer quantity of resources in addition to

the complex nature of train scheduling, requires an APS system that can autonomously detect the best possible solution in addition to alternatives.

Scheduling issues exist in different application domains and it is an ever-growing issue in terms of complexity and cost. Various aspects of scheduling have been investigated, such as, personnel (or staff) scheduling (Taylor, 2004), production and inter-facility transportation scheduling (Foulds, 2000) and railway scheduling (Wren & Wren, 1995). A number of approaches have been used for scheduling, for instance, Genetic Algorithms (Foulds & Johnson, 2000); Fuzzy-Logic Algorithms (Wooldridge, 2002); Integer Programming (Bond & Gasser, 1988); and Hybrid Integer Programming with Heuristics (Watson, 2001). In general, these conventional approaches are efficient when developing schedules in a static environment; however, these methods do not claim to reach optimal schedules, only near-optimal solutions at best. Furthermore, the ability to deal with dynamic and changeable environments has not been achieved yet. These conventional methods (including the CART/WMA applications) allocate resources to demands by following pre-determined algorithms, thus operating in a strictly defined sequential manner. When dealing with large quantities of resources and demands that frequently change, the time required to accomplish the allocation is rather excessive because after every change the allocation process begins from the beginning. When the frequency of changes is high, the optimisers tend to oscillate, consequently never obtaining an optimal solution. Centralised Intelligent systems are somewhat better because they are powered by heuristics and are therefore faster, yet, they are still inadequate when dealing with repeated changes or adjustments (Knoors, 2002; Winston, 2004). Consequently, rather then these conventional approaches or advanced analytics, it is a Multi-Agent system that will provide the most suitable BI technique to investigate for the realisation of the investigations objectives. Multi-Agent systems and intelligent agents have been explored in detail in the literature review. Multi-Agent systems have been applied successfully to a variety of problems, such as electronic commerce, and energy management. The logistical problems of the LUL provide an exceptional example of an environment that is uncertain, complex and dynamic, hence ideal for agent exploration (Kelle & Akbulut, 2004).

A core reason for the growing success of Multi-Agent technology is its potential to cope with high complexity problems that are in essence distributed or require peer-to-peer processing. From the technical point of view, the inherent distribution of a problem allows for natural decomposition of the system into agents that are assigned specific individual tasks, these agents can then interact so as to achieve a desired global functionality. In addition, the scalability of systems operating in highly complex domains can be improved by choosing amongst specific coordination models that harmonize agent activities with respect to the Multi-Agent system's task. From the economic point of view, Multi-Agent systems may tackle even the most complex problems with an acceptable degree of performance, with a lower cost than traditional solutions (Bellifemine, 2007). As discussed in the Literature Review and Appendix A, there are a number of algorithms and techniques for dictating agent communication and decision making protocols. An Anytime Algorithm (see Appendix A for further details) can continually calculate until an answer is required, the algorithm may return differing result depending on when it is halted. In a communication environment, for the co-ordination to be guaranteed, it is necessary for the algorithm to be deterministic. Thus, terminate after a fixed number of runs. By determining an 'upper limit' or

boundary for the algorithm, a degree of consistency in the results can be maintained. This consistency, can be achieved by defining a limit in CPU time, thereby, ceasing calculation once the agent's allocated time limit expires. Hence, if agents allocating demands and resources in an environment, like the LUL, if each agent was to apply an Anytime Algorithm, searching to maximise its utility, it would aim to find the most suitable resource for it user, thus it would be necessary to ensure that the agent allocating demands is able to effectively ensure that resources are assigned to the most suitable agent and suitable alternatives found for the others.

It has been established that it is viable to distribute resources using agents. Furthermore, by implementing a utility function, a discreet method of differentiating between various agents is provided. However, a Multi-Agent system applied in a dynamic environment such as LUL will need to not only assign resources such as track, train and drivers, but also discover a non-collision optimal route to provide schedules. As a result, within the LUL environment there is a necessity for real-time planning, reactive and pro-active agents, that are able to operate in a large distributed, dynamic environment. As investigated this can be realised through the use of a utility function as a weighting mechanism and an Anytime Algorithm, to ensure that the system is able to provide answers at any point at which they are required. However, the agents must still be able to find the best paths within the network. This will enable the agent to plan from one point in the domain to another, whilst detecting possible conflicts. One technique that is popular for real-time intelligent agents in computer games, which can be utilised within LUL is path-planning. Path-planning is applied within video games, since a game-agent (character) will be required to navigate around a terrain they have no knowledge of. However, these agents must automatically observe collisions within a certain range, especially since the obstacles and agents are both continually 'moving' (Li et al, 2008). In addition these patterns must be committed to memory, real-time since the agent will often be competing or collaborating with human players (Koenig, 2004; Shoham & Leyton-Brown, 2008).

To solve the planning problems the environment is segregated into cells. Each section of route in LUL can similarly be segregated into cells, and a section of route occupied would, therefore, become blocked to another agent. Hence, the Agent requiring that cell must wait until that cell is made available or search for an alternative route. The utility function can be employed at this stage with each agent that requires the cell being assigned a utility. This would ensure that the most suitable agent would have the highest utility, providing an optimal allocation of resources. For this approach the agent would be required to start the path search from the initial cell to the goal cell, incrementally. This approach is known as an agent-cantered search. An agent-centred search can be implemented by restricting the search to the terrain around the cell. The agents determine the local search space; perform a search to conclude the best path and moves to that cell. This process is iterated, until the destination has been reached (goal-cell). Conventionally, an agent-centred search is a real-time heuristic search and stores a value in memory for each state that it encounters during planning and uses techniques from asynchronous dynamic programming to make them more informed and avoid endlessly cycling. 'Learning Real-Time A*' (LRTA*) is probably the most popular real-time heuristic search method. In LRTA* the values of a state approximate the goal distances of the states. They can be initialised with a heuristic approximation of the goal distances to focus planning toward the goal cell. LRTA* could viably be extended by assigning a utility

function to each of the states in addition to the distance to the goal thereby allowing the control agent to assign priority (Champandard, 2008). Anytime Algorithms and LRTA* are techniques which can be employed to realise the objectives of this investigation have been explored. As illustrated by figure D.4, the objectives defined during the data investigation phase must be assessed to determine their viability. Having reviewed the techniques, and the data, it is evident that due to the size and complexity of the resources that LUL network. The objectives must be redefined, and focused. As a result this investigation will only study the Piccadilly line of the LUL network.

By focusing the investigation on a subset of the LUL network, this study can be more targeted and the performance of KDDS-BI more intimately scrutinised. Hence, the objectives of the investigation can be redefined:

- Investigate an intelligent agent–based system for optimal allocation of LUL resources, focusing upon the Piccadilly line.
- Include the possibility of duel trains of the same route to mimic, alternative LUL lines.
- Provide an intelligent agent–based system for the scheduling of LUL train drivers.
- Investigate the capability of the proposed systems to provide decision support.
- Ensure that the proposed system is effective and robust, thereby capable of providing reliable decision support.

Having redefined the objective and focused the study, a conceptual model for the scheduling and planning of resources using BI strategies can be explored. Initially, the various constituent agents the Multi-Agent system will consist of, must be investigated. The requirements of a LUL orientated Multi-Agent systems are that of a 'bargaining system'. Consequently, there will be a number of agents that represents resources within the system or provide services. Agents representing users or requirements will aim to find agents that best serve their needs. Hence, each agent fulfils a role within the system. To successfully discover the roles of these agents, the conventions of the Gaia methodology can be explored to define the roles of the agents. The Gaia methodology has been reviewed in the Technical Review (section 4.2.2).

Roles are abstract constructs, which are utilised to conceptualise and understand the system (Juan et al, 2002). The roles, therefore, represent a set of entities that can occupy the same position in a reoccurring structure (Jade, 2008). Although, not conventional to software engineering, the Gaia methodology views the system as an organisation, thus allowing for roles to be defined and used to understand the system. A role is defined by four attributes:

- *Responsibilities:* determine functionality, thus a key attribute associated with role. Responsibilities are typically divided into two types:
    - *Liveness properties:* describe those states of affairs that an agent must bring about, once certain environmental conditions have been met.
    - *Safety properties:* are invariants, and ensure that an acceptable state of affairs is maintained across all states of execution.

- *Permissions:* are the 'rights' that are associated with a role. The permissions thus, identify the resources that are available to that role in order to realise its responsibilities.
- *Activities:* these are computations that are associated with the role that the role can carry out, without interacting with other roles.
- *Protocols:* these define the manner in the role can interact with other roles.

The Multi-Agent system to be investigated is one which is based upon the concept of demand and supply. As a result, the system will have a number of agent types, which represent role players in the system: 'Seller (Supply) agent', 'Buyer (Demand) agent', 'Contractor agent', 'DF Agent'[23]:

- *Seller Agents*: These are agents that will represent the users, thus selling contracts. The primary aim of these agents is to acquire resources required by there users. One or more Seller agents can represent each user; this depends upon the scope of the users requirements.
- *Buyer Agents*: These are agents that provide services, and help accomplish objectives. Upon receiving a call for bidding by the contractor agent, a Buyer agent must determine there suitability for the task and level of tasks that the agent is already committed to, prior to placing a bid to provide services.
- *Contractor Agents*: These agents extend the functionality of the Seller agent, and provide the means through which the Seller agent will require resources. Representing a group of Seller agents, the primary task of these agents is to interact with the DF agent to acquire addresses of agents suitable for the seller's requirements. Having located suitable Buyer agents the contractor agent will contact these agents with a call for bids from agents that are willing to provide their services. Upon receiving bids from interested agents the Contractor agent will sort the agents by suitability, assigning the task to the most suitable agent.
- *DF (Directory Facilitator) Agents*: The DF agent acts as a yellow pages service. All agents entering the system must register with the DF agent. This permits the functionality to locate any agent, with a simple request to the DF agent. In a distributed system such as a cluster, a DF agent may reside on each system, thus when searching for services, the DF agents can communicate amongst the peer-DF agents to locate agents providing services.

These agents and their corresponding roles can be presented through Gaia conventions as illustrated by code-table D.1 through to code-table D.4.

1	Role Schema:	Seller Agent
2 3	Description:	Represents users and aims to acquire the relevant resources required to complete a task. Sells contracts to work.
4 5	Protocols and Activities:	Informed of requirements, send message to Contractor, receive address of suitable agent, initiate bargaining with agent
6 7	Permissions:	*Read:* Requirements Addresses from contractor

[23] Contractor agent & DF agent is a notation specified by FIPA and is a requirement for a FIPA-compliant system.

8 9		Write: Message to contractor Negotiations with Buyer agent (collaborate/compete)
10 11 12 13 14 15 16	Responsibilities: Liveness: Safety:	 Seller = (Requirements.MessageContractor.ReceiveAddress.Negotiate)+ • Send message to contractor after getting requirements. Ensure that a certain load level (amount received) for the resources is fulfilled.

Code-table D.1: Role specifications of Seller agent.

1	Role Schema:	Buyer agent
2 3 4 5	Description:	Provides a service or resource. In order for a Buyer agent to perform a service or provide a resource, it must bid for and buy a contract from a Seller agent.
6 7 8	Protocols and Activities:	Receive bid, Asses bid, send bid to Contractor, decline contract, initiate bargaining with Seller agent if successful
9 10 11 12	Permissions:	*Read:* Offer from Contractor Write: Message to contractor Negotiations with the supply agent (collaborate/compete)
13 14 15 16 17 18 19	Responsibilities: Liveness: Safety:	 Buyer agent = (Receive bid.Asses bid.send bid to Contractor | decline contract.[initiate bargaining])* • Only bid for contract if load level allows Check load level (amount required) by Seller agent before initiating communication.

Code-table D.2: Role specifications of Buyer agent.

1	Role Schema:	Contractor agent
2 3 4	Description:	Represents a group of Seller agents. It is the task of the Contractor agent to find the most suitable agent for the Seller agents it represents
5 6 7 8	Protocols and Activities:	Receive request for agents from Seller, send message to DF Agent, receive Buyer agent addresses, send message to Buyer agents, receive bids, asses bids, send message to winner Buyer agent, receive confirmation, send ID of wining Buyer agent to Seller.
9 10 11 12 13 14 15	Permissions:	*Read:* Request from Seller Addresses form DF Agent Bids from Buyer agents Confirmation of Acceptance from Buyer Write: Message to DF Agent Message to Buyer Agent Asses bids

16		Message to Seller Agent
17 18 19 20 21 22 23 24 25	Responsibilities: Liveness: Safety: 	Contractor = (Receive Request. Send Message to DF Agent.Receive Buyer Agent Addresses.Send message to Buyer agents*. Receive bids*.Asses bids.Send message to winner Buyer agent||receive confirmation. Send ID of wining Buyer agent to Seller)* • Receive confirmation from Buyer agent Compare the Seller load level and Buyer agent load level to ensure suitability.

Code-table D.3: Role specifications of contractor agent.

1	Role Schema:	DF (Directory Facilitator) agent
2 3 4	Description:	Provides a yellow pages service. Each agent entering the system must register with a DF Agent. Enables all agents to locate one another by the roles they play within the system.
5 6 7	Protocols and Activities:	Receive request for agents from contractor, search for required agent addresses, add addresses to list, send list in message to Contractor.
8 9 10	Permissions:	*Read:* Request from Contractor Write: Addresses of Required Agents Message to Contractor
11 12 13 14 15 16 17	Responsibilities: Liveness: Safety:	DF Agent = (ReceiveRequest.search for required agent.[add addresses to list]*.send list in message to Contractor)*. • [Requirement from contractor = Service provided by Agent] = add to list Ensure that each Seller agent provides a required load level.

Code-table D.4: Role specifications of contractor agent.

Once all roles within the systems have been identified, consideration must then be provided to the structure of the agents and how they shall communicate to ensure that the agents are capable of fulfilling the requirements placed upon them by the user and the system. The design principles are introduced as following, which ensure that they are capable of fulfilling their system objectives and provide a conceptual model for the actions. Since the contractor and DF agent are FIPA-specified, only the Buyer and Seller agents need be modelled:

- *Buyer Agents*: These are agents that will attempt to acquire resources for their users; this can take the form of an agent representing a train that may require a resource such as a section of track (route), a driver or any other relevant resource. The Buyer agents will find both a primary match (Full matching) and reservation match (Partial matching), to ensure that the next best alternative is declared at all times.
- *Seller Agents*: The Seller agents are the agents that will represent the resource to ensure that this resource is provided to the most suitable Buyer. The Seller agent however can refuse to assign the resource to a Buyer agent, this ability to refuse, is explicit to Multi-Agent system, and indicates a key difference between Multi-Agent system and Object-orientated systems.

Since the DF agent and Contractor agent are necessary to maintain the integrity of the system and manage its functions of the Buyer and Seller agents. It is the Buyer and Seller agents that will be doing the core of the negotiations. The structure of these agents can be defined through code table D.5 & code-table D.6. However, it is the sum of the negotiations that determines the level of intelligence of the agent systems. The mechanisms that intelligent agents employ to communicate have previously been explored (Literature Review; section 2.4.2). Prior to further exploring the possible negotiations of the agents, it is imperative to determine that the agents can recognise the content of the communication.

	BUYER AGENT
1	A new resource becomes available / A current resource becomes unavailable,
2	
3	Initialise Agent and set requirement criteria,
4	**DO** Search in DF Agent for all matching flags,
5	**IF** Any matching flags,
6	**THEN** set no of matches = n,
7	**DO** Set Search = 0,
8	Search resource criteria,
9	**IF** Resource meets any requirements,
10	**THEN** Set as Full Match
11	**IF** Full Match already set,
12	**THEN** Compare Full Match with new resource,
13	**IF** New resource meets greater number of requirements,
14	**THEN** Set as Full Match,
15	**ELSE** Set as Partial Match,
16	**IF** Partial Match already set,
17	**THEN** Compare Partial Match with new resource,
18	**IF** New resource meets more requirements,
19	**THEN** Set as Partial Match,
20	**ELSE** Reject Resource,
21	Search +1,

22	**WHILE** Search < n,
23	**ELSE** Terminate agent,
24	**WHILE** No new resource becomes available / A current resource becomes unavailable.

Code-table D.5: Pseudo-code for Buyer agent.

	SELLER AGENT
1	A new resource becomes available,
2	Initialise Agent and set requirement criteria,
3	Register Agent with DF Agent and set Flag,
4	Search DF agent for number of Buyer agents,
5	**DO** Set number of Buyer Agents = n
6	Set Agents Contacted = 0
7	Wait for Buyer agent to contact,
8	Compare requirements,
9	**IF** Buyer Agent meets requirements,
10	**THEN** Assign resource to Buyer Agent,
11	**ELSE** Reject Offer,
12	Agents Made Contact + 1,
13	**WHILE** Agents Made Contact < n,
14	Terminate Agent.

Code-table D.6: Pseudo-code for Seller agent.

The ability to communicate is only valid if the agents are able to effectively understand one another. For efficient communication between agents to be possible, it is essential that all agents within the system, operate and communicate through an agreed set of concepts. This approved set of terminology in known as an 'ontology'. In addition to the terms and conventions which are used, the ontology describes the domain within which the agent is operating, thereby eliminating misunderstandings/conflicts between the agents. This ensures that both the sender and receiver are able to understand these messages due to the ontology specifying the explicit specifications of the meanings, concepts and relationships specific to the domain. To reduce ambiguity many agent communication languages explicitly state the name of the ontology implemented (Farber, 1999). An analogy can be applied to illustrate the necessity for an ontology:

"In the event that an agent is buying a particular engineering item from another agent. To make a sale possible the buyer needs to unambiguously specify to the seller the desired properties of the item, such as its size. The agents thus need to be able to agree upon how 'size' is defined, in addition to what terms such as 'inch' or 'centimetre' mean.

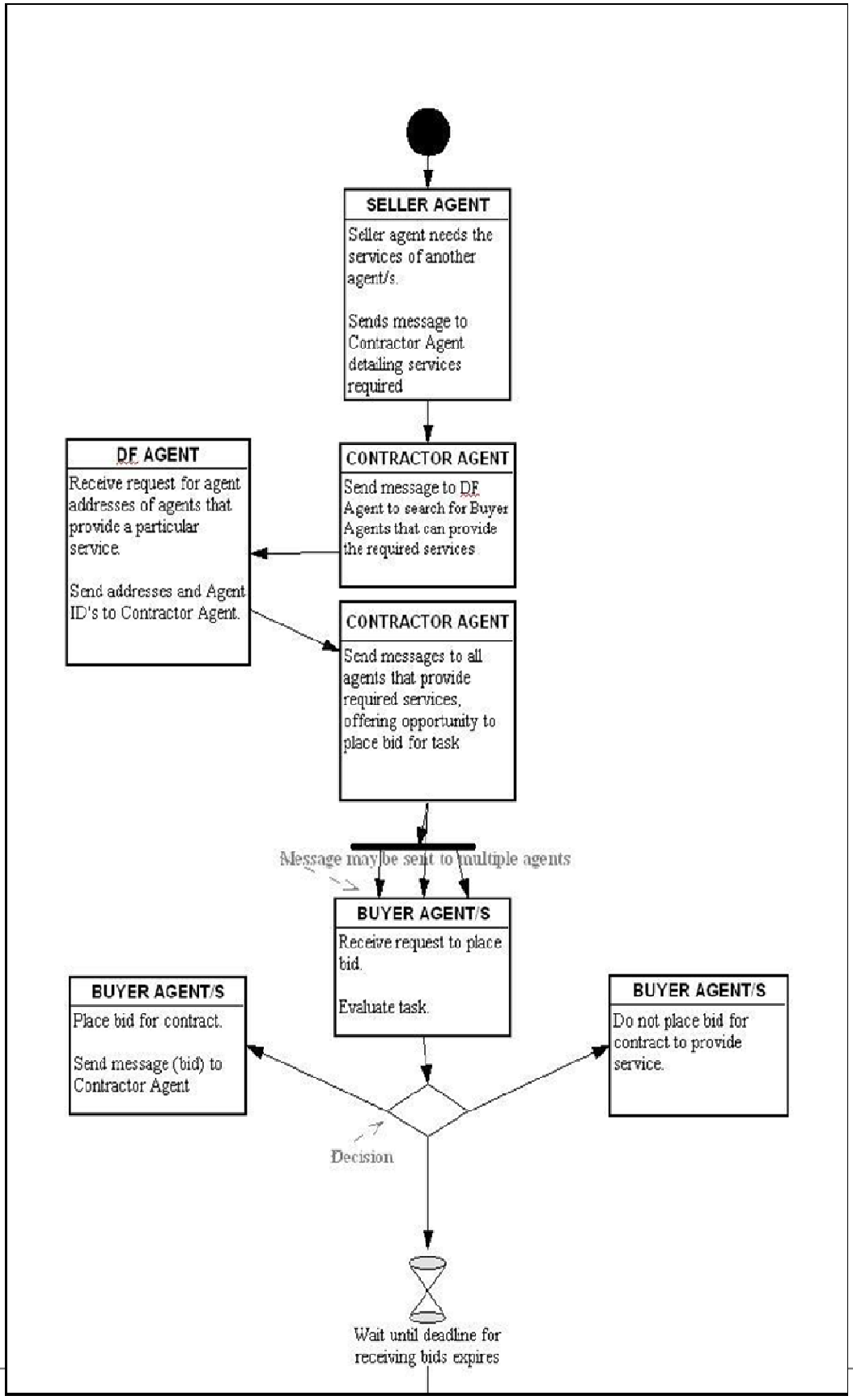
Consequently, a typical Multi-Agent system ontology, essentially contains the taxonomies of classes and subclasses for the objects and attributes, coupled with definitions of relationships between these entities. The main purpose of an ontology is to enable communication amongst different systems in a way that is independent from the individual systems' technology, information architectures and application domains. The key element of an ontology is a vocabulary of basic terms and precise specification of how those terms are defined. Yet an ontology is more than an agreed vocabulary; it provides fundamental constructs that are leveraged to develop a higher-level of knowledge. The component terms must be selected with great care, ensuring that the most basic foundational concepts are defined and specified. Formal techniques are used to define relationships between each of these terms. These defined relationships provide the semantic basis for the chosen terminology. Besides being a taxonomy or classification of terms and contributing to the semantics. The ontology also includes the relationships between these terms. It is these relationships that enable the expression of domain-specific knowledge. The ontology enables agent-based systems to simultaneously "interoperate without misunderstanding" and "retain a high degree of autonomy, flexibility and agility" (Jennings et al, 1998; Ossowski, 1999). Hence, the ontology can be regarded as a '*formal definition of a body of knowledge'*. The ontology furthermore describes the environment (business, physical, economic etc.) within which agents operate. The ontology is normally constructed as a semantic network describing the, subject domains and specific applications. The network connects fundamental concepts like objects, processes, properties, attributes and relations, representing the structure of the relevant knowledge (Wooldridge, 2002). Yet the ontology only ensures interoperability of agents within a system and not with those of a different system. For this purpose a set of agent standards are required to ensure that all agent-based systems be developed to a certain specification. There are a number of relationships that must be specified within an agent systems' ontology with regard to LUL. Upon reviewing the details, requirements and specifications, in addition to the role models, the terms within an ontology can be explored. The first task was to define classes of objects and relationships between objects that are of interest. Key classes of objects and relationships are illustrated below:

- *Route*: A railway line on which a train moves. The initial and terminal stations define the route.
- *Section*: A segment of the track that connects two stations.
- *Location*: A station at which the train must stop. Locations and sections compose a route.
- *Driver*: A resource that must be assigned to a route.
- *Carriage*: A resource that must be assigned to a route.

Multi-Agent systems match resources to demands. These matches can be either full or partial. As new demands and resources are made available agents perform re-matches if required. The matches may be the result of competition or collaboration amongst the agents. A number of demand- supply relationships have been created:

- SECTION ←→ ROUTE
- ROUTE ←→ LOCATION
- LOCATION ←→ SECTION
- CARRIAGES ←→ DRIVER
- DRIVER ←→ ROUTE

These relationships illustrate the resources that will be negotiated between the Buyer and Seller agents. Provided the ontology has established the definitions and requirements for the resources, the agents will then be able to negotiate with one another to acquire or sell these resources. These negotiations can be encapsulated in a data flow diagram (figure D.4). The data flow diagram, illustrates the flow of information through the system, thereby illustrating a typical agent-negation process. The flow of information through the system implies a conventional linear format. However, it is the necessity that the system is able to determine which agent can best serve the purpose and the negotiations that are implemented to resolve problems which results in an agent approach being the most suitable.


SELLER AGENT
Seller agent needs the services of another agent/s.
Sends message to Contractor Agent detailing services required
CONTRACTOR AGENT
Send message to DF Agent to search for Buyer Agents that can provide the required services
DF AGENT
Receive request for agent addresses of agents that provide a particular service.
Send addresses and Agent ID's to Contractor Agent.
CONTRACTOR AGENT
Send messages to all agents that provide required services, offering opportunity to place bid for task
Message may be sent to multiple agents
BUYER AGENT/S
Receive request to place bid.
Evaluate task.
BUYER AGENT/S
Place bid for contract.
Send message (bid) to Contractor Agent
BUYER AGENT/S
Do not place bid for contract to provide service.
Decision
Wait until deadline for receiving bids expires

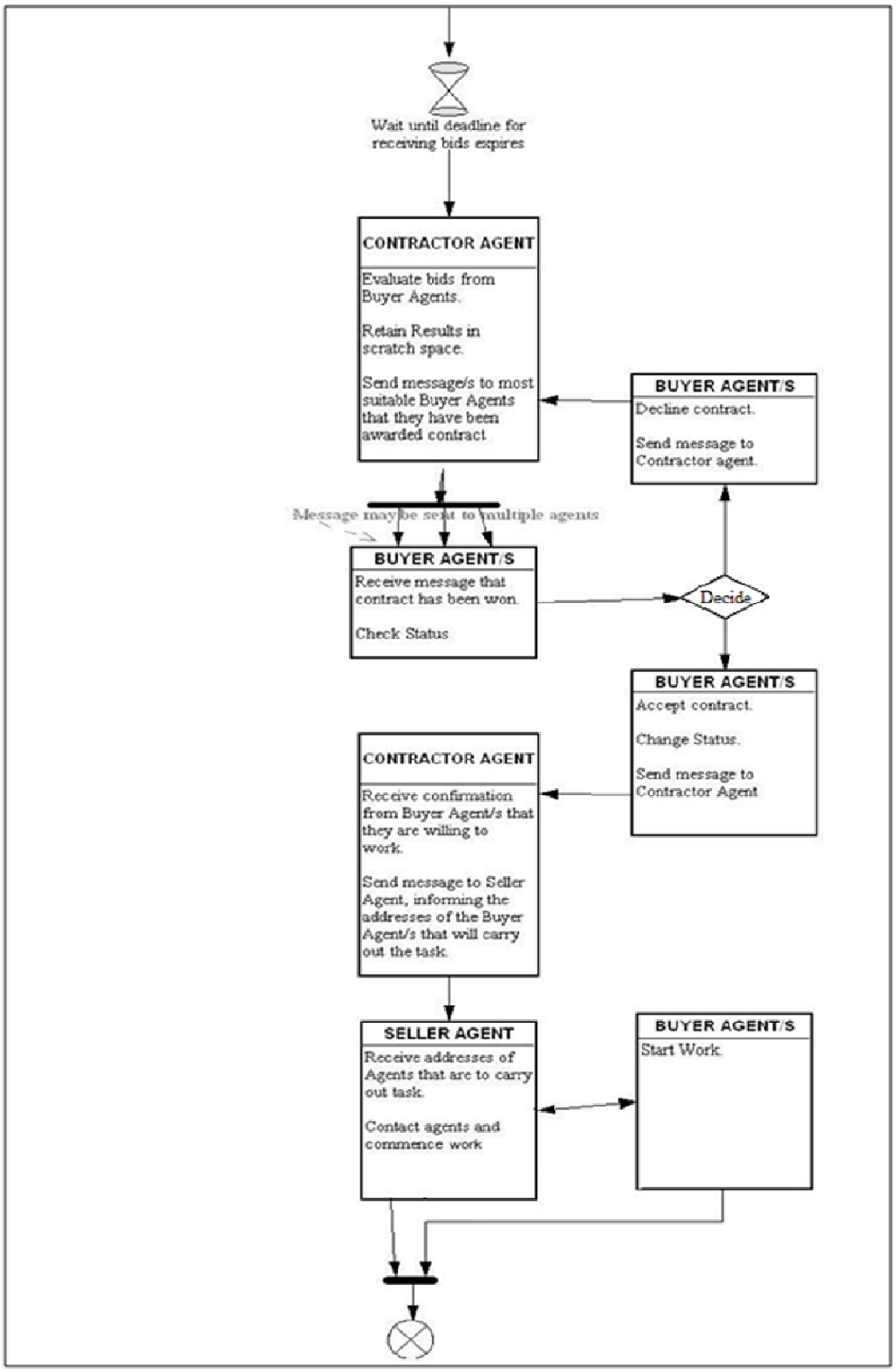

Wait until deadline for receiving bids expires
CONTRACTOR AGENT
Evaluate bids from Buyer Agents.
Retain Results in scratch space.
Send message/s to most suitable Buyer Agents that they have been awarded contract
BUYER AGENT/S
Decline contract.
Send message to Contractor agent.
Message may be sent to multiple agents
BUYER AGENT/S
Receive message that contract has been won.
Check Status
Decide
BUYER AGENT/S
Accept contract.
Change Status.
Send message to Contractor Agent
CONTRACTOR AGENT
Receive confirmation from Buyer Agent/s that they are willing to work.
Send message to Seller Agent, informing the addresses of the Buyer Agent/s that will carry out the task.
SELLER AGENT
Receive addresses of Agents that are to carry out task.
Contact agents and commence work
BUYER AGENT/S
Start Work.

Figure D.4: The flow of data through the system when searching for agents to perform on a single task.

The intelligence associated with an agent system is from an emergent process. Thus, it is the sum of all the interactions that creates an intelligent system, rather than one aspect. For this purpose it is paramount that the tasks be distributed to the correct agents, as a result:

- The system will initiate with an agent (seller-agent) requiring a task to be completed, this task request is forwarded to the Contractor agent.
- The Contractor agent will search the DF agent for suitable agents (buyer-agents) that are able to provide the required services.
- Once the Contractor agent has compiled a list of suitable agents, it will send a message to all agents providing the required services.

  The message will consist of:

  - Agent ID.
  - Task ID.
  - Salience; the level of priority the task should be given.
  - Task type/service required.
  - Grade of required Agent.
  - Estimated time to complete task. (This is a weighting that is based upon the number of resources required to complete the task).
  - Deadline, the time by which the buyer agent (agent providing service, thus bidding for contract) must respond.
- Upon receiving the message the buyer-agent must decide whether it wishes to bid to carry out the task.
- The Contractor Agent will therefore receive a bid from all agents wishing to carry out the task. The Contractor Agent must now decide upon the most suitable agent for the task.
- The bid will consist of a message detailing:
  - Task ID.
  - Agent ID.
  - Agent Grade (skill level).
  - Number of other tasks the agent has already committed to).
- When deciding which agent to assign a task a number of factors must be considered. The ideal agent to award the contract will be a buyer-agent that:
  - Has previously performed similar tasks.
  - Has the least number of allocated tasks, therefore would be able to quickly complete the task.
- The contractor agent must, therefore, decide on the agent/s that are best suited to the task and assign the task to that agent.

Consequently, the data flow diagram (figure D.5) can be further explored using the algorithm pseudo-code illustrated in code-tables D.6 & D.7. Unlike code-table D.5, the Buyer agent can be defined in more technical terminology, including a weighting function ('grade').

	BUYER AGENT
1	Receive a list of task requests R = (r_1, r_2, r_3, ……..r_n);
2	Set response list = 0;
3	For each request $r_i \in$ **R** do
4	Evaluate Task(r_i, grade);
5	If grade > **Ω** then
6	response list = response list + r_i;
7	End for.
8	Sort the response list in ascending order by estimated time to complete task
9	
10	For every $r_i \in$ response list do
11	Generate a bid b_i;
12	Send b_i to Contractor Agent;
13	End for;
14	

Code-table D.6: Pseudo-code for Buyer agent, including a weighting function.

	CONTRACTOR AGENT
1	While System time < deadline for task do
2	Receive Bids B = (b_1, b_2, b_3, ……..b_n);
3	
4	Set Agent Bid List = 0
5	For each request $b_i \in$ **B** do
6	Evaluate Bid (b_i, Agent ID, grade);
7	Agent Bid List = Agent Bid List + (b_i + Agent ID + grade);
8	Sort bids by Agent ID in descending order of grade.
9	End For;
10	Search Agent Bid List
11	Assign task to Agent with highest grade and lowest load level.
12	End While;

Code-table D.7: Pseudo-code for Contractor agent.

The pseudo-code algorithms will enable the agents that supply a certain service to receive tasks from contractor agents. The contractor agent will only receive back bids from agents that have the skill to complete the task. The

contractor agent will then be able to assign the task to the agent that has the greatest skill level and the lowest load level, which will result in a system that has few idle agents, this requires that the task be assigned to the agent that is most suitable thus ensuring that the processors are able to balance the work-load. The roles of the agents have been studied, in addition to the mechanisms employed to define terminology, communicate and extend this communication to negotiate for resources. However, the final aspect of the conceptual model is to examine how the BI solution (the Multi-Agent system) can be tested. A fully tested solution will ensure valid results, which can be analysed for reliable decision support. To ensure that the Multi-Agent system is tested thoroughly, it will be subjected to a number of tests. These tests will aim to discover the performance of the Multi-Agent system under various conditions. As a result the test will consist of examining the systems with:

- A number of scenarios, which replicate the conditions under which the system would be expected to perform and provide decision support.
- A *large* number of agents that achieve tasks via a *significant* amount of communication. Such a scenario is the result of agents needing to collaborate to achieve an objective, thus communicating to continually update one another of their status.
- A *large* number of agents, who will require *limited* communication to achieve their goals. Such scenarios are common when the agents are competing thus aiming to fulfil their own requirements without considering those of peer agents.
- A *small* number of agents that achieve tasks via a *significant* amount of communication.
- A *small* number of agents, who will require *limited* communication to achieve their goals.

Analysing the performance of the Multi-Agent system by varying the level of communication and number of agents with a view to invoking failure will ensure that the test scenarios are able to effectively investigate the systems performance. Since the aim of the tests is to cause the system to fail, if the system passes these tests it can be accepted that the conclusions are scientifically valid (Binder, 2000). Given the limited nature of the domain within which the solution must operate, namely a desktop PC of a decision maker the technical requirements must be assessed. The software whilst capable of handling large data sets must not require significantly large processing power. For this reason, a mid-specification Desktop PC has been selected as a platform:

- Intel Celeron 440 Processor (2 GHz. 800 MHz FSB, 512 KB Cache),
- Windows Vista Home Edition,
- 1 GB RAM,
- 80 GB HDD.

The specification of the PC has been selected to ensure that results can be obtained without necessitating a complex hardware infrastructure. Furthermore, the relatively low cost of the PC will reflect the resources available to the majority of decision-makers. Thereby, providing a suitable solution that can be integrated in many (if not all) working environments at a relatively low cost.

D.2.3 Development

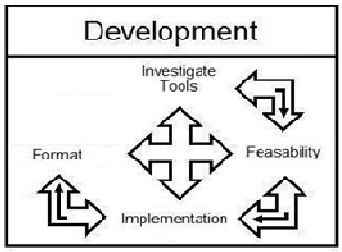


Figure D.5: Development stage of KDDS-BI.

Having explored the requirements, techniques and specifications that must be observed if the objectives of this investigation are to be attained, in addition to a conceptual model that will provided greater insights into the nature of the data permitting the resources of an organisation to be optimally allocated. It is now possible to examine various applications that can facilitate in the realisation of the objectives. Thus enabling the raw data to be analysed and informative discoveries uncovered that will permit decision makers to ensure that resources are optimally performing, resulting in a competitive edge over competitors. There are various software solutions that are available from a number of vendors, which have been reviewed in Appendix A: Section A.6-Table A-2. Once the various software packages that have been identified in Appendix A: Section A.6-Table A-2, had been carefully considered and tested it was discovered that due to the size and complexity of the problem. A combination of commercial and open source software applications would provide the functionality that could be further investigated to realise the objective of this study. Consequently, it is a combination of Magent-A i-Enterprise suite and the Jade platform that will be explored in combination to realise the investigation objectives. Table D.2 illustrates a subsection of Appendix A: Section A.6-Table A-2 and illustrates the key features of the selected intelligent agent solutions.

Magent-A	Intelligent Enterprise (i-Enterprise) suite	Commercial	i-Enterprise is Magent-A's fully integrated system designed to compete directly with products offering comprehensive support for all enterprise-wide activities, such as ERP. All systems agents within i-Enterprise are interconnected on the 'peer-to-peer' principle. Hence, e-Commerce Systems communicate with the Real-Time Logistics System to resolve delivery problems and the Logistics Systems agree with the e-Commerce System to issue tender to employees on the internal corporate market, or to assemble a team for a new project. All these capabilities result in synergetic increases in the intellectual capacity, flexibility and efficiency of i-Enterprise and the enterprise itself. All constituent components of i-Enterprise System can be delivered separately and integrated with existing software
Telecom Italia Lab (Tilab)	JADE Framework	Open-source	Jade (Java Agent DEvelopment Framework) is a software framework fully implemented in Java language. It simplifies the implementation of Multi-Agent systems through a middle-ware that claims to comply with the FIPA specifications and through a set of tools that supports the debugging and deployment phase.

			The full FIPA communication model has been implemented within Jade and its components have been fully integrated: interaction protocols, envelope, ACL, content languages, encoding schemes, ontologies and, finally, transport protocols. The transport mechanism, in particular, is like a chameleon because it adapts to each situation, by transparently choosing the best available protocol. Java RMI, event-notification, HTTP and IIOP are currently used, but more protocols can be easily added. Most of the interaction protocols defined by FIPA are already available and can be instantiated after defining the application-dependent behaviour of each state of the protocol. SL and agent management ontology have been implemented already, as well as the support for user-defined content languages and ontologies that can be implemented, registered with agents, and automatically used by the framework. Jade is being used by a number of companies and academic groups, both members and non-members of FIPA, such as BT, CNET, NHK, Imperial College, IRST, KPN, University of Helsinky, INRIA, ATOS and many others.

Table D.2: Subsection of Appendix A: Section A.6-Table A-2.

An open-source solution such as the Jade platform can be explored to originate a BI solution that will permit the scheduling of LUL drivers, referred to as BIDS (Business Intelligence-Driver Scheduler). The Magent-A i-Enterprise suite will be scrutinised to determine train timetables and schedules. Since the objective of determining schedules for trains is inherently complex and requires a large number of agents. Magent-A will provide the functionality to explore this problem using a high-granularity Multi-Agent system for the scheduling of train timetables within the LUL network, this Multi-Agent BI solution will be referred to as BITS (Business Intelligence-Timetable Scheduler). A high-granularity Multi-Agent system is one that is saturated with a large quantity of agents that are able to interact with each other and users. Furthermore, the intelligent agents exhibit self-organisational properties. That is, they can autonomously change relationships between constituent agents with a view to improving system performance, altering between competitive and collaborative behaviour. This trait enable a Multi-Agent system to display emergent behaviour, as the successful completion of their tasks is subject to the decision and action of other agents (Kendall et al, 2000). Such a system can be more efficiently explored through the Magent-A i-Enterprise suite since the capabilities of path-planning and decision–making are ‘semi-automatically’ coded within the agents once accurately defined through the suite. However, the ontology and specifications for the environment must be studied and implemented for a suitable solution. This approach of developing a solution around the ontology is illustrated in figure D.6.

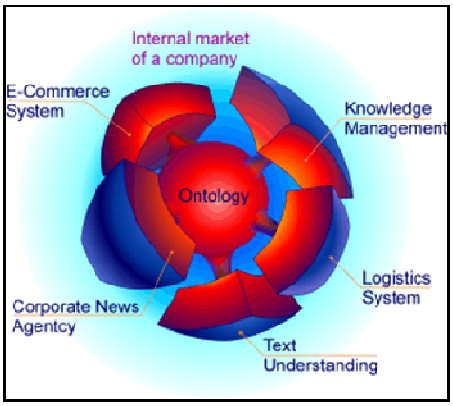


Figure D.6: The structure of Magent-A i-Enterprise.

The Magent-A i-Enterprise suite consists of Magent-A Multi-Agent Engine, Ontology and Interfaces. Furthermore agent negotiations logs and environment details can be explored, which illustrate the Magent-A approach. The logical architecture of the Magent-A Engine (figure D.7) includes the virtual world, its runtime system, interface system and extensions. If applications are to be run, a subject domain ontology and application ontology are required.

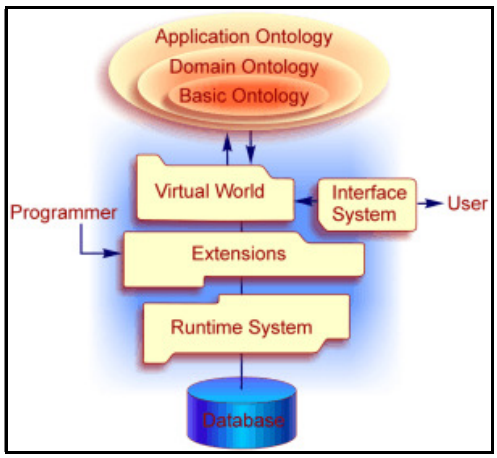


Figure D.7: Typical logical architecture of a Magent-A Engine.

A virtual world is a model of a subset of the real world. Traditionally Multi-Agent system virtual environments are simple, containing only agents that communicate by sending messages. The virtual worlds in a Magent-A architecture however, contain scenes that can contain any object type, its properties and functions. Agents are able to perform actions on these objects. Within virtual worlds (populated by physical and abstract entities), scenarios are created, actions performed on objects and responses observed in accordance with rules imported (defined) from corresponding sections of the real world. Having loaded the virtual world, users can construct different scenarios describing how this world works, using the interface system. The Engine architecture provides the ability to generate both simple agents (those which do not make use of ontologies) and intelligent agents (which employ an ontology for reasoning and decision-making).

Runtime systems includes virtual parallel machines for running virtual worlds and supporting concurrent operation of agents; a subsystem for communication support; modules for ontology support; the dialogue management subsystem and service subsystem. In addition, the Magent-A Engines enable creation of intelligent agents as virtual creatures with physical and mental 'bodies'. A physical body of an agent possesses certain sensors and executive mechanisms, allowing agents to "visualise" and 'hear' situations, 'feel changes in an environment' and 'sense a touch'. These devices considerably decrease the cost of perceiving a scene. Additionally, a physical body can observe the correctness of reproducing a world scene, perform planned actions, and respond to stimuli. Agent's mental body handles scenes in abstract worlds, works out scenarios and analyses results. In contrast, the driver scheduler will be investigated through open-source software. Although, the driver scheduler could have been developed as an extension of the ontology investigated and developed in BITS. Since this objective does not require a high-granularity Multi-Agent system, it was discovered that unnecessary load was placed upon the system and resulted in greater memory consumption then necessary. Furthermore this approach will permit the analysis of both a Multi-Agent system consisting of a large number of agents, and that of a Multi-Agent system with limited agents. Jade is a software framework fully implemented in Java language. It simplifies the implementation of Multi-Agent systems through a set of tools that supports the debugging and deployment phase, comply with FIPA specifications. The agent platform can be distributed across machines (which not even need to share the same OS) and the configuration can be controlled via a remote GUI (figure D.8). This GUI is launched from the command line:

'java jade.Boot –gui'

Jade creates multiple containers for agents, each of which can be on the same work station or individual networked work stations. Together, a set of containers forms a platform. Each platform must have a Main Container which holds two specialised agents called the AMS agent and the DF agent (figure D.9).

- The AMS (Agent Management System) agent is the authority in the platform. It is the only agent that can create and terminate other agents, terminate containers, and shut down the platform.
- The DF (Directory Facilitator) agent, implements a yellow pages service which advertises the services of agents in the platform so other agents requiring those services can find them.

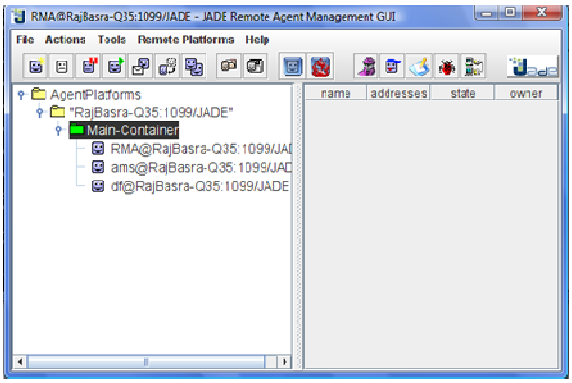


Figure D.8: Jade remote management GUI.

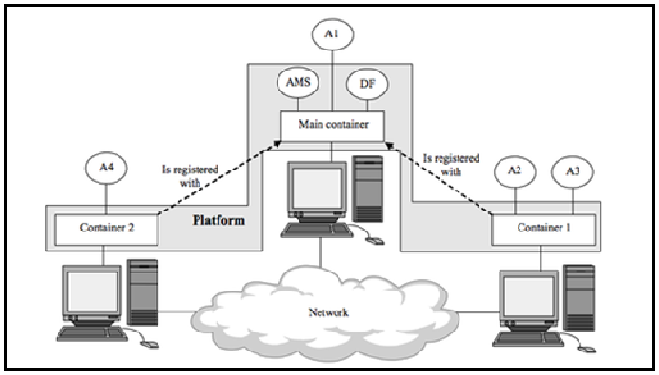
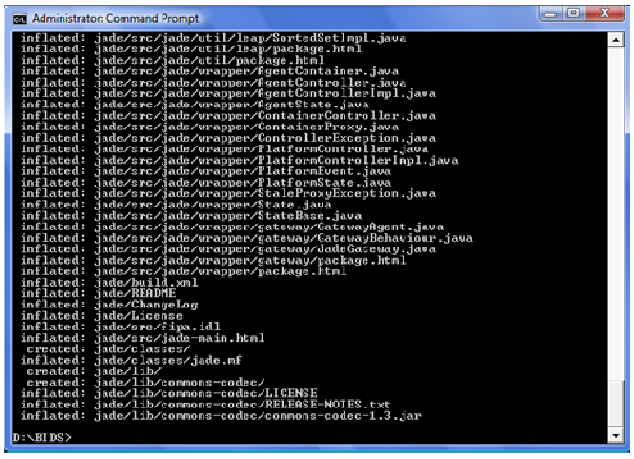


Figure D.9: Jade platform.

The platform configuration can furthermore be changed at run-time by moving agents from one machine to another one, as and when required. The only system requirement is the Java Run Time (version 1.4 or later). The communication architecture offers flexible and efficient messaging, where Jade creates and manages a queue of incoming ACL messages, private to each agent; agents can access their queue via a combination of several modes: blocking, polling, timeout and pattern matching based[24].

The combination of open-source and commercial intelligent agent solutions will provide the means through which a BI solution can be investigated for the realisation of the objectives of this study. Initially, however the solutions must be implemented so that they may be further explored. Since the approach selected consists of a combined open-source and commercial platform. These platforms will have to be individually implemented prior to investigation. Once installed the Jade source files must be extracted, this can be achieved via the jar tool 'jar -xvf' from the command line (figure D.10). In addition to class file extraction, the 'Java CLASSPATH' must be set (figure D.11). This ensures that the correct libraries are used in the event that the agent code is compiled or executed.



[24] http://jade.cselt.it/ Accessed: November, 2008.

Figure D.10: Jade installation from MS DOS command line.

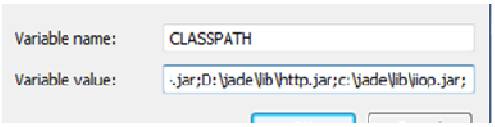


Figure D.11: Setting system Class path.

Upon successfully configuring the combination of selected platforms, the solution could be formatted to fully integrate the BI solution. Since the proposed BI solution will be in the form of a Multi-Agent system. Jade will be explored as the mean through which to schedule drivers to trains (BIDS). Whilst i-Enterprise will provide the basis upon which to investigate a scheduler for train timetables (BITS). Consequently, the formatting of the data is achieved through the design and creation of the ontology, the agents and a simulation of the target domain.

Initially, BIDS can be explored. For this to be successfully achieved agents that represent the drivers and agents attempting to schedule a driver to a particular time slot must be developed. BIDS agents will be based upon the model developed in the data modelling phase of KDDS-BI. As discussed, the agent negotiations will be based upon that of a bargaining system. BIDS agents attempting to schedule a driver are essentially a 'buyer of resources'; since they are acquiring (buying) through bids, the resources (drivers). Thus, the BIDS scheduler agent will attempt to assign the driver with a corresponding shift time and the highest grade. BIDS agents representing the driver embody a 'seller of the resources'. These agents, consequently, attempt to get the driver they represent assigned to a shift. In addition a GUI will be developed through Java Swing. The BIDS GUI will allow a user to input the details of a driver, thereby creating an agent to represent the driver. Although the agents and negotiations (conducted using ACL) are based upon the design explored during the data modelling phase, BIDS agents implement the Jade classes (Jade, 2008):

- *Behaviour*: This abstract class provides an abstract base class for modelling agent tasks, and it sets the basis for behaviour scheduling as it allows for state transitions (i.e. starting, blocking and restarting a Java behaviour object). This class is executed to initiate or respond to messages.
- *One shot behaviour*: This abstract class, models atomic behaviours that must be executed only once and cannot be blocked. Hence, the: 'done()' method, always returns 'true'. This behaviour class is explored to ensure that a new driver is added to the roster, once entered into the system.
- *Ticker Behaviour*: This abstract class implements a cyclic task that must be executed periodically. The ticker behaviour is explored to ensure that the agents will periodically search for potential drivers. As a result, with each 'tick' (a pre-determined duration) the behaviour is iterated to satisfy the requests received from the BIDS agents representing drivers.

In contrast to BIDS, the train timetable scheduler BITS has been developed using the i-Enterprise platform. However, this approach not only requires the declaration of agents. But furthermore, requires the development of a suitable ontology. In addition to a representation of the real-world, this representation of the target domain can

be explored as a simulated environment within which the agents can conduct communication to negotiate train journeys.

Since the ontology forms the basis upon which the system operates, therefore, must include all proposed objects and their properties, in conjunction with the relationships that inter-relate these constituent components. Studying the conceptual model, each train within the LUL network is required to be modelled along a route. For each *'Line/route'* there will be a set number of requirements, which in turn can form a basis upon which the ontology is created:

- Each route has a number of stations at which it must stop, as well as segments of track that connect the stations.
- Each station has trains that stop at it and the route that it is located on.
- Each train must be assigned to a route and stations, as well as have carriages and a driver.

Examining, these requirements and constraints a suitable ontology can be developed. Initially a descriptive ontology must be defined (figure D.12). The descriptive ontology defines the entities and objects within the simulated environment the 'virtual world'. Hence, these entities and objects must be defined (figure D.13) and assigned icons that will be used to represent the object in the virtual world (figure D.14). Once declared and assigned icons, attributes must be designated to the objects; these can be selected from a variety of data types (figure D.15).

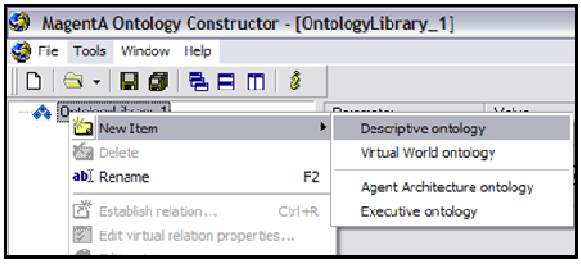


Figure D.12: Defining a new descriptive ontology.

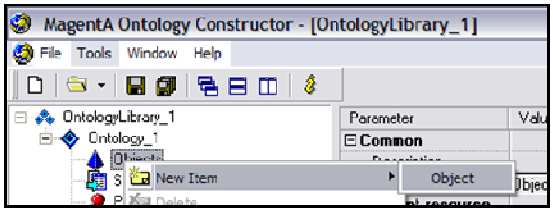


Figure D.13: Defining objects within the descriptive ontology.

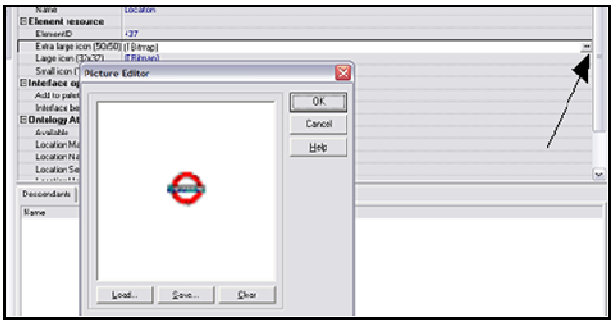


Figure D.14: Assigning images to represent objects.

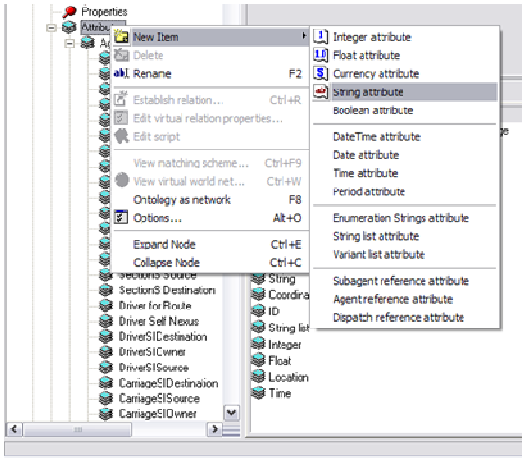


Figure D.15: Object-attribute declaration.

With the attributes and entities have been defined and declared, 'Demand Agents' and 'Resource Agents' must be created (figure D.16). These agents will represent that requirements of users (demands) and resources of LUL. Once the required agents have been created, the following stage is to create relationships between these components (figures D.17 & D.18). This will permit the declaration of the matching conditions and decision-making machine attributes that will allow the agents to accurately communicate. The decision-making machine contains the conditions and scripts that dictate the possible behaviours and communication of the agents in addition to creating templates for matching relationships. The decision-making machine is the instrument through which all information of the descriptive ontology is passed to the agents that are then implemented in *scenes* in the *virtual-world* in order to mimic real-world situations.

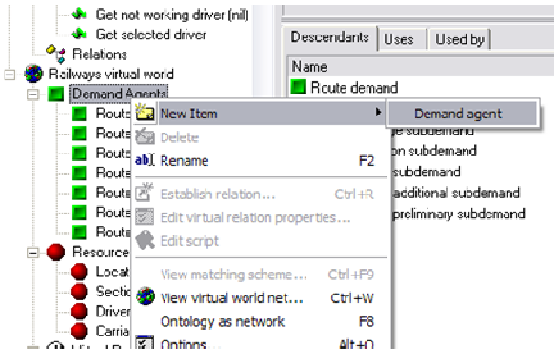


Figure D.16: Development of agents.

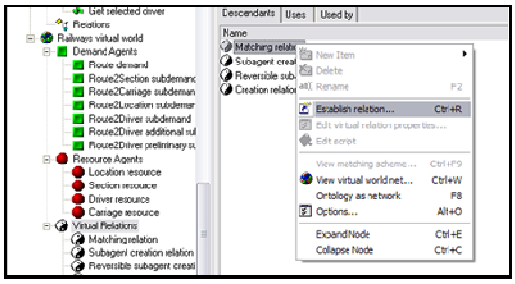


Figure D.17: Defining Agent relationships.

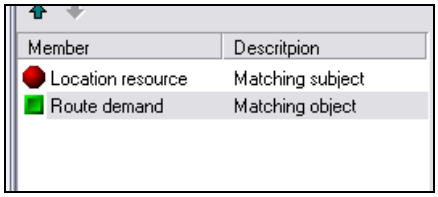


Figure D.18: Agent relationships.

The descriptive ontology will be used to intermediate between the user and the system, providing a platform for the virtual world. The virtual world provides a simulated environment in which the schedules for trains, in addition to possible scenarios can be simulated. Hence, the scenarios form the core of the BI solution, since it is these scenarios that will provide decision makers the opportunity to simulate constraints and develop schedules that can be used to support the decision making process. Consequently, the circumstances simulated, referred to as 'scenes', not only simulate the real-world, but furthermore, specify the agents and the actions that the agents are to take. These scenes can be created in advance and saved or implemented at the time-of-use. Although, the possible number of scenes that can be generated is immeasurable, since the number of possible situations that can be encountered in the real world is infinite. To realise the objectives of this investigation three ontology scenes will be formulated, explored and studied:

1. *Rollback in Scheduling*: Demonstrates the capability of BITS to readjust the schedule being produced to give priority to a more important demand or resource in the event a lower priority resource is required to be accommodated. Thereby providing insight in to how the agents can renegotiate in the event that demands or resources available change once a solution has been proposed.

2. *Dynamic Rescheduling*: Demonstrates the systems ability to react to the addition of a newer higher priority line. Previously scheduled services must adjust their schedules to accommodate the new service safely. While discovering stations with facilities that permit overtaking with the least amount of disruption. Hence, react in real-time to unexpected changes in demands or resources available.
3. *Recovery*: Demonstrates the systems ability to find alternative resources in the eventuality that resources are made unavailable. Furthermore, an alternative cross-proposal is submitted to the user. Hence, react in real-time to unexpected changes whilst ensuring that the 'next best' alternative is calculated.

Figure D.19: Palette defining entities in the virtual world.

The scenes must be declared in the virtual world through the descriptive ontology. Thus, the ontology must be loaded into the virtual world interface. The interface is awkwardly known as the 'physical world', since this is the 'physical' representation of the domain simulated within the 'virtual world'. Within the physical world the palette (figure D.19) provides access to the ontology entities, in this event consisting of:

- *Route*: 
- *Location*: 
- *Drivers:* 
- *Carriages:* 

Thus, the palette extends the ontology to provides the means through which the entities can be declared within the physical world, e.g. to create a station click on 'Location ->  ', then click on the point on the physical world on the proposed location of the entity. Once the required number of stations has been declared, they must be inter-connected. Click on the lines creation button, then left click on the origin station, hold the mouse button and release once over the next station. In the dialogue box select 'Section' (figure D.20). 'Section' provides the entity that represents track, whilst 'route' determines the direction of travel.

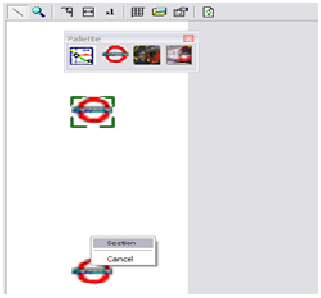


Figure D.20: Physical world.

A full representation of the real-world must then be constructed in this manner, to permit an accurate simulation environment for agent negotiation. As declared in the ontology, if a schedule is to be created, a driver and designated number of carriages must be added. Clicking on the initial station and the final station opens a dialogue box. This will allow the user to set the properties of the line (figure D.21). However, if properties of the agents are to be changed at a later stage, be this of the route or any other agent. Right click on the agent, and choose *'Actions ⇒ Show Personal Inspector'* (figure D.22).

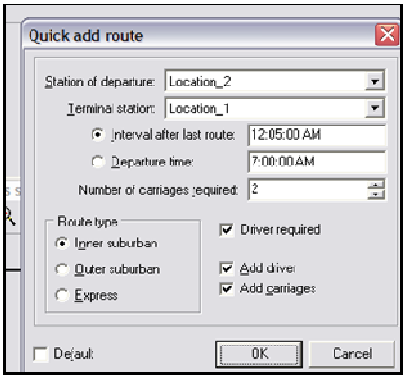


Figure D.21: Line properties.

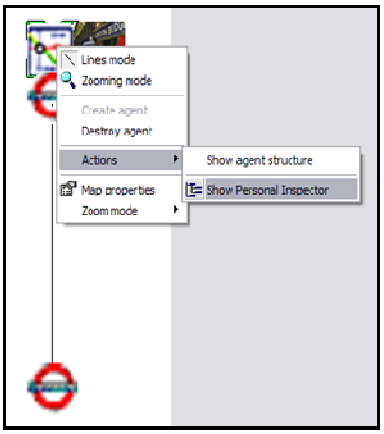


Figure D.22: Agent properties.

The property inspector for 'Route' also allows you to set the time segments that separate each station/location. This allows the delay between each station to be specified (figure D.23). The delay is set to simulate the average time for which a train will remain at a station prior to re-commencing the route.

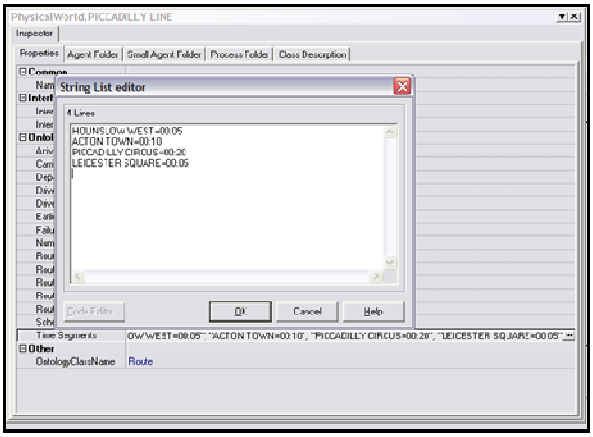


Figure D.23: Average duration of a time for which a train remains at a station.

With the correct specifications established and implemented within the system, whilst adhering to the defined constraints, this process will enable the specified scenes to be animated. The BI solution will then be able to provide the user with a schedule and report that will be free from any conflicts and thus saving on logistical time and effort.

D.2.4 Decision Support

The 'Development' phase of KDDS-BI has provided the opportunity to discover a suitable approach, and investigate and create a representation of the LUL network, within which agents can negotiate. The proposed BI solution BIT and BIDS can accordingly be analysed to interrogate the LUL data (resource information) to facilitate the agents to negotiate the optimum allocation of resources, so that this information can be exploited for decision support. The initial stage of the 'Decision Support' phase of KDDS-BI is to gather the output by applying the BI techniques to the dataset (figure D.24). Thus, both BIDS and BITS will both be further explored, by assigning drivers to schedules, in addition to the scenarios defined in the development phase of KDDS-BI, permitting the analysis of the proposed BI solution to realise the objectives. These results can then be analysed to extract valuable knowledge.

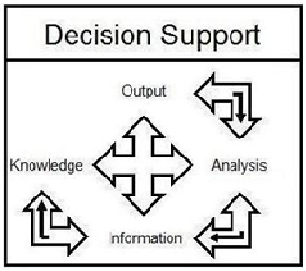


Figure D.24: Decision Support stage of KDDS-BI to provide business decisions.

The initial analysis was conducted through BIDS. The objective of BIDS is to deploy agents that will attempt to schedule the most suitable driver, hence the driver with the highest grade to the explicit shift they represent. In BIDS the user via a command line parameter will inform each scheduler-agent of the time slot that requires a driver. The scheduler agent will then negotiate with the DF agent and AMS, to provide details of agents representing drivers in an attempt to acquire the most suitable driver (resource) available to occupy the predetermined shifts. Messages exchanged by BIDS agents have a format specified by the ACL language defined by FIPA, as discussed in the literature review. This format comprises a number of fields and in particular:

- The sender of the message.
- The list of receivers.
- The communicative intention (also called 'performative') indicating what the sender intends to achieve by sending the message. The performative can be REQUEST, if the sender wants the receiver to perform an action, INFORM, if the sender wants the receiver to be aware a fact, QUERY_IF, if the sender wants to know whether or not a given condition holds, CFP (call for proposal), PROPOSE, ACCEPT_PROPOSAL, REJECT_PROPOSAL, if the sender and receiver are engaged in a negotiation, and more.
- The content i.e. the actual information included in the message (i.e. the action to be performed in a REQUEST message, the fact that the sender wants to disclose in an INFORM message …).
- The content language i.e. the syntax used to express the content (both the sender and the receiver must be able to encode/parse expressions compliant to this syntax for the communication to be effective).
- The ontology as discussed this is the vocabulary of the symbols used in the content and their meaning (both the sender and the receiver must ascribe the same meaning to symbols for the communication to be effective).
- Some fields used to control several concurrent conversations and to specify timeouts for receiving a reply such as; 'conversation-id', 'reply-with', 'in-reply-to', 'reply-by'.

Once all shifts have been entered into the system, the details of drivers can be submitted. Since the drivers represent a more dynamic resource. Consequently, subject to greater changes, the details of the driver are entered via the BIDS GUI (figure D.25). The motivation for the GUI for the entry of driver details is that this could be achieved by drivers entering their details into the system at terminals distributed throughout the LUL network. This will enable users/drivers to populate the local roster with their details, their grade and for the times that they are willing and more significantly able to commence their shift.

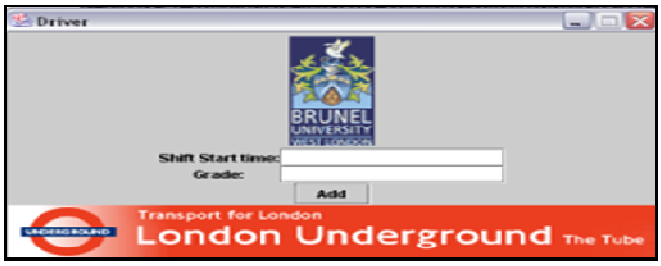


Figure D.25: GUI through which driver details can be entered.

Once all driver details have been added the agents representing the drivers (driver-rep agent) register with the DF agent and AMS ready initiate negotiations. Driver-rep agent incessantly waits for requests from scheduler agents. Upon receiving an invitation for an offer, the driver rep-agent will check the roster for any suitable drivers, if no drivers are available the scheduler-agent will be sent a refusal (figure D.26), in which case the scheduler-agent will then periodically request all driver-agents to provide an offer.

Figure D.26: No drivers available.

Figure D.27: Driver found and allocated to shift in roster.

Upon receiving an offer the scheduler-agent shall accept the offer from the highest graded driver that is available for that shift, thus issuing an order to assign the driver. The details of successful agent acquisitions (matchings) are then output to the command line, permitting decision makers to create rosters for the drivers (figure D.27). Once the driver rep-agent receives an assign driver order, the agent shall serve it by removing the driver from the roster, thereby making the resource unavailable. The system is dynamic and flexible enough to cope with multiple agents that each explicitly represents a driver. However, in the event that a more suitable driver becomes available, BIDS will renegotiate to assign this driver to the schedule, providing decision makers with a flexible system through which to explore various combinations for rosters. The scheduling agents can as a result assign drivers and potential 'next best alternative' drivers in the event that a driver becomes unavailable. Whilst continually searching for more optimised combinations, although as explored in the modelling phase a 'time-out' limitation can be applied, or the agent made to terminate once a successful resource has been assigned.

BIDS provide a forthright Multi-Agent system, which employs a limited number of agents to conduct negotiations and assign drivers. The system exhibits autonomous behaviour, therefore could be operated independently from human intervention. Due to the high-risk environment and bearing in mind the objectives of this investigation, BIDS provides autonomous and dynamic decision support, enabling the exploration of various resource allocation scenarios within the domain of driver scheduling. The objective of BIDS, as a result is to provide decision support, not autonomously execute changes to the environment (assign drivers to shifts), as is often the case with Multi-Agent systems. In contrast to BIDS, BITS provides a far more complex high-granularity Multi-Agent BI solution. Consisting of a large number of agents all of which simultaneously negotiate to satisfy their objectives BITS is capable of generating timetables for the LUL network, whilst permitting decision makers the opportunity to run simulated scenarios, to explore 'what-if' possibilities. In addition to this the interaction between agents can be viewed and details of how the agents were able to negotiate the allocation of resources. In order to demonstrate the capabilities and opportunities of BITS, 3 scenarios were constructed to satisfy the objectives of this investigation. In addition, to explore potential situations in which a dynamic approach to scheduling would provide the means through which to optimally allocate resources. Furthermore, schedules are generated for all resources, a well as usage level for each station and segment which can be viewed as reports or visual representation in the form of a graph. The reports are generated as a HTML page, while the generated schedules of services and drivers can be exported to a spreadsheet if required.

The initial scenario investigated; 'Rollback in Scheduling', demonstrates the capability of BITS to readjust the schedule being produced to give priority to a more important demand or resource in the event a lower priority resource is required to be accommodated. Thereby providing insight in to how the agents can renegotiate in the event that demands or resources available change once a solution has been proposed. Initially the details of the train must be declared to the system. Figure D.28 illustrates the properties of an 'underground express' departing from Heathrow station to Leicester Square. The express train represents the potential to investigate the use of express trains, or a train travelling along the same section from another route, i.e. numerous routes in LUL network overlap. Once BITS is initiated the agents will negotiate to discover a suitable schedule. Figure D.29 represents the virtual world and depicts the agent negotiations required to generate the schedule (figure D.30). In the virtual world (figure D.29) the dotted-blue lines are successful matches between resources (track, carriages, drivers, stations etc) and demands, whereas, the solid-pink lines represent alternates, hence, the 'next-best alternative'.

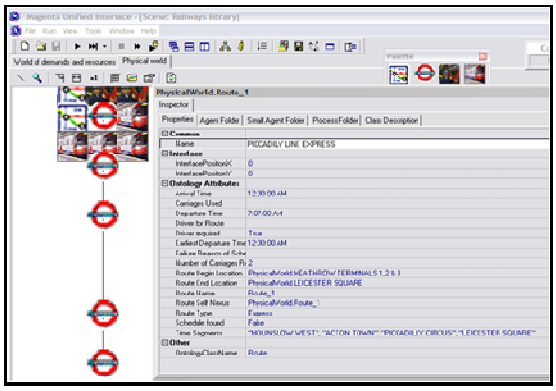


Figure D.28: Details of train to be scheduled.

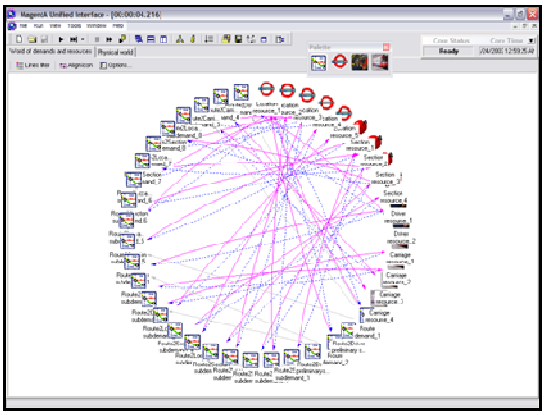


Figure D.29: In the virtual world the dotted-blue lines are successful matches between resources (track, carriages, drivers, stations etc) and demands, whereas, the solid-pink lines represent alternates, hence the 'next-best alternative'.

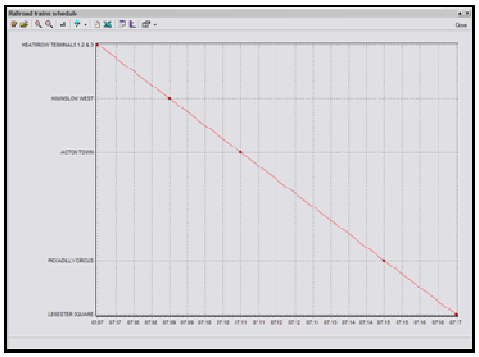


Figure D.30: Schedule for train detailed in figure D.28.

In the event a lower priority train is introduced (figure D.31). The agents must renegotiate to ensure the best possible matching (figure D.32). The result of this renegotiation is that a new schedule generated (figure D.33). However, prior to this, there is a conflict that must be resolved. The new train and the Express train will simultaneously require the same resource (track), between Acton Town and Piccadilly Circus, hence, create a

conflict. In such an event, the Express train must overtake; this requires a station with sufficient facilities. Thus the lower priority train is delayed at 'Hounslow West' to allow the higher priority express to use the track (figure D.34). The agent negotiations can be further examined by double-clicking on an agent to view the 'decision-making machine' (figure D.35). The decision making machine details the conditions and scripts that dictate the possible behaviours and communication of the agents in addition to creating templates for matching relationships. In addition, the details of the scene and information of resource usage can be viewed as a report (figure D.36).

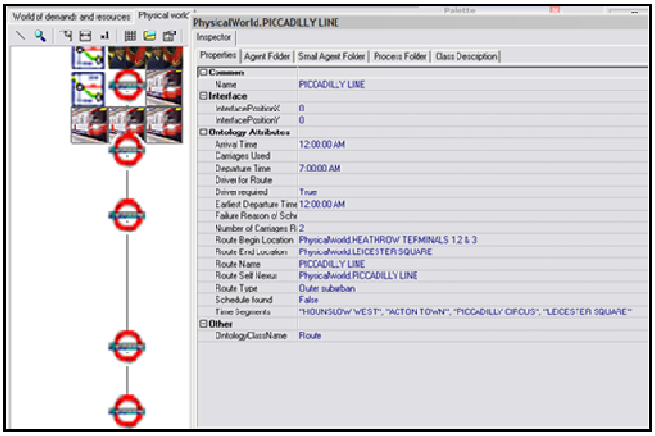


Figure D.31: Details of a lower priority train entered.

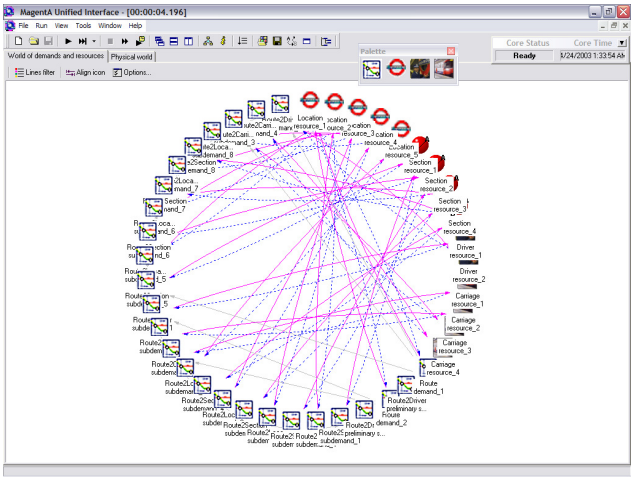


Figure D.32: Re-negotiation amongst agents to accommodate the new demands and resource allocation.

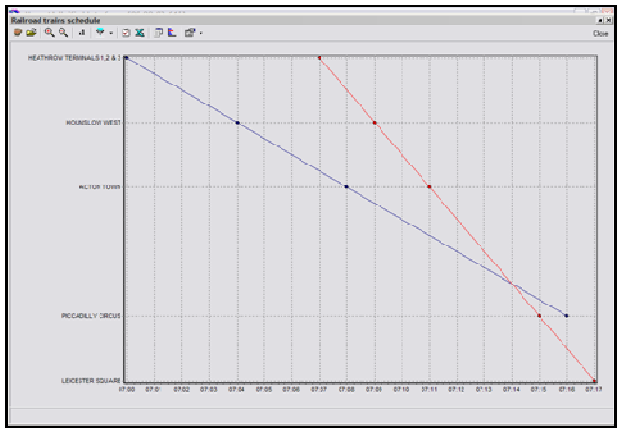


Figure D.33: New schedule generated with conflict detected at '07:14', between Acton Town and Piccadilly Circus.

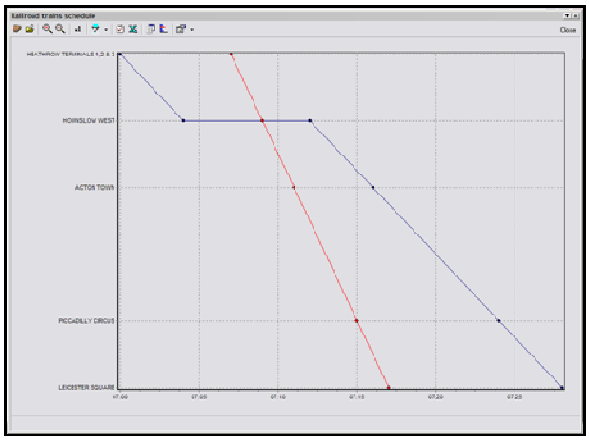


Figure D.34: Schedule modified to resolve conflict. Lower priority train remains at Hounslow West, permitting higher priority train to occupy resources first.

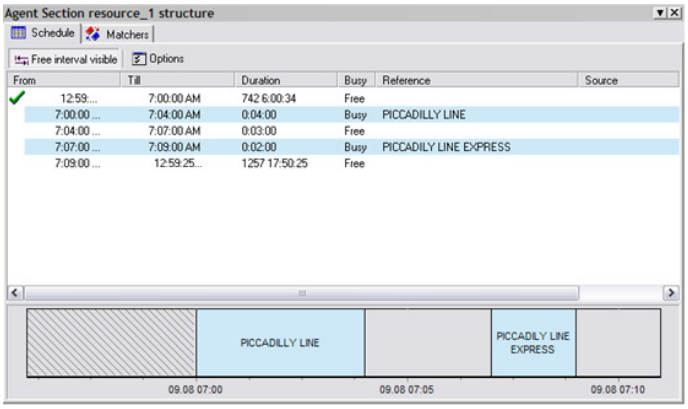


Figure D.35: Resource usage for a section of track.

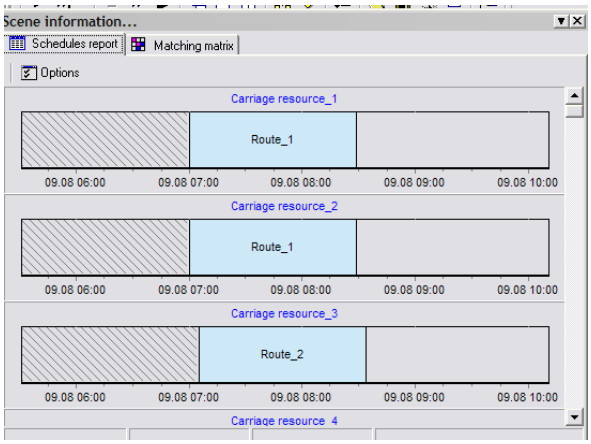


Figure D.36: Resource usage for carriages.

The second scene explored; 'Dynamic Rescheduling', demonstrates the systems ability to react to the addition of a newer higher priority line. Previously scheduled services must adjust their schedules to accommodate the new service safely. While discovering stations with facilities that permit overtaking with the least amount of disruption. Hence, react in real-time to unexpected changes in demands or resources available. To demonstrate this opportunity BITS must react to the addition of a 'High-priority Service', this requires current trains on the same line to dynamically alter schedules. Each Service/route must have a driver and carriage assigned to it. The stations 'Osterley', 'Barons Court' and 'South Kensington' have been defined to allow overtaking. Using these specifications three services are scheduled with an interval of five-minutes between departures (figure D.37). The schedule for these services is illustrated in figures D.38 as a graph and D.39 as a report.

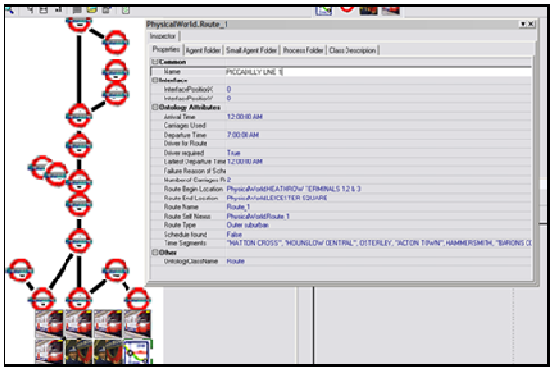


Figure D.37: Initial schedule for three trains departing at 5 minute intervals is created.

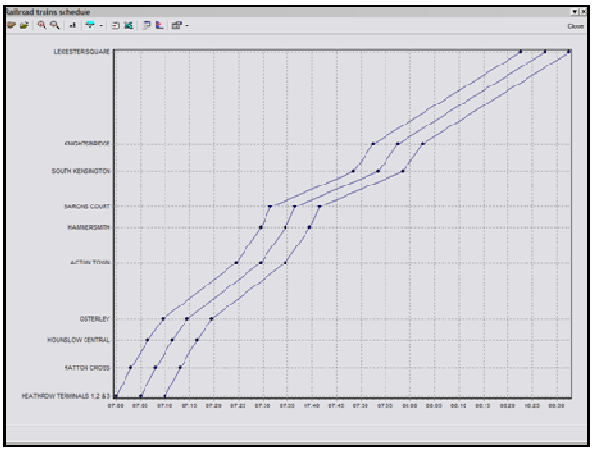


Figure D.38: Schedule for services as a graph.

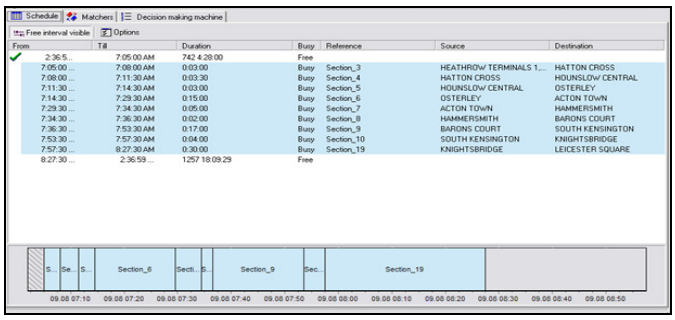


Figure D.39: Schedule for services as a report.

In the event that a new train, with higher priority, then those previously scheduled is introduced in to the timetable (figure D.40). The agents must re-negotiate and provide this service with access to resources. As a result the re-negotiation provides a new schedule (figure D.41). As a result the lower priority trains are retained at 'Osterley', 'Barons Court' and 'South Kensington' to avoid conflicts, permitting the faster train to overtake. The 'Show load levels' button:  , on the trains schedule dialogue box, will illustrate the load levels for each station (figure D.42). This informs the user of the usage level of each station and section.

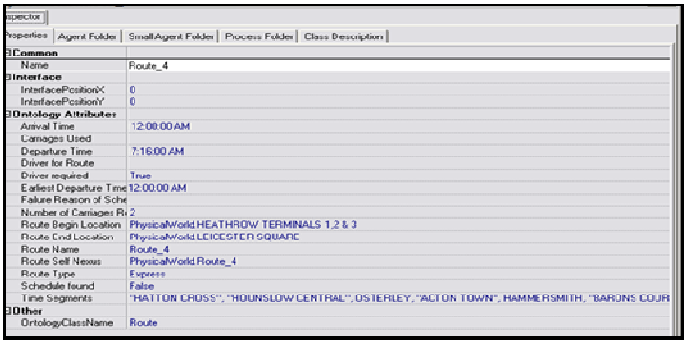


Figure D.40: Higher priority train introduced into schedule.

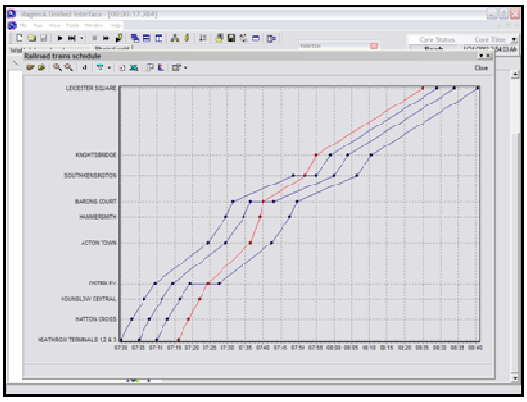


Figure D.41: Schedule modified to resolve conflicts. Lower priority trains remain at Osterley, Barons Court and South Kensington, permitting higher priority train to occupy resources.

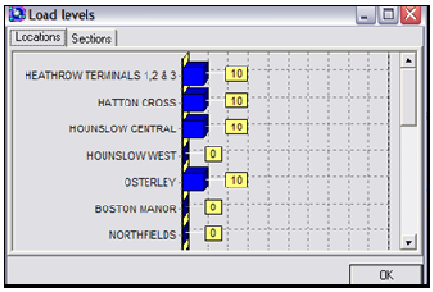


Figure D.42: Load levels for each station, illustrating the usage level of resources. Since the higher priority train does not stop at all stations, the bypassed stations illustrate a load level of 0.

The final scenario investigated demonstrates the systems ability to find alternative resources in the eventuality that resources are made unavailable ('Recover'). In addition, an alternative cross-proposal is submitted to the user. Illustrating the ability of BITS to react in real-time to unexpected changes whilst ensuring that the 'next best' alternative is calculated. Hence, this scenario represents one that is essential within the LUL network. In the event that any resource becomes unavailable (faulty track, train, etc.), it is imperative that BITS agents are capable of negotiating for alternative resources and generate a cross-proposals to the user. Figure D.43 illustrates the details of a train on the Piccadilly Line. With a driver who commences his shift at 2 a.m., with the scheduled shift terminating at D a.m. Figure D.44 illustrates the schedule generated by the agent negotiations. In addition the details of the drivers shift may be exported to a spreadsheet (figure D.45).

A train can be scheduled for the return journey using the same resources. The agents will as a result negotiate and insert the new train scheduled in to the timetable (figure D.46). However, since the train departs at a time once the driver has concluded his shift the resources, bar the driver are re-used. The introduction of a second driver with a suitable shift is illustrated by the generated report in figure D.47, in this instance the report has been exported to a HTML output. If however, the driver is made unavailable (figure D.48) and since the criteria for a valid route insists upon the allocation of a driver the resulting schedule will display an error message and

attempt to cross-propose the user with alternative resources, in this instance driver 3, whom is valid for that time allocation (figure D.49).

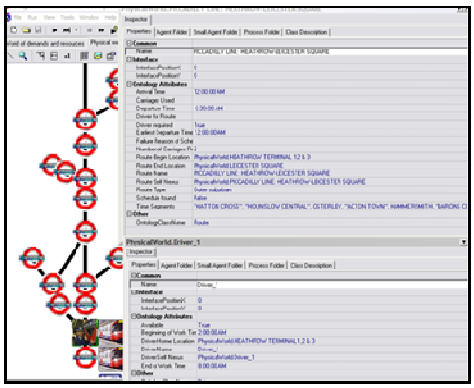


Figure D.43: Schedule for a train on the Piccadilly Line. With a driver for which is scheduled to work from 2 am until 8 am.

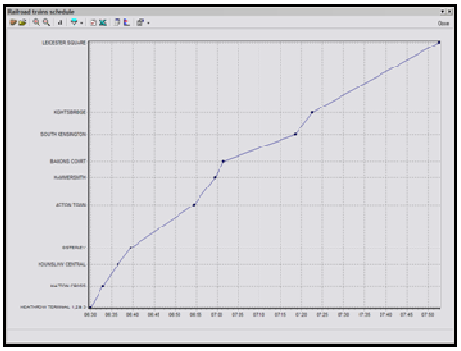
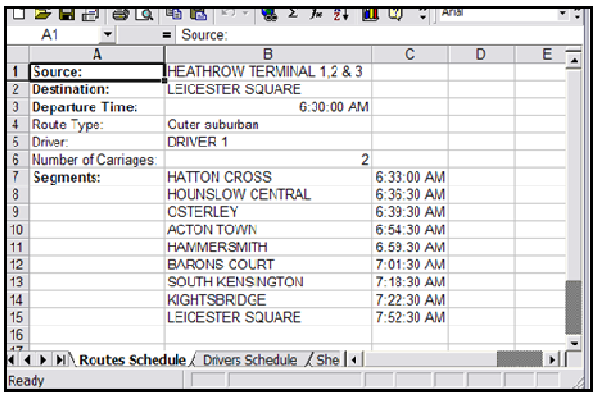


Figure D.44: Schedule generated for the train route.



	A	B	C	D	E
1	Source:	HEATHROW TERMINAL 1,2 & 3			
2	Destination:	LEICESTER SQUARE			
3	Departure Time:	6:00:00 AM			
4	Route Type:	Outer suburban			
5	Driver:	DRIVER 1			
6	Number of Carriages:	2			
7	Segments:	HATTON CROSS	6:33:00 AM		
8		HOUNSLOW CENTRAL	6:36:30 AM		
9		OSTERLEY	6:39:30 AM		
10		ACTON TOWN	6:54:30 AM		
11		HAMMERSMITH	6:59:30 AM		
12		BARONS COURT	7:01:30 AM		
13		SOUTH KENSINGTON	7:13:30 AM		
14		KIGHTSBRIDGE	7:22:30 AM		
15		LEICESTER SQUARE	7:52:30 AM		
16					

Figure D.45: Details of the drivers shift exported to a spreadsheet.

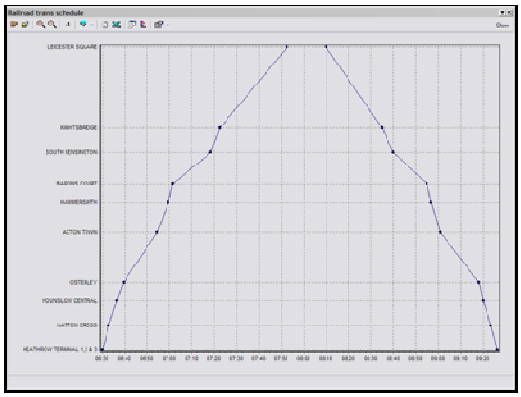


Figure D.46: Updated schedule, including details of the train for the return journey.

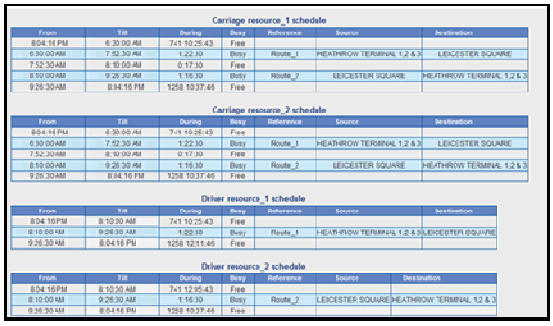


Figure D.47: HTML based report for resource utilisation.

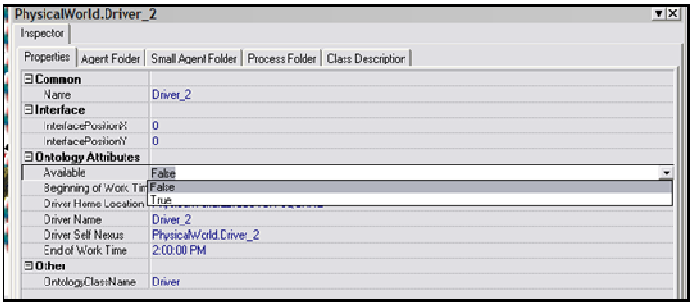


Figure D.48: Driver made unavailable.

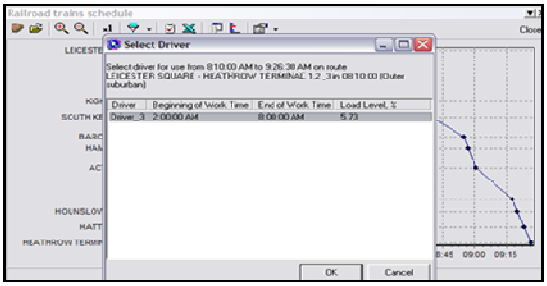


Figure D.49: Alternative Driver proposed.

To investigate the applicability of KDDS-BI within the domain of resource allocation two intelligent agent based BI solutions have been explored; BIDS and BITS. Furthermore, as permitted with KDDS-BI there has been a degree of analysis as the output has been collected, so that information can be extracted and transformed in to knowledge for decision support. BIDS provided a low granularity multi-agent system to schedule drivers to shifts. BIDS agents still posses a great degree of massaging and communicative flexibility, demonstrated by the dynamic ability to search for resources while ensuring the resource acquired results in optimum allocation. The system was tested using a combination of:

- Limited shifts and a high quantity of drivers.
- Limited shifts and a limited quantity of drivers.
- A high number of available shifts and a high quantity of drivers.
- A high number of available shifts and a limited quantity of drivers.

BIDS performed well under these circumstances, displaying the ability to effectively allocate resources. Although explored as a means through which to allocate drivers to shifts, the BIDS can readily be extended to allocate any organisational resource. In contrast to BIDS, BITS provides a more complex intelligent-agent based BI solution. The ability to schedule trains, in multiple events has been a complex problem for which intelligent agents provide a novel solution. The scheduling of trains, due to its complexity can be compared to a supply-chain, hence the use of intelligent agents that are not mobile. In contrast, BITS agents reside in the user's workstation/network; are more secure; developed in highly efficient language; thus enabling them to be considerably more efficient. The number of agents in BITS, as with any Multi-Agent system will vary, depending upon the task. However, since in the potential number of agents simultaneously negotiating, especially in the created scenes, is significantly large, BITS is considered a 'high granularity' solution. In addition, BITS demonstrates the advantages of using a multi-agent approach, while permitting schedules to be generated and calculate usage levels of each of the resources. Furthermore, these details can be exported into an excel file. It is however, the 'ontology scenes' which provide true insight into the performance of the system as a decision support tool. These scenes represent simulations of real-world opportunities and dilemmas. In order to demonstrate the capabilities of the system three ontology scenes were constructed:

- *Recovery*: Tests the systems ability to negotiate for alternative resources in the eventuality that current resources become unavailable or superior alternatives become available. Altering the shift of a driver tested the system, so that if no driver was available, nor where any carriages assigned for the return route. The system with absolutely no delay in execution time was able to designate carriages located at the station to the route and report to the user the error of no driver and list alternatives.
- *Rollback in Scheduling*: Evaluated the systems ability to adjust the schedule of a new train to account for a previously defined schedule with a high priority and allow for overtaking at stations where resources allow with minimal delays. The system efficiently created new schedules and allowed for reports to be generated.
- *Dynamic Changes to Schedule*: Demonstrated the ability for a number of previously defined low priority trains to adjust their schedules upon the addition of a high priority train and allow overtaking at

relevant points bearing in mind safety implications. The system was able to implement these alterations while ensuring that the trains did not exceed defined speeds and allocated the time segments accurately.

Since, the system performed well, whilst providing suitable output and analysis of the scenarios, which can be explored for potential 'what-if' schedules, or even implemented to generate 'active' schedules. It can be deemed that, BITS provides a valuable BI solution through which resources can be allocated optimally within an organisation.

D.3 Conclusion

This case study has endeavoured to investigate the applicability of KDDS-BI to discover valuable information to support the decision making process within the field of 'managing organisational resources'. Managing organisational resources was selected since having investigated various aspects of marketing. Managing organisational resources forms a major aspect of 'operations management', which is another basic function of an organisation (Stevenson, 2007) as defined in the introduction of this research. Hence, scrutinising this aspect of an organisational activity not only provides insight into the applicability of BI strategies across an organisation, but furthermore, the capabilities of KDDS-BI to structure such an investigation. Consequently, the investigating BI to provide solutions and KDDS-BI to structure these investigations provides far greater opportunities then that of just a marketing tool, but rather one which can be explored to manage the supply chain.

Due to organisations competing on a global level, in previous; untapped markets, these organisations are required to amass an ever-increasing amount of resources. However, increasing capacity in this manner can result in an organisation falling prone to diseconomies of scale. It is, therefore, imperative that to remain competitive organisations manage and distribute these resources, be they physical, financial or human, effectively and optimally. Optimally allocating resources throughout an organisation can exponentially increase the level of efficiency. In order to analyse the performance of KDDS-BI in such circumstances, the LUL rail network was scrutinised. LUL was selected due to the vast number of resources that form the network. Furthermore, the formidable task posed by scheduling these resources is one which requires significant resources, does not permit real-time analysis, or the exploration of 'what-if' options. Consequently, any deviation from original plans, which is regularly encountered in an environment as dynamic as the LUL network, results in significant delays. Since LUL is a public service, the 'profit' or value added is measure not financially, but rather in the level of service provided. Hence, there is significant incentive to investigate the integration of technology that can improve the quality of the service, especially in light of the limitations of conventional enterprise resource planning/advanced scheduling and planning applications.

In contrast to data that has been collected through the course of daily organisational activity, the dataset in this event consists of the resources that an organisation possesses and must manage. Therefore, the LUL network was analysed to identify the major resources that must be routinely distributed. Upon investigating these resources it was determined that BI strategies, such as those consisting of an intelligent agent approach provides

a solution which is both novel and innovative. Intelligent agents provide an effective means through which dynamic, complex and unpredictable conditions can be effectively managed. Furthermore, this area of investigation provided further opportunity to analyse the performance of KDDS-BI, in the event it is required to provide structure to investigations other then those involving marketing and advanced analytics. Upon analysing the data (resources) and requirements of the LUL network, it was determined that the investigation would focus upon a subset of the network, this would permit the analysis of the diverse and flexible nature of KDDS-BI to facilitate such exploration , in addition to the performance of the selected BI strategy. The results of the analysis can be extrapolated to encompass the entire network; however, this was beyond the scope of this investigation. Having modelled the structure of the agents, in addition to determining the most effective means through which the constituent agents would conduct negotiations. It was ascertained that a combination of open-source and commercial platforms would provide the most effective solution. Furthermore, this combinatorial platform provided the opportunity to investigate and analyse a solution consisting of both; limited agents, and a vast number of agents. As a result, two solutions were investigated and discovered:

- BIDS: This facilitated the allocation of human resources (drivers) to scheduled shifts.
- BITS: This facilitated the allocation of all resources that a train route consists of.

BIDS, provided insight into the process of agent negotiation, whilst determining the most appropriate driver to schedule to a shift. BITS in contrast provided a far more complex system, providing the dynamic scheduling of all resources to provide timetables for trains. The core facet of BITS, is the ontology, which defines the criteria of all negotiation, in addition to the attributes of the various entities/resources. Once an appropriate ontology had been explored and discovered, it was possible to further examine the ability of the agents to negotiate for the defined resources. This was realised through the investigation of a three ‘scenes’. These scenes simulated real-world scenarios, thereby permitting the analysis of an intelligent agent approach to organisational resource planning. Analysing the performance of the proposed novel and innovative solutions provided a basis upon which to asses the performance of KDDS-BI as a framework to provide decision support for managing organisational resources. The calculated time-tables and schedules provided significant results, in addition to facilitating ‘what-if’ scenarios. These results can be presented to decision-makers/managers, to enable them to more effectively allocate resources throughout an organisation. The LUL network was selected, since this was a complex domain and therefore represented a challenging environment. However, the process of allocating resources throughout an organisation or planning in advance the optimal allocation is one which can be realised through the same method of investigation.