Investigation of the visual aspects of business intelligence

Niall Cunningham

Dublin Institute of Technology, brendan.tierney@dit.ie

Follow this and additional works at: http://arrow.dit.ie/scschcomdis

Part of the Computer Engineering Commons

Recommended Citation
http://arrow.dit.ie/scschcomdis/9

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License
Investigation of the Visual Aspects of Business Intelligence

Niall Cunningham

A dissertation submitted in partial fulfilment of the requirements of Dublin Institute of Technology for the degree of M.Sc. in Computing (Knowledge Management)

November 2008
I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Knowledge Management), is entirely my own work and has not been taken from the work of others and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute’s guidelines for ethics in research.

Signed: __________________________

Date: 03 November 2008
ABSTRACT

As the need for Irish firms to move into knowledge economy increases knowledge management is being more important. Among its central aims is for knowledge creation and development to enhance knowledge and inspire innovations. One of the ways knowledge management achieves these aims is through the development of knowledge management tools for the firm.

Among these tools is the option to use business intelligence to enhance knowledge in the firm by exposing hidden knowledge and building on it. In particular data mining has been highlighted as being effective within business intelligence and knowledge management in discovering hidden knowledge.

This dissertation investigates these systems and looks at how visual data mining in this area is enhancing knowledge creation throughout the firm and can give firms a competitive advantage over rivals. It considers the difficult decision for knowledge management practitioners at selecting the correct tool and provides a framework which can be apply in different situations. The dissertation concludes with analysis of the results and future work in the area.

Key words: Knowledge management, Knowledge management tools, Business Intelligence, Data Mining, Visualisation, Evaluation Framework, Visual Data Mining Tools, Knowledge Creation
ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my supervisor Brendan Tierney for his valuable advice, insights and guidance throughout this project.

I would also like to thanks to Brian Mac Namee for his help regarding the data mining tool section of this project.

Finally I would like to thank my family, friends and girlfriend Siobhan for their support during the year.
# TABLE OF CONTENTS

**ABSTRACT** .................................................................................................................................................. 2

**TABLE OF FIGURES** .................................................................................................................................. 9

1 **INTRODUCTION** .................................................................................................................................... 12
   1.1 PROJECT INTRODUCTION ................................................................................................................. 12
   1.2 BACKGROUND .................................................................................................................................... 14
   1.3 RESEARCH PROBLEM ....................................................................................................................... 15
   1.4 INTELLECTUAL CHALLENGE ............................................................................................................ 17
   1.5 RESEARCH OBJECTIVES .................................................................................................................. 18
   1.6 RESEARCH METHODOLOGY ........................................................................................................... 19
   1.7 RESOURCES ...................................................................................................................................... 20
   1.8 SCOPE AND LIMITATIONS .............................................................................................................. 21
   1.9 OBJECTIVES OF THE FRAMEWORK ............................................................................................... 22
   1.10 ORGANISATION OF THE DISSERTATION ....................................................................................... 23

2 **KNOWLEDGE MANAGEMENT** ................................................................................................................ 1
   2.1 INTRODUCTION ................................................................................................................................... 1
   2.2 DEFINING KNOWLEDGE .................................................................................................................... 1
   2.3 DEFINING KNOWLEDGE MANAGEMENT ......................................................................................... 1
   2.4 OBJECTIVES OF KNOWLEDGE MANAGEMENT ............................................................................. 2
   2.5 KNOWLEDGE DEVELOPMENT .......................................................................................................... 3
   2.6 BUSINESS INTELLIGENCE AND DATA MINING .............................................................................. 4
   2.7 CONCLUSION ..................................................................................................................................... 7

3 **DATA MINING** ....................................................................................................................................... 9
   3.1 INTRODUCTION ................................................................................................................................... 9
   3.2 KNOWLEDGE DISCOVERY ................................................................................................................ 9
   3.3 DEFINING DATA MINING ............................................................................................................... 11
   3.4 DATA MINING APPROACHES ........................................................................................................... 13
      3.4.1 Predictive Modelling ................................................................................................................ 15
      3.4.2 Database Segmentation ........................................................................................................... 15
      3.4.3 Link Analysis ........................................................................................................................... 16
      3.4.4 Deviation Detection ................................................................................................................ 17
   3.5 THE IMPORTANCE DATA MINING ................................................................................................. 17
   3.6 DATA MINING MODELS .................................................................................................................... 18
      3.6.1 Decision Trees and Rules ......................................................................................................... 19
      3.6.2 Linear Regression .................................................................................................................... 20
   3.7 CONCLUSION ..................................................................................................................................... 22

4 **ORGANISATION OF THE DISSERTATION** .............................................................................................. 23
## VISUALISATION

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>4.2</td>
<td>Defining Visualisation</td>
</tr>
<tr>
<td>4.3</td>
<td>Benefits of Visualisations</td>
</tr>
<tr>
<td>4.4</td>
<td>Visualisation Theories</td>
</tr>
<tr>
<td>4.5</td>
<td>Information Visualisation</td>
</tr>
<tr>
<td>4.6</td>
<td>Visual Analytics</td>
</tr>
<tr>
<td>4.7</td>
<td>Knowledge Visualisation</td>
</tr>
<tr>
<td>4.8</td>
<td>Framework for Knowledge Visualisation</td>
</tr>
<tr>
<td>4.9</td>
<td>Function Type Perspective</td>
</tr>
<tr>
<td>4.10</td>
<td>Knowledge Type Perspective</td>
</tr>
<tr>
<td>4.11</td>
<td>Recipient Type Perspective</td>
</tr>
<tr>
<td>4.12</td>
<td>Visualisation Type Perspective</td>
</tr>
<tr>
<td>4.13</td>
<td>Steps for Visualising Data</td>
</tr>
<tr>
<td>4.14</td>
<td>Insight Gained</td>
</tr>
<tr>
<td>4.15</td>
<td>Conclusion</td>
</tr>
</tbody>
</table>

## VISUAL DATA MINING

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>5.2</td>
<td>Defining Visual Data Mining</td>
</tr>
<tr>
<td>5.3</td>
<td>Visual Data Mining Benefits</td>
</tr>
<tr>
<td>5.4</td>
<td>Types of Visual Data Mining</td>
</tr>
<tr>
<td>5.5</td>
<td>Visual Data Mining Exploration</td>
</tr>
<tr>
<td>5.6</td>
<td>Tasks to be Visualised</td>
</tr>
<tr>
<td>5.7</td>
<td>Visualisation Technique Classification</td>
</tr>
<tr>
<td>5.8</td>
<td>Data to be Visualised</td>
</tr>
<tr>
<td>5.9</td>
<td>Visualisation Techniques</td>
</tr>
<tr>
<td>5.10</td>
<td>Standard 2D/3D Display</td>
</tr>
<tr>
<td>5.10.1</td>
<td>Box Graphs</td>
</tr>
<tr>
<td>5.10.2</td>
<td>Line Graphs</td>
</tr>
<tr>
<td>5.10.3</td>
<td>Pie Graphs</td>
</tr>
<tr>
<td>5.10.4</td>
<td>Distribution and histogram graphs</td>
</tr>
<tr>
<td>5.10.5</td>
<td>Bar Chart</td>
</tr>
<tr>
<td>5.11</td>
<td>Geometrically Transformed Displays</td>
</tr>
<tr>
<td>5.11.1</td>
<td>Parallel Co-ordinates</td>
</tr>
<tr>
<td>5.11.2</td>
<td>Scatterplot Matrices</td>
</tr>
<tr>
<td>5.11.3</td>
<td>Projection pursuit</td>
</tr>
<tr>
<td>5.12</td>
<td>Icon Displays</td>
</tr>
<tr>
<td>5.13</td>
<td>Chernoff Faces</td>
</tr>
<tr>
<td>5.14</td>
<td>Dense Pixel Displays</td>
</tr>
</tbody>
</table>
8 EXPERIMENTATION AND EVALUATION ................................................................. 107

8.1 INTRODUCTION ........................................................................................................... 107
8.2 EXPERIMENTATION ..................................................................................................... 107
8.3 DATASETS ..................................................................................................................... 108
8.4 USER PROFILE ............................................................................................................. 109
8.5 BUSINESS QUESTIONS ............................................................................................... 110
8.6 VISUAL DATA MINING TOOL SELECTION ................................................................. 110
8.7 SAS ENTERPRISE MINER TOOL .................................................................................. 111
8.8 RAPIDMINER TOOL ..................................................................................................... 111
8.9 SPSS TOOL ................................................................................................................... 112
8.10 APPLYING THE FRAMEWORK ................................................................................... 112
8.11 CRITERIA SELECTION ............................................................................................... 112
8.12 TASKS ......................................................................................................................... 115
8.13 PHASE 1: FUNCTIONAL TESTS .................................................................................. 116
8.14 PHASE 2: USER TESTING .......................................................................................... 116
8.15 PHASE 3: USER INTERVIEWS ..................................................................................... 117
8.16 EXPERIMENT REVISIONS ......................................................................................... 117
8.17 EVALUATION .............................................................................................................. 118
8.18 PHASE 1: EXPERT USERS FUNCTIONAL RESULTS .................................................... 118
8.19 PHASE 2 AND 3: FEEDBACK AND INTERVIEWS ....................................................... 122
  8.19.1 The first impression of the tools ............................................................................. 122
  8.19.2 Is the data understandable from the overview in each tool ................................... 123
  8.19.3 Difficulty of creating a regression model .............................................................. 124
  8.19.4 Are the results understandable in each tool ....................................................... 125
  8.19.5 How to the decision trees compare ..................................................................... 125
  8.19.6 Are classification explanations quickly understood ........................................... 126
8.20 PHASE 3: USER INTERVIEWS ................................................................................... 127
  8.20.1 What were the good and bad things about the tools ........................................... 127
8.21 SELECTION OF THE BEST TOOL AND ANALYSIS OF RESULTS ......................... 127
8.22 CONCLUSION ............................................................................................................ 128

9 CONCLUSION ............................................................................................................... 130

9.1 INTRODUCTION ........................................................................................................... 130
9.2 RESEARCH DEFINITION AND RESEARCH OVERVIEW ........................................... 130
9.3 CONTRIBUTIONS TO THE BODY OF KNOWLEDGE ............................................... 131
9.4 EXPERIMENTATION, EVALUATION AND LIMITATION ........................................... 132
9.5 FUTURE WORK AND RESEARCH .............................................................................. 132
9.6 CONCLUSION ............................................................................................................. 133

BIBLIOGRAPHY ............................................................................................................. 135
**TABLE OF FIGURES**

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>STRUCTURE OF THE DISSERTATION</td>
</tr>
<tr>
<td>2-2</td>
<td>ZEILLER’S (2005) BUILDING BLOCKS OF KNOWLEDGE MANAGEMENT</td>
</tr>
<tr>
<td>2-4</td>
<td>THE BUSINESS INTELLIGENCE KNOWLEDGE CYCLE FROM WHITE (2005)</td>
</tr>
<tr>
<td>3-6</td>
<td>NEGnevtsky’s (2004) EXAMPLE OF A DECISION TREE FOR A NEW PRODUCT FOR HOMEOWNERS</td>
</tr>
<tr>
<td>3-7</td>
<td>FITTING A REGRESSION LINE TO A SET OF DATA POINTS HAN AND Kamber (2006)</td>
</tr>
<tr>
<td>4-8</td>
<td>THE SCOPE OF VISUAL ANALYTICS Thomas and Cook (2006)</td>
</tr>
<tr>
<td>4-9</td>
<td>THE FOUR PERSPECTIVES OF THE KNOWLEDGE VISUALISATION FRAMEWORK AND EACH OF THEIR COMPONENTS</td>
</tr>
<tr>
<td>4-10</td>
<td>INTERACTIONS BETWEEN THE SEVEN STAGES Fry (2007)</td>
</tr>
<tr>
<td>5-11</td>
<td>A GENERAL VISUAL DATA MINING FRAMEWORK DESIGN (SCHULZ, NOCKE AND SCHUMANN 2006)</td>
</tr>
<tr>
<td>5-12</td>
<td>A VISUAL DATA MINING FRAMEWORK DESIGN WITH REPRESENTATIVE MODULES (SCHULZ, NOCKE AND SCHUMANN 2006)</td>
</tr>
<tr>
<td>5-13</td>
<td>CLASSIFICATION OF INFORMATION VISUALISATION TECHNIQUES KEIM (2001)</td>
</tr>
<tr>
<td>5-14</td>
<td>CONFIGURATION OF A BOX PLOT (Bakker, A., Biehler, R., Konold, C. 2004)</td>
</tr>
<tr>
<td>5-15</td>
<td>TIPPETT’S WARP BREAK DATA (Bakker, A., Biehler, R., Konold, C. 2004)</td>
</tr>
<tr>
<td>5-16</td>
<td>MULTIPLE LINE GRAPH OF THE CAR DATA SET (Grinstein et al, 2001)</td>
</tr>
<tr>
<td>5-17</td>
<td>PIE CHARTS COMPARING MALE VERSUS FEMALE PERSONAL CELL PHONE USE RUMSEY (2007)</td>
</tr>
<tr>
<td>5-18</td>
<td>STACKED HISTOGRAM OF PRODUCTS SALES BY YEAR Dix, Beale and Wood (2000)</td>
</tr>
<tr>
<td>5-20</td>
<td>THE PARALLEL CO-ORDINATES PLOT OF THE VEGETABLE OIL DATA SET (Xu, Hong and Song, 2007a)</td>
</tr>
<tr>
<td>5-21</td>
<td>(A) SHOW AN ORIGINAL SCATTERPLOT MATRIX. THE DATA POINTS IN THE DIAGONAL PLOTS FORM A STRAIGHT LINE (B) SCATTERPLOT MATRIX WITH HISTOGRAMS. THE DIAGONAL PLOTS SHOW THE HISTOGRAM OF EACH DIMENSION (Cui, Ward and Rundensteiner 2006)</td>
</tr>
<tr>
<td>5-22</td>
<td>THE PARALLEL COORDINATES PLOT DISPLAYS THE US CENSUS DATA OF THE 50 STATES. COLOUR IS USED TO POINT OUT WHICH PARTY WON THE 2000 PRESIDENTIAL ELECTION</td>
</tr>
<tr>
<td>5-23</td>
<td>FIVE EXAMPLES OF CHERNOFF FACES KEIM (2005)</td>
</tr>
<tr>
<td>5-25</td>
<td>ARRANGEMENT OF ATTRIBUTE AREAS FOR DATA WITH SIX ATTRIBUTES KEIM (2005)</td>
</tr>
</tbody>
</table>
TABLE OF TABLES

TABLE 3-1  DATA MINING OPERATIONS AND ASSOCIATED TECHNIQUES (CABENA ET AL 1997) ........ 14
TABLE 6-2  COMPUTATIONAL PERFORMANCE CRITERIA (COLLIER ET AL 1999) .......................... 78
TABLE 6-3  FUNCTIONAL CRITERIA (COLLIER ET AL 1999) ...................................................... 78
TABLE 6-4  USABILITY CRITERIA (COLLIER ET AL 1999) .......................................................... 79
TABLE 6-5  ADDITIONAL TASK SUPPORT CRITERIA (COLLIER ET AL 1999) ............................... 79
TABLE 7-6  KNOWLEDGE MANAGEMENT QUESTIONS WHICH WILL BE ASKED TO EVALUATE THE TOOL .... 94
TABLE 7-7  SHOWS THE SOFTWARE QUALITY QUESTIONS WHICH WILL BE ASKED TO EVALUATE THE TOOL 96
TABLE 7-8  SHOWS THE USABILITY QUESTIONS WHICH WILL BE ASKED TO EVALUATE THE TOOL ...... 97
TABLE 7-9  SHOWS THE VISUALISATION QUALITY QUESTIONS WHICH WILL BE ASKED TO EVALUATE THE TOOL.............................................................................................................................. 98
TABLE 8-10 KNOWLEDGE MANAGEMENT QUESTIONS WHICH WILL BE ASKED TO EVALUATE THE TOOL 113
TABLE 8-11 SHOWS THE SOFTWARE QUALITY QUESTIONS WHICH WILL BE ASKED TO EVALUATE THE TOOL ........................................................... 113
TABLE 8-12 SHOWS THE USABILITY QUESTIONS WHICH WILL BE ASKED TO EVALUATE THE TOOL .114
TABLE 8-13 SHOWS THE VISUALISATION QUALITY QUESTIONS WHICH WILL BE ASKED TO EVALUATE THE TOOL........................................................... 114
TABLE 8-14 FUNCTIONALITY TOOL MATRIX PART1 .............................................................. 120
TABLE 8-15 FUNCTIONALITY TOOL MATRIX PART2 .............................................................. 121
1 INTRODUCTION

1.1 Project Introduction

As Ireland continues to witness manufacturing jobs disappear to countries with lower cost bases, efforts to move to a knowledge economy are taking on a new urgency, McGuire (2004). The challenge is to capitalise on innovations in research and development that will create wealth both in terms of jobs and the economic health of the country to counter these job losses.

The Irish government’s (2006) strategy towards encouraging this economy has been to launch the strategy for science, technology and innovation as part of the national development plan. In order to implement this strategy 3.8 billion is being invested up until 2013 (Finfacts 2006). This massive investment as outlined by O’Leary (2006) underpins the critical importance the government sees in creating a knowledge economy, in order for Irish firms to remain competitive against the emerging eastern European countries.

For these firms to be competitive Begley (2008) explains the vital resource of people’s knowledge has to be managed effectively to encourage the development of cutting edge ideas that will lead to new products and services. This knowledge management concept is new to most Irish firms who have yet to recognise it as a means to enable talented groups of people to maximise innovation by feeding off intellectual stimulation and motivation.

Firms working to align their knowledge management strategies to meet these new economic challenges, face the prevalent issue of how to make knowledge more visible throughout the organisation (Eppler & Burkhard 2004b; Sparrow 1998). There are many strategies to achieve this goal such as encouraging employees to share knowledge improving business processes.
However, as many employees can testify that with the increasing addition of new and pertinent information, it is being increasingly common for users and researchers to experience information overload (Keim et al 2008; Chen 2005; Eick and Karr 2002). Myers (1991) refers to this problem as being:

“Drowned in information while being starved for knowledge and distracted from wisdom.”

This statement infers that in firms which gather and store massive amounts of customer, product and research, data valuable knowledge is not being analysed or utilised, Wang and Wang (2002). Therefore, with the clear potential to give the firm new ideas, encourage innovation, and provide information for more informed decisions, it is worth investigating using knowledge discovery in these databases to enhance and achieve knowledge management goals.

Marghescu (2003) explains that the most effective knowledge discovery process is to share out the analytical results in a format appropriate for business users. For example, by using a combination of understandable language and visualisations, resulting in the knowledge being used effectively (Denrell, Fang & Winter 2003). Using these visual analysis tools to mine for data the firm can determine which customers to target (who), what these customers want (what) and can decide how to satisfy their needs, Markides (1998).

With the potential benefits of using visual data mining tools as part a firm’s knowledge management strategy, there is an urgent need for an approach for selecting the most effective and appropriate tool, Rhyne (2006). The only way to address this problem is through the development of a framework of evaluation methodologies and benchmarks which can be applied to any firm. The first part of this project therefore focuses on the current state of visual data mining, and then develops evaluation frameworks which can be applied to measure any tool. This framework will then be implemented and refined by applying it to several tools to show how it would be used as part of a knowledge management strategy.
1.2 Background

According to O’Leary (2006) Ireland’s economy’s success in the 1990s is attributed to high-tech, export orientated industries that were attracted to these shores. Intel, Apple and Hewlett Packard established manufacturing bases in Ireland. Along with the pharmaceutical industry which has become a prominent part of the economy with 13 of the top 15 pharmacy global companies now having operations in Ireland. At this time, Ireland’s ability to attract these large scale foreign investments was due to the large pool of educated, relatively low cost, English speaking labour, with access to the growing E.U. market and a favourable tax regime.

These unique competitive advantages have since been gradually eroded with labour costs now almost matching US averages. Other eastern European countries have also emerged with the same attractive packages for foreign direct investment. (Manufacturing wages noted to be 96% of the US average). Consequently, continued investment into Ireland with the same types of investments will be difficult.

This has also meant the manufacturing industries are leaving in droves. The Industrial Development Authority highlighted this trend and is now selling the Irish mind to overseas companies. In order for Irish firms to remain competitive in this climate, Ireland urgently needs to manage its knowledge and encourage new ideas, products and services.

As Alavi & Leidner (2001) explains, knowledge management is a dynamic continuous task that offers theories, strategies and methods to manage knowledge in organisations with the aim to help an organisation to compete by being more innovative, effective and thus more profitable Stewart (1997).
One of the biggest challenges to knowledge management is the ever increasing amount of information being gathered by firms. The only way around is to offer useful, friendly and effective products for business people, who are searching for data and information and need to retrieve it as efficiently as possible (Marghescu 2003; Soejarto 2003).

Consequently, when developing analytical tools, the characteristics and needs of the users must be kept in mind. This follows the current trends in data mining to increasingly target and design applications specifically for end users. To also make it easy, convenient and practical to explore very large databases for organisation and users. This is without requiring them to have experience in data analysis and to provide highly specialized solutions rather than creating new data mining tools Fayyad and Uthurusamy (2002)

The focus of users of data mining tools is fast access to data, real time interaction with the system and high quality information Marghescu (2003). It makes sense to question, which is the best-suited visual data-mining tool available. Yet to date there is no unified way to evaluate these tools with only several different evaluation approaches for different components that make up a visual data-mining tool (Burkhard, 2004; Collier et al 1999; Marghescu, 2003; Grinstein et al, 2001)

1.3 Research problem

This dissertation tackles this challenging problem by developing a framework to provide direction and decision in evaluating and selecting a data mining tool for an organisation’s sector and background. In particular the research will look in depth at evaluating visualisation data mining tools. Previous work at evaluating software, data mining and visualisation tools will be investigated and used to create a visual data mining framework. However no studies to the knowledge of this author have looked at combining these different methodologies into a single unified framework. After the framework has been developed it will be applied to several visual data mining tools to evaluate and compare the tools using a number of criteria from the framework for marking the evaluating the visual data mining tool.
The research will demonstrate the importance of visual data mining tools, the reasons for evaluating these tools (Grinstein et al, 2001; Fayyad 1997), how to evaluate them, the benefits and drawbacks of each and how best to find the most suitable tool. This will not be a comprehensive review of commercial tools but instead provides a method for evaluating tools and as a point of reference for selecting the best visual data mining tool for the particular problem.

The aim of the research project is to create an evaluation framework, for knowledge management advisers and business users to provide criteria on choosing a suitable for visual data mining tool for an organisation. This dissertation builds on other work undertaken on software evaluation, evaluating data mining tools (Collier et al 1999), evaluation of visualizations and techniques and an article by Marghescu (2003) to develop a generic complete evaluation framework to assess visual data mining tools.

The completed evaluation framework will have a number of uses such as:

1. Providing researchers and business users with a tool to evaluate different visual data mining tools. Enabling the person to compare diverse tools and determine which is the most suitable for an organisation.
2. Providing a complete set of criteria which can be used to evaluate data mining tools. Ensuring those evaluating tools consider all options.
3. An approach to evaluate existing visual data mining tools in use, providing a way to identify problems with the tool and identify user interactions which are limiting the strategic effectiveness of the knowledge.
4. Providing a way to benchmark new visual data mining tools against existing commercial tools on the market identifying shortcomings (Mazza and Berrè 2007; Grinstein et al, 2001).
5. Acting as a benchmark for visual data mining tools, enabling higher standards of quality and changing the focus from improving data mining algorithms and visual algorithms to improving the expression the knowledge to normal business users and decision makers.

In order to develop this evaluation framework an in depth literature review of knowledge transfer, perception, data mining, data visualisation, knowledge
visualisation, software evaluation methodologies and previous approaches taken to visual data mining evaluating will be examined.

With understanding of these areas the evaluation framework will be developed. It will then be applied against three commercially available products to determine the effectiveness of the framework and to discover the most effective solution for the problem. As part of the application of the framework a data set will be used on three different tools with the visualisation results being analysed on their ability to transfer knowledge. A survey will also be done on a group using the tools to give different perspectives on the visualisation tools effectiveness and the ease of extracting knowledge and gaining insights into the data.

1.4 Intellectual challenge

There are several intellectual challenges for this dissertation including the following:

- Investigating the current state of data mining and visualisation
- Researching visual data mining techniques
- Investigating the link between knowledge, visualisation and data mining
- Investigating the current state of software evaluation
- Becoming familiar at other methodologies for evaluating data mining
- Becoming familiar at other work done on visualisation evaluation
- Understanding the gap in the evaluation methodologies
- Devising a framework for evaluation of visual data mining from other work
- Creation of a metric to determine the success of the framework
- Becoming familiar with commercial visual data mining tools
- Reflecting on how well the framework performs on commercial tools
- Determining shortcomings, advantages and future work

The project also requires understanding between the link between problem solving supported technology and the logic, reasoning and knowledge of the area required by the user. The problem is determining measures on how the quality of what is an
intuitive and interactive display that can communicate analytic results through meaningful visualisation and representations (Keim 2008).

1.5 Research objectives

The intended use of the framework is to create a measurable guideline on visual data mining tools. The framework will initially be formed from previous documentation and added to with informal interviews, questionnaires and observations of important criteria in determining the best visual data mining tool. This can then be applied to different tools to measure their success in helping to achieve knowledge management objectives.

This framework in particular deals with projects where knowledge management practitioners apply visual data mining tools to make hidden knowledge visible for decision makers. The completed framework can then be applied by practitioners to allow them to apply a standard to thoroughly measure a tools success. Such a framework can also be used on to measure existing tools and could be used a basis for a standard on visual data mining quality.

The following objectives will be achieved throughout the dissertation and will contribute to the overall outcome:

1. An in depth literature review will be done on knowledge transfer, knowledge visualisation, data mining, data visualisation, visual data mining, techniques to evaluate software, data mining evaluation techniques, visualisation evaluation techniques and any previous work done previous into evaluating visual data mining tools.

2. Show the benefits of improved knowledge transfer in organisations and how data mining with visualisation can help organisation meet strategic objectives of being more competitive and innovative.

3. A literature review will also be done on several commercial products available using the completed framework, including details of setup, features, how data is processed by the each product.
4. The framework for evaluating visual data mining tools based on the previous work into the different areas, which is focused on optimising knowledge transfer. This framework will be generic allowing it to be applied to any organisation.

5. An assessment approach for evaluating how well the evaluation framework performs and an assessment on the performance of the framework, looking critically at the result identifying gaps and future work which needs to be done to advance the evaluation framework further.

1.6 Research methodology

Both primary and secondary research will be used during the creation of this dissertation. The secondary research program will take the form of an extensive literature review of Business Intelligence, visualisation and data mining tool evaluation. There will be various different sources used to complete the literature review including the following:

- Journals (ACM, IEEE)
- White papers
- Conference proceedings
- Newspapers
- Company websites
- Broadsheet newspapers

The primary research that will be used is a survey on the evaluation of several data mining visualisation tools to a small test group. The survey will be used to establish the following:

- Highlight the weakness and strengths of visual data mining tools
- Show which visualisations are most effective in passing on knowledge from data mining results
- Evaluate the users understanding based on a number of criteria including the following:
  - Timeliness (how quickly results were displayed from large datasets)
The knowledge gained through analysing the results of the survey will be used as feedback into the evaluation framework of visual data mining tools. Structured interviews will also be used to gain insight from experienced users of data visualisation tool users and experts in the field of data mining visualisation.

1.7 Resources

The project will have several deliverables including the following:

- A literature review of visualisation, data mining, software evaluating methodologies and of previous work in this area will provide an understanding to the background of the area, demonstrate the need and provide a basis for the creation of the framework to evaluate visual data mining tools.
- The evaluation framework which is developed from analysing previous work on the subject will provide business users and researches with a way to evaluate visual data mining tools in order to determine which will be most effective in the organisation. The framework can then be applied to evaluate any analytical tool and also can be used to benchmark other tools and provide a way to set standards in visual data mining.
- An approach to assessing the evaluation framework will be provided as a way to critique the developed evaluation framework
- A background and other information on the tools used to apply the framework including information on the common datasets used how each product handled the data and information on the results given from the tools.
- A survey of each of the tools results and of the framework giving different perspectives. This is important as each individual user, from their own experience, views and beliefs will interpret and take different knowledge from the same data mining visualisation. Therefore it is important to access the product that a group with different domain knowledge be used.
• The results from structured interviews with people involved in visual data mining and their thoughts on the problems and challenges in the area. This may include developers, analysts or business users who use these products.

The availability of several resources is vital to the successful completion of this dissertation. The following is a list of these resources:

• Library facilities
Access to IEEE and ACM Journals is essential, along with access to InfoVis and KDD conferences via the Dublin Institute of Technology’s library facilities. It is very important that they are accessible from home.

• Computer with network access
The availability of a computer with network access in DIT to connect to the Oracle resources such as the database, discover tool and SAS’s business intelligence tools.

• Internet access and email
The availability of internet access and usage of email facilitated the completion of this research project.

• Access to a users
Access to a number of users for testing who have several years IT experience using visual data mining tools

• Access to experts
Access to a number of experts for feedback on the completed system.

• Guidance from supervisor
Regular contact with the project supervisor is of crucial importance in completing the research.

1.8 Scope and limitations
The aim of the research project is to create an evaluation framework for visual data mining tools that remains usable in the spite of the continually changing state of the market (Collier et al 1999; Marghescu, 2003). This framework will act as a guide for analysts to select the correct business tool and to get the maximum return from these system tools. The research only concentrates on evaluating the visual aspects of data mining tools and does not cover issues of how effective data cleaning and preparation or algorithms are at modelling the data.
Visualisation and visual analytics is taken to mean an iterative process that involves information gathering, data pre-processing, knowledge representation, interaction and decision making in the context of data mining. The ultimate goal is to gain insight of the problem at hand which is described by vast amounts of scientific, forensic or business data from heterogeneous sources.

Such requirements and limitations are:

- The framework will use several works of previous methods for evaluating data mining tools, visualisation and visualisation techniques to create the framework
- A small group of 6 will be used for evaluating the methodology
- The users will have a level of expertise on dataset to easily gain insights from the data
- Only a single dataset will be used on several data mining tools to evaluate the tools
- Three tools will be evaluated to verify the success of the framework

The main reasons behind these requirements and restrictions are to ensure the project does not run over time and is clearly defined.

### 1.9 Objectives of the Framework

The framework will provide a definitive measure for a visual data mining tools success. It will provide different criteria which will be weighted by the practitioner on the importance to give a final rating for a tools success. It will therefore give managers more confidence on acting on knowledge discovered with the tool. With the most effective tool chosen it will communicate the results in a way which decision makers can understand. More importantly with tangible proof of the tools success the cost associated with that part of the knowledge management project is justified and continued investment supported.

This is based on the hypothesis that current underinvestment in knowledge management is based on immeasurable project deliverables. It is also based on decision makers not making more extensive use of knowledge discovered using these
tools as a good source of knowledge. Thirdly the hypothesis is based on the lack of unified scientific measures for these tools which means that quality can range greatly. In the longer term with higher quality standards tools the deficiencies can be recognised and can result in better knowledge tools. At the moment it is this lack of knowledge on the quality of the tools which limit their improvement and their success at meeting user’s needs.

1.10 Organisation of the dissertation

The remaining chapters of this research project are organised is shown in figure 1-1.

![Figure 1-1 Structure of the dissertation](image-url)
The chapters are laid out as follows:

**Chapter 2**: Gives an introduction to knowledge management, business intelligence and describes in detail the data mining link between the two. This acts as a good primer in understanding the context of the content on data mining.

**Chapter 3** goes into more detail into data mining, the types, concepts and techniques of the area and how data mining operates and emphasises the benefits. This leaves a gap in understanding the visual aspects of visual data mining tools which is tackled by the next chapters in depth analysis on visualisation.

**Chapter 4** covers the area of visualisation, the different types of areas of visualisation and how data mining and knowledge management can benefit from exploiting visualisations benefits. With such a wide and diverse area it also explains the needs for evaluations to select the best tool.

**Chapter 5** covers the work done on visual data mining and visualisation to discuss visual data mining tools and there close connection to knowledge management objectives.

**Chapter 6** looks at the work done to select the best visual data mining tool and explains why a framework to evaluate these tools are needed.

**Chapter 7** shows the development of an evaluation framework for visual data mining by describing characteristics of a good visualisation, testing criteria, along with details of the revisions which were necessary while developing the tool.

**Chapter 8** is the practical application of this framework to select the best data mining tool from the available software and with the results and the revisions made to the framework from the results. This includes selection of datasets, identification of business questions and selection to the tools to evaluate. The data is then scored through the use of the framework by users. This chapter also gives an analysis of the results.
Chapter 9 summarises and concludes the dissertation. It reviews the aims and objects of the research project as well as the difficulties encountered during the research. The chapter also lists the additions made to the current body of knowledge and discussed the possibility of future research and the challenges that will need to be overcome to advance visual data mining.
2 KNOWLEDGE MANAGEMENT

2.1 Introduction

This chapter explains what knowledge management, its objectives is, describes the importance of knowledge development and shows how data mining is the most effective business intelligence tool which can be used to enhance a knowledge management system. It then outlines how knowledge management is used with data mining itself and then summaries the chapter. This chapter begins with defining what knowledge to ensure the aims of knowledge management are clearly defined.

2.2 Defining Knowledge

In order to understand knowledge management and how visual data mining tools can enhance knowledge management, it is necessary to have a clear understanding of knowledge. Data as represents the raw facts without meaning. Information is what is obtained when data is organised in a meaningful context, while knowledge is characterised as the meaningful organisation of information Zack (1999).

2.3 Defining Knowledge Management

“Knowledge management can be defined as the process for acquiring, storing, diffusing and implementing both tacit and explicit knowledge inside and outside the organisations boundaries with the purpose of achieving corporate objectives in the most efficient manner” – (Watanabe & Senoo 2008)

“Knowledge management is the name of a concept in which an organisation consciously and comprehensively gathers, organises, shares and analyses its knowledge in terms of resources, documents and people skills” - (Techtarget 2008)
“Knowledge management is defined tactically by all kinds of management activities that promotes the use of knowledge circulation process” - (Lee, Lee & Kang 2005)

Watanabe and Senoo (2008) and Techtarget (2008) both describe knowledge management as the combined process of acquiring, storing, diffusing and implementing knowledge. This definition introduces the concept of tacit knowledge, developed by Nonaka (1995) which is stored in people’s heads such as skills, know-how built up from experience. It also introduces the concept of explicit knowledge which is physically stored in some medium such as books, video and audio. The objective for organisations is to leverage these different resources efficiently to further the corporate objectives.

The second definition is by Lee, Lee and Kang (2005) also describe knowledge management as a process saying it involves all kinds of tactical management activities which it suggests promotes the use of the knowledge circulation process. This process is the combination of knowledge creation, accumulation, sharing, utilisation and internalisation.

2.4 Objectives of Knowledge Management

When firms take on implementing knowledge management they usually have one of three objectives for doing so such as developing a knowledge intensive culture, building a knowledge infrastructure or making knowledge visible (Alavi and Leidner 2001; Davenport and Prusak 1998). In order to develop knowledge intensive culture knowledge sharing would be encouraged and the firm would proactively encourage people to seek and offer knowledge. Those firms concentrating on the knowledge infrastructure encourage communities of practice where people are given space, the time and the necessary tools to encourage interaction and collaborate. The third objective is to make knowledge visible and show the role of knowledge in the firm by using maps, yellow pages and hypertext tools.
To achieve these objectives, knowledge management is split into tasks which are best represented through Zeiller’s (2005) building block of knowledge management taken from Probst (Probst, Raub and Romhardt 1999:1998), see figure 2-2.

The core tasks as viewed in figure 2-2 are knowledge identification, knowledge acquisition, knowledge development, knowledge distribution, knowledge preservation and use of knowledge. With the tasks of knowledge measurement combined with knowledge goals to get the success of the core tasks and to set the goals. Each core task must be looked at separately but one of the most crucial tasks for innovation and new product and service development is knowledge development.

### 2.5 Knowledge Development

When dealing with data we usually mean transforming raw data into useful knowledge or building on knowledge to create new knowledge or ideas. According to Marghescu (2003) in order to transform raw data and develop knowledge it is necessary to eliminate all the insignificant details, obtain meaningful categories and put the results in a form that makes them easy to understand and comprehend. As only a user can use their experience and knowledge to do this task, they need to decode the results which may take the form of reports, tables or graphics. These users must be aware of information content and meaning in order to order the data. Through this process
the user can identity important content and change data into information. Then only then as Spiegler (2000) points out, citing Churchman (1972), that

“Information becomes knowledge when it adds insight, abstractive value and better understanding”

Due to the importance of how knowledge is interpreted and analysed the role of knowledge management systems can become critically important. As Burkhard (2004) explains with users having limited time, attention and capacity to review data it is vital that the right system is available. Such a system must provide information depth with a trade off between overview and details.

Knowledge management systems are I.T based systems developed to enhance the organisation processes of knowledge creation, storage/retrieval transfer and application (Alavi and Leidner 2001). Bowman (2002) describes the structure and features of these systems as including text and multimedia search and retrieval, knowledge mapping, personalisation, collaboration and messaging.

To develop a complete knowledge management system it is necessary to use a combination of available tools in the software market (Ngai and Chan 2005). Among those tools available are business intelligence tools. Business intelligence tools centre on fully utilising massive data to help organisations to gain competitive advantages (Wang and Wang 2008) and improve the use of information and knowledge available to the organisation.

2.6 Business Intelligence and Data Mining

Business intelligence tools enable organisations to understand their internal and external environments through the systematic acquisition, collation, analysis, interpretation and exploitation of information (Chung, Chen and Nunamaker 2005). These include tools for rule modelling, data profiling, data warehousing, online analytical processing and data mining. In particular it is the data mining element which connects knowledge management and business intelligence. As, it has been shown that
Data mining is a knowledge management process that involves human knowledge (Brachman et al. 1996).

The major aspect of data mining that can enhance knowledge management is mainly to extend human knowledge (Wang and Wang 2008). For example, given a sales database, data mining can reveal the consumer purchase patterns previously unknown to the firm. Brohman (2006) explains this in more detail through the use of his view of the business intelligence value chain in figure 2-3 that specifically defines the insight generation process in data mining. This model developed and validated with data analysts and decision.

Figure 2-3 Brohman’s (2006) business intelligence value chain and the modes of knowledge creation

According to Brohman (2006) there are four primary stages that define all data mining process models: initiation, analysis, evaluation and deployment. The analysis stage consists of preparing and cleaning the data, analysing the data and producing a report to summarise results. In the evaluation stage the results are assessed for accuracy and relevancy and interpreted in relation to the business problem. In the final stage deployment the results are used to make a decision and monitored to learn from the success or failure.
In the same figure Nonaka’s (1995) four modes of knowledge creation: socialisation, combination, externalisation and internalisation are integrated are shown to fit data mining. Through socialisation in the early stages of the data mining process decision makers and data analysts share experiences and perspectives that may lead to the generation of new knowledge. In the combination stage the data miner transforms explicit knowledge in the data warehouse into new explicit knowledge by merging, categorising, reclassifying and synthesizing data related to the task at hand.

Once analysis is complete the data analyst uses externalisation to summarise the finding in a report and makes recommendations to the business problem. This report is given to the business manager to evaluate who searches the report for coded evidence to support the beliefs, paradigms and viewpoints he applied to develop the business problem or translate the business problem to a data mining task.

Finally internalisation opportunities exist in the deployment and follow up stages of the data mining process. In which the deployed solution is analysed through manager’s internalisation of the information in which the manager makes a decision on the result of the analysis and interpretation completed.

This concept of knowledge management and business intelligence integration is also discussed further in White (2005), where he provides a flowchart that articulates the use of business intelligence in the knowledge management context for decision making. The flowchart model shown in figure 2-3 illustrates the involvement of collaboration and interaction between the knowledge workers for socialization.
The knowledge management cycle is shown in figure 2-4 from White’s(2005) perspective of business intelligence puts performance measures into a business context improving the business decision making and action taking processing because results become actionable. These performance measures in creates business information which can be embedded in enterprise portal pages, documents, spreadsheet, presentation, audio, video and email.

2.7 Conclusion

This chapter defined knowledge and knowledge management. It outlined knowledge management objectives and the importance of knowledge development. It detailed out business intelligence tools can be used to enhance knowledge management and how data mining in particular can enhance knowledge in the firm.

Although this chapter detailed data mining briefly as being part of business intelligence it could not give the depth of knowledge which will be needed to understand visual data mining tools. For this reason the next chapter on data mining explains its importance and its potential benefit if used as part of knowledge management within an organisation.
3 DATA MINING

3.1 Introduction

This chapter builds on chapter two’s brief introduction to data mining in knowledge management by first explaining the field supporting data mining called knowledge discovery. It then defines the importance of data mining in our competitive climate and in growing our knowledge economy.

The rest of the chapter provides the principles of data mining to enable the reader to gain enough understanding to appreciate the reasons behind visual data mining tools importance. This chapter concludes with a conclusion to review this chapter. As already state this chapter begins with introducing data mining’s supporting field of knowledge discovery.

3.2 Knowledge Discovery

"Knowledge discovery could be defined as a set of techniques coming mainly from the area of artificial intelligence but also borrowing important building blocks from other fields such as statistics and databases " - Grobelnik and Mladenic (2005)

The above definition gives a good understanding of overall what knowledge discovery is meant to mean however more it is very general. More specifically a according to Bradley, Fayyad and Mangasarian (1998) who first coined the phrase knowledge discovery refers to the processes of discovering useful knowledge from data They originally defined these processes to be data preparation, data cleaning, variable selection as well as the data mining process.
Since then it has evolved as a new discipline that meets the need for knowledge derived from data by creating using automated method from large datasets, Klosgen and Zytkow (2002). The knowledge discovery processes have also been refined by Elmsari and Navathe (2004) to be data cleansing, data transformation or encoding, data mining and the reporting and display of the discovered knowledge. There are other views of these processes however with Klosgen and Zytkow (2002) arguing for a more business view suggesting that the processes involve steps that are both iteratively and interactively performed. They classify these processes into business problem analysis, data understanding and preparation, search for knowledge, solution to the business problem and deployment and practical evaluation of the solution.

These differences on views on the important processes in Knowledge discovery in databases, drove the development of industrial organisations to come together to develop the standard processes called Cross Industry Standard Process for Data Mining (CRISP-DM) process list according to Han and Kamber (2006) and Crisp group (2008). In the CRISP-DM model of processes for knowledge discovery the Crisp group (2008), defined the process as shown in figure 3-5.

Figure 3-5 Phases of the Crisp-DM process model, Crisp group (2008)
This process begins with defining the business problem and its objectives identifying the data for mining and assessing data quality, as the availability of data is not the same thing as the appropriateness of data for mining, see figure 3-5. As seen in figure 3-5 the data may be dirty with errors and inconsistencies which must be first cleaned at the data preparation process which may involve transforming the data by replacing misspelled values with correct ones, identifying unusually (outlier) values, creating attributes or dropping data if necessary. The knowledge extraction process continues with data mining being performed on the data to produce a data mining model, which is then evaluated for quality and relevance to the problem objectives (Hornick, Marcade and Venkayala 2006; Keim 2005).

Although not shown in the models in figure 3-5, there are other important processes in KDD such as the pattern evaluation, identifying the truly interesting patterns representing knowledge based on interesting measures and knowledge representation, where visualisation and knowledge representation techniques are used to present the mined knowledge to the user, Han and Kamber (2006). The central purpose of knowledge discovery is to support data mining.

### 3.3 Defining Data Mining

“Data mining is extracting or mining knowledge from large amounts of data” - Han and Kamber (2006)

“Data mining uses a variety of data analysis tools to discover patterns and relationships in data that can be used to make reasonable accurate predictions. It is a processes not a particular technique or algorithm.” – Edelstein (2008)

“Data mining is the exploration and analysis of large quantities of data in order to discover meaningful patterns and data” - Berry and Linoff (2004)

The different definitions given above for data mining show a general consensus that data mining is the extraction of knowledge from large amounts of data by identifying and analysing interesting patterns in data. Han and Kamber (2006) define it in terms of
extracting information and knowledge from data. This is backed up by Edelstein’s
definition with a further clarification that data mining is not a set of algorithms or a
simple task but a process. That in order to discover knowledge there may be several
iterations and verifications on the hypothesis or predictions.

Effectively, data mining aims at revealing knowledge about the data under
consideration. Knowledge takes the form of patterns within the data that embody our
understanding of the data according to Redpath and Srinivasan (2003). To achieve this
algorithms are applied to masses of unanalysed data and by applying different
algorithms identify associations or patterns which has business applications. For
instance by identifying groups of customers more likely to buy a product, the sales
campaign can be more targeted and avoid a more costly broad campaign. To spot these
patterns the user has to have an understanding of the data in order to guide intelligent
agents to apply methods, to identify important variables among irrelevant data and

As defined by Edelstein, data mining itself is a process that is almost always iterative
where the domain expert on the data using the data mining tool guides the analysis and
manipulation of the data and by trying out different algorithms and viewing the results,
developing the patterns into a model Edelstein (2008). This model simply represents
an algorithm or set of rules that connects a connection of inputs to a particular target or
outcome, Berry and Linoff (2004). The interesting patterns that are identified from
data mining will be either statistical or predictive models of the data and the
relationships between them found using historical data and applying that model to new
data using a combination of machine learning, statistical analysis, modelling
techniques and database technology (Fayyad and Uthurusamy 2002; Berry and Linoff
2004). Hornick describes a model in more detail in the following description:

“The model is applied to data to predict individual behaviour (classification and
regressions) segment a population (clustering), determine relationships within a
population (clustering) as well as to identify the characteristics that most impact a
particular outcome (attribute importance)” - Hornick, Marcade and Venkayala (2006)
The finished model effectively summarises the subsets of the data that can be predicted using it to predict customer behaviour Hornick, Marcade, and Venkayala (2006). Actually determining what patterns are interesting from the data is harder to define as this depends on the purpose of the business solution. In order to identify the business solution it is necessary to look at the business problem that data mining aims to solve which is tackled by the field of Knowledge Discovery in Databases (KDD). Effectively data mining is the central step in KDD by applying algorithms and verifying hypothesis Klosgen and Zytkow (2002). However business standards have defined supporting processes such as data preparation and reporting.

### 3.4 Data Mining Approaches

In data mining the knowledge discovered is in the form of patterns within the data that embody the understanding of the data. These patterns are often also referred to as structures, models or relationships. These patterns found within the data cannot be separated from the approaches that are used to find those patterns because all patterns are essentially abstractions of the real data. The approach used is called an abstraction model or technique. The approach chosen is inherently linked to the pattern revealed Redpath and Srinivasan (2003).

Depending on the purpose of the knowledge to be discovered and the intellectual, economic and business interests, there are four main approaches associated with data mining techniques which may be employed including predictive modelling, database segmentation, link analysis and deviation detection, (Bradley, Fayyad and Mangasarian 1998; Berry and Linoff 2004; Klosgen and Zytkow 2002; Cabena et al 1997; Connolly and Begg 2002) as seen in table 3-1, along with the data mining techniques which are most often used with the approach.
The data mining techniques shown in table 3-1 are specific implementations of the data mining operation. Each operation has its own strengths and weaknesses which is why data mining tools provide a choice of data mining techniques to implement an operation and different data mining techniques may be used on the same data to get possible result (Connolly and Begg 2002; Redpath and Srinivasan 2003).

The data will have an inherent structure but it is not possible to describe it directly. Rather a pattern is an attempt to describe the inherent structure by using a particular approach. Patterns are best understood in terms of the approach used to construct them. For this reason the patterns are often discussed in terms of how they are arrived at rather than stating the data has a pattern in some absolute sense Redpath and Srinivasan (2003). Several of these techniques are described in more detail in the following sections.

<table>
<thead>
<tr>
<th>Operations</th>
<th>Data Mining Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive Modelling</td>
<td>Classification</td>
</tr>
<tr>
<td>(Classification/Prediction)</td>
<td>Value Prediction</td>
</tr>
<tr>
<td>Database Segmentation</td>
<td>Demographic clustering</td>
</tr>
<tr>
<td>(Affinity Grouping, Clustering)</td>
<td>Neural Clustering</td>
</tr>
<tr>
<td>Link Analysis</td>
<td>Association Discovery</td>
</tr>
<tr>
<td>(Description and profiling)</td>
<td>Sequential Pattern Discovery</td>
</tr>
<tr>
<td></td>
<td>Similar Time Sequence Discovery</td>
</tr>
<tr>
<td>Deviation Detection</td>
<td>Statistics</td>
</tr>
<tr>
<td></td>
<td>Visualisation</td>
</tr>
</tbody>
</table>

Table 3-1 Data Mining Operations and Associated Techniques (Cabena et al 1997)
3.4.1 Predictive Modelling

Predictive modelling is one of the most common data mining tasks and usually involves classification, estimation or prediction. In classification, the task is to build a model that can be applied to the unclassified data in order to classify it and any future data (Han and Kamber 2006). Estimation deals with predicting a continuous valued outcome and prediction is similar to the other two except the records are classified according to some predicted future behaviour or value Berry and Linoff (2004).

All three involve classifying data into subsets, examining the features of a newly presented object and assigning it to one of a predefined set of classes Berry and Linoff (2004). Using this data it is possible to predict future values and estimate the value of a single variable such as a new customer signup Klosgen and Zytkow (2002). Similar to humans using experience, historical data can be used to create the model which is then tested on new previously unseen data to determine the prediction accuracy and performance characteristics.

The techniques used for predictive modelling includes decision trees, regression and nearest neighbour techniques. In some cases neural networks and link analysis is used for classification or to improve the classification rate (Connolly and Begg 2002). The technique is often employed for customer retention management, credit approval, cross selling and direct marketing.

3.4.2 Database Segmentation

Database segmentation involves affinity grouping and clustering where the question being asked has not been identified and the objective is simply to find patterns and group data into subsets without a specific target or objective. This usually involves clustering which is the task of segmenting a heterogeneous population into a number of more homogenous subgroups or clusters. Unlike classification there are no predefined groups that records must be associated, the data is grouped together on the basis of self similarity, Berry and Linoff (2004).
Database segmentation uses all of attributes of the data to create a group which may include redundant and irrelevant features. This means it may reduce the classifications success. The actual process of segmentation involves taking the allowable data inputs, using methods to calculate the distance between records and using the results to segment the data, Connolly and Begg (2002). The technique is often employed in customer profiling, direct marketing and cross selling.

3.4.3 Link Analysis

Link analysis centres on finding associations, describing and profiling data mining techniques where the aim is to creating groups and understand subsets of data by building up a profile of a customer or understanding or determining group characteristics, Klosgen and Zytkow (2002). There are data mining techniques of link analysis association discovery, sequential pattern discovery and similar time sequenced discovery.

Association discovery finds items that imply the presence of other items in the same event, Connolly and Begg (2002). A simple approach is to use link analysis in order to generate links also known as associations between sets of data Berry and Linoff (2004). So for instance when a customer buys cereal the person will also buy milk. Sequential pattern discovery finds patterns between events such that the presence of one set of items is followed by another set of items over a period of time such as understanding a customer’s behaviour. In a similar manner sequence discovery is used to discover links between two set of data that are time dependant. Link analysis can also involve profiling the customer to understand customer behaviour or explanations. Decision trees are a powerful tool for profiling customers Berry and Linoff (2004). Among the uses for link analysis include in market basket analysis to identify cross selling opportunities, direct marketing and stock price movement (Connolly and Begg 2002; Redpath and Srinivasan 2003).
3.4.4 Deviation Detection

Deviation detection usually uses statistics and visualisation to identify unusually values known as outliers such as patterns which do not agree to some previously known expectation or norm and looking at discovering more about these groups (Connolly and Begg 2002; Han and Kamber 2006). The term outlier can be defined as:

“An observation that appears to deviate markedly from other members of the sample from which it occurs” – (Barnet and Lewis 1994)

Outliers are identified by examining the main characteristics of objects in a group. Objects that deviate from this description are considered outliers. An analysis of this deviation may find that it was caused by erroneous procedures, data measurement and collection or by interesting patterns. Deviation detection is important for fraud analysis and helping to detecting small important data subsets of users.

3.5 The Importance Data Mining

The number of datasets in business, official statistics, government, science and elsewhere is estimated at tens of millions and is rapidly increasing with the volume of data expected to double every twenty months, Klosgen and Zytkow (2002). Among the reasons for this increase include the computerisation of business processes, historical data being archived and various regulatory compliance measures requiring organisations to keep large amounts of historical data. With this amount of data corporate executives want to put this costly asset to good use, Hornick, Marcade, and Venkayala (2006). This is why more companies are recognising a need to use this data which is why many companies are coming up with a data strategy and are investigating using data mining Davenport (2006).

Today’s business landscape is also becoming highly competitive. Product margins are typically low due to increased competition and commoditization. Those savvy businesses who have long taken advantage of advanced analytics and data mining used
it to give them an edge in the marketplace. Those that do not leverage data mining are unlikely to realise their revenue and profit potential Hornick, Marcade, and Venkayala (2006). Data mining is becoming more common now because of the growing number of applications storing data, the growing interest in customer relationship management and the increase of availability in data mining software. Another reason for companies move to data mining is that traditional data analysis is being more time consuming for an analyst. This analysis may involve reformulating models and transforming the data which exceeds the limitations of available database packages Klosgen and Zytkow (2002).

### 3.6 Data Mining Models

Once the data mining operation has been selected and data mining technique is determined, it is possible to select data mining problem types which can be applied to the prepared data. For example for predictive modelling and classification of data, decision tree modelling could be applied to analyse the data. Generally however, the discovery of knowledge happens at different levels of granularity so several models may be applied to gain an understanding of the data with one of several models being used for specific portions of the data or to answer specific questions, Han and Kamber (2006). According to Connolly and Begg the selection of which model to apply on the data is also based on:

“the suitability for certain input data types, transparency of the data mining output, tolerance for missing data, transparency of the data mining output, tolerance of missing variable names, level of accuracy possible and increasingly the ability to handle large amounts of data”(Connolly and Begg 2002).

In the definition the transparency being referred to is the ability of a user to understand how or why a given model makes certain predictions, some algorithm produce such models while other are treated as black boxes Hornick, Marcade, and Venkayala (2006).
Another characteristic of models is supervised and unsupervised learning. In supervised learning the algorithm requires its source data to contain the correct answer for each record. This allows some algorithms to make corrections to a model ensuring as many answers are correct as possible. The correct marked predictor attributes columns supervise the learning process by pointing out the mistakes when the model fails to predict the correct target attribute or outcome. Examples of supervised learning include decision trees, neural networks and regression.

Unsupervised learning on the other hand requires no such knowledge to determine a correct answer. It merely looks at the data and applies an appropriate algorithm (Han and Kamber 2006; Hornick, Marcade, and Venkayala 2006). It is often necessary to combine models to approach and solve data mining problems (Berry and Linoff 2004). The following section briefly describes the most common models in more detail.

3.6.1 Decision Trees and Rules

A decision tree is created by partitioning a large data set into subsets, and then partitioning each of the subsets, until the subsets cannot be partitioned further. In keeping with the tree metaphor, the original data set is the root node, the subsets are nodes, and the un-partitioned subsets are leaves. Branches from a node are the subsets created by partitioning a node (Barlow and Neville, 2001; Soukup and Davidson 2002; Klosgen and Zytkow 2002). The purpose of building a decision tree is to partition a large heterogeneous group of things, usually people into smaller, homogeneous groups, Barlow and Neville (2001).

An example of a decision tree can be seen in figure 3-6, where each divided dataset represents a node with the top node being called the route node. The tree always starts from the root node and grows down by splitting the data at each level into new nodes. The top root node contains the entire data set, with all subsequent child nodes containing respective subsets of the set. All nodes are connected by ranches with the nodes at the end of the branches being called terminal nodes or leaves Negnevitsky (2004). During tree construction attribute selection measures are used to select the
attribute that best partitions the data set into two distinct classes (Han and Kamber 2006; Negnevitsky 2004).

A typical tree as shown on figure 3-6, in this case the decision tree is trying to identify the most likely to respond to the promotion of a new product, such as a new bank service. Typically this task is performed by determining the demographic of homeowners that responded to a similar product in the past. As seen in the figure 3-6 their representation of acquired knowledge in tree form is intuitive and easily to assimilate by people (Negnevitsky 2004; Han and Kamber 2006).

![Decision tree example](image)

Figure 3-6 Negnevitsky’s (2004) example of a decision tree for a new product for homeowners

Decision trees are easily converted to classification rules and do not require any domain knowledge or parameter settings, can handle high dimensional data and are therefore ideal for exploratory knowledge discovery. The learning and classification steps of decision tree induction are simple and fast and generally have good accuracy (Han and Kamber 2006; Soukup and Davidson 2002).

### 3.6.2 Linear Regression

Linear regression is a commonly used statistical technique for fitting observations or points in n dimensions with the target variable y, Elmsari and Navathe (2004). Linear
regression works by finding the best straight line to fit two attributes so that one attribute can be used to predict another , Han and Kamber (2006). Multiple linear regressions are an extension, where more than two attributes are involved and the data is fit to multidimensional surface.

Effectively it is about a set of points on a graph where a straight line is applied to them by using an algorithm function. The algorithm iterates over the data collecting statistics and then determines the coordinates of the line that best suits the set of two dimensional points. This is illustrated in figure 3-7. The model applied is expressed as the equation $y = mx + b$ where $m$ is the slope and $b$ is the $y$ intercept. Using the model consisting of $m$ and $b$ is sufficient to make predictions for $y$ given a value of $x$. For example if $m=2$ and $b=5$ then $x$ (or age) $=25$, then the value of $y$ (or income) $=55$ Han and Kamber (2006).

![Figure 3-7 Fitting a regression line to a set of data points Han and Kamber (2006)](image)

Regressions value becomes evident when the data is not in two dimensions but a hundred, thousand or several thousand. The attributes of regression models may include numerical values or consist of categories that are numbers or strings, Han and Kamber (2006). Regression algorithms such as linear regression have a number of benefits such as it can be used on both sparse and skewed data very well. Many non-linear problems can be converted to linear problems by performing transformation on the predictor values, Han and Kamber (2006).
3.7 Conclusion

In this chapter data mining and techniques, types, models and algorithms were introduced. The aim was to give an understanding of what data mining is used for and how it is applied to data to discovery knowledge. Several of the main models were discussed briefly such as the decision tree, linear regression, neural networks and genetic algorithms which will be useful in understanding concepts in later chapters.

Data mining was defined as the process to understand and convert raw data into useful information and explain in relation to its part within knowledge discovery. The main processes of knowledge discovery were discussed along with a description of the Crisp-Dm methodology. The data mining operations were discussed which can be used were explained in relation to the purpose of the knowledge discovery and main uses of data mining were discussed as being predictive modelling, segmentation, link analysis and deviation detection shown with corresponding data mining techniques used. The data mining models were discussed along with several data mining methods including decision trees and rules and linear regression. This chapter formed the basis of understanding data mining however visualisation must be also understood to appreciate visual data mining.
4 VISUALISATION

4.1 Introduction

This chapter starts by introducing the interdisciplinary and fragmented visualisation sciences, including a definition of visualisation, information visualisation, knowledge visualisation, the benefits of visualisation and the techniques used for visualisation. This chapter will provide a good foundation on the fundamentals of visualisation which is important in order to evaluate visual data mining. This chapter begins by looking defining visualisation.

4.2 Defining Visualisation

“Visualisation tools graphically display data to facilitate better understanding of its meaning. Graphical capabilities range from simple scatter plots to complex multi-dimensional representations” - Two Crows (2008)

“Visualisation systems traditionally focus on building graphical depictions of relationships among information in a human comprehensible form” - Yang, Rundensteiner, and Ward (2007b)

“Visualisation is defined as the process of exploring, transforming and viewing data as images to gain understanding and insight into data,” - Wojtkowski (2002)

Visualisation and visualisation systems in the context of data mining have many definitions some of which are provided above. Using these definitions a general definition of visualisation can be defined as being the graphical representation of data or concepts with the aim of making the data easily interpreted to gain knowledge and insights. It is this knowledge and insight which is the main purpose behind visualisation for exploration and reporting (Rhyne 2006; Soukup and Davidson, 2002).
The human comprehensible format of information relationships defined by Yang gives users a better understanding of the data. This means that the uses can learn some facts not easily discovered without the graphical depiction or the users’ knowledge to some facts can become deeper or more precise in a short time. (Yang, Rundaenstein, and Ward 2007a; Wijk 2005)

4.3 Benefits of Visualisations

Visualisation is a major asset when learning from data as it supports the reflection phase of the wheel of learning, and supports the user's task of comparing the behaviour of the process against the user's hypotheses (Keim 2002; Keim et al 1996). Visualisation also supports the ladder of inference, to explicate analyst’s thoughts both to himself, and to other persons.

At the first step of the ladder, the analyst can explicate from the visualised data, which points and variables he has selected for his study. At the second step, addition of meanings, the user can use, for example, histograms to justify his interpretation hat a value is high. In the third step the user makes assumptions Laine (2003). It can also facilitate the understanding of complex models that arise in data mining, Barlow and Neville (2001) and often leads to immediate insight in complex phenomena, interesting patterns, or the detection of outliers (Vliegen, Wijk and Linden 2006). In addition to overall patterns and trends, visualizations often reveal interesting local phenomena representing local patterns (Card and Rao 1995; Furnas 1986). Several studies have proven the power of visualisation with regard to these functions for example as discussed by Burkhard (2005a):

- Miller reports that a humans input channel capacity is greater when visual abilities are used Miller (1956)
- Our brain has a strong ability to identify patterns, which is examined in Gestalt psychology (Koffka 1935; Ellis1938).
- Visual imagery (Kosslyn 1980; Shepard and Cooper 1982) suggests that visual recall seems to be better than verbal recall. From this it is clear that humans
have a natural ability to use images but it is not clear how images are stored and recalled

- Several empirical studies show that visual representations are superior to verbal-sequential representations in different tasks (Larkin and Simon 1987; Glenberg and Langston 1992; Bauer and Johnson-Laird, 1993; Novick, 2001)
- Instructional psychology and media didactics investigate the learning outcome in knowledge acquisition from text and picture (Mandl and Levin 1989), or (Weidenmann, 1989) explores aspects of illustrations in the learning process.

Burkhard (2005a) a prominent figure in knowledge visualisation discusses in one of his papers how visual representation helps with various functions and provides this list which is useful to address emotions, to illustrate relations, to discover trends, patterns, or outliers, to get and keep the attention of recipients, to support remembrance and recall, to present both overview and detail, to facilitate learning, to coordinate individuals, to motivate people and to establish a mutual story, or to energize people and interactions.

4.4 Visualisation Theories

Tufte (1982; 1990) and Bertin (1983) are the most authoritative authors in the area of visualisation theories and techniques and have laid the foundations for exploratory data visualisation techniques. Their research has provided recognised rules for layout, colour composition and attribute mapping for visualisation of data with inherent 2D/3D-semantics. Defining for instance, that when using graphics for exploratory analysis the viewer should be guided by words to read the design Tufte (1982). Tufte defines characteristics of the best visualisation designs to be:

“Intriguing and curiosity-provoking, drawing the viewer into the wonder of the data, sometimes by narrative power, sometimes by immense detail, and sometimes by elegant presentation of simple but interesting data” - Tufte (1982)

In the case of visualising statistical graphics such as those used in data mining Tufte defines the purpose of visualisation as communicating complex ideas with clarity,
precision, and efficiency (Tufte 1982). To act as a guide for good visualisations, Tufte (1982; 1990) defines the five principles of graphical excellence which a graphic should have; these are a well-designed presentation of interesting data. These are that the data needs to consist of complex ideas communicated with clarity, precision, and efficiency, giving to the viewer the greatest numbers of ideas in the shortest time with the least ink in the smallest space, be nearly always multivariate and require telling the truth about the data.

Along with these five excellence principles, Tufte (1990) also create six principles of graphical integrity necessary to ensure visualisations tell the truth about the data and accurately represent the data. These were:

- The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented.
- Clear, detailed, and thorough labelling should be used to defeat graphical distortion and ambiguity.
- Writing out explanations for the data on the graphic itself, and labelling important events in the data.
- Show data variation, not design variation; in time-series displays of money, deflated standardized units of monetary measurement are nearly always better than nominal units.
- The number of information variable dimensions depicted should not exceed the number of dimensions in the data.
- Graphics must not quote data out of context.

These principles are needed because visualisation is not good by definition. Simply putting results into a graph does not mean it will be intuitive to an analyst (Wijk 2005).
4.5 Information Visualisation

Scientific visualisation was the basis and starting point for the new field of information visualisation which has its roots in statistical graphics, user interface design and cognitive science. In contrast to scientific visualisation the focus is on abstract data which lacks natural representation such as financial data, genomic data or transaction data as discussed by Burkhard (2005a). There are several established definitions of information visualisation such as:

“The communication of abstract data, relevant in terms of action through the use of interactive interfaces” - (Keim, 2006)

“the research field that encompasses the use of interactive graphical representations to support exploratory analysis of multidimensional data, so that analysis of complex data can benefit from the human ability of detecting patterns on images and domain knowledge” - (Artero and Oliveira. 2004)

“The use of computer supported, interactive, visual representations of abstract data to amplify cognition” - Card, Mackinlay and Shneiderman (1999)

Information visualisation differs from visualisation as it is computer supported, interactive and in terms of data mining designed to represent large amounts of multidimensional data in real time as discussed by Burkhard (2005a). The information visualisation has several aims such as

- creating interactive ‘windows’ to the ‘data’ Redpath and Srinivasan (2003) helping knowledge workers in the creation, verification and presentation of knowledge as well as sharing knowledge with others (Zeiller 2005; Fayyad and Uthurusamy (2002)
- helping to explore abstract data and to create new insights (Burkhard, 2004)

However Keim puts forward a simpler view of information visualisation aims in data mining as being for presentation, confirmation and exploration. He explains that the
choice of a presentation technique depends largely on the user. When information visualisation is used for confirmation analysis, one or more hypothesis about the data serves as a starting point. The process can be described as goal oriented examination of these hypotheses, where the visualisation either confirms the hypothesis or rejects it. Exploratory analysis as the processes of searching and analysing databases to find implicit but potentially useful information is a difficult task. The reason is because at the beginning the analyst has no hypothesis about the data. Tools as well as understanding are needed for the interactive and generally undirected search for structure and trends (Keim, 2006).

The results of information visualisation are computer applications which allow users to interactively explore abstract data with visual methods, ideally in the sequence as proposed by Shneiderman’s visual information seeking mantra Bederson and Shneiderman (2003; 1996) of:

“Overview first, zoom in and filter, then show details on demand”

However, the transformation of data into meaningful visualisations is not a trivial task that will automatically improve through steadily growing computational resources. Very often, there are many different ways to represent the data under consideration and it is unclear which representation is the best one. State of the art concepts of representation, perception, interaction and decision-making need to be applied and extended to be suitable for visual data analysis (Keim et al 2008)
4.6 Visual Analytics

“Visual analytics is more than only visualisation. It can rather be seen as an integral approach combining visualisation, human factors and data analysis” - (Keim et al 2008)

“(Visual analytics is) the science of analytical reasoning facilitated by interactive visual interfaces” - Thomas and Cook (2005)

The two definitions of visual analytics given above describe it as being mainly concerned with the analytical reasoning facilitated gained through the use of interactive visual interfaces Thomas and Cook (2005). It is different from other areas of visualisation as it integrates methodology from information analytics, geospatial analytics, and scientific analytics, see figure 4-8. In this diagram the scope of visual analytics are shown, with how they overlap to create the field. From this diagram in figure 4-8 the human factors such as interaction, cognition, perception, collaboration, presentation, and dissemination play a key role in the communication between human and computer as well as the decision making process and form a central role in this discipline.

![Figure 4-8](image)

Figure 4-8 The scope of visual analytics Thomas and Cook (2006)

The production refers to the creation of materials that summarise the results of an analytical effort, presentation as the packaging of those materials in a way that helps
the audience understand the analytical results in context using terms that are meaningful to them, and dissemination as the process of sharing that information with the intended audience Thomas and Cook (2006). In matters of data analysis, visual analytics furthermore profits from methodologies developed in the fields of data management and knowledge representation, knowledge discovery and statistical analytics (Keim et al 2008).

Visual analytics is an iterative process that involves information gathering, data pre-processing, knowledge representation, interaction and decision making. The ultimate goal is to gain insight in the problem at hand which is described by vast amounts of scientific, forensic or business data from heterogeneous sources. On the one hand, methods from knowledge discovery in databases, statistics and mathematics are the driving force on the automatic analysis side, while on the other hand human capabilities to perceive, relate and conclude turn analytics into a very promising field of research (Keim 2006). In many ways visual analytics and information visualisation is a precursor to knowledge visualisation.

4.7 Knowledge Visualisation

Knowledge visualisation is a new discipline in visualisation proposed by Eppler and Burkhard (2004a) combining knowledge management principles with information visualisation, visual analytics, cognitive art, communication science and information architecture, Zeiller (2005). The aim of the knowledge visualisation is on the effective transfer of knowledge as discussed by Burkhard (2005a). It has growing support in its field Zeiller (2005). Burkhard defines it as

“The use of visual representations to improve the creation and transfer of knowledge between at least two persons” - Burkhard and Meier (2004)

Knowledge Visualisation thus designates all graphic means that can be used to construct and convey insight Eppler and Burkhard (2005). It looks at how knowledge in contrast to information has to be reconstructed by each individual. The field depends on the successful use of visualisation techniques in knowledge-intense processes,
where knowledge has to be re-constructed by each individual. This process happens through communication and interaction with explicit information verbal or visual. In contrast to information visualisation knowledge visualisation concentrates on as discussed by Burkhard (2005a):

- the recipients, by customising the visual formats to the needs and backgrounds of the different stakeholder.
- on the knowledge types that explicit information, such as know-why or know-how the process of communicating this knowledge by use of one or more visualisation methods from information design, information architecture, information art, information visualisation, or other fields.
- Tackling the difficult problems of information overload, misinterpretation and misuse, Burkhard (2004)

While other visualisation fields concentrate on developing new visualisation techniques knowledge visualisation aims at linking existing visualisation techniques to relevant and predominant problems. To do this knowledge workers need to collect and structure existing visualisation techniques, identify and systemise key problems in knowledge intense processes and try to link the most promising visualisation technique with to individual problems. Only if there is no method that seems promising will knowledge visualisation researchers invents a new and customised method to solve the problem. Thus knowledge visualisation solution orientated in bridging the gap between proposed ideas and real world needs as discussed by Burkhard (2005a).

Beyond the mere transport of facts, knowledge visualisation aims to transfer insights, experiences, attitudes, values, expectations, perspectives, opinions and predictions, in a way that enables someone else to re-construct, remember and apply these insights correctly. In terms of format, knowledge visualizations rely on indirect communication that triggers sense making activities in the viewer and motivate him or her to complete the picture him- or herself. In this way knowledge visualisation can augment knowledge-intensive communication between individuals, for example by relating new insights to already understood concepts, as in the case of visual metaphors. This visual communication of knowledge is relevant for several areas within knowledge management as discussed by Burkhard (2005a).
4.8 Framework for knowledge visualisation

In order to implement knowledge visualisation Burkhard developed a knowledge visualisation framework to outline the perspectives which need to be considered for an effective transfer and creation of knowledge through visualisation. These perspectives centres around four questions of what is the goal of using a visualisation method, what type of knowledge needs to be visualised, who is being address and what is the best method to visualise this indicator, Burkhard (2004a; 2005b). These perspectives are illustrated in more detail in figure 4-9.

![Table of perspectives and their components]

Figure 4-9 The four perspectives of the knowledge visualisation framework and each of their components

Each perspective is discussed in more detail in sections 3.8.1 to 3.8.4.

4.9 Function type perspective

The function perspective distinguishes the purpose of visual representations for information processing based on research in perception and neuroscience, Ware (2000). The six functions with social, emotional and cognitive functions are summarised in the CARMEN Acronym, Burkhard (2005c) and Eppler and Burkhard (2005):

- Co-ordination, visual representations help to coordinate individuals in the communication process
• Attention, they allow getting the attention by addressing emotions to keep attention and to identify patterns, outliers and trends.
• Recall, they improve memorability, remembrance and recall
• Motivation, they inspire motivate, energise and activate viewers
• Elaboration, they foster the elaboration of knowledge in teams
• New Insights, they support the creation of insights by embedding details in context, showing relationships between objects, or after effects

4.10 Knowledge type perspective

According to Buckhard (2005b; 2005c) the knowledge perspective aims to identify the type of knowledge that needs to be transferred. He identifies five types of knowledge that are found in the knowledge management literature referring to Alavi and Leidner (2001) which differentiates them into:

• Know what, declaration of facts
• Know how, procedural knowledge, as knowing how things are done
• Know why, experimental knowledge, as knowing why things occur which captures underlying cause and effect relationships and accommodates exceptions
• Know where, orientation knowledge where information can be found
• Know who, individual knowledge as knowing the expert

4.11 Recipient type perspective

The recipient type perspective aims to identify the target group and the content of the recipient. The recipient can be an individual, a team, an organisation or a network of subjects, Burkhard (2005b). By knowing the context and the educational, emotional and cultural background of the recipient audience is essential for finding the right visualisation method for the transfer of knowledge. From a business perspective, graphic and information design (Bertin, 1983; Tufte 1990) do not focus on this perspective.
4.12 Visualisation type perspective

The last perspective is the visualisation type perspective structures the visualisation methods into seven main groups that are devised from the practices of architects as mentioned by Burkhard (2004; 2005c). In this part Burkhard discusses how:

- Sketches can help to quickly represent the main idea, presenting key features, supporting reasoning and arguing and allowing room for interpretations.
- Diagrams are abstract, schematic, representations used to explore structural relationships among parts by denoting functional relationships. Diagrams can explain casual relationships, reduce the complexity to the key issues, structure and display relationships.
- Maps represent individual elements such as roads in a global context such as a city and illustrate overview and detail, relationships among items, structure information through spatial alignment and allow zoom in and easy access to information.
- Images are impressive, expressive or represent reality. They catch the attention, inspire, address emotions, improve recall, and initiate discussions. Images are instant and rapid, instructive and facilitate learning. Visual metaphors support recall, lead to surprising effects, support reasoning and communications.
- Objects exploit the third dimension and are haptic. They help to attract recipients, support learning through constant presence and allow integrating digital interfaces. Objects exploit the third dimension and are haptic. They help to attract recipients, support learning through constant presence and allow integrating digital interfaces. Objects in space are helpful for example for information points, knowledge fairs, or exhibitions. Interactive visualisations allow to access, explore and make sense of different types of digital information. Interactive visualisations help to facilitate, enable interactive collaborations across time and space and allow to represent and explore complex data or to create new insights.
- Stories are imaginary visualisations that are efficient in disseminating knowledge across time and space. The use of stories allows analyst to transport
a illustrated mental image by using spoken or written language. They help to establish a shared vision, a mutual story, which motivates and activates individuals.

4.13 Steps for Visualising Data

In order to combine the various fields of graphic design with computer science, statistics, the visual design principles behind data representation must be understood as part of a single process. The process of understanding the data begins in many cases with a set of numbers and a question. The following steps form a path to answer this question (Fry 2007; Ware 2000):

- Acquire, obtain the data whether from a file on a disk or a source over a network
- Parse, provide some structure for the data meaning and order it into categories. The data needs to be broken down into its individual parts.
- Filter, remove all but the data of interest. This step involves filtering the data to remove portions not relevant.
- Mine, Apply methods from statistics or data mining as a way to discern patterns or place the data in mathematical context
- Represent, choose a basic visual model such as a bar graph, list or tree
- Refine, Improve the basic representation to make it clearer and more visually engaging
- Interact, add methods for manipulating the data or controlling what features are visible

Sometimes all of these steps will be used while other times four will need to be used. The initial format of the data determined by how it is acquired or parsed will often drive how it is considered for filtering or mining. The statistical method might drive the initial presentation. Therefore the final representation reflects the results of the statistical method rather than a response to the initial question.
Figure 4-10 shows the stages in order and demonstrates how later decisions commonly reflect on earlier stages. Each step of the process is inextricably linked because of how one step affects another. The connections between the steps illustrate the importance of the individual or team in addressing the project as a whole.

The mining, representation, refining and interaction stages in figure 4-10 need to be given more consideration than earlier steps in order to address the limits of human perception. These limits are especially true for multidimensional data where the number of dimensions can be over a dozen. In these multidimensional cases the large number of axes needed to create the display tends to overcrowd the important data, limiting the value of the plot for detecting patterns or other useful data. In this case the aim is to select some dimension of a data set in order to create a visualisation from which relevant information can be extracted. By identifying attributes that are significant in order to reduce dimensionality, improving efficiency of visualisation and the accuracy of algorithms for classification in visual data mining (Badjio and Poulet 2005).

Although an optimal subset of attributes is not necessarily unique, visualisations of more than a dozen attributes are useable for visual data mining, which necessitate attribute reduction. However to decide which attributes to select preliminary investigation into the data must be done by taking account of different tradeoffs between performance and complexity (tolerating lower performance in a model that also requires less features.)
4.14 Insight Gained

Insight is commonly stated as the goal of information visualisation and visual data mining by many authors Rhyne (2006). Insight is either directly obtained from a set of created visualisations or through confirmation of hypothesis as the results of automated analysis methods (Keim et al 2008). However the definition for insight remains vague making successful insight difficult to evaluate. In most cases it is equated with insight into tasks such as finding extreme values or the answers to questions which were being investigated.

Therefore it is useful to identify essential characteristics of insight and then consider whether measurements methods capture those characteristics. The important characteristics of insight include it being Rhyne (2006):

- Complex – insight is complex, involving all or large amounts of the data in a synergistic way, not simply individual data values
- Deep – insight builds up over time, accumulating and building on itself to create depth. Insight often generates further questions and hence further insight.
- Qualitative – insight is not exact, it can be uncertain and subjective and have multiple levels of resolution
- Unexpected – insight is often unpredictable, serendipitous and creative
- Relevant, insight is deeply embedded in the data domain, connecting the data to existing domain knowledge and giving it relevant meaning. It goes beyond data analysis to relevant impact.

4.15 Conclusion

This chapter outlined the principles of visualisation, its benefits and the process of visualising data. It discussed insight and introduced knowledge visualisation. These components are necessary in order to understand visual data mining and will be used at later stages to analyse and evaluate visual data mining tools. Among some of the topics discussed included defining visualisation as:
Visualisation systems traditionally focus[ing] on building graphical depictions of relationships among information in a human comprehensible form” - (Yang, Rundaensteiner and Ward 2007a)

Information visualisation and visual analytics were defined which form the basis of visual data mining. Along with knowledge visualisation which can be used as a framework for evaluating how well tools are able to transfer knowledge. The next chapter will talk about visual data mining and visualisation techniques used in visual data mining.
5 VISUAL DATA MINING

5.1 Introduction

This chapter investigates visual data mining the combination of data mining and visualisation fields to glean knowledge and deeper insight from large amounts of data (Keim et al 2003; Wong, 1999). This chapter look at visual data mining and also discusses approaches to evaluation. This chapter begins with defining visual data mining.

5.2 Defining Visual Data Mining

“Visual data mining discovers implicit knowledge from large data sets using data and or knowledge visualisation techniques” - Han and Kamber (2006)

“Visual data mining can be seen as a visual hypothesis generation and verification process: visualizations allow the user to gain insight into the data, formulate new hypotheses, then verify them via visual data mining methods” - (Chen et al 2007)

“visual data mining tries to integrate visual information retrieval and browsing techniques in one content-based similarity measurement process” - Eidenberger (2004)

The definitions above for Visual Data Mining (DM) describe it in several different ways. Generally visual DM can be described as a new field of study combining the power of information visualisation, visual analytics and data mining in order to discover novel and interpretable patterns with the help of the human perception abilities, Chen (2007). To achieve this visual DM combines automated data mining methods and computer aided, interactive visual techniques to allow analysts and knowledge workers to discover, interact and represent knowledge from data (Shneiderman. 2001; Chen et al 2007; Schulz, Nocke and Schumann 2006). This
combination makes it a highly attractive and effective tool for the comprehension of data distributions, patterns, clusters and outliers in data, Han and Kamber (2006).

The vision of visual data mining according to Wong stems from the following principles of simplicity, user autonomy, reliability, reusability, availability, and security, Wong (1999). Wong defines this simplicity in a number of forms such as usability, simplicity to learn with intuitive and friendly input mechanisms as well as instinctive and easy to interpret output knowledge. Simplicity also refers to a simple way to retrieve or recall means a customised data structure to facilitate fast and reliable searches and simple to execute a task in a minimum number of steps to achieve results Wong (1999). Kreuseler defines the aims of visual DM as enhancing the knowledge discovery and transfer process through the use of graphical representations of data mining results and processes and though the combination of visual and computational approaches to data exploration and insights into the data (Andrienko, Groth and Grinstein 2008; Kreuseler, Nocke and Schumann 2004). While Marghescu defines the aim of visual DM within the context of the data information knowledge loop as transforming raw data into information that the user can find useful, novel and interesting for the scientific or business work, Marghescu (2003).

The main difference between visual DM and other forms of data mining is it ability to allowed the analyst to use their experience to guide the data mining and knowledge discovery process. Enabling analysts to get an overview of the data with visualisation, helping to introduces insights, preference and biases in the early stages of the data mining lifecycle to reduced its overall computation complexity and reduce the set of uninteresting patterns in the product. Even more useful may be the new insights developed by the analysts concerning the quality and implications of the decisions made by the data mining process. Finally VDM as a mechanism to enable analysts to monitor the data mining process, inputs and process and interact with the data mining process may influence the way decision are made about to explore the data and ultimately what knowledge is gained (Badjio and Poulet 2005; Ganesh et al 1996).
5.3 Visual Data Mining Benefits

Visual DM has a number of major advantages. According to Eidenberger (2004) it provides a rich visualisation of retrieval space where in contrast to traditional approaches the user is provided with a variety of information details and different views on the data. The rich information includes histories on all revenant aspects of earlier retrieval steps. Secondly even though the user interface does offer rich information, the interaction metaphors used are intuitive, easy to learn and powerful. Thirdly Visual DM offers more degrees of freedom to the analyst to explore the data further being investigated. Finally it allows for the easy integration of human centred similarity measures that can be applied to large dataset collections or just to fractions of them.

Visual DM is particularly useful in exploratory data analysis and exploring large databases. In this situation little is known about the data and the exploration goals are often vague. Since the user is directly involved in the exploration process, shifting and adjusting the exploration goals is automatically done if needed. In addition to this direct involvement by the analyst the main advantages of visual data mining over automatic data mining techniques are (Badjio and Poulet, 2005):

- Visual data exploration can easily deal with highly non-homogeneous data
- Visual data exploration is intuitive and requires no understanding of complex mathematical or statistical algorithms or parameters

Visual data mining can also improve confidence in the results, so that the KDD process is not just a black box giving more or less comprehensible results. The quality of results is improved by the use of human pattern recognition capabilities. If the analyst is the data specialist, then domain knowledge can be used during the whole processes and not just during the interpretation of results and computer devices can display vast amounts of information using various techniques (Badjio and Poulet 2005). In addition, visual presentations can be very powerful in revealing trends, highlighting outliers, showing clusters, and exposing gaps in data (Poulet 2004; Shneiderman 2001)
Visual data mining is particularly useful when little is known about the data and when goals are indistinct, as analysts’ guide the exploration process they can easily shift or adjust the goals as needed, (Keim 2004). Many real world problems, including intrusion detection, fit into this category, Toeh et al (2004). Using such an exploration approach analysts can hypotheses about complex, high dimensional, very large dataset using the interactive data presentation and query resources to allow domain experts to quickly examine ‘what if’ scenarios while interacting with multivariate visual displays, Oliveira and Levkowitz (2003). The drawback is that visualisation do not scale well when handling millions of data items, which is one of the reasons for a lack of capacity with current visual data exploration tools.

One of the final benefits is based on a synthesis of automatic and visual data mining which often leads to better results, a higher degree of satisfaction and confidence in the findings, For instance presenting data in an interactive, graphical form often fosters new insights, encouraging the formation and validation of new hypotheses to the end of better problem-solving and gaining deeper domain knowledge (Keim, Panse and Sips 2003) and provides the ability for intuitive feedback for data analysis and supports decision making activities (Shneiderman 2001). The data can also be filtered directly and instantly, with multiple complementary perspectives being displayed simultaneously, highlighting subsets of data simultaneously Fry (2006); Heidenburger, 2004).

5.4 Types of Visual Data Mining

A significant number of papers describing visual DM approaches found fall into two categories. Either they use visual data exploration systems or techniques to support a knowledge extraction goal or a specific mining task or they use visualisations to display the results of a mining algorithm such as a clustering process, decision tree or classifier and thus enhance user comprehension of the results (Oliveira and Levkowitz 2003; Wong, 1999).
A third visual representation type is to use visualisation in creating the model of data along the steps of an analytical mining algorithm or along the steps of the whole knowledge extraction process with the goal of supporting users in the process of interacting with the algorithm (Oliveira and Levkowitz (2003); Han and Kamber (2006); Costa, D. D. and Venturini, G. 2006). The last type is the overview data visualisation approach where the visualisation is used to give a general overview of the data these four types of visual DM are explained in more detail:

- **Overview data visualisation**

  Usually in visual data mining the analyst wants to start by getting an overview of the data (Keim et al 2003). The datasets can be visualised at different levels of granularity or with different combinations of attributes or dimensions without any data mining algorithms being applied to give an overview of the data. This data can be presented in a number of ways for example as box plots, 3-D cubes, data distribution charts, curves, surfaces or link graphs Han and Kamber (2006). From this overview of the data an analyst can identify interesting patterns or groups in the data and can focus one or more of them and drill down and access details of the data (Keim et al 2003).

- **Data mining result visualisation**

  The result element of visual data mining deals with the actual visual presentation of data mining algorithms results or knowledge obtained from data mining in visual forms. For example as scatter plots and box-plots, as well as decision trees, association rules, clusters, outliers or generalised rules Han and Kamber (2006).

- **Data Mining Process Visualisation**

  The visualisation of the data mining algorithm processes and datasets in visual forms showing how the data is extracted, cleaned, integrated or pre-processed and mined. It may also show which method is selected for data mining and where the results are stored Han and Kamber (2006).
• Interactive visual data mining

Visual data mining exploration usually takes a three step process of overview first, zoom and filter and then details on demand. The analyst first take an overview of the data, identifying interesting patterns or subsets of data and focusing on one or more of them. It is this interaction which speeds up visual data mining and allows the analyst to drill down into the data and combine the other types of visual data mining (Keim et al 2003; Heidenburger, 2004).

5.5 Visual Data Mining Exploration

The visual DM process can be seen as a hypothesis generation process, where visualisation of the data allow the user to gain insight into the data and come up with new hypothesis (Keim, 2004). The verification of these new hypotheses can then be done using via visual data mining or by automatic means using statistics or machine learning (Eidenberger, 2004; Oliveira and Levkowitz 2003; Badjio and Poulet, 2005). In addition to providing analysts with potential these new insights into the data and direct involvement with the data visual data allows the analyst to (Keim, 2004):

• Easily deal with highly non-homogenous and noisy data
• Intuitive understand and requiring no understanding of complex mathematical or statistical algorithms or parameters.

This makes knowledge discovery in the data faster, increases confidence in the result and can provide better result where automatic algorithms usually fail. These benefits are achieved in visual data mining by integrating visual information retrieval combined with the browsing techniques in one content based measurement process (Eidenberger, 2004) and trigger re-computation of algorithms with different attributes. This process can be seen more clearly in figure 5-11 (Schulz, Nocke and Schumann 2006).
Figure 5-11  A general visual data mining framework design (Schulz, Nocke and Schumann 2006)

In the proposed framework for developing a visual data mining tool in figure 5-11 the major components of the tool are explained which is useful for understanding how the tool works. During the pre-processing stage data which has been imported can be cleaned and filtered for later stages to remove missing values or irrelevant data. The calculation of descriptors tries automatically to gather enough metadata to enable the user to determine fitting mining and visualisation techniques to represent results. The algorithmic kernel is the core of the tool and does the work of analysing data involving extracting substructures, clustering or decomposing a visualisation into more detail and calculating a graphical layout for the resulting data. The interaction on the gained graphical representation is used for the actual visual exploration of the dataset. At this stage the post processing can be done to further manipulate and query the data set interactively through the visualisation and to write back those changes to the database (Schulz, Nocke and Schumann 2006). This high level overview of the visual data mining process is better understood by reviewing the implementation of this framework and the related features of each stage as shown in figure 5-2.
5.6 Tasks to be visualised

The mission critical visual data mining tool must enable the user to perform diverse analytical tasks such as Thomas and Cook (2005):

- Understanding past and present situations quickly and events that created current conditions
- Identifying possible futures and warning signs
- Monitoring current events or warning signs for events.
- Supporting the decision maker in times of crisis

These tasks will be conducted through a combination of individual and collaborative analysis often under extreme time pressure. The tool must also enable hypothesis and
scenario based analytical techniques, provide for the analyst to reason based on the available evidence. The tools purpose is to enable the analyst to discover the unexpected whether by detecting unexpected relationships are missing. The tool must also focus on capturing the discoveries and the results of the analytic process and on making them available to others as a basis for future understanding Thomas and Cook (2005).

5.7 Visualisation Technique Classification

As Negash (2004; 2007) states it is the actual visualisation technique which exploits the human capabilities to interpret and extract information from data, provides an overview of the data stores and amplifies cognition. One approach used for actually selecting the visualisation technique to use is by Keim (2002; 2001) who bases it on three criteria of Keim:

- the data type to be visualised

The data type to be visualised may be one dimensional data, such as temporal data, two dimensional data such as geographical maps, multi-dimensional data such as relational tables, text and hypertext such as news articles and web documents, hierarchies and graphs such as telephone calls and, algorithms and software such as debugging operations.

- the visualisation technique

According to Keim(Keim et al 2003) the visualisation technique used may be classified as standard 2D/3D displays, such as bar charts and x-y plots, Geometrically transformed displays such as hyperbolic plans (Walter & Ritter 2002) and parallel coordinates Inselberg and Dimsdale (1990), icon based displays such as chernoff faces and stick figures, pixel displays such as the recursive pattern (Ankerst, Keim and Kriegal 1995) and circle segments (Ankerst, Keim and Kriegal 1996), and Stacked displays, such as treemaps (Johnson and Shneiderman 1991; Shniederman 1992) and dimensional stacking Ward (2002).
The interaction technique used which allows analysts to directly navigate and modify the visualisation as well as subsets of the data for further operations. Examples include Dynamic Projection, interactive filtering, interactive zooming, interactive distortion, interactive linking and brushing. Each visual data mining tool in this way may be designed to support a different data type or may use a combination of visualisation and interaction techniques. (Keim et al 2003). These three visual classification criteria can be seen in figure 5-13, each visualisation technique can potentially be used in conjunction with an interaction and distortion technique or data type to be visualised Keim (2001).

![Figure 5-13  Classification of information visualisation techniques Keim (2001)](image)

The following three sections discuss data to be visualised, the visualisation technique and interaction and distortion technique in more depth.
5.8 Data to be visualised

In visual data mining the data usually consists of a large dataset, that presents multiple attributes to be mapped also known as dimensions such as an observation, measurement or transaction (Pillat et al, 2005; Keim 2002). These dimensions could be information on customer properties, ecommerce transactions or sensor data. The data from these sources may be one dimensional, two dimensional, multi-dimensional or may be more complex data types such as text, hypertext or hierarchy graphs. One of the most important parts of visualisation understanding the dimensions in data. The different types of dimensions include:

- One dimensional data

One dimension data usually has one dense dimension. A typical example of one dimensional data is temporal data. With each point in time one or multiple values may be associated for example for stock prices.

- Two dimensional data

Two dimensional data has two distinct dimensions. A typical example is geographical data which has the dimensions of longitude and latitude. X-Y plots are a typical way of showing two dimensional data with maps being a special type of x-y plot for showing two dimensional geographical data. Although both one and two dimensional data may seem easy to visualise if the number of records to be visualised is too large, temporal axis and maps quickly get flooded making it difficult to understand and interpret the data.

- Multidimensional data

In many cases the data set will have more than three attributes and therefore will be challenging to visualise as people live in a three dimensional world and have no of higher dimensional space. In these cases there are too many objects to visualise at once and also too many dimensions to visualise at once Toeh and Ma (2005).
Given the 2-D nature of visualisation, some ordering or organisation of the dimensions must be assured. This organisation can have a major impact on the expressiveness of the visualisation. Different ordering of dimensions can reveal difference aspects of the data and affect the perceived clutter and structure on display. Thus completely different conclusions can be drawn based on each display (Peng, Wad and Rundensteiner, 2004).

Since no simple mapping is possible more complex visualisation techniques need to be used such as the parallel coordinate technique. Parallel co-ordinates displays each multidimensional item as a polygonal line which intersects the horizontal dimensional axes at the position corresponding Keim (2002).

- Text and hypertext

There is some data which cannot be visualised dimensionally such as text and hypertext. As they cannot be easily described with numbers and therefore most standard visualisations cannot be applied. In most cases a transformation is required to transform the data into description vectors is necessary before visualising the data. This is done by using word counting which is often combined with a principle component or multidimensional scaling.

- Hierarchies and graphs

The interdependencies between pieces of information are widely represented in graphs. A graph consists of a set of objects called nodes, and connections between these objects called edges.

- Algorithms and software

This aims at translating algorithms and software coding into visualisations to support a programmer debugging code for example by visualising errors.
5.9 Visualisation Techniques

Until recently some most well known techniques used for visualising data sets have included x-y plots, scatter plots, line plots, box plots and histograms. Although these are useful for data exploration they are relatively limited to small and low dimensional data. Within the last decade however a large number of modern visualisations have been developed allowing visualisations of larger multi-dimensional datasets sets without two or three dimensional semantics (Marghescu, 2003; Keim 2002; Card, Mackinlay and Shneiderman 1999; Spence, 2000; Ware 2000).

The following discusses various techniques used in order to implement a specific visualisation system, and discusses how they can even be combined into hybrid visualisation techniques.

5.10 Standard 2D/3D Display

Spatial data can be presented in 2D views that provide information about only two dimensions or 3D views that provide information about 3D structure (Tory et al 2006); Keim (2002). The two views are effective for different tasks with 2D views often being used to establish precise relationships, whereas 3D views are used to gain a qualitative understanding and to present ideas to others Springmeyer, Blattner and Max(1992). The properties of 2D graph drawings are they are:

- Planarity (no cross lines)
- Orthogonality (only orthogonal lines)
- Grid property (co-ordinates of vertices are integers)

The aesthetic properties or optimisation goals are:

- Minimal number of lines crossing
- Optional display of symmetries
- Optional display of clusters
- Minimal number of bends in polyline graphs
• Uniform distribution of vertices
• Uniform edge lengths

Some example of these types of techniques includes box graphs, line graphs, pie charts, distribution and histogram charts and bar charts.

5.10.1 Box Graphs

Box graphs are powerful visualisation for comparing distributions, providing a compact view of where the data is centred and how it is distributed over the range of a variable (Bakker, A., Biehler and Konold 2004; Tukey 1977). From each visualisation five values from a set of data can be conveniently seen at once, the extremes, the upper and lower hinges and the median, see figure 5-14 McGill, Tukey and Larsen. (1978).

Figure 5-14  Configuration of a Box Plot (Bakker, A., Biehler, R., Konold, C. 2004)
An example of a box plot can be seen in figure 5-15 showing Tippett’s warp break data (Bakker, A., Biehler, R., Konold, C. 2004) show the range of breaks which occur.

Figure 5-15    Tippett’s Warp Break Data (Bakker, A., Biehler, R., Konold, C. 2004)

5.10.2Line Graphs

A line graph is a way to summarise how two pieces of information are related and how they vary depending on one another. Line graphs are usually shows the values of one column on the horizontal x axis against another column y axis within an x and y coordinate system. Line segments connect at adjacent points from the values of the data column. Numbers along the side of the graph show the scale.

The data values for the x axis can be either discrete or continuous. If the data values are discrete the values become labels for successive locations on the axis. The data
values for the y axis must be continuous. Often line graphs are used to demonstrate time series trends Soukup and Davidson(2002).

To accommodate multi dimensional data sets, multiple line graphs are displayed in a multi line graph, see figure 5-16. The diagram shows a car dataset comparing different car make types and distinguishing them by using different coloured lines (Grinstein et al, 2001).

![Figure 5-16 Multiple line graph of the car data set (Grinstein et al, 2001)]

### 5.10.3 Pie Graphs

A pie chart is a circle graph divided into pieces, each displaying the size of some related piece of information making up some whole. The unique column names for the values become the labels for the slices of the pie, while the continuous column values are summarised into the contribution (Soukup and Davidson 2002). An example of two pie charts can be seen in figure 5-17, where one pie chart of women’s phone usage is compared against men’s phone usage Rumsey (2007).
Figure 5-17 Pie charts comparing male versus female personal cell phone use Rumsey (2007)

5.10.4 Distribution and histogram graphs

Distribution and histogram graphs use an aggregation method to illustrate the data distribution of a single dimension. To achieve this data is partitioned into several sub ranges with a minimum and maximum value. Each sub-range corresponds to a bin, decided by the number of data points whose value falls in the range associated with that bin. The number of bins is the bin count which influences the usability and effectiveness of a histogram (Cui, Ward, and Rundensteiner, 2006; Soukup and Davidson(2002). An example of a stacked histogram can see seen in figure 5-18, showing the fruit sales from 1992-1997 Dix, Beale, and Wood (2000)

Figure 5-18 Stacked histogram of products sales by year Dix, Beale and Wood (2000)
5.10.5 Bar Chart

Bar charts are among the most simple presentation graphic, which show only aggregated values such as the total sales for each month as seen in figure 5-19 Keim and Hao (2007); Fayyad, Grinstein and Wierse (2002a). This aggregation makes for easy comparison between the groups Soukup and Davidson(2002).

However bar charts usefulness is limited if the analyst is interested in relationships between the different attributes such as product price, number of orders, and quantities. The reason for this limitation is that multiple bar charts for different attributes do not support the discovery and correlation of interesting subsets, which is one of the main tasks in mining transaction data Keim and Hao (2007).

![Pixel bar chart showing sales transaction distribution on the x-axis: month, on the y-axis ordering: price; colour: price Keim and Hao (2007)](image)

Figure 5-19

5.11 Geometrically Transformed Displays

Geometrically transformed display techniques are used to find interesting transformation of multidimensional data sets. The group of geometrically display techniques range from well known parallel coordinates, scatterplot matrixes, landscapes, projection pursuit and projection views (Keim 2002; Keim, 1997).
5.11.1 Parallel Co-ordinates

In parallel coordinates each dimension corresponds to an axis, which are organised into vertical lines (Rundensteiner et al 2007). Data is represented by connecting a set of points on each axis which is scaled from the minimum to the maximum value of the corresponding dimension. Using this method it is possible to visualise a high number of dimensions, making it easier to spot correlations between variables in the data set.

The main difficulty to directly applying parallel coordinates to a large dataset is that the level of clutter present in the visualisation reduces the amount of useful information one can perceive (Keim 2001; 2005) and the scaling to the maximum and minimum for each variable makes it difficult to access directly correlations between a pair.

The parallel coordinate’s plot of the vegetable oil data is shown in figure 5-20. From the structure of the resulting display it is possible to draw conclusions for the relationship between corresponding data values (Xu, Hong and Song, 2007a; Hao et al 2007)

![Figure 5-20](image)

Figure 5-20 The parallel co-ordinates plot of the vegetable oil data set (Xu, Hong and Song, 2007a)
5.11.2 Scatterplot Matrices

The 2D scatterplot is a classic technique to display the relationship between two variables. It is generated by plotting points or markers whose positions are by two variables. It makes it easy to understand the joint behaviour such as clusters, outliers, trends and correlations exhibited by the points in a scatterplot. The number of points which can be plotted is almost limitless. The 2D scatterplot can be extended to three, four five or more dimensions by animation, different colours or interactions. However generally a grid of 2D scatter plots is the standard means of extending the scatterplot to higher dimensions (Hao et al 2007;Keim 2005).

In the scatterplot matrix in figure 5-21a show an original scatterplot matrix, thirty two dimensions are projected onto 16 scatterplots, where each pair of dimensions has two scatterplots showing their relations. This is useful for looking at all possible interactions or correlations between dimensions. The scatterplot matrix is symmetrical with the data values for a particular dimension shown in the respective row on the y-axis and in the respective column on the x axis (Cui, Ward and Rundensteiner, 2006). The scatterplot matrix has good interpretation ability and is easy to see the bi-variant relationships of the data (Xu, Hong and Song, 2007a).

Figure 5-21 (a) Show an original scatterplot matrix. The data points in the diagonal plots form a straight line (b) scatterplot matrix with histograms. The diagonal plots show the histogram of each dimension (Cui, Ward and Rundensteiner 2006)
5.11.3 Projection pursuit

Projection pursuit is a recent statistical method aiming at solving the problem of identifying structure in high dimensional data. The visualisation achieves this effect by projecting data onto a low dimensional space making it possible to spot structure between two data classes (Corchado, MacDonald, Fyfe 2004; Fayyad and Uthurusamy 2002).

The drawback to this technique is it is not always clear how to get useful information from the linear combinations of the dimensions because of the lack of easy interpretation of the X and Y axis (Fayyad and Uthurusamy 2002; Grinstein et al, 2001), see figure 5-22 for an example of parallel coordinate visualisation.

Figure 5-22 The Parallel Coordinates plot displays the US Census data of the 50 states. Colour is used to point out which party won the 2000 presidential election.
5.12 Icon Displays

An iconographic display is a graphical representation visualising high-dimensional data by letting each coordinate dimension of a parameter or attribute of an entity such as a pixel icon or glyph and displaying a number of these entities at one on the screen. These displays can integrate several dimensions at once and therefore represent high dimensional datasets (Grinstein et al, 2001; Keim 2001).

There are two types of glyph and icon visualisation the first are displays where certain dimensions of the dimensional data set are mapped to certain features of the glyph or icon such as Chernoff faces. The second type of glyph and icon visualisation have glyphs or icons backed together in a dense display with textures representing features of the data set, for example shaping coding or colour icons (Grinstein et al 2001; Keim 1997)

5.13 Chernoff faces

Chernoff Faces developed by Chernoff (1973) display data using cartoon faces by relating different variables to a selection of facial features. The technique was developed using the idea since facial characteristic of real faces are used that they may be particularly useful for people given their heightened sensitivity to facial structure and expression (Lee, Reilly and Butavicius 2003).

The facial features generally used in Chernoff faces are shown in figure 5-23; these are then mapped to different dimensions in a multidimensional data set (Morris, Ebert and Rheingans 2000). An example of a Chernoff visualisation is shown in figure 5-24 Keim (2005)

![Figure 5-23 Five examples of Chernoff faces Keim (2005)](image-url)
5.14 Dense Pixel Displays

The idea behind dense pixel techniques is to map each dimension value to a coloured pixel and group the pixels belonging to each dimension into an adjacent window, see figure 5-25. Since in general dense pixel displays have only one pixel for each data value the technique allows the visualisation of a vast amount of data values from anywhere up to a million values (Keim et al, 2007; Keim 2002; Keim 1996).
Dense pixel techniques use different arrangements for different purposes. However with the appropriate arrangement of the pixels the resulting visualisation provides detailed information on local correlations, dependencies and hotspots. All pixel oriented techniques split the screen into multiple windows for each attribute. Inside the windows the data values are arranged according to the given overall sorting which may be data driven or query driven. This can allow functional dependencies and other interesting relationships between attributes to be detected by relating corresponding regions in the multiple windows (Keim and Kriegel 1996).

5.15 Recursive Pattern Technique

The recursive pattern technique displays as many items of data as is possible at the same time with the amount only limited by the number of pixels of the display. When dealing with multidimensional data without 2D or 3D semantics there is a major problem with finding a way to display meaningful arrangements of data onto the screen Fayyad, Grinstein and Wierse 2002b; Keim et al 2006).

Even if the data has a natural order to it such a time dimension there are still many possibilities in arranging the data. For instance the data could be arranged left to right or top down in a column by column fashion. However when done pixel wise the result does not provide useful results. One solution is to organise the pixels into small groups arranged to form a global pattern. The strategy corresponds to a two step approach with a first order pattern formed by grouping the pixels and a second order pattern formed by the global arrangements. This process can be iterated up to an arbitrary level forming a general recursive scheme. The pattern is based on a simple back and forth arrangement. First a certain number of elements are arranged from left to right, then below backwards from right to left and so on(Keim 1995; Ankerst 2001). Figure 5-26 a, b and c show three ways that the recursive pattern can be used.
A real example of a recursive pattern technique being used is shown in figure 5-27 where the stock prices for the Dow Jones, Gold, IBM and US Dollar are shown for seven years. The several vertical bars correspond to the seven years and the subdivision of the bars into the 12 months within each year. The coloring maps high stock values are shown in light colours and the low attributes are shown in dark colours. Using this visualisation it is possible to see that the gold price was very low in the fifth year and that the US Dollar exchange rate was the highest in the eight month of the second year (Ankerst 2001).
5.15.1 Circle Segments

The circle segment technique is another approach to visualizing large high dimensional datasets. The idea is to represent the whole dataset by a circle which is divided into segments one for each attribute. Within the segment each attribute value is visualized by a single coloured pixel. The arrangement of the pixel starts at the center of the circle and continues to the outside a line orthogonal to the segment halving line in a back and forth manner. The rationale behind this approach is that close to the centre all attributes are close to each other enhancing the visual comparison of their values (Xu, Hong, Song, 2007b; Keim 2005; Ankerst 2001).

In figure 5-28 these segments are shown arranged with the attributes. An example is shown in figure 5-29 of 20 years of fifty different stock prices in the Frankfurt Stock Index. The light represents the high stock prices and dark colours low ones. Thus light circular regions correspond to high stock prices of different stocks at the same time. It can easily be seen that most stocks prices have a very similar trends where a few show a different progression (Ankerst 2001; Keim 1997)

![Illustration of the Circle Segments technique for 8-dimensional Data](image)

Figure 5-28 Illustration of the Circle Segments technique for 8-dimensional Data (Ankerst 2001)
5.16 Stacked Displays Hierarchical

Stacked displays are designed to present data partitioned in a hierarchical fashion. In the case of multidimensional data, the data dimension to be used for partitioning the data and building the hierarchy have to be selected appropriately (Kiem, 2005). The usefulness of the resulting visualisation depends largely on the data distribution of the outer nodes and therefore, the dimensions which are used for defining the outer node system have to be selected carefully (Merino et al 2006; Keim 1997). Two examples of stacked display hierarchical dimensional stacking and tree mapping.

5.16.1 Dimensional Stacking

Dimensional stacking is a 2D or 3D point projection of the data where dimensions are embedded within other dimensions. The stacking divides a 2D grid into sets of embedded rectangle, representing categorical dimensions or attributes of the data. Two outer dimensions are placed along the X and Y axis and each additional pair of dimensions is embedded into the outer rectangles until all dimensions are incorporated (Keim 2005; Langton, 2006; Fayyad and Uthurusamy 2002; Grinstein et al 2001).
The dimension of the dataset is first turned into a number of bins. Then two dimensions are defined as the horizontal and vertical axis creating a grid display. Within each box in this grid this process is applied again with the next two dimensions. This process continues until all dimensions are assigned.

The inner most dimensions are named the fastest dimensions because along these dimensions two small bins immediately next to each other represent two different ranges of the dimensions. For dimensional stacking the bins within which data points fall are shown as filled squares.

The data points within a group share similar attributes in many features. Thus this view helps the analyst to search for groupings in the dataset as well as to detect subtle variances within each group of data points. The other data points are considered outliers may also be readily perceived if most data falls within a small number of groups. An example of cluster reduction with dimensional stacking is shown in figure 5-30. Eight adjacent neighbors are used in the calculation. In figure 5-30a and b shown the difference in data order after the change. The dataset is based on the display having optimal ordering starting with petal length, width, sepal length, sepal width. Using this method there are fewer islands and the filled bins are more concentrated (Peng 2004).

![Figure 5-30 Dimensional stacking visualisation for Iris dataset. (a) Represents the data with original dataset, and (b) shows the data with clutter reduced (Peng 2004).](image)

5.16.2 Tree mapping

The tree map visualisation method maps hierarchical information to a rectangular 2D display in a space filling manner. Using the method sections of the hierarchy containing more information is allocated more display space while portions of the
hierarchy which are less important to the specific task at hand can be allocated space. The tree map partitions the display space into a collection of rectangular boxes representing the tree structure (Lü and Fogarty, 2008; Tu and Shen 2007; Johnson and Shneiderman 1991; Bederson and Shneiderman 2002). An example of a tree map is shown in figure 5-31 Keim (2005), showing how a countries budget is spent.

Figure 5-31 Shows an example of a tree map Keim (2005)

5.17 Interaction and Distortion Techniques

In order to effectively explore the visualisations developed it is necessary to use interaction and distortion techniques. An interaction technique enables the analyst to directly interact with the visualisations and dynamically change the visualisations according to the exploration objectives. Interaction techniques also provide a means to relate and combined multiple independent visualisations. Distortion techniques allow the analyst to focus on visualisations details while preserving an overview of the data. The basic idea of distortion techniques is to show portions of the data with a high level of detail while showing other portions of the data with a lower level of detail. The terms dynamic and interactive, depends on whether the changes to the visualisation is made automatically or manually (Kiem, 2002; Rodrigues 2006). The following is the several different types of interaction and distortion techniques:
• Dynamic Projection

The idea of dynamic projection is to change the projections in order to explore a multidimensional dataset. The number of possible projections is exponential in the number of dimensions, i.e., it is intractable for a large dimensionality. The sequence of projections shown can be random, manual, precomputed, or data driven (Keim 2002; Keim 2001; Keim and Kriegel 1994).

• Interactive filtering

When exploring large datasets, it is important to interactively partition the data sets into segments and focus on interesting subsets. This can be done by a direct selection of the desired subset through browsing or by the specification of properties of the desired subset through querying. However browsing techniques are very difficult for very large datasets and querying often does not produced the desired results (Keim 2005; 2002).

• Interactive Zooming

Zooming us a well known technique which is widely used in a number of application. When dealing with large amounts of data, it is important to present the data in a highly compressed form to provide an overview of the data but allow a more detailed display of the data. Zooming means displaying the data objects larger and also means that the data representation automatically changes to present more details on higher level zooms. The objects can be represented by a single pixel on a zoom level, as icons on an intermediate zoom level and as labelled objects on a high resolution. This approach allows an initial view to allow the analyst to detect patterns, correlations and outliers in the data set. In order to explore a region of interest the user can zooming with the result that the effected rows are displayed in more detail or possibly in textual form , Keim (2005;2002).

• Interactive Distortion
Interactive distortion techniques support the data exploration process by preserving an overview of the data during drill-down operations. The basic idea is to show portions of the data with a high level of detail while others are shown with a lower level of detail. Popular distortion technique is hyperbolic and spherical distortions, which are often used on hierarchies or graphs, but may be also applied to any other visualisation technique, Rodrigues (2006) and Keim (2005; 2002).

- Interactive Linking and Brushing

There are numerous possibilities to visualize multidimensional data each with strengths and weaknesses. Linking and brushing overcomes the shortcomings of an individual techniques. This makes it possible to detect dependencies and correlations. Interactive changes made in one visualisation are automatically reflected in the other visualizations (Cui Ward and Rundensteiner 2006; Keim 2002).

Typical examples of visualisation techniques which are combined by linking and brushing are multiple scatterplots, bar charts, parallel coordinates, and maps (Artero and Oliveira 2004; Keim 2002).

- History Mechanism

This function is critical for interactive exploration since it allows the user to quickly back out of changes and try a different exploration path without having to rebuild the desired starting point from scratch. Support for undo also promotes more experimentation and exploration as there is no fear of losing work done thus far. Therefore, exploration typical includes different undo and redo steps to choose mining techniques as well as appropriate parameterisation to achieve the desired results. This includes storage of actions and storage of the system state (Kreuseler, Nocke and Schumann 2004; Artero and Oliveira 2004; Stolte, Tang and Hanrahan 2002).

5.18 Defining the interaction process

Before setting the criteria for the evaluation of a visual data mining tool it is necessary to get an overview of the interaction process. This can then also be used to divide the framework into manageable pieces. The interaction process can be seen in figure 5-32,
in which the tasks of the user and visual data mining tool are defined. This diagram extends the interaction process defined by Marghescu (2004), replacing the tasks and define the roles of user and tool.

This interaction flow recognises the use of domain knowledge with a certain goal in mind in identifying important information along with the essential tasks which must be performed during data analysis. This is important as it leads to breaking the questionnaire and feedback for the framework into groups of how effective the tools are at achieving the goals, exploration, hypothesis generation, analysis and the insight gained. The diagram also shows the primary functionality of a visual data mining tool and desirable attributes.

![Visual Data Mining Interaction Process Diagram](image)

**Figure 5-32  Visual data mining interaction process**

### 5.19 Conclusion

This chapter discussed visual data mining and visualisation techniques along with the characteristics of visualisations and rules set down by authoritative authors in this area and discussed the visual data mining interaction process. The next chapter builds on this chapter and introduces evaluating visual data mining tools.
6 EVALUATING VISUAL DATA MINING TOOLS

6.1 Introduction

The chapter looks research and other bodies of work surrounding evaluation of visual data mining tools. Theses area includes the following:

- knowledge management tool measurements (6.2)
- ISO software evaluation (6.3)
- usability(6.4)
- data mining tool evaluation (6.5)
- visualisation evaluation (6.6)
- evaluating visual data mining tools (6.6)
- desirable attributes for visual data mining tools (6.11)
- testing approaches (6.12)

With an understanding of these different evaluation methodologies it will be possible to apply and defend criteria decisions, testing techniques and to avoid the same pitfalls other authors have made. To begin this evaluation process investigation begins with looking at the important knowledge management tool criteria.

6.2 Knowledge Management Tool Measurements

The area of knowledge management system and tool evaluation was investigated as part of developing a evaluation framework for visual data mining tools. In particular the work of Jennex and Olfman (2004) was particularly useful as parts of their success effectiveness model of knowledge management can be applied to data mining tools.

According to Jennex and Olfman (2004) a success knowledge management system and therefore each tool should perform one of the functions of knowledge creation, storage/retrieval, transfer and application well. They used studies to determine that the largest determining factors in determining a knowledge management tools success is the cost, functionality and vendor. In order to measure this they have subdivided these
criteria further two three levels of details and decision alternatives that affect the ultimate selection of the tools. They also noted collaboration and communication as being deciding factors. Other authors have highlighted training and education of the tool as being important factors (Ngai and Chan 2005).

Another methodology for measure knowledge management tools were knowledge management performance index (Lee, Lee & Kang 2005) used to evaluate management performance. It uses the basic assumption that knowledge may be viewed from an unified perspective; it circulates in the organisation creating knowledge assets and influencing organisation performance. It has multifaceted characteristics such as state of mind, object, having access to information, or potential for influencing future action. With a unified perspective of knowledge make the following assumptions (Lee, Lee & Kang 2005):

1. KM activities result in knowledge circulation processes, there are five components: creation, accumulation, sharing, utilization and internalisation of knowledge.

2. Km is defined tactically by all kinds of management activities that promote the use of knowledge circulation processes

3. A firm can increase its flexibility and adaptability in a rapidly changing business environment by focusing on the efficiency of km activities

**6.3 ISO Software Evaluation**

Another approach to measure visual data mining tools is to apply standardised methodologies designed to evaluate software. The International Organisation for Standardisation (ISO) developed ISO 9126 standard for evaluating software quality (ISO 2004). The standard splits evaluation to external and quality in use metrics.

The external approach divides the aspects of its quality into the six sub criteria of functionality which it calls reliability, usability, efficiency, maintainability and
portability. These sub criteria are further sub divided so that the low level characteristics can be used to evaluate each external quality element (Zeiss et al 2007; Abran et al 2003), which can be seen in figure 6-23.

The other evaluation quality metric for ISO 9126 refers to the quality perceived by the end user executing the software product in a specific context (Abran et al 2003; Hyatt and Rosenberg 1996). It breaks the quality of use into the three criteria of effectiveness, efficiency and satisfaction, see figure 6-33. Unlike the external quality metric it concentrates on the interaction between the user and a product while completing a task in a technical, physical, social and organisational environment. This is shown in more detail in figure 6-34 showing the quality in use environment and how the observation is used to measure the quality. As seen in the diagram the user interacts with the product and the satisfaction along with the performance and efficiency is measured (Wixon and Wilson 1997; Bevan 1995). Using this approach the use is the different context and actually user response can be measured by seeing how users react in different situations.
The useful content from this is both the functional approach to the capabilities of the tool along with actual user testing to measure how the tool performs in different scenarios. The quality of use approach also has the potential of being modified to become more relevant to knowledge management measurement by measuring satisfaction and effectiveness can be measured in relation to collaboration, knowledge creation and the knowledge value. It alternatively could be used to access necessary training or skills to operate a product, assess the environment such as the supporting business processes.

Although both knowledge management measurement methodology and software quality methodology have provided good insights into methods to measure visual data mining other user measurement techniques need to be investigated. In the knowledge management area user knowledge and usability is extremely important to ensuring knowledge creation is being done efficiently.
6.4 Usability

“Usability is most often defined as the ease of use and acceptability of a system for a particular class of users carrying out specific tasks in a specific environment”

- Holzinger (2005)

The usability or easy as use as Holzinger (2005) defines is testing how easy or acceptable a system on users in an environment. As knowledge workers will be using the visual data mining tool on a daily basis ensuring the most effective tool is very important. It is generally accepted that there are five essential usability characteristics should be part of any software project. These include (Nielsen, 1994):

- the learn ability, so the user can rapidly begin working with the system,
- efficiency, enabling a user who has learnt the system to attain a high level of efficiency,
- memorability allowing the casual user to return to the system after a period of non use without having to relearn everything
- no catastrophic errors occur
- satisfaction making the system pleasant to use

Using these criteria Nielsen (1994) began developing sub criteria for these characteristics and developed a usability model as shown in figure 6-35 which can be applied and measured on software systems. Such as the percentage of tasks accomplished, errors and frequency so that quantitative measures could be created and software compared. A similar usability model was also generated a year before by Dix (1993) around creating these measurements.
Since Nielsen (1994) and Dix’s (1993) work measuring usability continues to be an evolving field (Winter, 2007; Abran, 2003). Recently these usability principles have also been developed to be applied to both data mining and visual data mining tools.

### 6.5 Data Mining Tool Evaluation

According to Redpath and Srinivasan (2003) in order to judge the effectiveness of data mining tools, criteria needs to be established which can be used to measure tools effectiveness. There have been several different perspectives on what are good characteristics of data mining tools. Carr (1999) defines these characteristics need to be defined around the tasks of providing:

- an overview of the entire dataset.
- zooming and focusing on certain data subsets.
- filtering by reducing down the dataset according to specified attribute values.
- details on demand, in which additional data provided by an action performed by the user.
• relationships providing the ability to relate data items with similar characteristics.
• history providing the ability to undo an action and show the steps performed up to a certain point.

In the same year Goebel and Gruenwald (1999), gave their characteristics to look for in data mining tools such as include evaluating the:

• Ability to access a variety of data sources
• Online/Offline data access, where the data is a snapshot of the database or live
• The underlying data model of the type of data stored
• Maximum number of tables/rows/attributes
• Databases size the tool can comfortably handle, system limitations
• Attributes types the tool can handle
• Query language, refers to the interface between the analyst, the knowledge and the data.

There were other authors who worked on the problem of evaluating data mining tool are such as King and Elder (1998), Abbott and colleagues (Abbott et al, 1998) and Collier and colleagues (Collier et al 1999). Like Goebel and Gruenwald (1999), Collier’ approach for evaluating data mining tools by creating an evaluation framework. The methodology was to use preselected criteria and use a scoring framework so that tools could be easily compared. The selected criteria determined from working on consultations were performance, functionality, usability and support. The performance criteria centred on the ability to handle a variety of data sources in an efficient manner. This criterion was sub divided with examinable questions showing how these criteria could be applied, which is shown in table 6-2.
Table 6-2  Computational performance criteria (Collier et al 1999)

The functionality criterion included the inclusion of a variety of capabilities, techniques and methodologies for data mining. In the same way as the performance criteria the functionality criteria was divided into sub categories and questions which could be measureable as in table 6-3.

Table 6-3  Functional criteria (Collier et al 1999)

The usability criteria, was meant to measure the accommodation of different levels and types of users without loss of functionality or usefulness. A good tool should guide the user toward proper data mining instead of data dredging aimlessly searching for an answer. As knowledge discovery is a highly interactive process a good tool will provide meaning full diagnostics to help debug problems, improve output and easily allow adjustments to models, see table 6-4.
Lastly the support of additional activities, such as being able to perform a variety of tasks such as data cleaning, manipulation, transformation, visualisation and other tasks which support data mining was used as criteria. The corresponding sub criteria can be seen in figure table 6-5.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Cleaning</td>
<td>How well does the tool allow the user to modify spurious values in the data set or perform other data cleaning operations?</td>
</tr>
<tr>
<td>Value Substitution</td>
<td>Does the tool allow global substitution of one data value with another (e.g., replacing ‘M’ or ‘F’ with 1 or 0 for unisex)?</td>
</tr>
<tr>
<td>Data Filtering</td>
<td>Does the tool allow the selection of subsets of the data based on user-defined selection criteria?</td>
</tr>
<tr>
<td>Binning</td>
<td>Does the tool allow the binning of continuous data to improve modeling efficiency? Does the tool require continuous data to be binned or is this decision left to user discretion?</td>
</tr>
<tr>
<td>Deriving Attributes</td>
<td>Does the tool allow the creation of derived attributes based on the inherent attributes? Is there a wide variety of methods available for deriving attributes (e.g., statistical functions, mathematical functions, boolean functions, etc.)?</td>
</tr>
<tr>
<td>Randomization</td>
<td>Does the tool allow randomization of data prior to model building? How effective is the randomization? How efficient is the randomization?</td>
</tr>
<tr>
<td>Record Deletion</td>
<td>Does the tool allow the deletion of entire records that may be incomplete or may bias the modeling results in some way? Does the tool allow the deletion of records from the population? If so, does the tool allow these records to be easily reintroduced later if necessary?</td>
</tr>
<tr>
<td>Handling Blanks</td>
<td>Does the tool handle blanks well? Does the tool allow blanks to be substituted with a variety of derived values (e.g., mean, median, etc.)? Does the tool allow blanks to be substituted with a user-defined value? If so, can this be done globally as well as value by value?</td>
</tr>
<tr>
<td>Metadata Manipulation</td>
<td>Does the tool present the user with data descriptions, types, categorical codes, formulae for deriving attributes, etc. If so, does the tool allow the user to manipulate this metadata?</td>
</tr>
<tr>
<td>Result Feedback</td>
<td>Does the tool allow the results from a mining analysis to be fed back into another analysis for further model building?</td>
</tr>
</tbody>
</table>

Table 6-5 Additional Task support criteria (Collier et al 1999)

The limits to using these criteria are that as visual data mining tools developed these methodologies have become less relevant. They do not take into account the visualisation benefits and interaction potential of tools which is why a review of visualisation evaluation material must also be done.
6.6 Visualisation Evaluation

Visualisations and visual interfaces are important parts of any visual data mining tool meaning it is necessary to investigate the quality of visualisations Keim (2004). Much of the criteria set forth by Tufte (1997) in chapter four is applicable to this area. However there has also been work done by Redpath and Srinivasan (2003) on developing more visualisation for evaluation. Their conclusions are that the important criteria identified include:

- perceptually satisfying presentation, where a judgement is made on how clearly features of the dataset are revealed based on how quickly they are learnt and understood.
- intuitiveness of the technique, of how well what is viewed is interpreted.
- ability to manipulate the display dynamically
- ease of use, of how efficiently the data can be displayed

To this point the evaluation criteria has centred on knowledge management, ISO standards, usability, data mining and visualisation. These do not cover the evaluation of visual data mining tools entirely which is why the next section investigates previous evaluation work in the area of visual data mining.

6.7 Evaluating Visual Data Mining Tools

The usability criteria of Nielsen (1994) and Dix’s (1993) as mentioned in 6.4 was used as the basis for the development of criteria to evaluate visual data mining tools. This work was done by at the beginning by Marghescu(2003). According to this initial view of visual data mining evaluation the major factors were preparing the data for usage, showing the data and using the data are the principle attributes. These deciding factors were then split into sub criteria to use for measurable feedback, see figure 6-36.
When defining the usability for visual data mining Marghescu and Rajanen and Back (2004) also point out the importance of user categorisation when evaluating visual data mining tools as different users will have different experience levels and therefore different perspectives on the system. There work identified the following user types:

- Data domain specialists, who according to the persons knowledge about the data, can select the best attributes or request the support of an automatic tool for attribute selection

- Data analysis specialist, who can be a statistician or a machine learning expert. The statistician expert can adequately use filter approach and determine the appropriate parameters for the initialisation of attribute selection algorithm.
The machine learning expert can perfectly initialise supervised classification algorithms used by the wrapper approach.

Marghescu, Rajanen and Back’s (2004) work was preceded by Marghescu’s (2003) a year earlier when she split visual data mining tool evaluation into the quality of the visualisation, interaction and information. Marghescu and Rajanen and Back (2004) define this quality to mean the proper representation of the data.

To develop the visualisation criteria they based some of the work on Shneiderman’s (1996) view of the data needing to have an overview of the data, details on demand and a filtering ability relation to the visualisation. Interesting Marghescu’s (2003) also identified the necessity for evaluating the analysis, reporting and knowledge transfer abilities but which is not obviously incorporated in the author’s selection criteria.

These criteria seemed to have been refined in Marghescu, Rajanen and Back’s (2004) in which they again centred on user testing to evaluate the quality of visual data mining tools. They stated that to accomplish certain goals and tasks the user employs the domain knowledge with the data available in databases. With the data mining systems acting as access to the database and visualisations representing an interface to the data stored.

Using figure 6-37 they explain that with a specific goal in mind users examine visualisation, interact with them and get some information. Therefore the user satisfaction and success of the data mining process depend on good visualisation and interaction with the information.

![Figure 6-37](image)

*Figure 6-37* The relationship between visualisation, interaction and information

Marghescu, Rajanen and Back’s (2004)
In this case a good visualisation is referred to as one which properly represents the data of interest, has adequate and practical initial setting and whose graphics conveys structures and content of the data. Accordingly identify Marghescu, Rajanen and Back’s (2004) identify that there needs to be a variety of exploration tasks to facilitate access to desired information and that visualisations make the user to think about the data and allow the results to be transferred to other applications.

These desirable attributes are made into the criteria for evaluating visual data mining tools which is referred to as the quality of use. The quality of use which is defined by Marghescu, Rajanen and Back’s (2004) as being the totality of features and characteristics of the tool that reflect on its ability to satisfy users needs. It therefore measures satisfaction with the tool and the tools features. The main features of the system that influence user behaviour and attribute are as mentioned above the visualisation of the data, user-system interaction and information obtained. The criteria are explained in more detail in the associated sub sections of quality of visualisation 6.6, quality of interaction 6.7 and quality of information in 6.8.

6.8 Quality of Visualisation

This criterion is concerned with evaluating the capability of the visualisation system to transform the input data and make it accessible to the user. The issues to visualisation are divided further into the following sub criteria:

- Initial settings, referring to the requirements on the input data format, degree of abstraction and the setting of the parameters for visualisation

- Data displayed regards the possibility to visual the data structure, data variation, data content and data comparison. The description, tabulation and decoration are also seen as important to evaluation

- Exploration tasks include the five visual tasks identified by Shneiderman(1996), such as overview, details of data, filter and details on demand
• Reporting functions, which allow the user to transfer results outside the application for various purposes. The concern being how satisfied is the user with the benefits from the visualisation.

### 6.9 Quality of Interaction

The second part of Marghescu, Rajanen and Back’s (2004) criteria towards evaluating visual data mining tools is the evaluation of the interaction with the system. They define the sub criteria for interaction as the achievement of an efficient, accurate, learnable and easy to use system. Similarly to the visualisation the criteria were expand into more practical measurements which could be used later in questionnaires. This involved breaking the interaction criteria into the following:

- **Ease of use**, standing for the characteristic of the system to be easy to control by the user and to provide the user with freedom of action (control and flexibility)
- **Learn ability**, affecting how easy and fast the user feels that they master the system to perform desired tasks
- **Accuracy**, reflecting the frequency and severity of system errors or failures
- **Efficiency** measures the degree to which users feel that the software helps them in their work
- **Supportability** regarding the users access to documentation and support when needed

### 6.10 Quality of Information

The third and final part of Marghescu, Rajanen and Back’s (2004) criteria evaluates the user satisfaction with the output of information provided by the system. They define the sub criteria for measuring information in terms of the richness of information, accuracy of the information presented, clarity and novelty. These criteria are explained in more detail in the following:
• Richness of information stands for completeness, usefulness and interestingness. It must correspond to users expectations needs and expectations

• Accuracy of the information regarding how precise, correct and consistent is is is with user knowledge

• Clarity of information, referring to when information is presented in a clear and understandable way, allowing users to interpret the data and make inferences

• Novelty of information reflecting the characteristic of it being new and up to date

Bajdjo and Poulet (2005) points out the other work in area of desirable criteria of visual data mining tools. They discuss Wong (1999) suggested qualities which for a good visual data mining system such as not imposing knowledge on its users but instead guiding them through the mining process allowing them to draw conclusions. Bajdjo and Poulet (2005) also cite Wong’s (1999) other desirable aspects such as providing:

• estimated error or accuracy of the projected information for each step of the mining process. This allows the user to consider this factor when analysing visualisations.
• being adaptable to a variety of systems and environments to reduce customisation effort, provide assured performance, and improve system portability.
• wide availability as the search for deeper insights of existing knowledge cannot be planned allowing an a company wide scale knowledge to be discovered and shared.
• security to protect the data, newly discovered knowledge and the users identify for social reasons.
However Wong states that these aspects stem from the principles of simplicity, user autonomy, reliability, reusability, availability and security Wong (1999). Simplicity in this case is usability, simple to learn with intuitive and friendly input mechanisms as well as instinctive and easy to interpret output knowledge. Wong (1999) also uses simplicity to refer to the way to retrieve or recall a customised data structure fast, reliable searches and the difficulty to execute a task in a minimum number of steps to achieve a result.

6.11 Desirable Attributes for Visual Data Mining Tools

Marghescu, Rajanen and Back’s (2004) work was a good first step in identifying criteria for visual data mining based. However there are other notable works on visual data mining which present alternative criteria to measuring visual data mining tools. Schulz, Nocke and Schumann (2006) stand out as presenting potential criteria which they present as desirable attributes for visual data mining tools. There criteria include the following:

- generality
  - Adaptability to different application backgrounds such as financial, social or academic.
  - Scalability to various users with various background knowledge.
- flexibility
  - Flexible control mechanism to select, connect and parameterise measures, mining algorithms and visualisation techniques such as using scripts or interactively using menus
- Visual queries, with a direct visual feedback
- Support to derive additional data, gain a deeper insight into data features
- Usability
  - Data abstraction to get easy access to different kinds of structures data sources independent of the internal and external storage format.
Acceptable reply times of calculations, where approximation techniques are considered in case of unfavourable runtime complexities or a lower upper bound given by a user.

Intuitive means of interacting with even complex mining methods.

Efficiency to handle large datasets and avoid screen storage space and temporal bottlenecks.

Screen efficiency to effectively apply the whole screen space displaying large structures.

Task orientation, determining if the user can fulfil all tasks to get the exploration target with correct focus and context provided, overview and detail abilities, brushing, history, sorting and filtering and zooming possible.

This diverse attributes along with the criteria and work from other fields make it extremely difficult to decide on the best visual data mining tool. Even Marghescu, Rajanen and Back’s (2004) model for visual data mining is too brief and does appear to consider all potential important factors in choosing a tool. Even with the correct criteria selected the next important stumbling block is deciding on the best way to apply the test criteria so that it is measured. To do this the testing approaches available must be understood.

6.12 Testing Approaches

With these criteria set as background to the testing the methodology used by authors evaluating these tools was investigated. Among the approaches discovered included using heuristic evaluation and cognitive walkthrough.

Heuristic evaluation is the most common informal method to evaluate usability. It involves having usability specialists to judge whether each dialogue or interactive element follows usability principles. The original approach is for each evaluator to inspect the interface alone. Only after the evaluation are the evaluators allowed to communicate and aggregate their findings. These restrictions are important to ensure independent unbiased evaluations (Holzinger 2005). To perform this type of testing
three to five expert evaluators are necessary, although less experienced people can perform heuristic evaluation but may lead to finished results (Mazza and Berrè 2007).

An alternative to the heuristic evaluation is to use a cognitive walkthrough, where a task oriented methodology is used with an analyst to explore the systems functionalities, Mazza and Berrè’s (2007). The cognitive walkthrough simulates step by step user behaviour for a given task. This method also emphasises cognitive issues such as learn ability by analysing the mental process required by the users. According to Holzinger (2005) this cognitive testing is the most indispensible. As it provides direct information about the system, the problems the encounter and the interfaces they use. The most common methods to support these testing methods are to get user to use thinking aloud techniques while performing tasks, field observations or through questionnaires.

The thinking aloud technique verbalises the users thoughts to enable the observer to see how the system is understood, making it easier to identify user misconceptions. This method is very useful as it provides a way of revealing why users do something, provides an approximation of how individuals use the system in practice, providing a excellent source of data. This technique comes with the disadvantage as being perceived as unnatural and taxing by the users, non analytical learners feel inhibited and can be time consuming since each user must be briefed as part of the preparation. In an ideal scenario field observation would be used to visit and observe the normal interaction between the tool and the user which ensures the information gained would be relevant. Another approach for getting feedback is to use questionnaires to study user’s interaction with their preferred features. This approach does not study the actual user interface only the people’s opinions which may differ from their experience, Holzinger (2005).

Yet another approach from Mazza and Berrè’s (2007) is to use quantitative studies consisting of an analysis of determinate hypothesis tested through direct measurements. The hypothesis could centre on the users performance in relation to a specific task or the number of trials required performing a specific task. Implementing this requires the definition of one or more variables related to the hypothesis and a metric associated with each of them. The actual evaluation is done by means of
controlled experiments. These consist of asking the user testers to run a task, performing some measurements using observation methods and completing the study with questionnaires or interviews.

The qualitative research involves the analysis of qualitative data which may be obtained through questionnaires interviews and observations of users using the system in order to understand and explain social events. This data is complementary to quantitative methods used in experimental studies for the ability it provides to analyse events from the point of view of the participants, which is largely lost when textual or analytical data are quantified. Interviews are another useful form of qualitative research that can be conducted with users, asking specific questions to elicit information about user’s impressions and general comments. Focus groups are another form of qualitative research that involve group interviews of individuals selected and assembled by researchers, who discuss and comment on, based on experience, the topic that is subject to research.

Another approach suggested by Mazza and Berrè’s (2007) is to use empirical evaluation methods involving real users in the study, although this usually performed on implemented system. However finding these users outside of the firm can be difficult and may not be practical. The alternative is to use non experts who are familiar with data mining but not experts at understanding algorithms used in data mining (Abbott et al, 1998). If taking this approach it is noted that the tools would need to use language a novice would understand and provide guidance for an expert user to be well received.

The last approach (Abbott et al, 1998) to apply the evaluation is to use several expert evaluators who are asked to rate the apparent strength of several tools in each category as judged from marketing material, reviews and experience. The scores are averaged for each category, weighted and summed to create a single score for each product. The top tools are then used in the second stage of the selection process and then rated based on the additional characteristics of experience in the domain, quality of technical support, the ability to export models as source code or ASCII text and tools were demonstrated by expert users, or other similar criteria. After this stage extensive hands on evaluation was used and the overall usability was then determined based on scoring.
the data. The results of performing this testing can then be done by using weighs in a numbers of ways such using a scorecard approach (Collier et al 1999) or using scales such as the six point one used by King and Elder (1998).

6.13 Conclusion

The chapter examined other bodies of work surrounding evaluation of knowledge management tools. In particular the criteria and approaches applied by other authors when evaluating tools. The evaluation criteria ranged from criteria which apply to knowledge management tools to data mining and visual data mining tools.

The other work in evaluating visual data mining tools from Marghescu, Rajanen and Back’s (2004) and Badjio and Poulet (2005) were reviewed. Other work evaluation work in related fields was also reviewed to ensure that all criteria and methodology for evaluating these tools were considered. Among them was work done in knowledge management tool measurements, ISO software evaluation standards, usability, data mining tool and visualisation criteria. The author also noted other work for desirable attributes for visual data mining tools which could be applied. The research also looked at how the criteria selected could most effectively be applied and other author’s perspectives on these techniques.

With an understanding gained a solid foundation of how to produce a unified view of evaluating visual data mining tools can be developed. While being able to justify criteria decisions, proven testing techniques and avoiding some of the same pitfalls. The next chapter will look at determining the best criteria for a framework in visual data mining and provide an approach to applying it.
7 VISUAL DATA MINING EVALUATION FRAMEWORK

7.1 Introduction

This chapter presents a framework designed to guide the evaluation of visual data mining tools. The framework is first shown as at a series of high level criteria and then shown in a more detail which can be applied. The completed framework can then be used by stakeholders who can then use the results to evaluate the tools. Prior to discussing how the methodology is created several benchmark issues need to be considered to avoid giving a biased result.

However non experts can be appropriate at times depending on who is available to participate. The advantages of using non experts include the application of recognised and accepted principles; intuitiveness; effective identification of major and minor and problems, ability or allow for unknown user needs, problem identification. The disadvantages include separation from actual end users

7.2 Framework Creation

The approach taken to developing the evaluation framework was to develop criteria, then questions, methodologies to apply the framework and finally the approach to evaluate the results. To develop the initial criteria research was done in chapter 6 in evaluating practices in related fields and then more specifically in visual data mining. This work forms a pool of criteria from which to develop a unified framework for evaluation. This is discussed in more detail in 7.3 along with the questions which can be taken from the sub criteria. The initial criteria selected will be refined later on based on feedback from user testing.

The next major part to this chapter is deals with providing a methodology to apply the select criteria for testing in 7.8. This will include who the users will be where, there experience, how they will be part of the testing, the stages which will be used, how the feedback will be recorded and will lead on to the final part of the framework.
The final part of the framework is how to best analyse the results and make conclusions on the tool in 7.10. The feedback taken from applying the framework will provide insights into refinements needed which can be documented and the changes taken on board.

The framework criteria is made up from previous work done in the field by other authors, user feedback from applying the criteria, interviews with expert data mining tool users and knowledge management expert Scott Brown working on the team that is applying knowledge management tools to Sun Microsystems over 20,000 staff. These different influences were the deciding factor in the selection of the criteria and approaches.

### 7.3 Criteria Selection

During the research into previous and related work it was noticed that there was overlaps on some of the criteria applied, along with criteria which was left out of Marghescu and Rajanen and Back (2004) but which could be extremely important factors. To ensure no criteria was missed the selection process was divided into knowledge management, software quality, usability and visualisation quality.

These criteria were further divided up into further sub criteria to develop quantitative and qualitative questions which can be used to compare tools. These criteria and sub criteria can be seen in figure 7-38. The knowledge management criteria were chosen as the core of the evaluation is to measure the effectiveness of a visual data mining tool as a tool for knowledge management. Therefore it is practical to apply several knowledge management criteria in selecting the best tool. The sub-criteria for the knowledge management criteria are detailed in more depth in section 7.4.

The software quality criteria are based on the ISO software standards along with work done by Marghescu and Rajanen and Back (2004) and other work from related areas. This criteria is concerned with measuring the effectiveness, functionality and
effectiveness of the tool. This criteria and sub-criteria are detailed in depth in section 7.5.

The usability criteria which is so important in almost all of the research completed looks the interaction and user satisfaction with using the tool. This criteria and sub-criteria are detailed in depth in 7.6. Finally the visualisation quality criteria is based on visualisation principle along with Shneidermans (1996) principles on overview and details on demand along with Tufte(1986) visualisation principles. The criterion are detailed in depth in section 7.7.

Figure 7-38  Showing the main framework and characteristics (Source author)

The criteria in figure 20 will now be detailed in depth in the following section starting with knowledge management criteria.
7.4 Knowledge management criteria

The knowledge management criterion was chosen to judge the tool with specific focus on knowledge transfer traits. This ensures that not only is the tool’s quality measured but also the quality of the tool at transferring knowledge to the user. The objective of the sub-criteria is to measure the user’s success of gaining insight with the tool and how well it facilitates knowledge sharing. This criterion can also be used to identify the processes and people which may need attention to ensure that the revealed knowledge is used effectively.

It is intended that the sub-criteria will aid in identifying the actionable knowledge from the data. This knowledge is evaluated in terms of knowledge gained or created, knowledge transfer ability to be able to transfer this knowledge to others. These measures look at measuring the tool’s ability to help to innovate and build upon existing knowledge. The sub-criteria include the following:

- knowledge gained, the insights or new knowledge gained from the tool
- knowledge transfer, the user ability to make inferences and extend knowledge and provide a format to allow it to be represented to others
- knowledge acquisition, measuring how difficult it is to interpret the data

<table>
<thead>
<tr>
<th>Investigates</th>
<th>Evaluation Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge gained</td>
<td>Is the knowledge gained actionable? How useful is the information gained? Is the knowledge reliable? Is it complete? Is it interesting? What was the saving from campaigns run? Does it meet management’s expectations or set goals?</td>
</tr>
<tr>
<td>Knowledge transfer</td>
<td>Is the knowledge gained easily presentable to users? Is there functionality for reporting? Is it difficult for users to understand the presented results? How is the discovered knowledge spread? Is the knowledge available when needed? Could the insight be easily understood by others</td>
</tr>
<tr>
<td>Knowledge acquisition</td>
<td>Is the knowledge understandable? Do users feel there is too much information on the screen? Does the visualisation have more text than graphics? What is the density of data graphics in the representations?</td>
</tr>
</tbody>
</table>

Table 7-6: Knowledge management questions which will be asked to evaluate the tool
These criteria were used as the basis of developing questions to measure each attribute see table 7-6. These criteria do not measure functionality or time how long these functions take which is why software quality criteria were also used.

7.5 Software quality criteria

The visual data mining tool is mainly evaluated by firms based on its ability for providing the functionality to meet a job. By using software quality as a measure of its cost, features and response times the important it can users several useful considerations in choosing a visual data mining tool. The four important sub criteria identified are:

- Functionality, ensuring the tools serves all the firms requirements, such as providing real time data
- Efficiency (timing), the time taking to complete a task or series of tasks
- Errors, measuring the reliability and confidence in the tools
- Software cost, the total overall cost

These criteria evaluation questions were developed which are shown in table 7-7.
<table>
<thead>
<tr>
<th>Investigates</th>
<th>Evaluation Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Functionality</strong></td>
<td>Is the tool suitable for the organisation’s data? Does it require expert users to get results? Does it have undo and redo actions? How does the tool’s features compare against other tools? Can the tool scale to large datasets and what is the limitation on the datasets size? Can the system work with other tools and allow data to be exported in different formats? Is the tool limited to specific platforms? Is it a server-client model for using the tool? Does it provide data cleaning, value substitution?</td>
</tr>
<tr>
<td><strong>Efficiency (Timing)</strong></td>
<td>What is the response time to query’s on large and small data? Is there a noticeable deterioration on timing performance on large datasets? Does the tool allow the user to change attributes to give less accuracy but in a faster time? Does the tool offer different levels for accessing and processing the data?</td>
</tr>
<tr>
<td><strong>Errors</strong></td>
<td>Does the tool run consistently without crashing? Does the tool fail to execute commands? Does the tool keep the user informed on the error rate of each task? Are the errors meaningful? Does the errors help the user to debug problems</td>
</tr>
<tr>
<td><strong>Software Cost</strong></td>
<td>What is the cost for the hardware and software? How much is training, support, yearly license, consultant costs?</td>
</tr>
</tbody>
</table>

Table 7-7 shows the software quality questions which will be asked to evaluate the tool

### 7.6 Usability

In almost every evaluation methodology usability is noted as a major factor in determining how effective a tool is at representing data and results. This also makes it critically important for in ensuring users concentrate applying their knowledge to the data. The usability can be broken down into the following sub criteria:

- effectiveness, measuring how effective the tool is at providing insights
- tasks complete, investigating how long it takes to complete a task
- user experience, identifying what the user thought of the tool
- accessibility, looking at if the tool depends on location
- learnability, measuring how well the interface, data and results are interpreted
- attractiveness, measuring how satisfied users are at using the tool

These criteria are shown with related questions in table 7-8.
Table 7-8 shows the usability questions which will be asked to evaluate the tool.

<table>
<thead>
<tr>
<th>Investigates</th>
<th>Evaluation Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effectiveness</td>
<td>Ability to answer business questions with tool? How many steps required completing a task? What is the time needed to create a good model? Are there large quantities of data pre-processed to obtain structures in the data? How effective is the tool at meeting goals? What are the numbers of tasks complete? How many are achieved?</td>
</tr>
<tr>
<td>User experience</td>
<td>What were the good characteristics recalled by users? What were the bad characteristics? How many times did the user express frustration? How would you rate the tools functions? How would you rate the tool overall?</td>
</tr>
<tr>
<td>Accessibility</td>
<td>How easy is it to use the tool? Can the tool be accessed from different locations?</td>
</tr>
<tr>
<td>Learn ability</td>
<td>Does the tool have an intuitive feel? How long does a person need to use the tool before being able to complete a task? Is much help needed to achieve this task? Is it difficult to learn new features while using the tool? Does this tool require a lot of training before seeing good results? What is the user satisfaction with the tool? Does the tool give adequate feedback? Does the user need to consult the help a lot?</td>
</tr>
<tr>
<td>Understandability</td>
<td>Is it easy to navigate? Can the results be easily understood by users?</td>
</tr>
<tr>
<td>Attractiveness</td>
<td>Is the data graphics simple and able to handle the complexity of the data? Do users like using the tool?</td>
</tr>
</tbody>
</table>

7.7 Visualisation Quality

The visualisation criterion was chosen as the tool is designed to be visual and this is also a major factor to evaluating knowledge management tools and in determining if knowledge is gained. This criterion was spilt into the following criteria:

- giving good overview, allows the user to get useful overviews of the data
- details on demand, allows the user to manipulate a visualisation
- properly represents data, accurate represents the data
- distortions, ensure that the data appears as the as it is in the database
- exploration ability, measuring the exploration ability
- layout, measuring the layout of the data
These sub criteria were turned into measurable evaluation questions which are shown in table 7-9 that describe the most important qualities.

<table>
<thead>
<tr>
<th>Investigates</th>
<th>Evaluation Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Giving good overview</td>
<td>Ease of expressing and integrating domain knowledge? Is there good flexibility in the visualisation?</td>
</tr>
<tr>
<td>Details on demand</td>
<td>Does the tool provide zoom in, zoom out and overview functionality? Is it possible to go into more detail directly from an overview visualisation?</td>
</tr>
<tr>
<td>Properly represents data</td>
<td>Does the representation describe the data accurately? Is there a capability of dealing with uncertain or incorrect dirty data? Is it easy to integrate the results with other software applications?</td>
</tr>
<tr>
<td>Distortions</td>
<td>Are there any noticeable distortions in representing the data visualisation?</td>
</tr>
<tr>
<td>Exploration ability</td>
<td>Is the tool properly developed to allow the user to explore the data and model freely? Does the tool have easy to understand parameters? Does the tool have easy to use parameters?</td>
</tr>
<tr>
<td>Layout</td>
<td>Is the decoration of the visualisation and description of the data suitable? Does it take away from the data or cause confusion? Does the tool attempt to show too much data or too little to understand the data? Is the labelling of the data with words and explanations appropriate? Are there any problem caused by the informal structure and information content?</td>
</tr>
</tbody>
</table>

Table 7-9 shows the visualisation quality questions which will be asked to evaluate the tool.

### 7.8 Applying the Methodology

The first step in determining which criteria and weightings to apply is to look at the firms preliminary requirements. Depending on the area, the knowledge of the users, the budget for the project and functionality required the deciding factor in selecting a data mining tool can change massively.

### 7.9 The testing phases

This section looks at how to apply the correct criteria for the firm, considering the users, how the testing will be performed and how to analyse the feedback. Firstly this section investigates which criteria to select. Using all the criteria questions generated is
impractical meaning tradeoffs need to be made to select the best criteria depending on the situation. Weighting can also be applied to the criteria depending how important each is to the firm. Once a subset of the criteria has been selected with related questions selected the next

As selecting a tool is difficult and there may be many potential tools on the market it can be necessary to reduce the tools to a manageable amount for testing. For this reason the suggested application of the framework is to use a three phase approach where in phase one most of the functionality and several software quality questions can be asked without user testing. The feedback can then use used to determine the final visual data mining candidates for user testing.

The second and third phase involve user testing where the user interacts and both uses and observes the tool in use and gives feedback on the interaction and visualisation abilities and knowledge creation abilities. Initially the plan included having separate user and observer roles and evaluating them from both perspectives. However during the course of applying the framework it was found that the best procedure was to give an introduction, demonstration and interaction with the tool with both user and observer both involved. The drawback to this testing is that the observer must become or be familiar with the tools to set up the testing with those would be involved in using the tool. It is considered that the users being tested will have some understanding of data mining. The second phase concentrates on using a questionnaire along with think out loud feedback to the observer. This feedback is recorded and then the user is then interviewed for further feedback after finishing using the tools to be tested so that comparisons between the tools can be made. These phases are shown in figure 7-39.

![Frame work implementation phases](Source: Author)
The following sections will discuss these three phases in more detail starting with phase 1: the functional tests.

7.10 Phase 1: Functional Tests

The functional tests measures how well the established tool is at meeting the functional requirements of the organisation. This includes features such as importing data from different sources, providing the range of standard algorithms to explore the data and giving a range of presentation facilities to enable decision makers to use the knowledge. These functional tests were put into an example functional test question sheet that can be found in appendix A for ease of implementation.

This phase works well as it can be done without the user to evaluate how well the tool suits the organisation's needs and how well it has performed at sharing knowledge. This approach also means that knowledge practitioners evaluating several tools for selection can use this step to reduce the range of available visual data mining tools to a small selection. The criteria questions selected for this phase are based on the criteria and sub criteria in table 7-39, which can be done without a user present. The knowledge management questions which need to be asked include the following:

- Is the tool designed specifically for the firm's domain?
- Is there good reporting functionality?
- Is there a good community available for collaboration on learning or for problems?

The functionality questions which will need to be asked include the following:

- Does the tool support methodologies such as CRISP-DM and Semma?
- Which OS platforms are supported?
- What hardware is required?
- What support and training is available?
- What are the costs involved?
• Does the tool depend on other software?
• Is the tool multilingual?
• Which algorithms are supported?
• Are all necessary functions available?
• Does the tool have security?
• Does the tool have a backup ability?
• Does the tool have a limit on the size of the data it can handle?
• Can the tool support many data source type which is available?
• Can the data export to other applications?
• What is the reporting ability supported?
• Does it have command line support, logging?

The visualisation questions which will need to be asked include the following:

• Did the tool provide a selection of overview options?
• Does the tool provide details on demand (zoom in/zoom out capabilities)?
• Is there a capability to deal with uncertain data visualisation?

The usability questions which will need to be asked include the following:

• Does the tool support a client server model?
• Does it support multiple users, if so how many
• Does the tool provide good initial configuration ability?

Once these questions are answered they can be measured using section 7.14 evaluating phase 1 functional results. Like the criteria a sub section of phase 1 questions can be selected when measuring comparing a lot of tools.

7.11 Phase 2: User Testing

The user testing stage involves getting users to interact with the tool and its results in order to evaluate the visualisation and interaction qualities of a tool with user feedback. This involves getting either experts or non experts at using the tools. The non experts can have knowledge about the data but have limited experience with data
mining and visual data mining tools. The advantage of this type of user is that they are unbiased towards any particular visual data mining tool. This conversely means that their ability to work with the tool, understand and get the most out of the tool is left to brief user testing.

The alternative is to use expert users who have ample experience in reporting and analysis and some data mining skills. With the usual downside being that typically as they have experience they are in high demand meaning they will generally not be available for long to provide insight into the tool. Expert users who have a background with other tools are often biased to those tools as they already understand them. However if the expert user is the user who will work mostly with the tool then the experts choice of tool should be given considerable weight. As obviously an expert user using a tool can start producing quality results immediately if given the same tool instead of having a slow learning curve.

In either case the best testers for the tools are those who will use the tool on a daily basis. This will ensure the new knowledge management tool can receive support from its users as they are involved, create champions to encourage the use of the tool, learning and collaboration. This approach is also useful as the user can immediately bring up concerns they may have about the tool, can tap into issues they have had in the past or enable the observer to identify necessary training for the user if a tool is selected,

The limitation to both approaches is that it requires the observer to be an expert in the tools in order to identify problems or mistakes. In the case of non experts they will need to be given an introduction and lead through the tool use process for each task with the user. If the knowledge management practitioner does not have experience of these tools a consultant will need to be brought on board with experience of one or more tools being tested. This consultant may bring obvious personnel bias towards a favourite tool but is preferable to inviting vendors to demonstrate and test users who have an agenda to push their own tool. Similarly proceeding without any observers who are experienced in data mining tools will lead to incorrect assumptions being made.
Yet using think out loud methods and questionnaires during the user visual data mining testing may still leave out important information which is not asked directly. This is why user testing should be followed up with a brief interview to get further thoughts on the user feedback.

**7.12 Phase 3: User Interviews**

The third interview phase was introduced to gather information about the user’s reactions to the tool by talking to them directly. The phase is used to gather more information than a questionnaire can provide and get more subjective reactions, opinions and insights into how people reason about the tasks given. For this interview stage a semi-structured approach is taken to allow questions to start the discussion while permitting the respondent the ability to provide additional information.

**7.13 Testing Approach**

Once the preliminary requirements are set and decisions on the criteria weightings and questions have been made the functional stage can be done. As mentioned the feedback is used to reduce the tool candidates to a manageable few which can be user tested. The approach for testing with users is to base the evaluation around tasks to explore the data, perform a number of tasks to provide the user with an impression of the tool and its visualisation, data mining and knowledge transfer properties. To do this a questionnaire needs to be devised based criteria for the situation selected before the first phase. The initial concept was this the user could fill in the questionnaire while tasks were being performed however during testing it was found that generally the for non expert users it was better to leave the questionnaire until all tools were demonstrated as it provided a measure of comparison between the tools.

The main parts to data mining which were identified in the data mining in chapter 3 are selecting business questions which need to be answered, understanding and preparing the data, searching for knowledge in the data based on the business question, determining the answer, deploying and practical evaluation of the solution. Using this
along with Fry (2007) work the main steps to data mining tasks can be identified. These include the following steps:

- identify business questions, to determine goals
- acquiring the data, selecting the data set to be analysed
- parsing the data, exploring the database and identifying important patterns or variables
- filtering the data, removing unnecessary variables, missing data or outliers which may skew the result
- mining the data with algorithms
- refining the algorithm to produce better results, changing the parameters or combining algorithms to get better results
- interacting and analysing results gain identify patterns and ultimately insight on the results

The early steps of identifying business questions, acquiring the data, parsing it and filtering the data usually take a long time to complete. It is therefore recommended is that most of these steps be already complete and are reviewed by the user being tested. This means the tester can quickly get an overview of these early tasks while concentrating on the use of algorithms, changing parameters and most importantly interacting with the tool and analysing the results.

**7.14 Evaluating Phase 2 User Results**

The first phase reduces the tools to those which have the functionality and features necessary. This testing however cannot identify usability problems thus it a second phase of user testing is required. With users insight the tools can be judged to evaluate how relevant is the tool for the firm, if deep or unexpected results can be found and identify training needs. The criteria questions selected for user testing with corresponding tasks will provide useful feedback to evaluate the tool on the user satisfaction. Like in phase one the process involves measuring and comparing quantitative and qualitative results between potential tools.
7.15 Dataset Selection

Ideally the dataset used will be from the actual datasets which will be used once the visual data mining tool is applied. Alternatively if not available similar datasets from this area should be used to get realistic feedback from users using the tool. By using relevant datasets which is recognisable to the tester the testing should be able to accurately measure how fully deploying one of the tools would affect the firm. Selecting a relevant dataset can be crucial in ensure the tool has the correct functionality and usability needed to explore the data and in measuring that the user can apply their knowledge to the tool effectively.

7.16 Refinements Made to the Framework

Once the first functional phase has been completed the results can be compared to other tool and limitations to the current tool investigated. After experimentation stage approaches or task be found not to be beneficial in analysing the quality of a visualisation in which case the framework will be modified to represent this change. When all phases are complete the results can be broken down into individual criteria and the sub criteria weighted based on the importance of the quality of the component to the organisation at meeting the organisations objectives.

7.17 Conclusion

This chapter explained the visual data mining framework, discussed its phases and showed how it would be implemented. The framework involves dividing the qualities of visual data mining tools into the criteria shown in figure 7-40 and then further sub dividing the criteria.
Once this was done the evaluation of the tool was split into three phases of functional testing, practical feedback and interviews in order to give enough feedback on the data to measure the tool. This will include who the users will be where, their experience, how they will be part of the testing, the stages which will be used, how the feedback will be recorded and will lead on to the final part of the framework.
8 EXPERIMENTATION AND EVALUATION

8.1 Introduction

To demonstrate how the evaluation framework can be used by knowledge management practitioners this chapter applies it to commonly used visual data mining tools. The framework is demonstrated with a broad range of user backgrounds experience in data analysis. This process also provides a method to further refine the framework.

The experiment will apply the evaluation framework to three visual data mining tools. The results and experiences can then be used to make alternations and refinements to the framework can then be feedback.

8.2 Experimentation

In order to test the visual data mining evaluation framework, datasets, users, business questions, must be selected which can provide feedback for the evaluation. The tools which will be tested then need to be specified. Once this is done the actual implementation of the visual data mining evaluation framework can be applied to determine the most suitable tool.

Evaluating the tools will involve selecting relevant criteria to measure the tool, specifying relevant questions to be used for later measurement. As outlined in the framework the three phases will then be applied to the tool by first evaluating the functionality, then user testing is done with interviews to test all of the criteria. The results will then be analysed to determine the best tool. The business problem selected for this testing is around identifying customers who are at risk of defaulting on their bank loans and potential customers who earn over fifty thousand dollars a year. To start this process the selected datasets used will be described with justification for its choice.
The selected business problem is based on a medium sized fictional financial bank who wants to use data mining as a knowledge management tool to enhance knowledge in the organisation. A financial firm implementing this was chosen as Rexer Analytics (Rexer 2008) reported that it is one of the most commonly used areas to use data mining. The business objective is to identify high risk customers who are likely to default on their loan in the future. Therefore the result could be used to decide if a credit loan should be given. The other objective is to look for potential new high value customers who earn over fifty thousand dollars in a year.

Each visual data mining tool tests were tested on a dual core laptop with three megabytes of ram running Windows XP and eighty gigabytes of hard drive space. The datasets which were used are detailed in section 8.3.

### 8.3 Datasets

To facilitate this analyse two datasets will be used to analyse the bank credit. The first dataset is taken from an anonymous German bank which gave which is available from the SAS (2008) data mining tool samples. The dataset gives details of customers with a label indicating if the user paid back there loan or not. The german.xls dataset has 20 attributes such as checking, duration, history, purpose, amount, and savings, if the customer is employed, the customers’ marital status, resident status, property, age and job. Along with these attribute is the good_bad label attribute to label good or bad loans. This value can is therefore used as the deciding factor to generate models to classifying future customers.

The second dataset is similar and is of credit given for car loans again with a label indicating if the loan went bad. This dataset is taken from the SPSS (2008) data mining tool samples. The dataset is called treecredit.xls has only six attributes including the customers age, income, number of credit cards, education level and car loans taken. An additional attribute called credit rating is also used to identify which have a good history of paying back loans.
Thirdly a dataset was taken from the UCI Machine learning repository (UCI 2008). It is called the adult data set and represents people who exceed fifty thousand dollars a year based on census data. It has fourteen attributes such as age, work class, education, marital status, occupation, relationship, gender, capital gains, capital lost, hours worked, nationality. The labelled attribute in this case is if the person over fifty thousand dollars.

8.4 User Profile

The users involved included both non expert users and expert users who have an understanding of visual data mining tool. The non experts involved included seven software quality engineers with computer science background, knowledge in usability testing, but no extensive experience in data mining or using data mining tools. The expert users involved four users who were familiar with data mining tools and had some experience using visual data mining tools. Neither group were experts in understanding statistical data which would have made the data a lot more relevant. This proved to be a problem with the experiment as without significant experience of chi square values they were unable to concentrate on understanding the statistical results which are extremely important. As more experienced users were not available the users had to be guided through the tool more to ensure these values were checked and understood.

As suggested in the framework for evaluating visual data mining, heuristic evaluation and cognitive walkthrough will be used with think aloud methods to get the users perspective. Originally the experiment would have use an observer role to evaluate the user unobstructed during the testing but this was impractical without allowing the users to spend several days learning the tool. Instead the observer gave an introduction to the data mining techniques which would be used then demonstrated and involved the user at different stages. The users were then tested on prepared datasets to minimise the time taken to evaluate each tool.
8.5 Business Questions

The business questions are based on the business objective of identifying customers who are at high risk of not repaying from the available bank credit datasets. The second question is based on the census data available of potential customers which are the high paid earners who will be the most profitable for the bank.

Generally however the business questions will be created by the firm’s management or a specifically created team for data mining. It is suggested that when using the tools the stakeholders or users who would use the tools feedback on a daily basis meet to collaborate and decide on the most important business questions. In the case of the bank firm the most important business question is deciding what can be gained from exploring the data and which model is best applied to the dataset. With these business questions the visual data mining tools which will be used must be provided.

8.6 Visual Data Mining Tool Selection

In order to evaluate the effectiveness of the framework and demonstrate how it can be implemented three visual data mining tools were selected. The selection of the tools has been based on popularity of the tool, available options and leaders in the area of databases and data mining. The most popular tools was based on the leading internet portal for knowledge discovery and data mining called Kdnuggets (2008) poll based on the most popular tool and Rexer, Gearan and Allen’s (2007),survey for the best data mining tool published by Rexer Analytics.

From this report the visual data mining tools chosen were SAS Enterprise Miner, Oracle and Rapidminer. The SAS Enterprise miner tool was selected based on it being one of the top five tools used in data mining, according to Kdnuggets (2008). The second tool selected to was Rapidminer, the number one data mining open source solution according to the same poll. The third tool was SPSS which appeared number one as the most popular data mining tool according to Rexer, Gearan and Allen’s (2007). These tools are analysed in more detail from second 8.7 to second 8.10.
8.7 **SAS Enterprise Miner Tool**

SAS (2008) is one of the largest software companies in the world which has over thirty years experience at producing data mining software. Currently SAS (2008) solutions are used at 45,000 sites in 109 countries including in 91 of the top FORTUNE Global 500 companies.

There product the SAS (2008) Enterprise miner tool has been developed to support the entire data mining process. The company’s website boasts that it can support virtually any modelling need. The tools main functions are arranged around the SEMMA methodology of Sample, Explore, Modify, Model and Assess. Under each of these categories Enterprise miner provides a number of tools to promote the process of data mining. The products customers give very positive feedback saying it offers cutting edge analytics and easy to use BI reporting in a single, integrated platform (DMreview 2008). This explains its place as being among the top five most popular data mining tool according to Kdnuggets (2008) poll. Another tool which they selected was Rapidminer.

8.8 **RapidMiner Tool**

RapidMiner (2008) is an open source tool which is one of the most widely used data mining and predictive analysis solutions worldwide. The company rapid-I controls and guides the development of the tool and ensures that Rapidminer is an enterprise capable solution. As such rapid-I provides the software in both free and commercial versions of the tool.

On its website it pushes its benefits of high performance, secure environment, low IT costs, reduced costs and risk and enterprise cable support and services. According to the Kdnuggets (2008) poll it is the most widely used open source data mining tool and the second most frequently employed software for data analysis.

According to RapidMiner (2008) the tool has thousands of users in more than 40 countries worldwide and has been downloaded over 330,000 times from 2005 to 2008. The tool provides 500 different modules for detecting patterns in data and for their 2D
and 3D visualisation. Like similar commercial tools Rapidminer support features such as importing common data formats of other data mining tools, Excel sheets, SPSS files, all common databases, and unstructured text documents like news texts, e-mail messages, web pages, web blogs, PDF documents, as well as time series and audio data.

8.9 SPSS Tool

SPSS (2008) Inc is another data mining tool developer and leading worldwide provider of predictive analytics software and solutions. The company has over 40 years experience at analysing data with over 250,000 customers worldwide with 1,200 employees in 60 countries.

According to the Kdnuggets (2008) poll they are the most popular data mining tool. They have several tools however this experiment will use SPSS 17 analytic tool. This tool is designed for survey, market research and direct marketing, planning and forecasting and quality improvement as well as reporting and ad hoc decision making.

8.10 Applying the Framework

The company background provided and the datasets, user profiles and visual data mining tools selected for evaluation the next step is selecting the criteria for evaluation. The next step is to select the criteria for testing and develop tests based on these criteria, set the tasks to evaluate these criteria, implement the three phases of functional testing, user testing and user interview and then analyse the results. The next section 8.11 discusses the selection criteria in more detail.

8.11 Criteria Selection

The criteria for evaluating the success of a tool are based on the situation and ultimately how effective each tool is at meeting business and knowledge management objectives. The businesses objectives identify were to identify high risk customers at a
risk of not repaying their loans and identifying the potential customers who will be most profitable for the bank. Using the four criteria of knowledge management, software quality, usability and visualisation quality several questions for each criteria were selected. The knowledge management criteria for later measurement selected can be seen in table 8-10.

<table>
<thead>
<tr>
<th>Investigates</th>
<th>Evaluation Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge gained</td>
<td>Is the knowledge gained actionable?</td>
</tr>
<tr>
<td></td>
<td>Is the knowledge reliable?</td>
</tr>
<tr>
<td>Knowledge transfer</td>
<td>Is the knowledge gained easily presentable to users and others?</td>
</tr>
<tr>
<td></td>
<td>Is there functionality for reporting?</td>
</tr>
<tr>
<td>Knowledge acquisition</td>
<td>Is the knowledge understandable?</td>
</tr>
<tr>
<td></td>
<td>Do users feel there is too much information on the screen?</td>
</tr>
<tr>
<td></td>
<td>Does the visualisation have more text than graphics?</td>
</tr>
</tbody>
</table>

Table 8-10 Knowledge management questions which will be asked to evaluate the tool

Several potential criteria questions were left out which were not relevant to the type of testing being used. The software quality criteria were used can be seen in table 8-11.

<table>
<thead>
<tr>
<th>Investigates</th>
<th>Evaluation Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functionality</td>
<td>Can the tool handle the organisations data?</td>
</tr>
<tr>
<td></td>
<td>Does it have undo and redo actions?</td>
</tr>
<tr>
<td></td>
<td>Can the tool scale to large datasets and what is the limitation on the datasets size?</td>
</tr>
<tr>
<td></td>
<td>Can results be exported to other applications?</td>
</tr>
<tr>
<td></td>
<td>Is the tool limited to specific platforms?</td>
</tr>
<tr>
<td></td>
<td>Is it a server-client model for using the tool?</td>
</tr>
<tr>
<td></td>
<td>Does it provide data cleaning, value substitution?</td>
</tr>
<tr>
<td>Efficiency</td>
<td>Is there a noticeable deterioration on timing performance on large datasets?</td>
</tr>
<tr>
<td>Errors</td>
<td>Does the tool fail to execute commands?</td>
</tr>
<tr>
<td></td>
<td>Are the errors meaningful?</td>
</tr>
</tbody>
</table>

Table 8-11 shows the software quality questions which will be asked to evaluate the tool

As it is not possible to get accurate quotes on the costs involved in setting up this system it is left out of being a deciding factor. The usability criteria questions which
will be used are shown in table 8-12. The visualisation criteria questions can be shown in table 8-13.

<table>
<thead>
<tr>
<th>Investigates</th>
<th>Evaluation Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effectiveness</td>
<td>How effective is the tool at meeting goals?</td>
</tr>
<tr>
<td>User experience</td>
<td>What were the good and bad characteristics recalled by users?</td>
</tr>
<tr>
<td></td>
<td>How many times did the user express frustration?</td>
</tr>
<tr>
<td></td>
<td>How would you rate the tool overall?</td>
</tr>
<tr>
<td>Accessibility</td>
<td>How easy is it to use the tool?</td>
</tr>
<tr>
<td></td>
<td>Can the tool be accessed from different locations?</td>
</tr>
<tr>
<td>Learn ability</td>
<td>Does the tool have an intuitive feel?</td>
</tr>
<tr>
<td></td>
<td>Does this tool require a lot of training before seeing good results?</td>
</tr>
<tr>
<td></td>
<td>What is the user satisfaction with the tool?</td>
</tr>
<tr>
<td>Understandability</td>
<td>Is it easy to navigate?</td>
</tr>
<tr>
<td></td>
<td>Can the results be easily understood by users?</td>
</tr>
<tr>
<td>Attractiveness</td>
<td>Is the data graphics simple and able to handle the complexity of the data?</td>
</tr>
</tbody>
</table>

Table 8-12 Shows the usability questions which will be asked to evaluate the tool

<table>
<thead>
<tr>
<th>Investigates</th>
<th>Evaluation Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Giving good overview</td>
<td>Is there good flexibility in the visualisation?</td>
</tr>
<tr>
<td>Details on demand</td>
<td>Does the tool provide zoom in, zoom out and overview functionality?</td>
</tr>
<tr>
<td>Properly represents data</td>
<td>Is there a capability of dealing with uncertain or incorrect dirty data?</td>
</tr>
<tr>
<td>Exploration ability</td>
<td>How would you rate the exploration ability the data and model freely?</td>
</tr>
<tr>
<td></td>
<td>Does the tool have easy to understand parameters?</td>
</tr>
<tr>
<td>Layout</td>
<td>Is the decoration of the visualisation and description of the data suitable?</td>
</tr>
<tr>
<td></td>
<td>Does the tool attempt to show too much data or too little to understand the data?</td>
</tr>
</tbody>
</table>

Table 8-13 Shows the visualisation quality questions which will be asked to evaluate the tool
These criteria are integrated into the functionality questionnaire in phase one and the user testing and interviews in phases two and three. With these criteria set it is possible to set the necessary tasks which will help answer the criteria questions.

8.12 Tasks

The purpose of using tasks was to guide the user to explore, interact and review results with the tool in a consistent user experience. The other intended reason was to ensure one part of the evaluation was not left out. After reading other evaluations it was decided that the best results would be gotten by keeping the tasks simple while still keeping them representative of the tools functionality as suggested by Chung, Chen and Nunamker (2005). This approach sacrificed demonstrating other functionality but previous evaluation work found that complex tasks caused the user to become confused and get an inconsistent understanding of the tool.

In this case the most important task is getting the user to understanding of the data and applies the most accurate model at classifying if customers will pay back loans. The second classification task is to determine potential customers who could provide a lot of value for the organisation. The tasks finally decided upon were to explore the data, investigate using a decision tree model on the data, investigate applying a regression model on the data, reviewing the performance of the models and evaluating the visualisations provided to understand the task. The two algorithms were selected as according to Rexer Analytics (Rexer 2008) poll the most commonly used algorithms are decision trees, regression and cluster analysis. These tasks are also representative of the knowledge creation steps and knowledge transfer of data mining was deemed most important. In order to measure these later steps the datasets imported into each tool and the other supporting tasks were done. With the same data was used to process all of the users.

One evaluation session was performed with each user. The users had been briefed about the evaluation in an introduction and would be required to evaluate the tool through a questionnaire. The initial time given to complete tasks was set as twenty minutes but actually would usually take forty minutes for the talk out loud and
questionnaire to be completed. Involving the user in several sessions as had been done in other work on evaluation was not practical. It was also decided that even with multiple sessions the user would need to spend a significantly greater amount of time to learn more about the tool.

8.13 Phase 1: Functional Tests

All the criteria from the functional testing framework are applied to the three tools except the costs involved because it not accurately be gathered without getting a quote from the product suppliers and licence, installation and training. These focus on functionality, visualisation and usability criteria asking questions about the supported methodology, OS platform, hardware required, support and training availability, dependant software. The functional questionnaire can be reviewed in appendix A.

8.14 Phase 2: User Testing

The users are given an introduction presentation lasting fifteen to twenty minutes in which the data mining concept which will be used are reviewed. The data mining techniques are also explained to ensure the user knows how the techniques works and has at least a limited user standing of the result. The user is then works with the observer to complete a number of tasks where the user comments during the process and provides feedback once the tool demonstrations are complete. The time taken to demonstrating the tools and get the feedback takes another twenty to thirty minutes and is recorded to allow the observer to guide the user through the data mining process. The user will be asked to speak during the test what the user is trying to do or thinking about doing.

The limitation with taking this approach is as the user is guided through the task they may find steps overly hard or easy as when the user hits a problem the observer helps to resolve it. Ideally time would be given to let the user troubleshoot these problems but this requires a lot more availability of the user to get trained up on the tool. The user questionnaire can be found in appendix B. This approach was taken after careful
consideration of user testing approaches other work has done (Santos et al, 2007; Carr 1999).

Only one user from my non expert group had six months experience with the Oracle visual data mining tools the rest are all quality engineers who are familiar with the concepts of data mining and usability testing. The expert group include people who are involved in data mining or reporting and analysis on a regular basis. The expert user includes users who have experience at using Weka(2008), Rapidminer(2008) and SAS(2008).

8.15 Phase 3: User Interviews

The third phase of user testing will be a followed by an interview straight after the user testing to get additional feedback from users by the tools. This approach ensures that the questionnaires limited questions do not miss any potential issues. The format of the interview will be a semi-structured approach allow questions to start the discussion and then allowing the user to elaborate. The questions will include try to get the users feelings on the tool and satisfaction. The questions for the interview can be found in appendix c. The users will also be asked separately for their comments on this approach of user testing and to discuss its strengths and weaknesses.

8.16 Experiment Revisions

Several changes had to be made to the dataset which originally would contain central statistic office data sets but after analysing them for several days it was realised that it would prove extremely difficult and time consuming to analyse and make this data usable. This concept around using this dataset would be based around the Irish governments need to learn more about determining characteristics in the economy but after investigation into the available online datasets it did not seem practical and would require a very deep understanding of statistics to get meaningful insights.
Another change was to the visual data mining tools to be tested which originally were going to make Oracle’s visual data mining tool part of the experiment. This was due to Rexer (Rexer, Gearan and Allen 2007) report indicating that it was an upcoming innovative tool. The decision to change to another tool was made as there was post install problems with getting the Oracle server it depended on working on the systems after several days of troubleshooting. The results of this testing is detailed in section 8.17.

8.17 Evaluation

This section looks at analysing the results from user testing and functional results. The functional results are detailed in section 8.18.

- Answers to questions, for some tasks
- Difficulty felt by the user as judged by him and by the observer
- Other additional observations considered relevant

The opportunity will also be given to give suggestions and make comments was given through open questions as well as during information conversation at the end of the sessions. The demographic data concerning age, gender and experience with reporting was also collected.

8.18 Phase 1: Expert users Functional Results

The functional tests feedback has been gathered from fact sheets from the tool vendors websites and from Intelligent enterprise (2008) unless otherwise stated. This feedback is based on the questionnaire on appendix A. The cost was based on estimates taken from online material and may differ based on the industry and company size.

The functionality feedback showed that RapidMiner, SAS Enterprise Miner and SPSS statistics server are very close on features and functionality as shown in figure 8-14 and figure 8-15. However there were noticeable differences, firstly on cost where the
SPSS statistics server required extra modules to be purchased for each algorithm. This makes SPSS look deceptively cheaper than other tools while seriously crippling the core algorithm functionality available. Adding basic algorithms such as decision tree and regression modules more than doubles the price and this is only for one user. This approach also encourages the company to provide less algorithm functionality to users, limiting the success and accuracy of algorithms. The most costly product is SAS Enterprise Miner at 92,600 Euro but again the only quote available was for an unlimited licence for multiple users making price comparison with the other tools difficult without getting a quotation from the vendor.

The third quotation is from RapidMiner for a medium sized firms licence for 10 people that cost 4000. This cost however only covers the basic costs and not the inevitable cost of training, support, consultancy, licence renewal and hardware cost. In a real situation these would be included along with the cost of hiring experienced data mining staff to work on the ongoing data mining project. Although the budget is be an important deciding factor in selecting a tool it is important to look at the other factors.

One of the next most important elements is the input and output formats available to each tool. Since all three datasets are available in excel format that is all that is required however the presentation formats are important as noted by those polled by Rexer Analytics (Rexer 2008). It is clear that SAS has the most visualisation functionality, with SPSS having a good selection of visualisation options.
<table>
<thead>
<tr>
<th>Criteria</th>
<th>RapidMiner</th>
<th>SAS Enterprise Miner</th>
<th>SPSS Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supported OS Platforms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-64-bit MS Windows</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>-All 32-bit MS Windows</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>-Linux/BSD/UNIX</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Cost of system</td>
<td>5,000 Euro</td>
<td>92,600 Euro</td>
<td>3496 Euro **</td>
</tr>
<tr>
<td>Input and export supported file types</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-Csv</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>-Excel</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>-Spss</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>-ODBC</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>-SAS</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>-Access</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>-dBase</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>-Text files</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>-Xml</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>-PDF</td>
<td>Yes*</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>-HTML</td>
<td>Yes*</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>-RTF/Word</td>
<td>Yes*</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Supported Algorithms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-Support Vector Machines</td>
<td>Yes</td>
<td>Yes</td>
<td>No **</td>
</tr>
<tr>
<td>-Decision Tree Learners</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>-Regression</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Data Pre-processing</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>-Filtering</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>-Sampling</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>-Handling missing values</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>-Feature selection</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>-Cross validation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 8-14 Functionality Tool Matrix Part1

* Corporate version  ** Optional module
<table>
<thead>
<tr>
<th>Criteria</th>
<th>RapidMiner</th>
<th>SAS Enterprise Miner</th>
<th>SPSS Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security</td>
<td>No *</td>
<td>Yes (on server)</td>
<td>Yes</td>
</tr>
<tr>
<td>Backup</td>
<td>No *</td>
<td>Yes**</td>
<td>No*</td>
</tr>
<tr>
<td>Client Server Support</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Multiple Users</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Designed for domain</td>
<td>No</td>
<td>No(Customisable)</td>
<td>No(Customisable)</td>
</tr>
<tr>
<td>Limited size of database</td>
<td>Limited by memory</td>
<td>Limited by memory</td>
<td>Not available</td>
</tr>
<tr>
<td>Dashboard support</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Visualisation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>-ROC Plots</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>-high dimensional visualisation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>-lift curve</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>ActiveX</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>GIF</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>JPEG</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>PNG</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 8-15    Functionality Tool Matrix Part2

* Mentioned but no details ** Options available

Another advantage both SAS and SPSS have over RapidMiner is the client server model which can be used enabling centralised management and security.

Another functionality test not included was testing if any tool was domain specific. It appears that all three tools can be customised to work for the organisation but none are designed specifically for a domain or user. However the customer base gives an indication that they do support financial processes RapidMiner having Bank of America, SAS with Banca Carige and SPSS with Lloyds TSB.

However these functional results are not enough to inform knowledge management practitioners which are the best tool. Each tool needs to be tested to identify deficiencies.
8.19 Phase 2 and 3 : Feedback and Interviews

The feedback was based around a number of questions:

8.19.1 The first impression of the tools

The expert’s first impression was that it the SAS tool seemed quite easy to use and that it has a lot of functionality but not so much that it put the user off. It came off that it was the industry standard with easy graphics and a big difference when displaying the decision tree and regression results, compared to the other two tools. Several people commented that they found it covers from very basic to complex, such as one testers comment saying:

‘that the tool is user friendly but has all of the complicated stuff in it if you need it’

Laoisa Gavey working in Accenture for the revenue commissioners with experience in reporting and data mining said SAS seemed like a very good tool for the financial area and could see how it could help with decision making.

In general RapidMiner was noted for having a “simple layout” and being “easy to use”. Abdul one of the non experts thought that it was geared towards more professional users and that from the point of power user, being able to go into xml was seen as a great benefit but was not good for a regular user. Both expert and non expert users liked the fact that RapidMiner was open source because it gave the impression that it could potentially changed for any task or opportunity.

Several people commented that the SPSS tool was poor and it got a bad negative reactions from the experts and non experts saying it was the worst of the three tools, that it was not user friendly and that it:

“Consumes time to analyse and setup”.
Another user said that SPSS looked simple to use but lacked the user friendly feel of SAS. There were also comments that it seemed to have less functionality than the other tools, was more difficult to read and interpret the results.

SPSS gave testers the impression that it was not user friendly, especially selecting the necessary variables at each stage when the other tools could automatically select the most important variables. One non expert compared the functionality between SAS in which it was possible to see who borrowed 2000 to 4000 dollars quickly in an age range, whereas SPSS did does have much of an overview of the data. However this perspective has to be weighed against these users not have experience with the SPSS tool which could explain their initial dislike for the tool. Certainly though SAS was seen as more of a commercial tool, finished and intuitive tool with RapidMiner also getting good reactions.

8.19.2 Is the data understandable from the overview in each tool

There were mixed reactions about RapidMiner’s overview ability with two saying it was possible to see the data easily with graphs and three saying that found it difficult and thought the visual explanations of the SAS and SPSS better.
RapidMiner’s use of large icons was mentioned by an expert tester as reducing the effective display of information and found the decision trees poor compared with uml tree in the SAS tool. SAS was seen as the most user friendly tool but it depended on what you want to do and noted that it is heavy on statistics like the first tool while in SPSS the decision tree produced was found to be too complicated. Although SPSS was noted as giving an easy excel type format to view the data. In order to access which was the best at giving an overview of the data each tester from the seven non experts and three experts was asked to rate each tools ability. The result is shown in figure 8-41 and clearly indicates that both experts and non experts felt that the SAS tool had the best overview ability with an average of 4 out of 5, compared with RapidMiner’s average at 2.75 and with the lowest score of SPSS’s 2.25.
Again in this the SAS enterprise Miner tool scores the best with RapidMiner being slightly behind most likely because of time needed to learn the tool. This could explain why SPSS scored as poorly as it gives a lot of customisability on each task which new users to the tool found difficult to interpret.

8.19.3 Difficulty of creating a regression model

There was difficulty shown at understanding the regressions results in SPSS while users with SAS and RapidMiner results found them easier to interpret.

“The simplest tool for regression was SAS; RapidMiner required the user to have too much knowledge to find options and too many expanding menus while SPSS was on par with SAS”

Another tester found SAS’s responsiveness very quick and liked the user was kept informed with:

“the green box around which current node is running”
In general since only log results on accuracy were presented no real features stood out for these tasks which lead to a lack of insightful feedback. With a deeper understanding of statistics the users may have commented more on the results.

8.19.4 Are the results understandable in each tool

In SAS both non experts and experts ranged on how understandable each tool was with seven testers saying SAS was understandable compared with five in RapidMiner and testers in SPSS. SAS was again highlighted as having good results and graphs and clear interface. However in comparison with other tools one person commented that graphically:

“RapidMiner was the worst for producing graphical understandable results.”

This was a drawback to the small user pool as without a large test pool it is difficult to determine if this is a considerable issue. Although it seemed that this user was used to windows type interfaces and was uncomfortable with the Java interfaces.

8.19.5 How to the decision trees compare

In was thought that in general all decision tree results looked similar in each tool although SPSS tool was pointed out for not having a usable result window to examine the decision and providing poor interaction with the decision tree. SAS was also pointed out as being the most intuitive. In RapidMiner an expert commented that it looked that the decision tree looked very pushed together and could not be presented to customers.
Are classification explanations quickly understood

The SAS tool was seen as usable following familiarisation but that it has too much functionality but for an inexperienced user it is be hard to use the all the necessary functions even with using this graphical interface. Another expert’s opinion said it was:

“Difficult to understand the data, i think again you have to know what you’re looking for and how to find things”

It was also comment that

“For the more professional user RapidMiner is sufficient but for a beginner the other tools can be picked up quickly and used”

In general most thought that SPSS was not good for explanations. This was confirmed when users were asked how to rate how difficult it was to explore the tool on a scale of one two five as shown in figure 8-42.

![Figure 8-42 Difficulty using tools overall](image)
The difficulty rating for SAS was the lowest at 2.7 with RapidMiner at 3.2 and SPSS at 3.35. Several testers commented that since the RapidMiner tool was open source it was expected that the user would know what they are doing so expected less of the software. Also testers thought that these were only there first impressions and with more time they could know more about what they are doing.

**8.20 Phase 3: User Interviews**

Several interview questions were asked to give more feedback on the tools:

8.20.1 What were the good and bad things about the tools

SPSS was seen as more simplistic in an old archaic manner by Laoisa Gavey being asked:

“to drag and drop different variables”

She thought this would make comparing models difficult. Rapidminer was seen as easier to use in general and that the decision tree made it much easy to understand the information. Due to the fact that SAS looked so advanced it made RapidMiner look to some not finished although there were three non exports who would choose RapidMiner as the best tool.

**8.21 Selection of the best tool and analysis of results.**

Clearly in user testing SAS performed the best followed by Rapidminer and finally SPSS. There are several changes which need to be made such as the testing time which was too brief to give an adequate introduction to the tools and the selection of the testers who did not have an in-depth understanding of the tools. These feelings were expressed by testers saying that more time was needed with the data before getting something out of data:

“Some of the results looked quite intimidating unless you knew what they meant”

“not too easy for uninitiated”
It was also noted that the questionnaire successfully focused the user on thinking about the tools again whereas when those being tested were asked direct questions they tended to give a very quick response.

Another consideration has to be that several in the group that thought of the tool from a quality and reporting perspective, that the users would need more of a background in statistics to consider that as an important factor. Users with a background in statistics was also considered important in order to appreciate of the accuracy and meaning of results, more experienced testers would be necessary.

Overall experimental results support the belief that visual data mining tools can help to reduce information overload and that the SAS tool was the best to do that. It also seemed that more experienced users need to be tested to get good feedback for this reason it would be beneficial to interview experienced users with a background in these tools to show potential problems with the tools which are missed in a brief review.

8.22 Conclusion

This chapter presented the experiment to test the framework and the results from this experiment from this the SAS was chosen as the best tool but realised that this work would need to be followed by interviews on experienced users to ensure that testing done at the preliminary stage does not miss anything.
9 CONCLUSION

9.1 Introduction

This chapter will summarise the work completed in earlier chapters and in the experimentation and results stage. This chapter also draws conclusions on the nature of the work and the area of visual data mining evaluation. To do this the research definition and research overview will first be discussed.

9.2 Research Definition and Research Overview

The main finding of this research was obtained through interviews experienced users, observation and feedback through evaluation and from online expert groups in data mining. It was found that an important aspect of selecting a visual data mining tool is the costs, the user experience of previous tools and the interaction and visualisation aspects. This includes the user interacting with the tool and how it is presented to others as it is one of the most important purposes of the tool. Especially with Rexer analytics (Rexer 2008) report saying that half of data mining is used in decision making. This feedback from these sources also presents good modifications suggestions to the tool.

Ideally a larger group of experience reporting testers would have been used who have been working regularly with visual data mining tools. Also this recognised that the users would be biased as interviews with experienced users showed they preferred the tool they were using and felt they would produce more effective results with that tool instead of relearning another tool. The functionality question was useful in insuring the knowledge practitioner considered other issues not directly related to the tools interface. The user evaluation was found to be very useful though in giving a far greater insight into how well the tools perform compared to just the vendors white papers.
Overall against other highly biased vendor sales team the visual data mining framework gave a balanced overview of all the features allows. The framework flexibility also allows it to be used in different environments and presents several ways in which the important criteria can be measured.

After researching this field all signs indicate the field of visual data mining will continue to grow at an even faster pace in the future, particularly in the area of knowledge management. Unlike several other knowledge management tools it results can be measured through successful marketing campaigns or financial savings and be proven to save firms money encouraging more investment into knowledge management tools and justifying the cost.

For medium to large firm’s visual data mining could potential be the biggest competitive advantage for firms over rivals as better knowledge on how to do business are revealed. This will become especially important as Irish firms other competitive advantages diminish with growing eastern European competition.

9.3 Contributions to the Body of Knowledge

This work presented an overview of the current knowledge in visual data mining and its part within business intelligence and knowledge management. It unified the diverse frameworks available on evaluating the tools and looked at related field to add to the criteria and approaches used for testing. The work then extended beyond previous work by using practical feedback from a knowledge management analyst, user feedback testing and interview, from expert data mining forums and from the experience of two users working on visual data mining tools.

The framework developed is usable by knowledge management practitioners as a tool for understanding what the potential benefits of using data mining as a knowledge management tool. It also can help justify the investment to management and providing a way to select the best visual data mining tool best suited for the firm.
9.4 Experimentation, Evaluation and Limitation

The limitation of the research was that using users who are not trained up on the tool or on data mining limited the usefulness of the feedback. As a knowledge management practitioner said the visual data mining tool is still only a tool and will only be as effective as the users are at getting results from the tools. Another issue was that the practical time to complete all tasks was between thirty to forty minutes with fifteen to twenty minutes spent giving an introduction to the tool before going through each tool. This limited testing time also restricted how much could be completed.

Another aspect is only two users who had experience in RapidMiner were interview although if practical users of the SAS and SPSS tool should have also been interviewed. This was made impractical by the lack to easy access to experienced users although for a practical implementation of the framework consultants with data mining experience should be included in the decision process to get a deeper understanding of the tools.

9.5 Future Work and Research

The future work involves modifying the framework again from the latest feedback and getting more experienced users involved in the decision making process. It would be beneficial to assess the cost of a data mining consultant and identify their role in selecting a tool. With this more user feedback would be needed and several applications of framework to company situations done to provide more confidence on in using the framework.

Another interesting area of future work is devising a general knowledge management tool framework for selecting tools and implementation considerations when applying these tools particularly how they will be integrated with existing users and business processes. This comes from the knowledge management analyst’s discussion of how they implement knowledge management tools. There is also an interesting area of looking at how information management affects knowledge management as it was inferred that information management is the necessary before knowledge management.
Finally there is a fascinating area of how users learn knowledge and interact with knowledge management tools. There is a lot of work which could be done into how best to interact with users and is very topical in knowledge management, visualisation and data mining area.

9.6 Conclusion

This chapter looks at the work done, conclusions drawn and possible future work in the area. In this dissertation it was shown that firms are facing an urgent need change to a knowledge economy. Knowledge management was shown to be the key to achieving this objective with knowledge management tools as being vital in helping with knowledge creation and innovation. In particular the use of business intelligence tools was discussed as part of the knowledge management tool set to enable knowledge creation and discovery of hidden knowledge.

To enable knowledge discovery and knowledge creation data mining was shown to be one of the potential knowledge management tools available to medium, to large organisations. Within this area visual data mining was discussed as a growly important tool in knowledge creation and its benefits were discussed over other tools. Through the investigation process of this area a need was recognised for a framework for evaluating and selecting the best visual data mining tool for knowledge management practitioners.

To do this, research on the areas of visual data mining, evaluation, visualisation, data mining and knowledge management was done to select the most effective criteria and identify how other work in the area was done. This was combined with user feedback, interviews with experienced users a knowledge management analyst to develop the final visual data mining framework. The framework was then tested by using financial three datasets from a fictional financial organisation which needed to rate if a customer could be granted a loan and finding potential customers.
The results of the framework were analysed and several issues found with the user group and with the framework which were intended to be tackled by adding interviewing experienced users about the tools to gain an insight on problems which may not be found in user testing. To do these interviews with experienced data mining tools users were done to identify important areas they looked for in the tool.

The revised and completed framework is a large step forward in helping knowledge management practitioners identify useful the most suitable data mining tool as part of a large system.
BIBLIOGRAPHY


Ahmad, K., Vrusias, B. 2004, 'Learning to Visualise High-Dimensional Data', Proceedings of the Eighth International Conference on Information Visualisation


Ankerst, M., Keim, D.A, and Kriegel, H. 1995, 'Recursive pattern technique for visualising very large amounts of data sets', proceedings of the IEEE Conference on Visualization


Barlow, T., Neville, P.  2001 ,‘Case study: visualization for decision tree analysis in data mining ’, Proceedings of the IEEE Symposium on information visualization


Bederson, B. and Shneiderman, B.  2003 , 'The Craft of Information Visualization: Readings and Reflections' Morgan Kaufman publishers

Berry, M. J. A. and Linoff, G. S. 2004,'Data mining techniques, for marketing sales and customer support, second edition', Wiley, p. 4-63

Bertin, J. 1983,'Semiology of graphics', Wisconsin press


Brachman, R., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G. and Simoudis, E. 1996,'Mining business databases',Communications of the ACM, Vol. 39 No. 11,


Brohman, M.K. 2006,'Knowledge Creation Opportunities in the Data Mining Process', Proceedings of the 39th Annual Hawaii International Conference on System Sciences, Volume 8

Burkhard, R. A. 2005a, ‘Towards a framework and a model for knowledge visualisation: synergies between information and knowledge visualisation', Knowledge and information visualisation, p. 238-255


Burkhard, R. A. 2004, 'Learning from Architects: The difference between knowledge visualization and information visualization', Proceedings of the eight international conferences on information visualisation IEEE


Card, S., Mackinlay, J. D. and Shneiderman, B. 1999, 'Reading in information visualisation: using vision to think', Morgan Kaufmann.


Connolly, T. and Begg, C. 2002, 'Database systems a practical approach to design implementation and management, third edition', Addison Wesley


Cui, Q., Ward, M. O. and Rundensteiner, E. A. 2006, 'Enhancing scatterplot matrices for data with ordering or spatial attributes', Proceedings of SPIE


Eidenberger, H. 2004 ‘Visual Data Mining’ Proceedings SPIE Optics East Conference


Eppler, M.J and Burkhard, R. 2004b ’Knowledge visualisation towards a new discipline and its field of application’, IEEE

Fayyad, U., Grinstein, G. and Wierse, A. 2002a,'Information Visualization in Data Mining and Knowledge Discovery', Morgan Kaufmann

Fayyad, U., Grinstein G. and Wierse, A. 2002b,'Evaluation of Visualisation Systems; Information Visualisation in Data Mining and Knowledge Discovery', Morgan Kaufmann

Fayyad, U., Shapiro, G. P. and Smyth, P. 1997,‘“From Data mining to knowledge discovery in databases”’, American Association for Artificial intelligence


Fry, B. 2007,'Visualising data, exploring and explaining data with processing environment', O’Reilly


Goebel, M. and Gruenwald, L. 1999 , 'A survey of data mining and knowledge discovery software ', SIGKDD Explorations. ACM SIGKDD


Han, J. and Kamber, M. 2006 , 'Data mining, concepts and techniques, 2nd edition', Morgan Kaufmann

Holzinger, A. 2005, Usability engineering methods for software developers, Interaction design and children, vol 48, issue 1 ,p 71 - 74


Hornick, M. F., Marcade, E., Venkayala, S. 2006 , 'Java data mining strategy, standard and practice a practical guide for architecture, design and implementation', Morgan Kaufmann


Keim, D., Hao, M. C. 2007, 'Value-Cell Bar Charts for Visualizing Large Transaction Data Sets', IEEE transaction on visualisation and computer graphics, vol 13, no. 4, July/August.

Keim, D. 2005,'scaling visual analytics to very large data sets', workshop on visual analytics

Keim, D., Panse, C., Sips, M. 2003,' PixelMaps: A new Visual Data Mining Approach for Analysing Large Spatial Data Sets', proceedings of the third IEEE International Conference on Data Mining


Keim, D. 2002,'Information Visualization and Visual Data Mining', IEEE transactions on visualization and computer graphics, vol8, no1 January.

Keim, D., 2001,'visual exploration of large data sets', Communications of the ACM


Markides, C. 1998,'Strategic Innovation in Established companies', Sloan management review, spring
Marghescu, D. 2003,' A conceptual framework for evaluating the quality of visualization tools in data mining', Electronic proceedings of the 26th information systems research seminar in Scandaninavia, Halikko Manor, Finland

Marghescu, D., Rajanen, M., Back, B. 2004,' Evaluating the quality of use of visual data mining tools', In proceedings of 11th European Conference on Information Technology Evaluation Amsterdam, Netherlands

Mazza, R., Berrè, A. 2007,'Focus Group Methodology for Evaluating Information Visualisation Techniques and Tools', 11th International Conference Information Visualisation

http://business.timesonline.co.uk/tol/business/article396330.ece


Miller, G.A. 1956, 'The magical number seven, plus or minus two: Some limits on our capacity for processing information', The Psychological Review, 63, p. 81-97.


Ngai and Chan 2005 ‘Evaluation of knowledge management tools using AHP’

Nielsen, J. 1994. 'Heuristic evaluation', Usability Inspection Methods, John Wiley & Sons,


Novick, L.R. 2001 , 'Spatial Diagrams: Key Instruments in the Toolbox for Thought',The psychology of learning and motivation


Poulet, F. 2004,'Towards visual data mining', International conference on enterprise information systems


Redpath, R. and Srinivasan, B. 2003,'Criteria for a comparative study of visualisation techniques in data mining', IEEE 3rd International conference on intelligent systems design and application, Tulsa, USA, ISDA


Shneiderman, B. 1996, 'the eyes have it: a task by data type taxonomy for information visualisation', proceedings of the IEEE symposium on visual languages, boulder/co step 3-6

Soukup, T., Davidson, I. 2002, 'Visual data mining, techniques and tools for data visualisation and mining', Wiley
Soejarto 2003,'Tough times call for Business Intelligence Services, an indisputable area of growth', Retrieved March 2007 from http://www.varbusiness.com


Spender, J. C. 1996a,'"Making knowledge the basis of a dynamic the theory of the firm"', Strategic management journal, special issues, pp 45-62

Spender, J. C. 1996b,'"Organizational knowledge, learning, and memory: three concepts in search of a theory"', Journal of organizational change management pp 63-78

Spiegler, I. 2000 ‘Knowledge Management: A New Idea or a Recycled Concept?’, Communication of the AIS, Vol. 3, article 14


Stewart, T. A. 1997,'Intellectual capital, the new wealth of organisations', Nicholas Brealey


Toeh, S. T. and Ma, K. 2005, 'Hifocon: Object and Dimensional Coherence and Correlation in Multidimensional Visualisation', ISVC.


Tory, M., Kirkpatrick, AE., Atkins, MS. Möller, T. 2006, 'Visualization Task Performance with 2D, 3D, and Combination Displays', IEEE transactions on visualisation and computer graphics

Tu Y. and Shen, H. 2007, 'Visualizing Changes of Hierarchical Data using Treemaps', IEEE transactions on visualisation and computer graphics, vol. 13, no. 6


Ware, C. 2000, 'Information visualisation - perception for design', 1st edition. Morgan Kaufmann publishers

Watanabe, R. M., Senoo, D. 2008, 'Organizational characteristics as prescriptive factors of knowledge management initiatives' Journal of Knowledge Management, volume: 12, issue 1, pages 21-36

Weidenmann, B. 1989, 'When good pictures fail: An information processing approach to the effect of illustration', in: Mandland Levin eds. Knowledge acquisition from text and pictures, Amsterdam, Elsevier,


Xu, Y., Hong W. and Song, J. 2007a,'Parallel Dual Visualization of Multidimensional Multivariate Data', Proceedings of the IEEE International Conference on Integration Technology

Xu, Y., Hong W., Song, J. 2007b,'Visual Pattern Recognition Method Based on Optimized Parallel Coordinates', Proceedings of the IEEE International Conference on Integration Technology


Zeiller, M. 2005, 'A Case Study Based Approach to Knowledge Visualisation', Proceedings of the Ninth International Conference on Information Visualisation
APPENDIX A: FUNCTIONAL TESTING

Tool name: __________________________

What OS platforms does the tool support? _________________________________
_____________________________________________________________________
_____________________________________________________________________

What is the average cost of the system? _________________________________

Which algorithms does the software support?

<table>
<thead>
<tr>
<th>Decision Trees/ Rules</th>
<th>Regression</th>
<th>Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistics (descriptive)</td>
<td>Visualization</td>
<td>Association rules</td>
</tr>
<tr>
<td>Sequence/Time series analysis</td>
<td>Neural Nets</td>
<td>SVM</td>
</tr>
<tr>
<td>Bayesian</td>
<td>Boosting</td>
<td>Nearest Neighbour</td>
</tr>
<tr>
<td>Hybrid methods</td>
<td>Genetic algorithms</td>
<td>Bagging</td>
</tr>
</tbody>
</table>

Which functions does the software support?

<table>
<thead>
<tr>
<th>Undo /redo</th>
<th>Filtering</th>
<th>Value substitution</th>
</tr>
</thead>
</table>

Does the system have security? _________________________________________

Does the tool have a backup ability? _________________________________

What is the limit to the size of the database system the tool can handle? ________

Can the tool handle different data sources?
_____________________________________________________________________
_____________________________________________________________________
_____________________________________________________________________
Can the tools export data to other applications?

Email ☐ Word ☐ Pdf ☐

What other reporting abilities are supported?

_____________________________________________________________________

_____________________________________________________________________

Visualisation

Did the tool provide a selection of overview options? _________________________

_____________________________________________________________________

Does it provide details on demand (zoom in/zoom out)? _______________________

_____________________________________________________________________

Is there capacity to deal with uncertain data? _______________________________

_____________________________________________________________________

Usability

Does it support a client server model? Yes ☐ No ☐

Does it support multiple users, if so how many? ____________________________

Does the tool provide good initial configuration ability? ______________________

_____________________________________________________________________

Knowledge management

Is the tool designed specifically for the organisation domain? ________________

_____________________________________________________________________

156
APPENDIX B: USER TUTORIAL

Introduction to Data Mining

Data mining is sorting through data to identify patterns and establish relationships. This can be done through association (by looking for patterns where one event is connected to another event) or classification involving classifying new data. For instance this information can be used to predict a mobile phone user’s usage in order to offer better packages.

Benefits of Data Mining

The main advantage is discovering patterns in huge amounts of data automatically for instance in order to predict buyer behaviour or measure if pharmaceuticals are successful on many variables.

Visual Data Mining Tools

As the name suggests visual data mining tools are applications which enable and simplify data mining by making algorithms, visualisations and presentation abilities to quickly develop and explore the data.

Purpose of Feedback

Before investing in a visual data mining tool, organisations need to ensure they have the best available tool which is suited to their task. This user feedback will identify the shortcoming of the tools from user experiences.
Purpose of Classification

The purpose of classification is to classify the data available and predict when a new entry is added if a condition will be true such as a person being able to pay back a bank loan.

Description of Datasets

The datasets include a German credit data set determining if a user has a good credit rating, determining if a person earns over fifty thousand a year based on his job and background and a third credit data set to determine if they can get a car loan.
APPENDIX C: TOOL QUESTIONNAIRE

_Gaining an overview of the data_

The tools were explained and an overview of the data is given. What is your first impression of each tool?

__________________________________________________________________
__________________________________________________________________
__________________________________________________________________

Do you find it difficult to understand the data from the overview in each tool?

__________________________________________________________________
__________________________________________________________________
__________________________________________________________________

If five is the best can rate the overview ability of each tool from 1 to 5

SAS : ________
SSS : ________
RapidMiner: ________

_Creating a regression model to investigate the data_

From review the regression task layout and results did you think this task is difficult in each tool?

__________________________________________________________________
__________________________________________________________________
__________________________________________________________________
Do you find the result understandable or good at showing the result in each tool?

_____________________________________________________________________

_____________________________________________________________________

**Changing a parameter in the decision tree**

After reviewing the decision tree models on the second dataset and change the decision tree depth to 4. Did you find it difficult to interrupt the results in any tool?

_____________________________________________________________________

_____________________________________________________________________

Did you find the explanations for the classification quickly understood?

_____________________________________________________________________

_____________________________________________________________________

**Review the visualisations**

After looking again at the visualisations displayed in the tools, for modelling and displaying results. What are your impressions of how usable each tool was?

_____________________________________________________________________

_____________________________________________________________________

If five is the best can rate the difficulty to use the tool of each tool from 1 to 5

<table>
<thead>
<tr>
<th>Tool</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS</td>
<td>______</td>
</tr>
<tr>
<td>SSS</td>
<td>______</td>
</tr>
<tr>
<td>RapidMiner</td>
<td>______</td>
</tr>
</tbody>
</table>
APPENDIX D: INTERVIEW QUESTIONS

Usability Questions

What did you think are the good/bad characteristics of the tool?

Did you feel you were concentrating on the tool or the data set?

Knowledge Management Questions

Does the manager have confidence in the results?

What was your overall opinion of the tool?

Did you find that for some tasks there was too much information on the screen?

In its current format could you present to others, do you think they would understand the data?