University of Fort Hare

Together in Excellence

TEXT DATA ANALYSIS FOR A SMART CITY PROJECT IN A DEVELOPING NATION

by

Aubrey J. Currin

TEXT DATA ANALYSIS FOR A SMART CITY PROJECT IN A DEVELOPING NATION

by

Aubrey Jason Currin
201001778

Dissertation

submitted in fulfilment of the requirements for the degree

Masters of Commerce

in

Information Systems

in the

Faculty of Management and Commerce

of the

University of Fort Hare

Supervisor: **Prof. Stephen Flowerday**

November 2015

Abstract

Increased urbanisation against the backdrop of limited resources is complicating city planning and management of functions including public safety. The smart city concept can help, but most previous smart city systems have focused on utilising automated sensors and analysing quantitative data. In developing nations, using the ubiquitous mobile phone as an enabler for crowdsourcing of qualitative public safety reports, from the public, is a more viable option due to limited resources and infrastructure limitations. However, there is no specific best method for the analysis of qualitative text reports for a smart city in a developing nation.

The aim of this study, therefore, is the development of a model for enabling the analysis of unstructured natural language text for use in a public safety smart city project. Following the guidelines of the design science paradigm, the resulting model was developed through the inductive review of related literature, assessed and refined by observations of a crowdsourcing prototype and conversational analysis with industry experts and academics. The content analysis technique was applied to the public safety reports obtained from the prototype via computer assisted qualitative data analysis software. This has resulted in the development of a hierarchical ontology which forms an additional output of this research project. Thus, this study has shown how municipalities or local government can use CAQDAS and content analysis techniques to prepare large quantities of text data for use in a smart city.

Keywords: smart city; public safety; participatory crowdsourcing; Text Mining (TM); Natural Language Processing (NLP); ontology

Declaration

I, Mr. Aubrey Jason Currin, hereby declare that:

- The work in this dissertation is my own work.
- This dissertation has not previously been submitted in full or partial fulfilment of the requirements for an equivalent or higher qualificationat any other recognised educational institution.
- I am fully aware of the University of Fort Hare's policy on plagiarism and I have taken every precaution to comply with the regulations.
- I am fully aware of the University of Fort Hare's policy on research ethics and I have taken every precaution to comply with the regulations.
- This study was supported by funding from the National Research Foundation (NRF) and International Business Machines (IBM). Any opinions, findings, conclusions or recommendations expressed in this research are those of the author(s) and do not necessarily reflect the views of the aforementioned institutions.

Signature:_____ Date: _____

Acknowledgements

I would like to thank my supervisor and mentor Prof Stephen Flowerday for his invaluable guidance, input, teaching and general commitment to the success of this research project.

I would also like to thank all the other academics and experts who contributed advice, guidance and motivation, including particularly the conversational analysis section of the research.

A special acknowledgement must be made to the National Research Foundation (NRF). Without their scholarship, I would not have been able to devote the amount of time required to complete this study. Other support services were provided by the Govan Mbeki Research and Development Centre (GMRDC) of the University of Fort Hare (UFH). I, as the author, acknowledge that the opinions, findings and conclusions or recommendations expressed in this study, and any publications resulting from it, are those of the author, thus the NRF, GRMDC and UFH accept no liability whatsoever in this regard.

I would like to thank all the family members and friends who supported and motivated me throughout the completion of this dissertation.

Lastly, I thank my Heavenly Father for bestowing upon me the talents and abilities that I have used to complete this research and the opportunities that he has afforded me along the path toward this destination.

Dedication:

I dedicate this dissertation to my wife Kate Rose Currin.

Kate, without you I would have never had the opportunity to pursue a tertiary education, nor would I have ever believed myself capable of it. Having you by my side means more to me than you will ever know.

Table of Contents

List of Figures

List of Tables

Definitions

Smart City – A city that utilises data and Information Communication Technologies (ICT) to improve the city's socio-economic development and the quality of life of its citizens (Schaffers et al., 2011).

Crowdsourcing - Utilising technologies that enable interactive information sharing, collaboration and interoperability in order to take a specific task or problem and invite the public or selected groups within the public to do it (Howe, 2006; Yuen, King, & Leung, 2011).

Business Intelligence – All relevant business information and the process in which an organisation acquires, analyses and disseminates information relevant to making business decisions (Lönnqvist & Pirttimäki, 2006).

Natural Language Processing (NLP) – Theories and techniques that address the problem of natural language communication with computers (Lehnert & Ringle, 2014).

Content analysis – A research technique for making replicable and valid inferences from texts by evaluating the saliency and frequency of specific words or phrases within a particular document or piece of text data in order to find keywords or frequently repeated ideas (Krippendorff, 2004).

Ontology – An explicit, formal specification in the form of a structured, taxonomical hierarchy of multiple terms and their shared conceptualisation within a certain domain of interest (Gruber, 1994; Jiang, Zhang, Yang, & Xie, 2013).

Acronyms

BI – Business Intelligence

DM – Data Mining

CRISP-DM – Cross Industry Standard Process for Data Mining

TM – Text Mining

IR – Information Retrieval

IE – Information Extraction

NLP – Natural Language Processing

BCMM – Buffalo City Metropolitan Municipality

IVR – Interactive Voice Response

SAPS – South African Police Service

DW – Data Warehouse

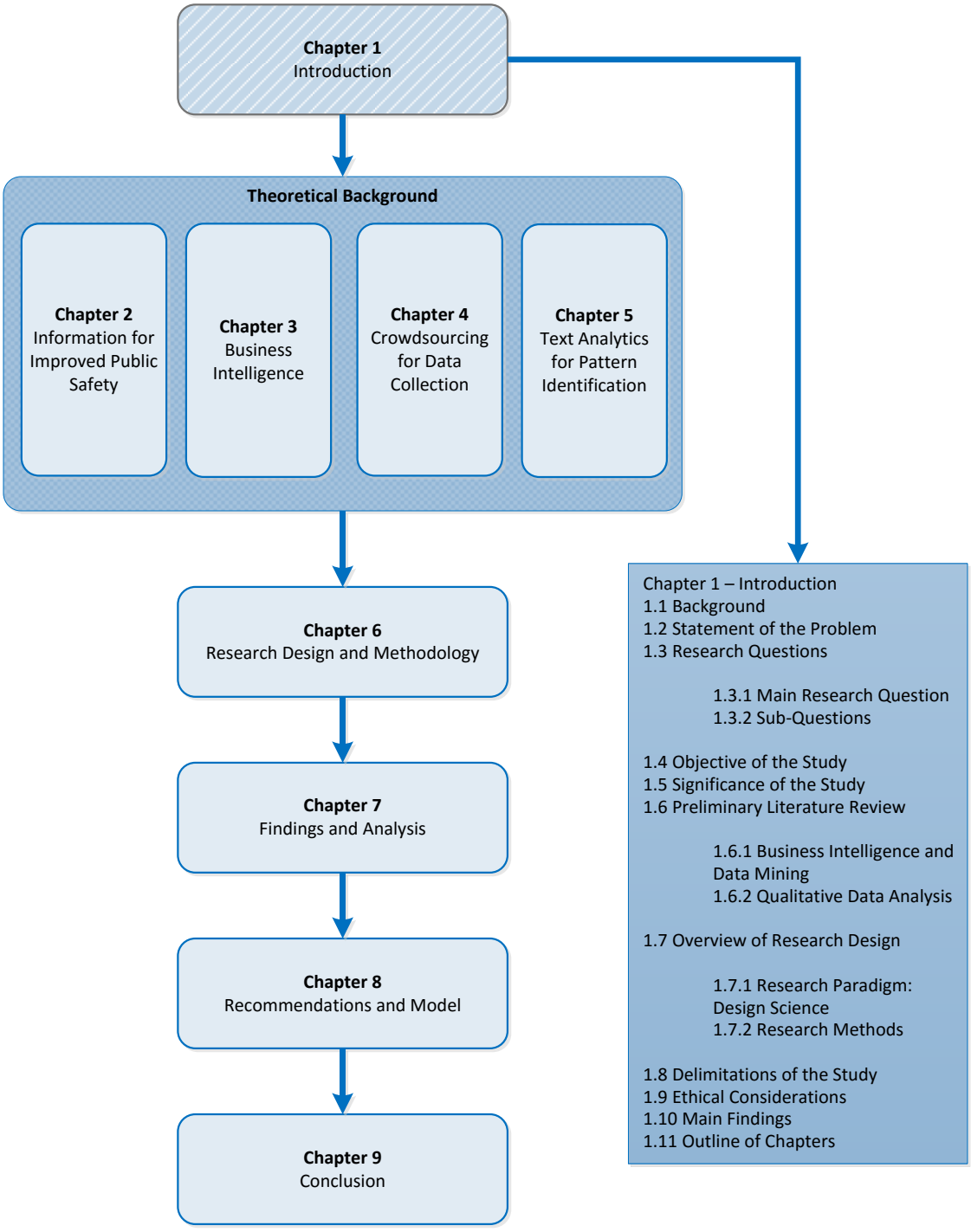
DSS – Decision Support Systems

QDA – Qualitative Data Analysis

CAQDAS – Computer Assisted Qualitative Data Analysis

OLAP – On-Line Analytical Processing

Chapter 1 – Introduction



1.1 Background

This century has seen large scale urbanisation occurring at an ever increasing rate resulting in a greater level of urbanisation than ever before (Dirks, Gurdgiev, & Keeling, 2010). Citizens have certain needs common to all members of the public, which they expect cities to meet (Komninos, Pallot, & Schaffers, 2013). These needs include basic services such as public safety, electricity, healthcare and education. Bhana, Flowerday, and Satt (2013) explain that although urbanization has benefits for cities and the people moving to them, the rapid development of this phenomenon has also lead to problems such as increased crime rates, strained resources and infrastructure, as well as difficulties in forecasting the requirements of services. In developing nations, public safety is usually one of the most urgent of these needs.

Many of these problems pertain to government services and resource allocation planning. This is not necessarily because city resources are scarce, but because they are not being managed effectively and efficiently (Exner, Zeile, & Streich, 2011). Municipalities can, however, be more proactive in targeting and deploying resources and services for the improvement of public safety if better information is available to them. Caragliu, Del Bo, and Nijkamp (2011) state that the urban performance of a city is dependent on its physical capital and, increasingly, also on the availability and quality of information communication and social infrastructure. In order to utilize the resources and manpower available to a city, they need to be fully monitored on a daily basis (Al-Hader & Rodzi, 2009). The collection of data on a city's environment will assist decision-making on resource management by identifying patterns in the data in order to forecast needs. Therefore it can be said that cities need to become "smarter".

Hence, similarly to how Business Intelligence (BI) makes use of data to inform the decision-making process within organisations, the smart city concept relies on data and ICT to improve the city's socio-economic development and quality of life (Schaffers et al., 2011). Data is therefore required as input to inform the smart city. In most previous cases, this data was obtained via automated sensors which collect quantitative data. Developing nations usually do not have the infrastructure or funding to implement a sensor based system, but they can take advantage of the dispersion of their citizens by collecting data via crowdsourcing.

Bhana et al. (2013) make a case for participatory crowdsourcing to be used for generating data using the crowd as sensors via the ubiquitous mobile phone. This option is more affordable for a city in a developing nation and provides rich qualitative data which can be much more informative. However, this means that the quantitative methods for data analysis used in sensor based networks are not applicable to the qualitative data obtained via crowdsourcing. Utilising qualitative data is not new, there is a long history of researchers using qualitative data, but its application in business has not yet developed to a large scale

BI had its initial focus on numerical data for the purpose of logistics and forecasts. Databases are populated with product, sales, customer, and supplier data (amongst others), with relational databases being one of the most popular formats (Nisbet, Elder, & Miner, 2009). Queries and calculations are then done with relative ease on numeric values and these summated values can be used in statistical analyses. Mitra, Pal, and Mitra (2002) explain that modern developments in software and hardware and the rapid digitisation (or computerisation) of business have led to vast amounts of data being collected and stored at an ever increasing rate, and "*as a result, traditional ad hoc mixtures of statistical techniques and data management tools are no longer adequate for analysing this vast collection of data"* (Mitra, Pal, & Mitra, 2002, p. 3).

The size of big data, however, is often not seen as the most pressing problem. In a white paper focused on big data and analytics published by SAS, Troester (2012) states that large organisations are inundated with terabytes and petabytes of data, while also predicting that storage capacities will soon be measured in exabytes, zettabytes and yottabytes. Troester (2012) continues to explain that the size or amount of data is, by many businesses, not seen as the biggest problem as the gradual growth in capacity and technology has allowed most organisations to make gradual changes to their hardware and software assets. White papers released by companies including SAS, Boldon James and Intel state that one of the biggest problems in modern data analysis (especially in big data analysis) is working with unstructured data, mainly in text format (Boldon James, 2012, 2014; Troester, 2012).

Gartner predicted that by 2015, 80 percent of information used in enterprises will be unstructured material from sources such as documents, e-mail, images, video and other

texts (Chiang, Goes, & Stohr, 2012). This has resulted in grave concern in the business environment about the lack of ability and methods for analysing unstructured text data (Kalakota, 2011). This unstructured data is further ballooned by social media combined with short and instant messaging, as companies endeavour to leverage these sites and applications for competitive advantage. Thus, it is evident that text analysis strategies for larger quantities of qualitative data (including "big data") are not only lacking for smart cities, but are lacking throughout the business world.

This study will therefore aim to develop a model for qualitative text analysis, but more specifically for the analysis of crowdsourced data for use in the enhancement of public safety in a smart city project.

1.2 Statement of the Problem

Bhana et al. (2013) explain that crowdsourcing is an effective method for the collection of qualitative data for a smart city as it takes advantage of the dispersion of citizens, thus allowing for the collection of a broad range of data over a short period of time.

This research project focuses specifically on one such project: the participatory crowdsourcing public safety smart city project in East London, South Africa. For the project to be of use, the data collected needs to be analysed correctly for presentation and use by the smart city's emergency and non-emergency units using as little time and effort as possible. **The research problem is that there is no specific method for the analysis of qualitative text reports for a smart city project in a developing nation.**

Crowdsourced data is qualitative rather than quantitative, which is a richer source of information. The analysis thereof, however, is much more time consuming and cognitively taxing, especially as the amount of data increases. There is a great deal of research work to show potential benefits of big data analytics and a motivation for businesses to implement it, but how exactly to analyze and extract meaningful patterns is proving much more difficult than anticipated. Brandel (2008) suggests that the specific approach to crowdsourcing is largely left to the individual organisation to decide what they are comfortable with. In order to improve public safety in East London, a new applicable model is required for the analysis of data gathered from the citizens of East London.

1.3 Research Questions

1.3.1 Main Research Question

- **How can natural language analytics be applied to qualitative public safety text data in a smart city project based in a developing country?**

In order to answer the main research question, the following sub-questions are addressed.

1.3.2 Sub-Questions

- **What information is extracted from the data to enable a smart city to improve public safety?**

The first sub-question sets a context for the required data by investigating the smart city concept with specific focus on public safety case studies. Further, the post analysis goals, strategies and requirements of the information are discussed in order to specify what the analysis will aim to achieve.

- **How can qualitative participatory crowdsourcing data be used for business intelligence?**

This sub-question aims to explore the characteristics (and implications) of crowdsourcing data obtained via an IVR and mobi site. Literature is reviewed in order to identify the positive and negative aspects of working with data of this nature. The concept of business intelligence will also be explored in order to ensure the usefulness of this data in a business context.

- **How can text analytics be used to identify patterns and trends in unstructured text data?**

The third sub-question considers different methods used for developing qualitative natural language data into useful information that can be used to improve public safety measures.

1.4 Objective of the Study

The objective of this study is the development of a model for the analysis of public safety natural language text data obtained from participatory crowdsourcing. This primary objective is divided into smaller secondary objectives as shown by the sub-

questions listed above. Answering the three sub-questions will provide the information required to inform a model that will guide analysts through the process of knowledge extraction from crowdsourced data for the improvement of public safety. The significance of using a model such as this is stated in the following section.

1.5 Significance of the Study

The significance of this study relates directly to the public safety of citizens in a smart city within a developing country. This study is aimed at a specific crowdsourcing, public safety prototype in the East London area. The local citizens, therefore, benefit from it. However, the model produced by this study is intended to be scalable and flexible enough to be usable as a template for other cities in a similar developing context. Further, generalisation may also allow the model to be used on additional smart city sub-areas, for example: health, education or transport. In terms of public safety, citizens can feel safer as well as experience an increase in general living standards. This study focuses on the correct and informative analysis of data obtained from participatory crowdsourcing, allowing the local citizens to contribute indirectly to the local municipalities' decision-making process for more accurate and appropriate resolution of public safety issues, thus giving them additional influence in their living conditions. Additionally, the model produced by this study will also be usable as a basis for text analysis strategies in business contexts.

1.6 Preliminary Literature Review

In the early 1990s, Howard Dressner first used the term Business Intelligence (BI) as an umbrella term for information and applications that support decision-making (Watson & Wixom, 2007). Over the last few years, BI has become a business priority with an ever increasing demand for better and more tailored solutions (Chiang et al., 2012).

1.6.1 Business Intelligence and Data Mining

Lönnqvist and Pirttimäki (2006) emphasise that businesses need to receive timely and effective information not only to succeed, but even just to survive. The term BI can refer to relevant business information or the process in which the organisation acquires, analyses and disseminates information relevant to making business decisions (Lönnqvist & Pirttimäki, 2006). Negash (2004) explains that BI has replaced terms such as: decision support, executive information systems and management information systems.

This places BI in the role of strategic informant to the businesses decision makers. Simply put, one can say that BI systems analyse data to inform decision-making.

Negash (2004) explains that BI transforms data – first into useful information and then into knowledge through human analysis. A large part of the knowledge extraction process falls within data mining (Mitra, Pal, & Mitra, 2002; Nisbet, Elder, & Miner, 2009; Tien, 2013). Basically, data mining involves pattern recognition and storing of data in structured relational databases that can be queried (Nisbet, Elder, & Miner, 2009). Wirth and Hipp (2000) argued in favour of the CRISP-DM (Cross Industry Standard Process for Data Mining) model and its usefulness in data mining projects. This process model is designed to be generic and flexible in order to be applicable to different industries and situations. The initial emphasis is on understanding the business (or problem space) and the data to be analysed.

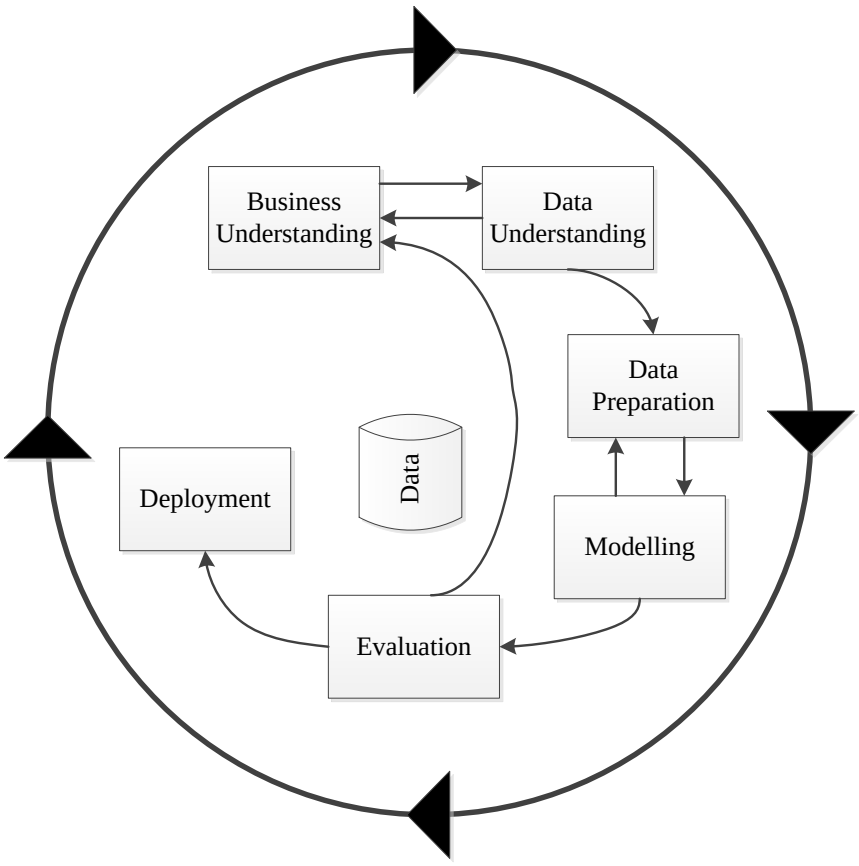


Figure 1.1: Phases of the CRISP-DM Model for Data Mining (Wirth & Hipp, 2000)

The CRISP-DM process depicted in Figure 1.1 has, however, mainly been used for the mining of structured data, as has been the main focus of BI until more recently with a heightening focus on text mining.

In their search for competitive advantage, organisations are becoming increasingly aware of the multitudes of data in text formats such as e-mail, documents, conversational records and social networks (Chiang et al., 2012). This data must, however, be correctly and appropriately analysed in order to be of any help to decision-makers. Leech and Onwuegbuzie (2008) compiled a compendium explaining the relationship between types of qualitative data analysis techniques and the source of the qualitative data to be analysed. The authors concluded that a more focused analysis using appropriate methods based on the data type, origin and purpose will lead to deepened understanding (Leech & Onwuegbuzie, 2008).

1.6.2 Qualitative Data Analysis

Weitzman (1999) summarises the qualitative data analysis process as following a path from questions to conclusions. Some of the main steps, amongst others in this iterative process, are identifying questions, the development of a coding scheme, coding chunks of data, reducing the data, and entering the data into displays in order to draw conclusions (Weitzman, 1999). Though this may seem a simple process, these steps are in fact quite general and can vary widely as there are many different ways of doing each one. Weitzman (1999) explains that exactly how this is done depends on the nature of the project and the questions asked.

Miner, Delen, Elder, Fast, Hill, and Nisbet (2012, p. 30) explain:

> *"Text mining and text analytics are broad umbrella terms describing a range of technologies for analysing and processing unstructured text data".*

Text mining in itself, however, is still a broad field spanning different areas and fields as depicted in Figure 1.2. Miner et al. (2012) describe seven main practice areas of text analytics which are: Search and information retrieval (IR); Document clustering; Document classification; Web mining; Information extraction (IE); Natural language processing (NLP), and Concept extraction. The seven text mining practice areas exist at the major intersections of text mining with its six related fields (Miner et al., 2012).

As indicated in Figure 1.2, Natural Language Processing (NLP) has been identified as the applicable text mining area for this study.

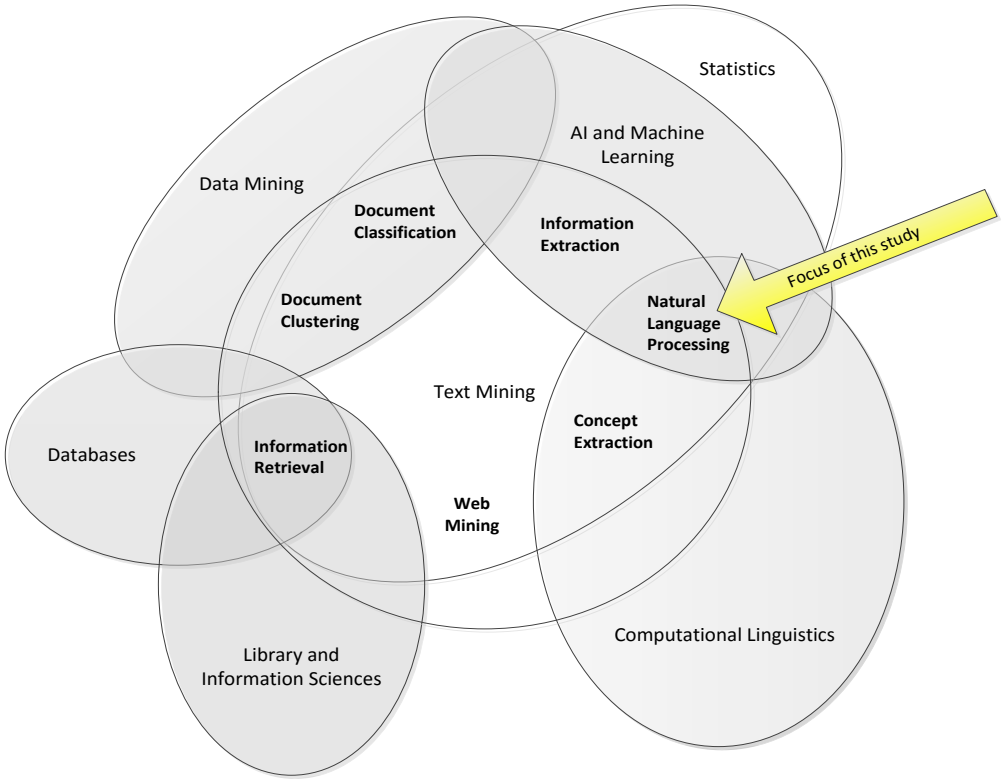


Figure 1.2: Venn Diagram-Seven Practice Areas of Text Analytics (Miner et al., 2012)

Miner et al. (2012) describe NLP as tasks involving low-level language processing and understanding. Chopra, Prashar, and Sain (2013) simply describe NLP as being everything a computer requires to understand and generate languages spoken by people. A more formal definition states NLP as being "concerned with theories and techniques that address the problem of natural language communication with computers" (Lehnert & Ringle, 2014, p. 13). The two latter definitions explicitly state the involvement of computers in NLP, but although Miner et al. (2013) do not explicitly include computers in this definition, by reflecting back to Figure 1.2 it is evident that these authors have placed NLP to include "computational linguistics" and/or "artificial intelligence and machine learning".

1.7 Overview of Research Design

This section provides an overview of the research design followed by this study; a more detailed research design follows in Chapter 6. Epistemologically, all research must have sound basis in existing knowledge and theory in order to ensure that sound, rigorous research is done. Olivier (2009) describes a research paradigm as an accepted

model or pattern that guides all research. Additionally, the use of an accepted paradigm helps to qualify a project as research due to it being based on well accepted methods (Olivier, 2009). The following section explains why design science is the paradigm followed by this study.

1.7.1 Research Paradigm: Design Science

> "*The design-science paradigm seeks to extend the boundaries of human and organizational capabilities by creating new and innovative artefacts*" (Hevner, March, Park, & Ram, 2004).

Hevner et al. (2004) explain that design science is a process that solves a problem. It aims to create a novel, useful artefact that is intended to be used to solve a real problem. This artefact must stand up to scrutiny, must be presented well, and be based on sound research. To ensure all the criteria are met, Hevner et al. (2004) provide seven guidelines for design science research. These guidelines will be followed throughout this study in the development of a useful model. It must be noted that for the purposes of this study, these guidelines will not be followed in numerical order. This study will apply these guidelines as depicted in Figure 1.3 and the descriptions that follow hereafter.

- **Guideline 1: Problem Relevance**

  From the background section of this proposal, it is evident that there is a lack of context sensitive guidance for the analysis of unstructured text data (especially when in larger volumes). This study has its focus specifically on the analysis of qualitative reports for the purpose of informing decisions pertaining to the improvement of public safety in the smart city.

- **Guideline 2: Research Rigor**

  The use of sound methods ensured the rigor of this study. All logical conclusions used to inform the model development in this study were based on the review of sound secondary literature. Input from experts and academics lend further rigour to this study through conversational analysis combined with content analysis and findings from prototype observation and assessment.

- **Guideline 3: Design as an Artefact**

  This study produced an artefact in the form of a model that will guide the analysis process of public safety crowdsourced data in text format.

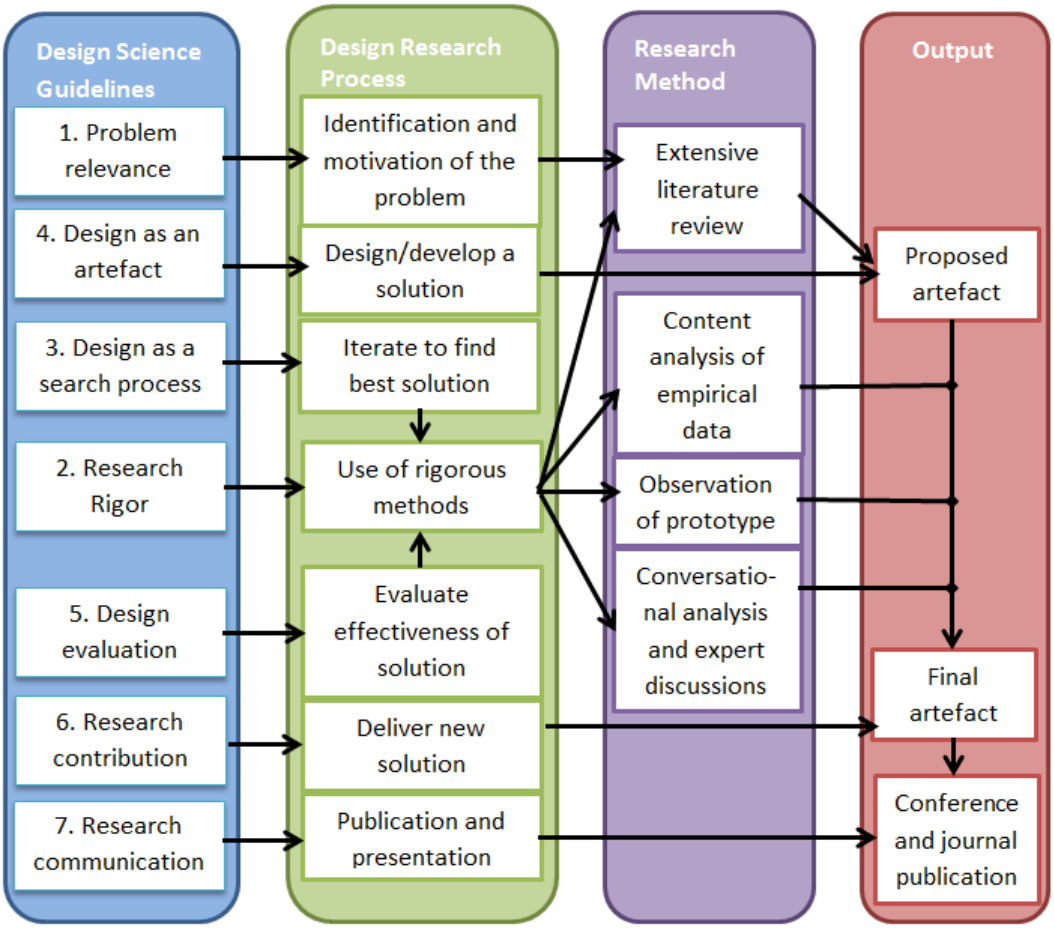


Figure 1.3: Application of Design Science Guidelines

- **Guideline 4: Design Evaluation**

  Expertise and opinion was sought from experts at various stages of this study for guidance, evaluation and refinement of the artefact. The proposed model was evaluated via application to a prototype.

- **Guideline 5: Design as a Search Process**

  The model produced from this study is the culmination of an iterative process of literature review, as well as the review and analysis of primary data. The most impactful problems and desired characteristics were searched for to help form the model. The model was then assessed and refined.

- **Guideline 6: Research Contributions**

  The development of an analysis model will allow the progression of the UFH public safety smart city project by ensuring that the reports obtained are analysed in a way

that will ensure useful information is obtained for the improvement of safety and living conditions in the smart city.

- **Guideline 7: Communication of Research**

  The findings will be published in academic journals and conferences and made available to the public for future research. The dissertation will be accessible through the library at the University of Fort Hare. IBM and Buffalo City Metropolitan Municipality (BCMM) will also be provided with access to this study for future research and developmental purposes.

Following these guidelines throughout the research process has ensured credible conclusions and productive, useful output. The following section introduces the methods used to achieve this.

1.7.2 Research Methods

Although this study is predominantly qualitative, it will include working with some quantitative methods as well. In order to achieve the goals set out by this study, a mixed methods approach was selected as the most appropriate approach to use due to the nature of the data that was used and the analysis process. Research approaches utilised in this study include the literature review, prototype observation, conversational analysis of expert discussions and content analysis of public safety reports.

1.7.2.1 Data Collection

Secondary Data:

Related literature was reviewed for the purposes of this study. The literature focused mainly on other smart city studies conducted in other locations, as well as crowdsourcing, BI and techniques for the analysis of qualitative text data. This information was obtained from books, articles, conference proceedings, white papers, methodologies, websites and other online publications.

Primary Data:

A prototype participatory crowdsourcing smart city system has been established with focus on public safety and has initially been targeted at the city of East London in the Eastern Cape Province of South Africa. Empirical data in the form of qualitative

observations were noted of the prototype's functioning and from its interaction with the model proposed in this study.

East London is part of the greater Buffalo City Metropolitan Municipality, which in the 2011 census was recorded as having a population of 1.4 million people (City Population, 2014; Statistics South Africa, 2012). The initial focus for the project was chosen to be only the 761 996 citizens of East London itself (East London, 2013; ECSECC, 2012) due to time, financial and geographical constraints, as well as wanting a smaller geographically concentrated area.

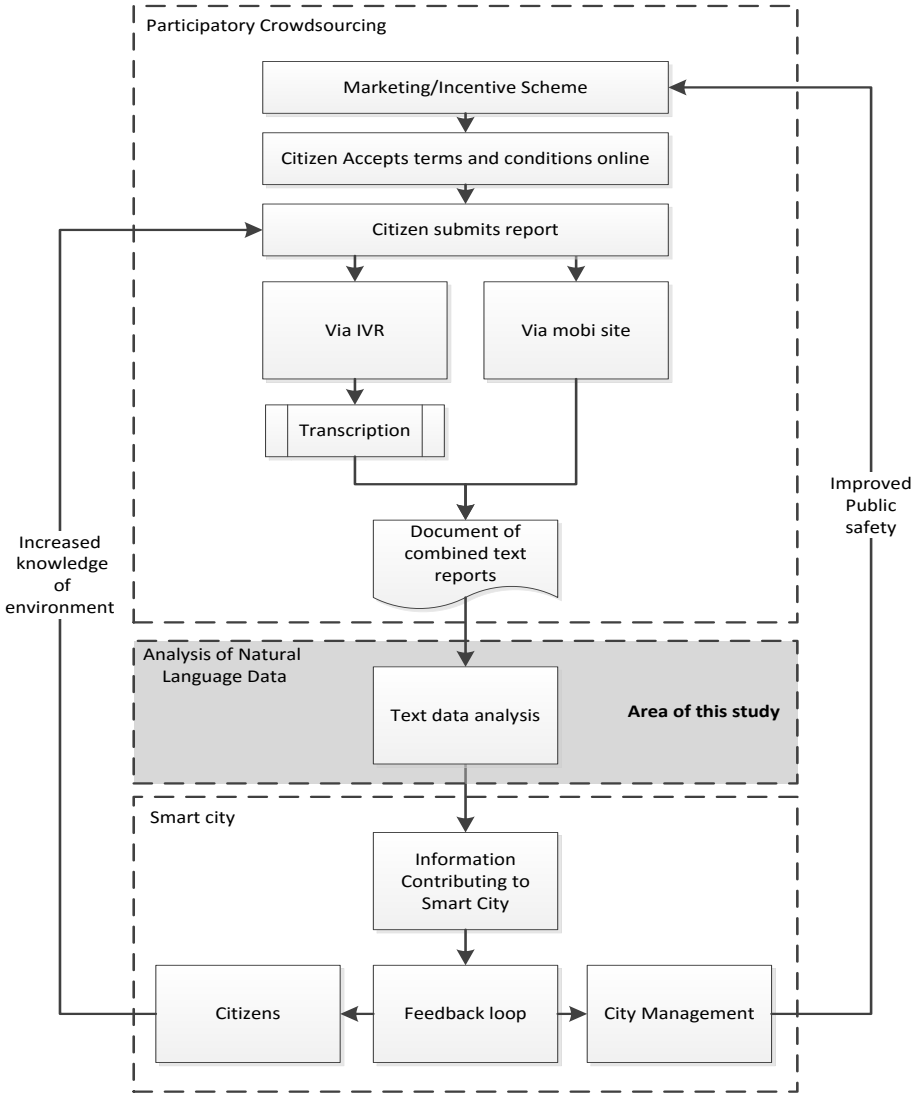


Figure 1.4: Flow of data from greater project to this study

The citizens of East London were requested to volunteer public safety reports via a telephonic IVR system or a mobile website (or mobi site) which has been built

specifically for the purpose of collecting these reports. These reports are also used as empirical data for this study.

An additional source of empirical data involved the conversational analysis of discussions with experts and academics. Regular on-going meetings with academics and leading experts in related fields (national and international) were held in order to help guide and inform this project during its progression as well as refining the final model. These meetings included, but were not limited to: workshops, discussions, status updates, and informal question and answer sessions.

1.7.2.2 Data Analysis:

Secondary Data

An extensive literature review was performed in order to form a theoretical foundation involved in the development of a model for the analysis of qualitative text data. The model was based on inductive logic and was further refined and tested using the empirical data collected.

Primary Data:

The analysis method that was applied to the citizen reports in this study is the Content Analysis method, which is "*A research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use*" (Krippendorff, 2004, p. 18). In order to give structure to the unstructured public safety reports, this study adhered to the guidelines as specified by Krippendorff (2004).

By using specific rules of coding, large amounts of qualitative data is reduced and can be quantified in order to make it easier for researchers to read and identify trends and patterns. Following the Content Analysis technique, the data itself is used to identify the groups that need to be coded; the same data is then coded according to the groups. This is an iterative process and the groups (or code list) can be amended as more become evident during the analysis process.

The Nvivo software package was used as a tool to code the reports for performing the content analysis. The instances and frequency of codes in the reports were grouped and quantified by Nvivo in order to reduce the data to a form which allows for easier pattern

identification and modelling. The software served to organise, manage and perform queries on the reports, but could not interpret results and findings. The Nvivo software queries were not in itself an answer to the research questions; ultimately it was up to the researcher to interpret the findings.

The functioning of the model as applied to the prototype and the public safety reports enabled the researcher to make certain observations. These observations were discussed and considered in the refinement of the model.

Conversational analysis of meetings with experts was conducted and the findings thereof were considered during the entire development of the analysis model. Conversational analysis is a method of data collection through social interactions which includes verbal and non-verbal cues (Goldkuhl, 2004). Regular meetings with senior researchers tempered findings and the conclusions. The input from experts was taken into consideration throughout this study and helped to ensure the usefulness and applicability of the model produced.

Obtaining the right data, applying the appropriate methods, as well as complying with the rest of the seven design science guidelines as depicted in Figure 1.3 thus ensured rigour and credibility of the study. Having outlined the data and methods included in this study, the following section will describe the focus of this study by presenting the delimitations.

1.8 Delimitations of the Study

This study made use of empirical data obtained through participatory crowdsourcing via an IVR and a mobi site. The use of participatory crowdsourcing entails humans acting as sensors by voluntarily reporting public safety issues they witnessed. This study is thus not about the use of automated sensors. The data obtained from the IVR system was transcribed into text format, matching the format of the mobi site data. This study therefore does not focus on analysing other formats of data such as graphics and video. The data in question is obtained from the smart city public safety prototype run in East London by the University of Fort Hare. Although East London forms part of the greater Buffalo City Metro, the reports are from East London only as the initial prototype area. The reports are thus only concerning public safety incidents that have occurred in and

around East London as reported by citizens of East London. The reporting format is in English and thus limits reports to English speakers with access to a mobile phone, telephone and Internet. As this study is aimed at the analysis of the data, it will not include adoption, participation, feedback processes, ensuring quality of data obtained, or ensuring the use of the information once analysed. The smart city concept spans a number of different areas including healthcare, transportation and education amongst others; however, for purposes of this research, the focus is on public safety data as part of smart living in a smart city.

1.9 Ethical Considerations

According to Punch (2006), it is the responsibility of the author to comply with academic integrity and honesty, which includes respecting all people at all times. Punch (2006) further lists a number of categories of ethical issues that researchers should consider. Of these categories, the following were seen to be applicable to this study: informed consent, anonymity, privacy, security and misuse of results. Further, ethical clearance was sought and approval was gained from the University Ethics Committee (UEC) ensuring that this study complies with guidelines and regulations of the University of Fort Hare and its ethics committee.

All participants were aware of the use of the information, remained anonymous, and were free to choose not to participate. The following section summarises the main findings of the research.

1.10 Main Findings

This study developed a model for the analysis of qualitative natural language public safety reports obtained via crowdsourcing for a smart city in a developing nation. An extensive review of literature led to the development of the Smart City Qualitative Data Analysis (SCQDA) model, based on principles from the CRISP-DM as discussed in Chapter 7.

The model was assessed by application to a prototype participatory crowdsourcing system. Observations from the assessment combined with findings from conversational analysis and expert discussions (discussed in Chapter 8) served to refine the model to its

final form as depicted in Figure 1.5. The SCQDA model is discussed in greater detail in Chapter 8.

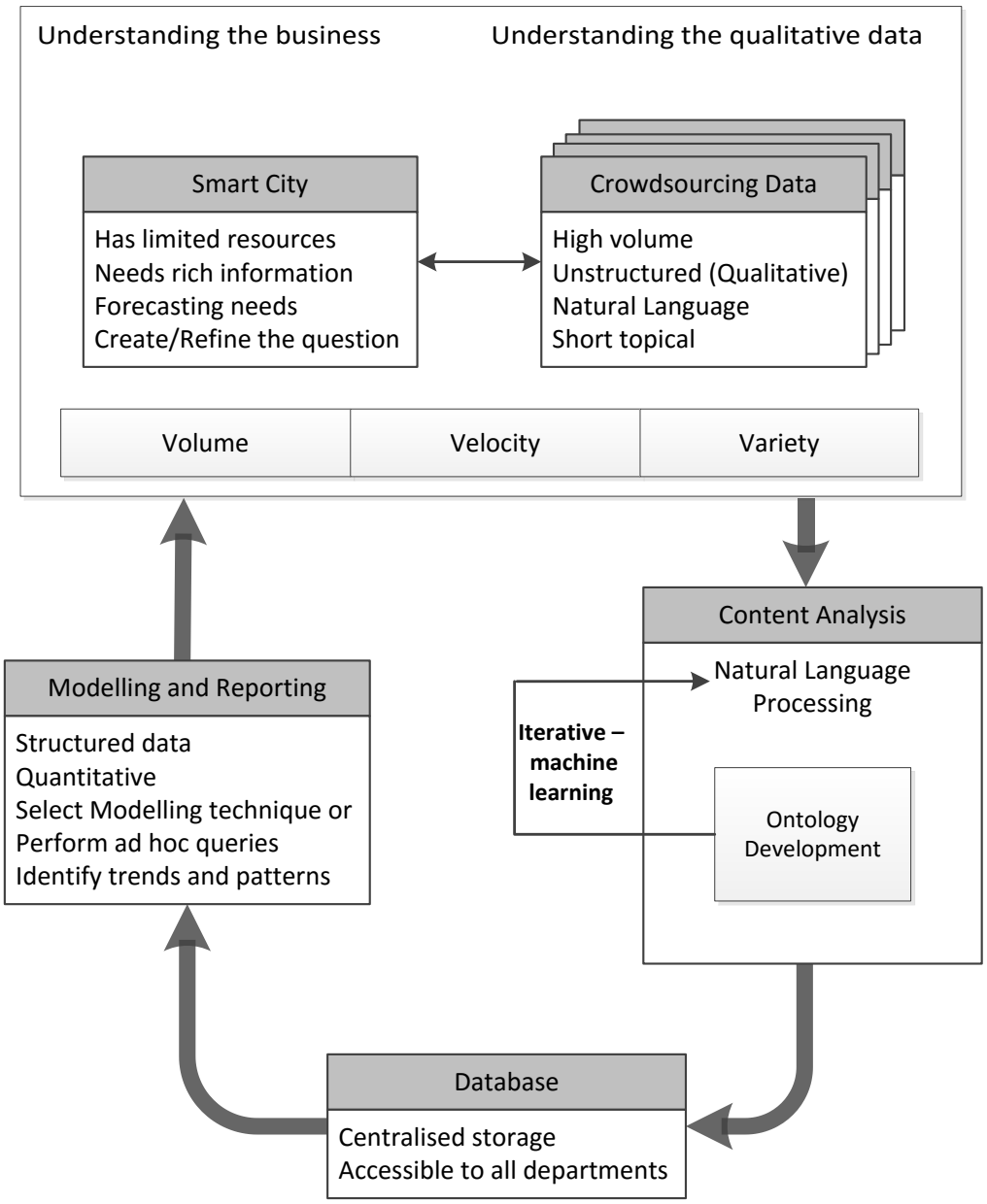


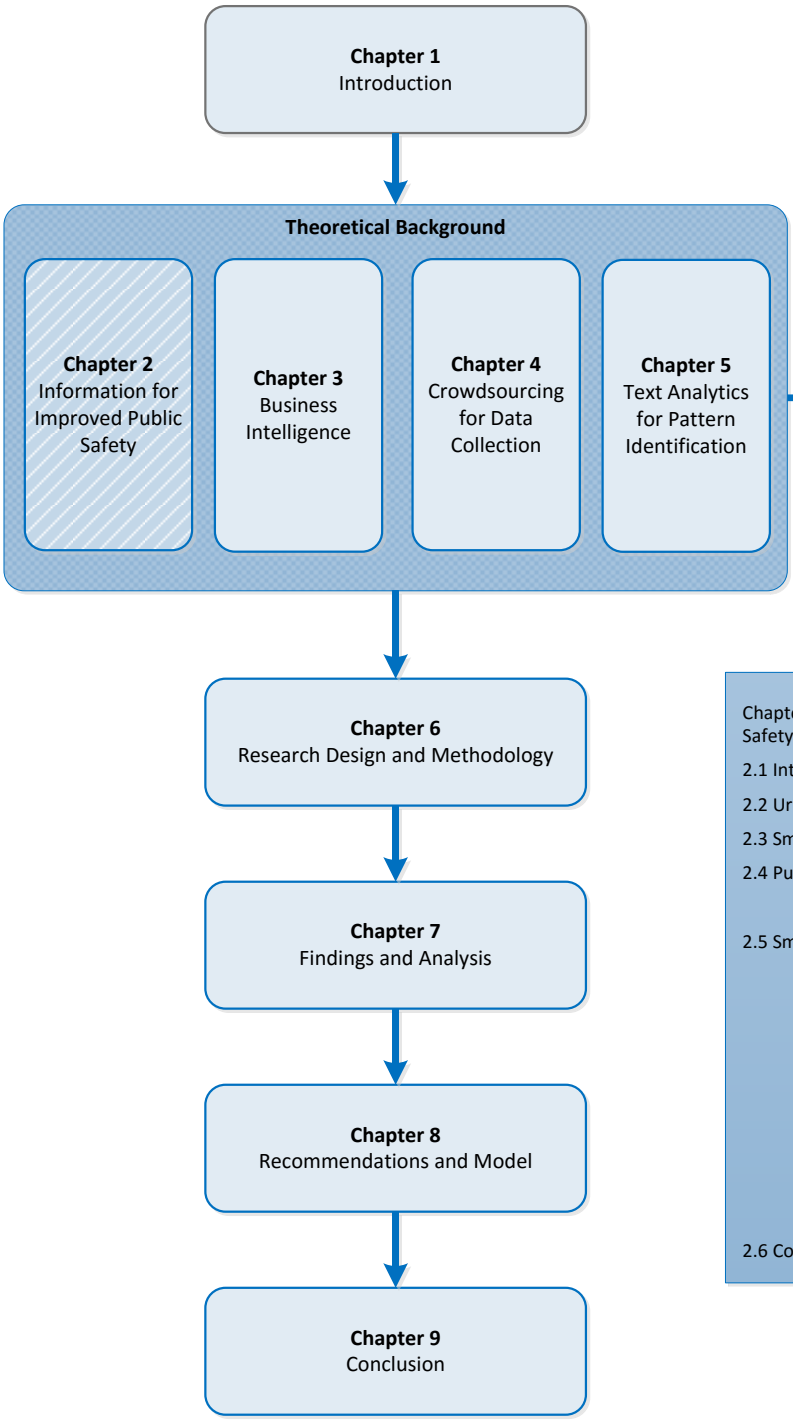
Figure 1.5: SCQDA model

At this stage it is important to note that the model begins with developing an understanding the business environment (the smart city) and the data to be analysed (qualitative participatory crowdsourcing data); this is achieved by literature review in Chapters 2 through 4. The next section provides an outline of the chapters in this dissertation.

1.11 Outline of Chapters

Chapter 1 is an introductory chapter which aims to familiarise the reader with the research problem and its setting. This chapter includes the background on the problem, how this study aims to address this problem, and the goals that this study sets out to achieve. The first chapter is followed by a comprehensive literature review spanning Chapters 2, 3, 4 and 5 in order to set a sound theoretical basis for the rest of this study. The literature covers topics as outlined by the three sub-questions outlined earlier in this chapter, including smart city information requirements, business intelligence, analysis implications of crowdsourcing text data, and current qualitative analysis techniques. Chapter 6 explains the research design and methodology used, as well as delimitations and ethical considerations. Chapter 7 contains findings, a discussion and analysis of research observations, and Chapter 8 presents the recommendations and final contribution (model) of this study. Chapter 9 concludes the study and provides suggestions for future research.

Chapter 2 – Information for Improved Public Safety



2.1 Introduction

The purpose of this study is the development of a model to guide the analysis of natural language data. In order to decide how to analyse data creatively and correctly, it is vital to first understand the environment in which the data is used and the purpose, goals and requirements of the information obtained from analysing the data. The model is generalizable to all smart city focus areas, but the initial context for development and testing is public safety.

This chapter explores the environment wherein a smart city improves public safety. For this purpose, this chapter discusses: **What information should be extracted from the data to enable a smart city to improve public safety?** In order to answer this question, the first section of this chapter explores and defines the concepts of urbanisation, the smart city, public safety as a category of a smart city, and the public safety respondents within East London. The second half of this chapter considers examples of how different smart city systems around the world have improved public safety in order to develop an understanding of what should be required of public safety information once it has been analysed.

2.2 Urbanisation and the Smart City

This century has seen large scale urbanisation occurring at an ever increasing rate resulting in a greater level of urbanisation than ever before (Dirks, Gurdgiev, & Keeling, 2010). Citizens have certain needs common to all members of the public, which they expect cities to meet (Komninos, Pallot, & Schaffers, 2012). These needs include basic services ensuring safety, electricity, healthcare and education. Citizens thus relocate to cities in search of better living conditions as well as prospects of employment.

In 1990, only 13% of the world's population lived in cities (Dodgson & Gann, 2011); this figure increased to 50% during 2007 – 2010, and the UN predicts that the urban population will be a clear majority by 2019 (Washburn, Sindhu, Balaouras, Dines, Hayes, & Nelson, 2009). The figure is expected to rise by 70% – 75% in 2050 (Bakıcı, Almirall, & Wareham, 2012; Nam & Pardo, 2011; Dodgson & Gann, 2011). Washburn et al. (2009) illustrate this rapid increase in urban population with a timeline as is shown in Figure 2.1.

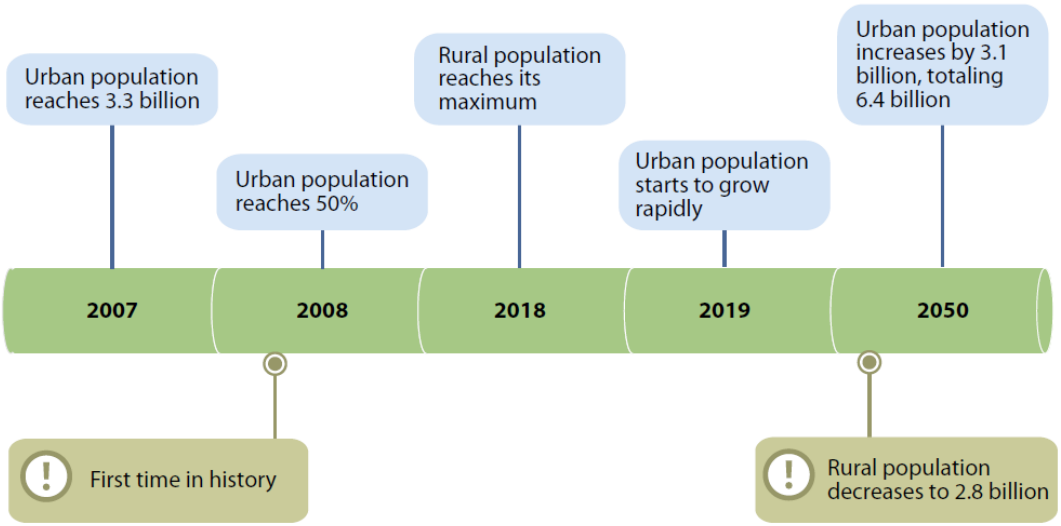


Figure 2.1: Urban Population Growth (Washburn et al., 2009)

It is evident that a large part of this timeline is predicted, but due to the use of statistics and estimates from the United Nations as well as corroboration by other authors mentioned above, this timeline can be accepted as reasonable, especially for the purpose of illustrating and emphasising the continuing rise in urban population.

This urbanisation provides a number of opportunities for cities. The citizens come to cities in search of better living conditions and employment opportunities, and the city in turn stands to benefit from the increased pools of talent, skills and labour that can be contributed by new citizens. Many cities, however, are not ready to deal with the additional strain on resources and infrastructure caused by rapid urbanisation. City resources cannot support all members of the public simply because the ratio of number of citizens to city resources (for example, water and electricity distribution) is high. This is not only because city resources are scarce, but because they are not managed effectively and efficiently (Exner et al., 2011).

Dirks et al. (2010) further explain that traditional resources are no longer the main focal point, but that the skilled, educated, creative and entrepreneurial population are also viewed as a city resource causing cities to compete with each other in improving the quality of life and opportunities on offer for the purpose of luring and retaining educated and talented citizens. This results in cities needing to address challenges including:

- Resource shortages
- Infrastructure deterioration
- Energy shortage
- Environmental issues
- Health concerns
- Unemployment
- Public safety

One approach to compete with other cities is to improve the quality of services. Other approaches involve competing by creating employment opportunities and innovation through research (Dodgson & Gann, 2011). Caragliu et al. (2011) state that urban competition involves investments in physical, social and human capital, but Information and Communication Technology (ICT) is viewed by many as being the key for improved planning and management of cities. A concept gaining increased popularity, that of the smart city, addresses the urbanisation issues discussed above by making use of information and other benefits of urbanisation. Bhana et al. (2013) describe how the smart city concept can be used to reduce problems associated with urbanisation by using information to address the management of resources and infrastructure. Cities can therefore improve the way they adapt to increased urbanisation by using principles of Business Intelligence, as discussed in Chapter 2, to improve the way resources are managed and to better inform the planning of further infrastructure expansion.

2.3 Smart City Defined

Advances in ICT, particularly the Internet, have empowered people as consumers of private goods; therefore, we now tend to expect the same quality from public service (Smart Cities background, 2013). Municipalities and local government are expected to function more like private businesses, with focus on customer centricity and information rich proactive planning. Thus, competitive and effective cities can benefit from implementing the principles of Business Intelligence to become smarter.

The term ‘smart city’ has been gaining popularity worldwide, but differing contexts have led to a range of labels being used for what can be seen as very similar concepts, examples include: digital city, intelligent city, high tech district and creative city (Allwinkle & Cruickshank, 2011). Thus, a clear definition is required to include cities

that may be overlooked. Conversely, there are also cities called "smart" that should perhaps not fall under the same label as others. More and more cities are claiming the label of smart city, but according to Hollands (2008), they do so without actually defining what a smart city is. According to Allwinkle and Cruickshank (2011), many of the cities claiming to be smart do so on the grounds that they make use of some form of ICT. Allwinkle and Cruickshank (2011), however, do not believe that this is enough to constitute the title.

Nam and Pardo (2011) argue that the use of cutting edge ICT alone is not enough to qualify a city as smart. Many definitions neglect the smart use, management, adoption and policy that enables smart ICT to contribute to the smart city (Nam & Pardo, 2011). Nam and Pardo (2011) therefore define a smart city as a city "with a comprehensive commitment to innovation in technology, management and policy" (p. 185).

Caragliu et al. (2011) hold that a definition for smart cities should also include the focus areas or systems which the smart city will influence. This sentiment is reiterated by Dirks et al. (2010) who explain that making a city's core systems smarter will result not only in better systems, but improved competitiveness as well due to the city becoming more attractive to skilled and innovative people who in turn will stimulate economic growth.

Washburn et al. (2009) state that a smart city is a city which applies Smart Computing technology to the seven critical infrastructure components and services. These seven infrastructure components and services are: healthcare, public safety, education, transportation, housing and real estate, city administration and utilities (Washburn et al., 2009).

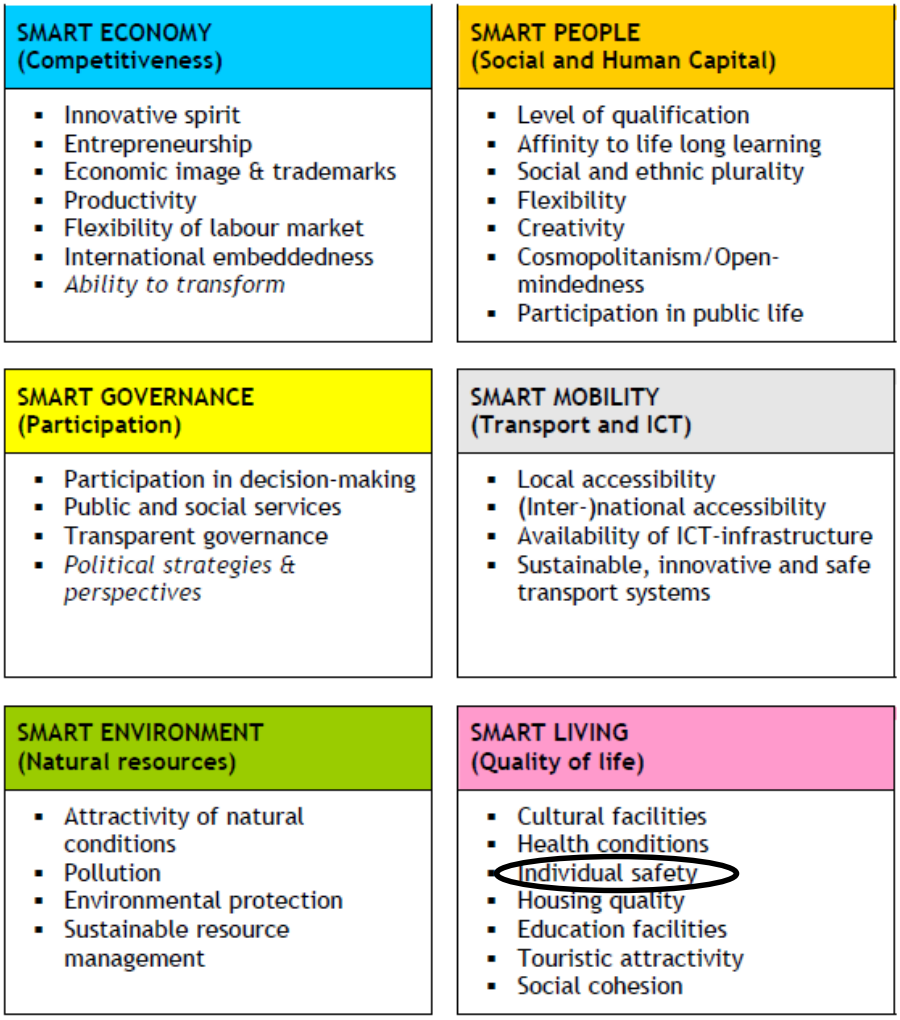


Figure 2.2: Characteristics and Factors of a Smart City (Giffinger et al., 2007)

Caragliu et al. (2011), however, choose to follow the six dimensions as identified by the Centre of Regional Science at the Vienna University of Technology for the identification and qualification of a smart city. These dimensions are listed as a smart economy; smart mobility; a smart environment; smart people; smart living and smart governance (Giffinger, et al., 2007). The areas are compared in Table 2.1.

Table 2.1: Smart City Contribution Areas

Authors	Washburn et al. (2009)	Giffinger et al. (2007)
Smart city contribution areas	Healthcare	Smart living
	Public safety	
	Housing and real estate	
	Education	Smart people
	Transportation	Smart mobility
	City administration	Smart governance
	Utilities	Smart environment
		Smart economy

Caragliu et al. (2011) explain their preference based on the firm grounding of these dimensions in the traditional regional and neoclassical theories of urban growth and development. The six dimensions are connected to theories of regional competitiveness, cost and economics of transport and ICT, natural resources, human and social capital, quality of life, and the participation of society members in cities. Giffinger et al. (2007) depicted the dimensions, their associated theory and sub-areas as shown in Figure 2.2. Table 2.1 shows how most of the areas reported are related, barring one: the smart economy.

It is, however, still difficult to come to one all-encompassing yet succinct definition for such a broad topic as the smart city. Bhana et al. (2013) reiterate the importance of viewing the smart city as one large organic system, as is also emphasised by Nam and Pardo (2011). This means that other parts of the smart city, such as: smart technology used, stakeholders, goals and areas of influence should all be considered. This allows for even greater variety in the smart city systems that can be implemented in various ways and places causing further complication in a singular all-encompassing definition. Bhana et al. (2013) therefore rather than defining the concept chose to conceptualise the smart city as depicted in Figure 2.3 incorporating all the aforementioned components.

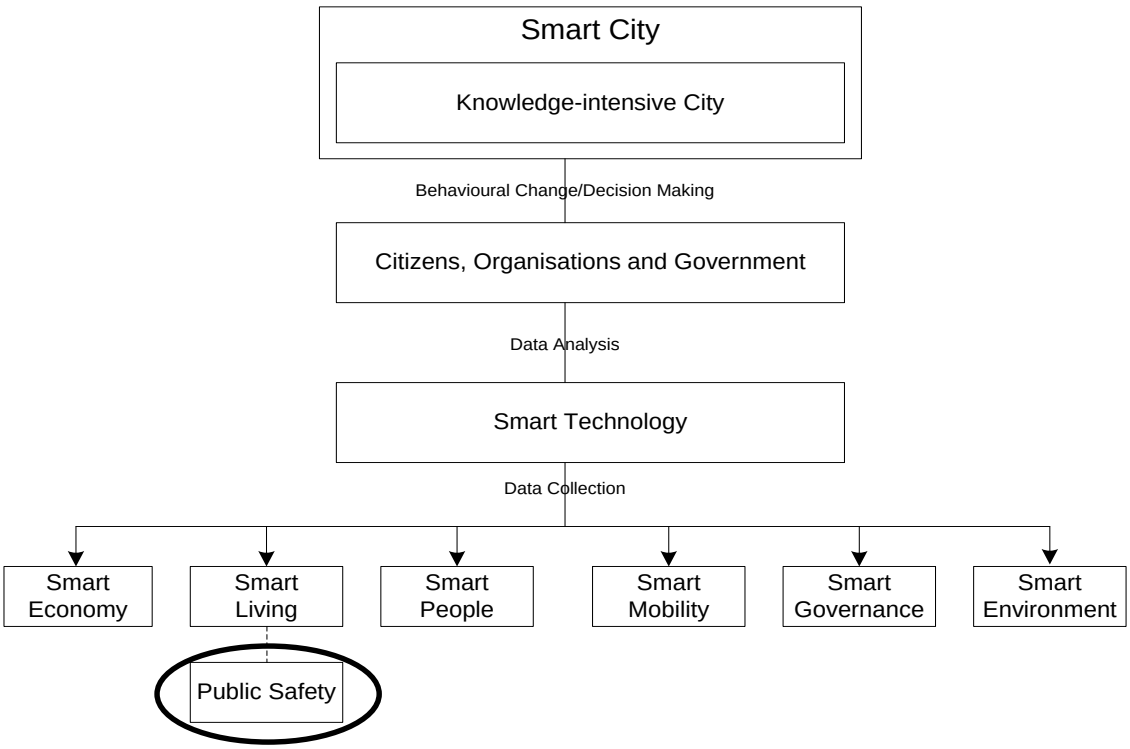


Figure 2.3: Conceptualisation of a Smart City (Bhana et al., 2013)

The six axes or dimensions of a smart city as given by Giffinger et al. (2007) can be seen at the bottom of Figure 2.3 with the addition of public safety as a sub-category within Smart Living. Public safety has been highlighted in this way as it is the focus area of this study. More specifically, this study is focused on the smart city public safety project prototype run by the University of Fort Hare in East London. The model developed in this study will be assessed and refined through application to the prototype and its reports. The rest of this chapter therefore focuses more closely on public safety and how other smart city systems have influenced public safety.

2.4 Public Safety within the Smart City

Smart Living as a characteristic of the smart city relates to the lifestyle of the citizens by addressing basic living needs. Smart Living includes the city's healthcare, safety and infrastructure such as entertainment facilities and housing (Giffinger et al., 2007). Public safety forms part of the Smart Living dimension of a smart city as can be seen in Figures 2.2 and 2.3. Public safety can be defined as any incident which creates or has the potential to create harm to citizens. Generally, public safety involves emergency response and crime reduction. This includes not only good policing, but also emergency services and disaster response. Therefore, public safety does not only involve criminal activity, as many people assume, but also forest fires and floods as well as road and infrastructure damage. Thus emergency units such as police and firefighters need to be able to respond to public safety incidents faster and over larger, more densely populated areas. This study will ultimately focus on a public safety smart city project conducted in the city of East London which is situated in the Eastern Cape Province of South Africa.

2.4.1 Public Safety in East London

East London is governed by the Buffalo City Metropolitan Municipality (BCMM). BCMM has an existing Department of Public Safety which forms part of the Directorate of Health and Safety (Department of Public Safety, 2014). The Department of Public Safety comprises three divisions:

1. Traffic and law enforcement
2. Fire and rescue services

3. Disaster management

Figure 2.4: East London (Google Maps)

The division of fire and rescue services has three subdivisions: operations, fire prevention, and training. The following fire and rescue services are provided:

Operations division	Fire prevention division
• extinguishing fires • rescuing people involved in motor vehicle accidents, high-angle rescues, diving and white water incidents • providing special services, such as pumping, fuel leakages • controlling and cleaning up hazardous materials	• inspection of buildings • inspection and registration of flammable liquid installations within area of jurisdiction of Buffalo City Municipality • fire and arson investigations • emergency evacuations • maintaining of fire water supply equipment • liaison with architects and fire engineers in respect of building plans • enforcement in respect of non-compliance with municipal by-laws, SABS regulations and the National Code of Practice

Note the training division conducts all internal and external training courses based on the services listed above.

BCMM has a comprehensive set of plans to deal with a variety of emergencies through its disaster management centre. Some of the disasters experienced to date include floods, fires, storms, shipping and aircraft accidents, and some incidents involving

hazardous materials. The disaster management division steps in when a community is affected by an event which it cannot cope with on its own. This division provides the following services:

Disaster Management Services
• risk and vulnerability assessments • prevention and mitigation • preparedness • response, relief and recovery • rehabilitation and reconstruction of infrastructure

Within the division of traffic and law enforcement, the traffic services enforce laws with regard to the National Road Traffic Act. Traffic services conduct learners and drivers tests and issue the appropriate licences as well as the registration and licencing of roadworthy vehicles. The law enforcement division is tasked with enforcing municipal by-laws and guarding municipal assets. This is separate and independent of the national police service. Functions of the division of traffic and law enforcement include the following:

Traffic services	Law enforcement (Metro Police)
• education about traffic safety • the erection and maintenance of road signs and surface markings • traffic law enforcement • issue of learners and driving licences • vehicle registration and licensing	• enforcement of municipal by-laws • protection of municipal assets • crime prevention

The South African Police Service (SAPS) respond to emergencies and crime on private property. The SAPS provide a range of services beyond crime prevention and the investigation of cases.

South African Police Service (SAPS)	
Prevention and investigation of: • crimes • missing persons • sexual offences	Administration of: • firearm and liquor licences • police clearance certificates • certifying documents • application of protection orders

It is evident that the SAPS generate a large amount of forms and other documentation in the execution of their duties. Much of this is still by manual processing of hard-copy documentation.

The information presented shows that BCMM has systems and responsibilities regarding public safety and public safety responders. It becomes evident that there are two major areas where smart city principles can bring great improvement. (1) The departments reported operate independently, keeping data and information sources separate with little interdepartmental communication. (2) The second point of concern is the lack of automation and digitization of information within these departments. Other smart city endeavours can serve as examples in possible improvements, thus the following section will explore some case studies from other locations.

2.5 Smart City Public Safety Case Studies

Many smart city public safety systems are looking to faster communication technologies to deliver information to police and fire departments closer to real time. This section will note how other cities in various countries have endeavoured to use information and communication technologies to improve public safety. The cases are then searched for an indication of which information may be useful for improving public safety in East London with additional focus on the goals and objectives which should be set for such information.

2.5.1 New York

In New York City, USA, crime was reduced by 27% through the application of smart city principles. Data is stored in a single central location which provides instant access to information to all officers (Washburn et al., 2009). The New York Police Department (NYPD) make use of near real-time dashboards that display a consolidated view of the reports to be addressed and the resources available to do so simultaneously (Washburn et al., 2009). This resource is accessible to the police officers through any Internet access point including mobile devices and public computers. The NYPD can thus manage its resources more efficiently and effectively with faster reaction times in emergency response situations.

The case of the NYPD exemplifies the benefits of having a single, up- to- date repository of information for all users to work from. This avoids miscommunications and information transfer delays by using a single source of data. Also to be noted is the mobility and flexibility of access to the information.

2.5.2 Perth

A large city in Western Australia, Perth reportedly spends a significant amount on policing and servicing, yet suffers from increasing crime rates in certain areas (Cozens & Grieve, 2011). Cozens and Grieve (2011) emphasize that a great deal of this can be attributed to the fact that city planners have little knowledge of crime patterns. In order to address this shortcoming, the city embarked on the gathering and analysis of applicable data. Efforts were focused on specific high crime areas after identifying the more "dangerous" places by initially focusing data gathering on specific areas or suburbs in order to isolate areas with higher crime rates. It was found that Northbridge was perceived as a particularly violent place after dark. Northbridge has a busy night-time economy made up of shops, restaurants and multiple nightclubs (Cozens & Grieve, 2011).

Crimes involving the area reported over a five year period (2005-2009) were examined and an increase of 71% in personal offences was discovered (Cozens & Grieve, 2011). It was also found that 75% of these incidents occurred between 08:00 p.m. Friday and 08:00 a.m. Sunday, confirming the hypothesis that most of the incidents were related to the weekend's night-time economy operating in the area (Cozens & Grieve, 2011).

Ultimately, different interpretations of the data were reported due to differing perspectives and motives. A police report focused on the concentration of incidents within certain time frames and thus recommended that licenced premises should be restricted to shorter hours of trade. The Western Australian Nightclubs association, however, focused on spatial concentration of incidents, recommending that the specific venues with the highest incident rates should be targeted for intervention. Specificity of the data must thus be considered. Data pertaining to an area or suburb can identify a general location with a high rate of incidents to focus on.

An important point highlighted by this study is that the analysis and findings of the data can be affected by factors such as the goals and purpose of the information. This study also shows that the time that incidents occur can be used to correlate patterns by itself or in conjunction with other information. Lastly, it is evident that more specific location data is important in order to identify exactly what size of area requires resources and/or services. Despatching resources to larger areas than required will be wasteful if the area analysed shows that a series of incidents are in fact concentrated in one small section.

2.5.3 Memphis

The Memphis Police Department (MPD) was experiencing a lack of success in stemming an escalating wave of crime using traditional policing methods. Being further constrained by ever tightening financial budgets, Larry Godwin, director of police services for the city of Memphis realised the need for more intelligent policing practices (IBM, 2011). The aim of the MPD was to focus their patrol resources more intelligently. In order to do so, they had to identify trends in crimes using a predictive tool that would allow precinct commanders to change tactics and redirect patrol resources with a twofold aim: 1) to prevent more crimes before they happen, and 2) to catch more criminals in the act. Through implementing these smarter approaches in their policing, MPD has made life safer for Memphis citizens by reducing the overall crime in the city by 30% (IBM, 2011). One of the contributing factors to the success of the new approach is stated to be the shunning of hierarchical thinking and the recognition of each individual's role in fighting crime, regardless of their position or ranking.

The system is based on a predictive model that combines data from multiple sources ranging from the pre-existing MPD documents and reporting system to video cameras that monitor the streets of the city. An important factor is that current and historical data are used in order to enable fast responses to current incidents but also make long-term changes. The system then identifies patterns within the data in order to reveal underlying crime trends allowing a deeper understanding of long-term factors (such as abandoned housing). It is reported that the key to success of the programme is discerning the difference between patterns that are merely "interesting" and those that are actionable (IBM, 2011). Re-analysing and re-applying of information should be an iterative process as changes in policing can lead to changes in the crime patterns. Police

director Godwin explained that the process can resemble a chess match, but that ultimately this match allows the MPD to make arrests that they would not have before.

IBM (2011) summarise the benefits of the MPD's predictive crime prevention practices as follows:

- 30% decrease in serious crime, with a 37% decrease in one of the specifically targeted areas
- 15% decrease in violent crime
- The amount of cases solved in the MPD's Felony Assault unit has quadrupled from 16% to almost 70%
- An ultimately improved ability to allocate resources in a budget constrained environment.

In this example it is seen that combining historical and current data can contribute to the capacity for predicting future crime trends. By identifying patterns and discerning which are important, crime trends can be identified not only for prevention, but also for gaining understanding into causative factors. Another factor to the success of this endeavour is making the information available to everyone who can make use of it regardless of hierarchies that can cause delays and incomplete information transfer.

2.5.4 Lancaster

The city of Lancaster, California was suffering from a lack of manpower in its public safety efforts. For a cost effective solution, Lancaster decided to use predictive analytics to enable better resource deployment. It was decided to gain a better sense of where crimes had occurred and where they were expected to occur in the future. Using statistical software to develop predictive models to use for crime prevention, the city succeeded in reducing its crime rate by about 35% over a three year period as compared to estimates from 2007 (IBM, 2012). Not only did this improve public safety and citizen confidence, but according to a comprehensive yet conservative Return on Investment (ROI) calculation by Nucleus research, the sheriff's department saved over $800 000 in operating costs in 2010 and 1.7 million dollars in 2011 by implementing their predictive analytics system (IBM, 2012).

The senior analyst employed by Lancaster city combined data from police reports with data from the municipality. Using software tools to analyse the data, he put particular emphasis on geographical data which was used to map the location of each serious crime instance.

A notable finding from this particular case is the initial difficulty experienced in modelling predictive algorithms. Analysts and trainers sought to build on work from other police departments in other cities, but found that predictive models and algorithms were not directly translatable for use in different cities as there was too much variation in causative factors of the crimes. It was found that further modelling was needed to align Lancaster's specific crimes, services and response systems with the software. Data experts aided Lancaster's analysts to transform its data into models that would produce accurate and useful information and would be reusable to model data in the future. Once other factors had been ruled out, it was found that time series analysis fit the best with crime in this particular city. This gave the sheriff's department insight into location and causes of older crimes and also allowed them to estimate how many to expect in the various cities over forthcoming months (IBM, 2012).

A lesson from this case is that there can also be tangible financial benefit in terms of cost saving in addition to the public safety benefits gained from using predictive analytics for public safety. Another lesson here is that there can be subtle differences in crimes in different places that may impact the analysis process. One model may not work everywhere, thus the solution needs to be tailored to the specific case.

2.5.5 São Paulo

The largest police force in Brazil, the São Paulo Military Police (SPMP), is responsible for visible police patrols, anti-rioting and preserving public order. A significant threat to the officers of the SPMP in executing their duties was organised crime syndicates that were able to access the radio system used by the SPMP in all their communications. The SPMP made the choice to switch to digital terminals which make use of encryption keys in order to secure their communications. The result of this change is a 60% reduction in the state's crime rate over a period of five years from 2006 to 2011 (Tait, 2012).

The case of São Paulo's military police highlights the importance of security and access control of public safety information. Allowing the public access to the public safety communication channels can result in the opposite effect of what is desired from a public safety smart city system. In the same way as public safety officers can strategically plan and allocate the use of its manpower and other resources, so too can criminals. Thus, information and communication for improved public safety must be informative and helpful, but it must also be secure and not harmful.

2.5.6 Joliet

Joliet, Illinois in the USA has realised a need for upgrading and extending their video surveillance systems. The Joliet Police Department previously had a standalone video surveillance network that monitored a few specifically chosen locations within the city (Cisco, 2010). A few cameras were monitored 24 hours a day by security officers and the rest of the cameras recorded footage that was then stored in case of future use (Cisco, 2010). This system was extremely time-consuming to manage.

The video tapes had to be exchanged on a daily basis and the tapes that had been recorded on were labelled and archived (Cisco, 2010). The stored footage was used in cases of accident, injury or other crime related incidents in order to assess exactly what happened, but this required someone to locate the correct video tape and review it until the event was found, which was only possible if the event occurred within the camera's field of view. A further delay experienced in managing video surveillance was that there was only one console from which to view the real-time video feed (Cisco, 2010).

The video surveillance system has since been upgraded and new digital cameras are installed as and when funds become available with the aim of providing surveillance throughout the city rather than just in selected areas. The upgrade includes networking all the cameras in order to allow central management of video feeds and recordings on a server located at the police headquarters. The archived recordings and live feeds are now accessible to authorised personnel from almost any location including home or police vehicle. As part of the installation, access to the video feeds was also given to other government agencies.

Public safety has been enhanced as explained by the Joliet Police network administrator: “Video surveillance cameras act as a force multiplier, because we can monitor more areas with the same number of officers. They help reduce crime, and also reduce citizen fear of crime” (Cisco, 2010, p. 3).

2.5.7 Arlington

In the American state of Virginia, Arlington County made public safety a top priority. Of the County’s approximately 200 000 residents, about 40% are below the age of 40 and are well versed in most modern communication technology, and thus were demanding more online government services (Cisco, 2006). The rest of the population, however, are diverse and many have limited network and Internet access or none at all. Arlington thus had to consider the implications of the digital divide and be mindful not to widen it by expanding their e-government services.

A more reliable fibre-optic based network was put in place to facilitate stable communication and data exchange between government agencies starting with their fire stations and then expanding to trade centres, water works and human services (Cisco, 2006). To further facilitate interagency collaboration, the network was extended to the neighbouring city of Alexandria, outside the county boundaries. The network carries multiple formats of data including voice video and text to enable the sharing of criminal databases, Geographic Information System databases, and more (Cisco, 2006).

A key part of the new network is the way in which first responders are kept informed when responding to emergencies – a mobile “crash cart” is sent out to the scene of an incident. The crash cart provides access to all formats of information via whatever the fastest network available in the area (Cisco, 2006). Commanders in any location thus have access to greater situational information and can make better decisions as and when required.

2.5.8 Madrid

The city of Madrid in Spain has invested in building a specialised dashboard for the coordination of public safety resources for the police, fire, highway, hotline and ambulance units. The city decided to focus mainly on improving surveillance by installing video cameras in selected areas around the city. The dashboard draws data

from surveillance cameras, traffic video feeds, map-based GPS data, as well as the status and location of personnel (Kaiserswerth, 2010). The success of this programme is clearly evident from the 18.3 percent decrease in crime recorded in its first year (Gutierrez, 2012).

From all of the case studies mentioned it is seen that time and area information related to the incident is vital. For insights into the situation at hand, one can look to any of these data categories individually or consider them in a combination; both ways can give insights. The importance of a central database has been highlighted as well as the need for all information user levels to have fast and easy access to the information. Thus, the model produced by this study must make allowance for multiple ways of grouping and modelling the data, flexibility in reporting, and minimise the time it takes to generate the required information.

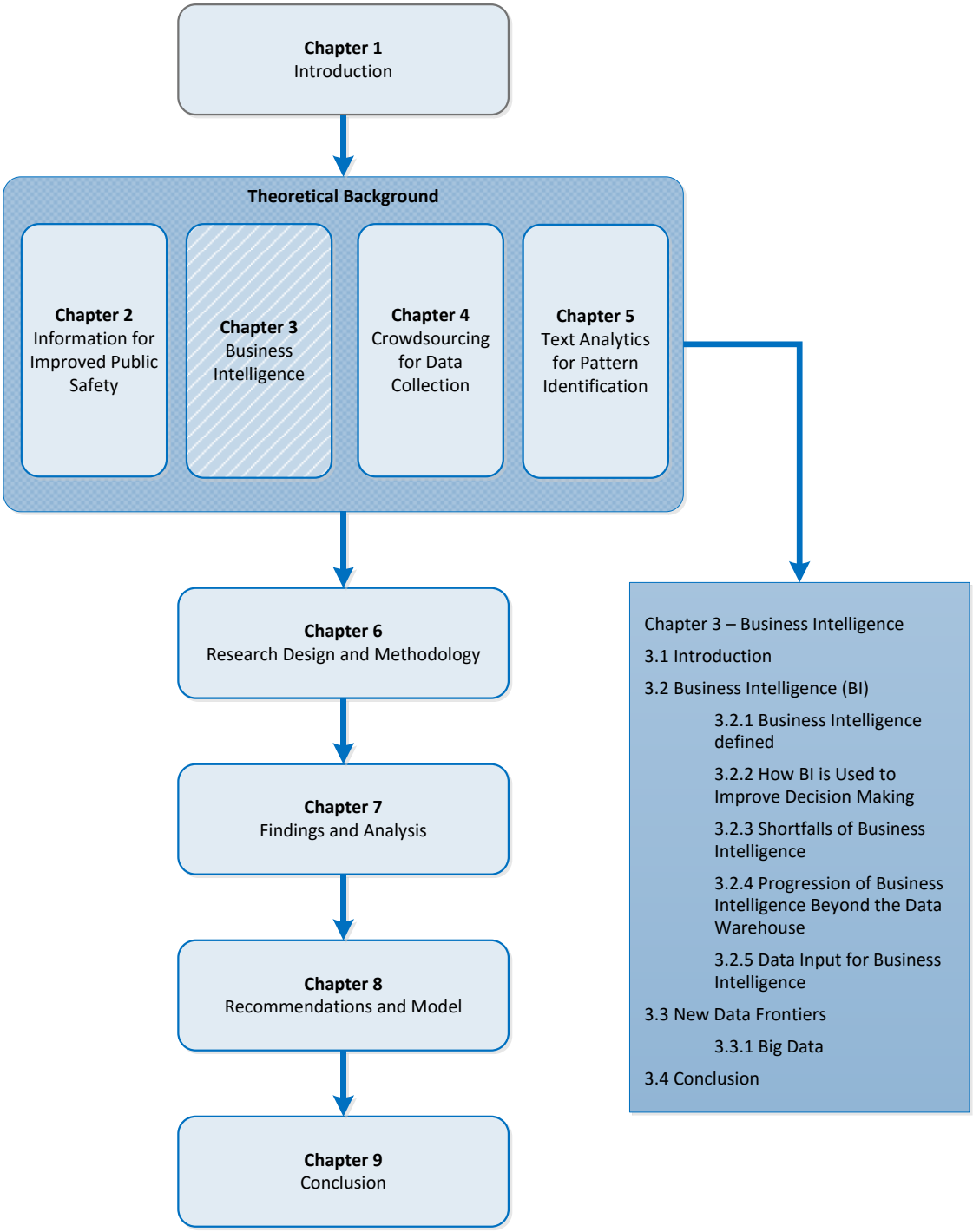
2.6 Conclusion

In order to better understand the analysis environment and goals, this chapter has explored the smart city concept, its six sub areas, and how they can be used to mitigate the problems related to increased urbanisation. These areas include: Smart Economy, Smart People, Smart Government, Smart Mobility, Smart Environment and Smart Living.

Public safety within the area of Smart Living has been an urgent concern in many initial smart city cases. From the summarised case studies described it is evident that increasing public safety and reducing crime by improving communication and/or access to information is indeed possible. In many reported cases, the authorities chose to install or upgrade video surveillance systems or automated sensors in order to gather data. Logically, this is a good option as the data is accurate and detailed, but as seen in Joliet, video data also has negative points including: cost, location and time wastage. Hours of footage can be recorded unnecessarily and often requires someone to search through the footage in order to isolate a specific section of a recording that may apply to a public safety event. It is also difficult to monitor everywhere simultaneously and incidents may occur out of the camera's line of sight.

This chapter also elucidated some of the important requirements of the post analysis data including the use of consolidated, singular, central storage points of information with multiple points of access provided to all authorised users from various levels and departments. For predictive analytics it is beneficial to use historical data in conjunction with current data, but analysts must remain conscious of the fact that different situations may require different models for this purpose. All the cases viewed show that time and location of the incident is important for analysis purposes, but geospatial information should include the greater area (or suburb) as well as a more specific targeted location such as an address, building or business for greatest impact. The next chapter explores current trends in BI in order to see what lessons the smart city can learn from smart business.

Chapter 3 – Business Intelligence



3.1 Introduction

Chapter 2 developed an understanding of the environment in which data analysis is to take place, the smart city, and what the ultimate goals of the analysis process are. To further the planning of the analysis process, one must also understand the nature of the data, how it is obtained, and how it is to be used.

The purpose of Business Intelligence (BI) is to ensure that an organisation's decision makers make well informed decisions to the benefit of the organisation. The same principle applies when considering that the smart city seeks to inform the city's decision makers to ensure improved utilisation of resources. In developing a strategy for data analysis for a smart city, it is therefore important to learn from the current BI environment.

This chapter explores the concept of BI including its past, current and future developments in order to apply lessons learned to the data analysis process for the smart city.

3.2 Business Intelligence

Lönnqvist and Pirttimäki (2006) emphasise that businesses need to receive timely and effective information not only to succeed, but even just to survive in the highly competitive modern economy. Viewing a city and its management as a business, one can relate this to the aim of a smart city. Thus, it can be said that a smart city is a city that utilises BI. One would then posit that traditional methods and models of BI should be applicable when implementing a smart city system.

3.2.1 Business Intelligence defined

As early as the 1970s, applications were employed for aiding in decision making. Watson and Wixom (2007) report that although decision support systems (DSS) first started aiding in decision making in the early 1970's, the term BI only became widely accepted after it was first used in the early 1990s. About ten years later in the late 2000s, the term 'business analytics' was introduced in reference to the analysis component which is so central to BI (Chen, Chiang, & Storey, 2012).

Negash (2004) explains that the term BI has replaced terms such as: decision support, executive information systems and management information systems. The term BI can thus, on a high level, refer to any relevant business information or, more specifically, the process in which the organisation acquires, analyses and disseminates information relevant to making business decisions (Lönnqvist & Pirttimäki, 2006). Negash (2004) goes a step further by combining data gathering, storage, knowledge management and analytical tools for a more comprehensive definition of BI. Hence the ideal of BI is to provide the right people with actionable information at the right time, in the right place and in the right format in order to assist the decision makers with managerial work (Negash, 2004).

More recently, other authors have combined the terms BI and analysis, advocating that analysis must be seen as more than a sub-section of BI (Chiang et al., 2012). Chiang et al. (2012) use the unified term BI&A (Business Intelligence and Analysis) to describe concepts and methods that use information to improve business decision making.

All definitions, however, share a similar focus and are based on analysing data and information. The shared purpose of BI is to help control and guide the data available to the organisation by condensing it into information and intelligence for managerial use (Lönnqvist & Pirttimäki, 2006).

3.2.2 How BI is Used to Improve Decision Making

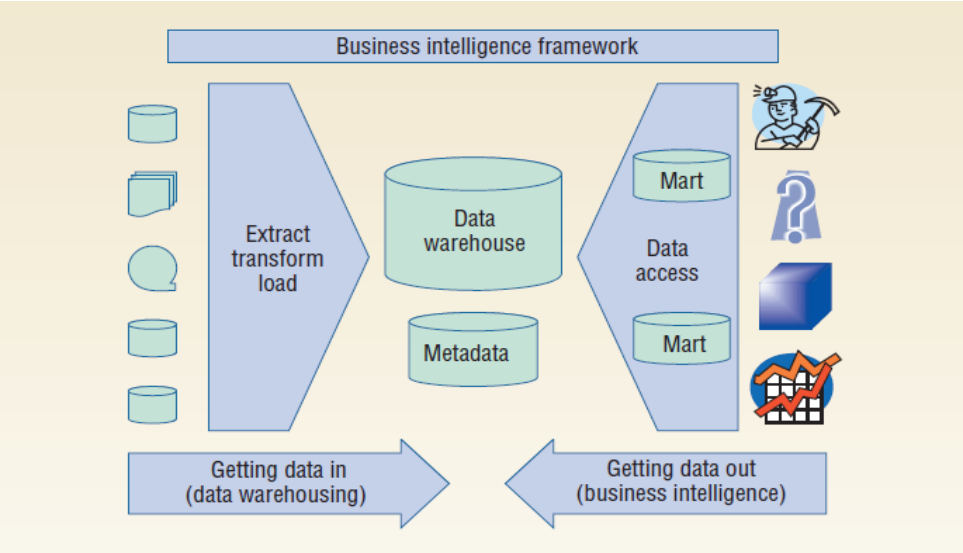


Figure 3.1: BI Framework (Watson & Wixom, 2007)

Historically, BI has been firmly rooted in the database management field relying on various data collection, analysis and extraction techniques (Chen et al., 2012). Watson and Wixom (2007) state that the BI process consists of two main activities: getting data in, also called data warehousing, and getting data out, which is also commonly referred to as BI, as depicted in Figure 3.1.

Irrespective of which definition one may prefer, it is obvious that BI requires input data. Data warehousing (DW), in this context, involves obtaining data from internal or external source systems and moving it to an integrated data warehouse (Watson & Wixom, 2007). This is usually in the form of a commercial relational database where tables are populated and linked to each other (Chen et al., 2012). The DW process involves a DW team who would transform the gathered data into a format which will be meaningful for DSS by matching and grouping related records by creating relations between database tables (Watson & Wixom, 2007). Watson and Wixom (2007) also report that obtaining the data is usually the most challenging part of BI as it can contribute up to 80 percent of the time and effort required, as well as leading to at least half of the unexpected project costs. Costs are generated by hardware requirements, poor data quality, and politics around data ownership (Watson & Wixom, 2007).

BI, in this context, refers to the data being taken from the data warehouse and used by applications and users to perform reporting, querying and predictive analytics for the purpose of enhancing decision making in order to produce business value (Watson & Wixom, 2007). Negash (2004) explains that BI transforms data – first into useful information and then into knowledge through human analysis; a large part of the knowledge extraction process falls within data mining (Mitra, Pal, & Mitra, 2002; Nisbet, Elder, & Miner, 2009; Tien, 2013). The techniques for analysis used in these databases are mostly grounded in statistical methods from the 1970s and data mining techniques from the 1980s (Chen et al., 2012). Current BI, however, is not without shortcomings.

3.2.3 Shortfalls of Business Intelligence

Although the BI concept described above does note that the data input phase can be labour intensive, the reporting and analytics of the output phase is largely left to users. This leaves one to make one of two assumptions: The reports and data that are retrieved

from the DW must be highly standardised and inflexible, or that a rather large amount of analytical knowledge is expected from the information users and decision makers in order to generate any ad hoc reports which they may require. This is perhaps due to differing business needs or a remnant of days past when a limited amount of data sources meant that data types and formats were much fewer, resulting in less work needing to be done prior to the use of the data.

This sentiment is reiterated by other authors who argue that current BI is still falling short in meeting the needs of many businesses (Berthold et al., 2010). Berthold et al. (2010) explain these shortcomings, generally stemming from inflexibility, as follows:

1. There is a lack of focus on the individual business user's needs – analysts and decision makers are often forced to rely on standard reports and predefined analytical content which often does not meet their needs in changing environments.
2. The information often lacks business contexts – business rules, best practices and business goals or strategies often have to be applied to the information after analysis, resulting in the users having to understand the data and the business information, which often needs to be retrieved from other sources.
3. There is often no collaboration – well grounded decisions are often a result of combining the expertise and opinions of multiple analysts. If collaborative decision making is the best method, BI can greatly benefit from combining with social software.

Hawking (2012) considers which characteristics to measure for assessing BI success. Considering a case study of a company developing its BI into a more mature phase, Hawking (2012) identifies the following as key characteristics for BI, which are often not met:

1. Having a single source of truth – all users should draw from the same information to avoid data discrepancies.
2. Business analysis across borders, processes and businesses – analysis as well as data should not be isolated in silos of departments, sections or even organisations.

3. Analysts move from data gathering to real business analysis – analysts should not be limited to working only on getting the data into the DW, but should be involved in the reporting and analysing the information output for business use.

From the above, one can infer that successful modern BI requires flexibility, contextualisation, collaboration, cross borders data and direct data-to-information analysis producing appropriate, timely results. The question must then be raised as to whether the DW architecture is still the best system for BI.

3.2.4 Progression of Business Intelligence beyond the Data Warehouse

During the 2000s, focus started shifting beyond the DW that companies were implementing, usually at great financial expense (Thomsen, 2003). Trends that developed in BI include Real-time BI, Business performance management, and pervasive BI.

Real-time BI has its focus on getting information to those who need it as quickly as possible. An example of this is Continental Airlines who combine data from flight manifests, reservation information, real-time flight data from the planes and current gate and departure times in order to have up- to- the- minute information on possible delays with flights and which passengers will be affected (Watson & Wixom, 2007).

Business Performance Management (BPM) makes use of scorecards and dashboards to summarise and display information regarding the organisation's performance with particular focus on time-critical operational processes (Golfarelli, Rizzi, & Cella, 2004). This information can then easily be compared to goals, benchmarks and previous performance for the purpose of maximising future performance (Golfarelli et al., 2004; Watson & Wixom, 2007).

Pervasive BI, also referred to as information democracy, promotes the spread of BI use to more users, thus enabling them to improve their work (Watson & Wixom, 2007). Developments such as web-based systems, dashboards and event- based triggers help the benefits of BI to spread through various organisational levels from management down to the sales teams (Watson & Wixom, 2007).

These and other BI technologies and applications, such as reporting dashboards, OLAP, interactive visualisation, data mining and scorecards, are considered as BI&A 1.0 and have been incorporated into commercial BI platforms by most IT vendors including Microsoft, SAP, Oracle and IBM (Chen et al., 2012).

Although BI is evolving and has already changed greatly over the last few years, it is evident that some of the pivotal concepts remain the same; timely use of relevant information to improve business value. If this has become universally available, making use of it will not create differentiation for a business, but merely keep up with the others. In the drive for competitive advantage, businesses are seeking to extend their data use to include new data sources that can give them additional insights for decision making. But different data leads to different processes and analysis.

3.2.5 Data Input for Business Intelligence

Although much of the focus has previously been on the data warehouse, Negash (2004) explains that it is only one of multiple systems that can be used as input for BI. As depicted in Figure 3.2, the author also includes Online Analytical Processing (OLAP), visualisation, Data Mining (DM), Decision Support Systems (DSS), Customer Relationship Management (CRM) marketing, Knowledge Management (KM), and Geographical Information Systems (GIS) (Negash, 2004).

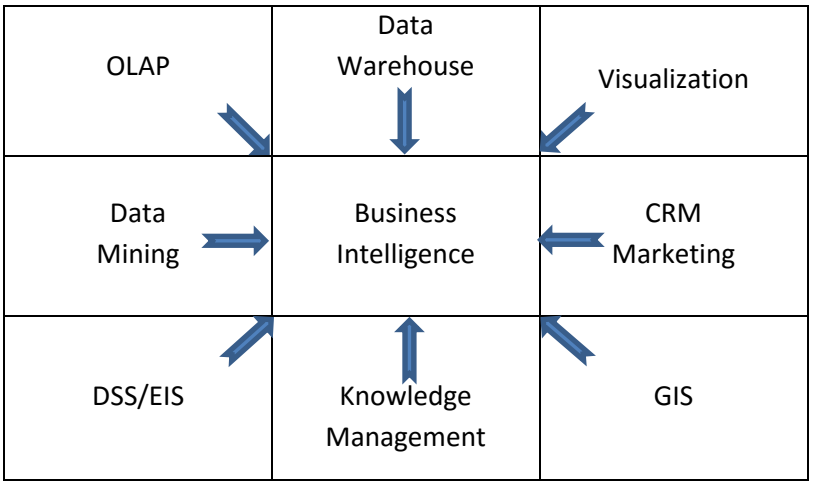


Figure 3.2: Relation of BI to Other Information Systems (adapted from Negash, 2004)

Alternatively, Langseth and Vivatrat (2003) state that the essential components of successful BI are:

- Real-time data warehousing
- Data mining
- Automatic learning and refinement
- Data visualisation

In stating the above, the authors also show that DW, DM and visualisation do not have to stand as individual inputs, as may be inferred from Figure 3.2, but can be combined into parts of one BI system. Moreover, DM is a key component in most forms of BI (Langseth & Vivatrat, 2003). Many other authors agree that analysing data, for any purpose (including BI), naturally involves DM as the largest part of the analysis process (Mitra, Pal, & Mitra, 2002; Nisbet, Elder, & Miner, 2009; Tien, 2013).

Most of the information systems seen in Figure 3.2 do exist in many of the larger organisations today. Small to medium enterprises (SMEs), however, do not necessarily have all of these capabilities at their disposal. Often the choice of which of the above systems to use depends on the data available. Different systems are designed to process different types of data in different formats. In the race to get ahead, most businesses are constantly looking for more data options which can be used to leverage some competitive advantage.

3.3 New Data Frontiers

Businesses are constantly looking for new ways in which to differentiate themselves from their competitors, and the competitors are constantly looking for ways to keep up. A great deal of focus in this regard has recently been on finding and making use of previously untapped data, especially as input for BI.

With modern developments in technology, data has become ubiquitous and cheaper than ever before. Mitra, Pal, and Mitra (2002) explain how modern developments in software and hardware and the rapid digitisation (or computerisation) of business have led to vast amounts of data being collected and stored at an ever increasing rate. A combined 2.5 quintillion bytes of data is created daily by businesses and individuals on topics ranging from the weather to personal opinion to retail transaction records and healthcare (Dinu & Iovan, 2014). Troester (2012) states that large organisations are

inundated with terabytes and petabytes of data, which will in the near future reach storage capacities measured in exabytes, zettabytes and yottabytes. The increased storage capacity of modern hardware combined with the availability of cloud storage has obfuscated the need to be selective in what data to keep when one can keep everything.

The ever increasing expansion of the Internet provides more and more data. Added to this are developments such as the mobile and semantic web as Internet alternatives and automated sensors that collect data continuously (Chen et al., 2012). The rapid expansion is hastened by the ability of individuals to create data freely through social media such as Facebook, Twitter, YouTube, blogging, etc., expanding web content exponentially (Fan & Bifet, 2013). This, added to the use of communication resources with never-ending interaction with information systems of multiple organisations, has given life to the concept of big data (Lebraty & Lobre-Lebraty, 2013).

3.3.1 Big Data

Big data was originally named as such describing data sets that are large enough to require supercomputers for processing it (Boyd & Crawford, 2011). Boyd and Crawford (2011) observe that although the quantities of data available today are immense, the modern desktop computer can handle vast sets of data. This is reiterated by Troester (2012) who explains that the size or amount of data is, by many businesses, not seen as the biggest problem due to the gradual growth in capacity and processing power of available technology which has allowed most organisations to make gradual changes to their hardware and software assets. Yet the use of the term big data persists, as the definition has evolved to include more than just size. The most popular definitions of big data categorise data according to the three V's of big data (volume, velocity and variety), although there is some disagreement among authors whether there should be three or actually four V's, adding Veracity.

3.3.1.1 What Constitutes Big Data?

McAfee, Brynjolfsson, Davenport, Patil, and Barton (2012) explain that the three key differences from big data as compared to other intelligence or analytics are 1) Volume, 2) Velocity and 3) Variety (three V's).

1) Volume: With around 2.5 exabytes of data generated daily since 2012, organiations have the opportunity to work with petabytes of data in one data set. This data is generated not only on the Internet, but organisations continually build internal data, especially larger corporations such as Walmart who generate more than 2.5 petabytes of data per hour from sales transactions alone. Attempts to name a specific range for data to qualify as big data have proven to be futile due to continuous growth, variation in the scope, purpose and sub-division of data, as well as different forms of analytics affecting data sets (Russom, 2011).

2) Velocity: The speed at which data is created and analysed has become a key factor in the agility of an information centric organisation. The drive for greater competitive advantage is bringing organisations ever closer to real-time or near real-time information use, depending on the purpose and nature of the data.

3) Variety: One of the main reasons why big data has become so big is the variety of sources contributing to it. Big data takes many different forms ranging from messages, videos and images posted on social networks to sensor readings, GPS signals from cell phones or sales transactions in stores, among many others. Many of the largest sources of big data are still relatively new. With the 2004 launch of Facebook, soon followed by Twitter and other social media tools, users have been able to create their own content faster than ever before by posting text, pictures and video clips. Add to this the ubiquitous smartphone (and other mobile computing devices) enhanced with Internet access as well as built-in sensors that can generate enormous streams of data associated with people, places and activities by user input and also automatically.

4) Veracity: The data may or may not be trustworthy or uncertain. As would be the case with a lay person's opinion posted online, which may or may not be acurate unless corroborated or substantiated (Vossen, 2014).

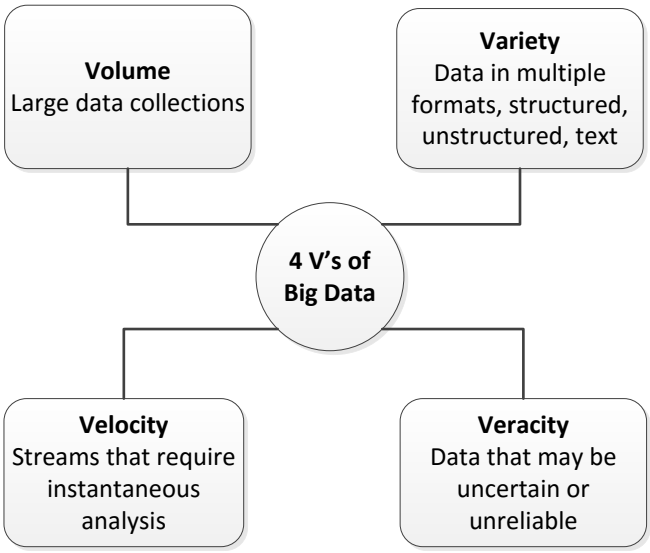


Figure 3.3: The defining 4 V's of Big Data (Vossen, 2014)

Thus it becomes difficult to discern exactly what qualifies data to be labelled big: must big data conform to the extremes of all three or even four V's, or is a data set big data if it conforms to any one of the V's? Are any of the V's more important than the others? One point concerning big data that business users and researchers alike do agree upon is the vast potential it has for application in uses such as BI, analysis of political opinion, epidemiology, industrial trends, emergency response, or the fight against criminality (Lebraty & Lobre-Lebraty, 2013).

3.3.1.2 Big Data and Business Intelligence

The face of BI is changing; big data is causing a rethink as to how things are done. The immense size and diversity of data available is overtaking the technology and techniques available to use it (Lebraty & Lobre-Lebraty, 2013), but it is also providing immense opportunities for business benefits.

Boyd and Crawford (2011) state that the notable aspect of big data is the interconnectivity that is possible, between pieces of data, that can be networked and connected in order to identify patterns leading to new insights about people, their relation to others, groups, or simply about the information itself (Boyd & Crawford, 2011). Boyd and Crawford (2012) state that "Big Data is less about data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets" (p. 663). To some extent this was already being done in the DW and relational databases which could then merely be upgraded with more powerful technology, but big data still

presents further challenges. It has been stated earlier in this chapter that volume is not the greatest issue concerning making use of big data. White papers released by companies including SAS, Boldon James and Intel state that one of the biggest problems in modern data analysis (especially in big data analysis) is working with unstructured data, mainly in text format (Boldon James, 2012, 2014; Troester, 2012).

3.3.1.3 Structured and Unstructured Data

Due to the variety of data sources available in big data, the input data for BI can be in a number of different formats. Data can be separated into two main groups: structured data and unstructured data. Structured data is any data format that fits easily and neatly into a relational database (Negash, 2004). This data is therefore usually in numeric formats and can be seen as quantitative data. Data that does not fit neatly into a database is then referred to as unstructured data; some examples are shown in Table 3.1 (Negash, 2004). Gartner predicted that by 2015, 80 percent of data used in enterprises will be unstructured material from sources such as documents, e-mail, images, video and other texts (Chiang et al., 2012). The focus is thus shifting away from the traditional database and DW practices as these are focused on structured data. The new area of focus is on analysing unstructured data in larger volumes.

Table 3.1: Some Examples of Unstructured Data (adapted from Negash, 2004)

Business interactions	Reports	Communications	Non text format
• Chats • E-mails • Marketing material • Presentations	• News Items • White papers • Research • Phone conversations	• Letters • Memos • Spreadsheet files • Word processing text	• Graphics • Image Files • Movies • Video clips • Web pages

As referred to in section 3.1, most structured data can still be used in the same way it has been for decades; relational databases on modern desktop computers and servers can handle vast quantities of data more than ever before, but the same analysis techniques will still work to transform the data into BI. The new opportunity available is to make use of the largely untapped resource of unstructured data. For the purposes of this study, the focus within unstructured data will be mainly on the text component thereof as it constitutes by far the biggest portion of unstructured data available to businesses. Table 3.1 shows many examples of unstructured data available to

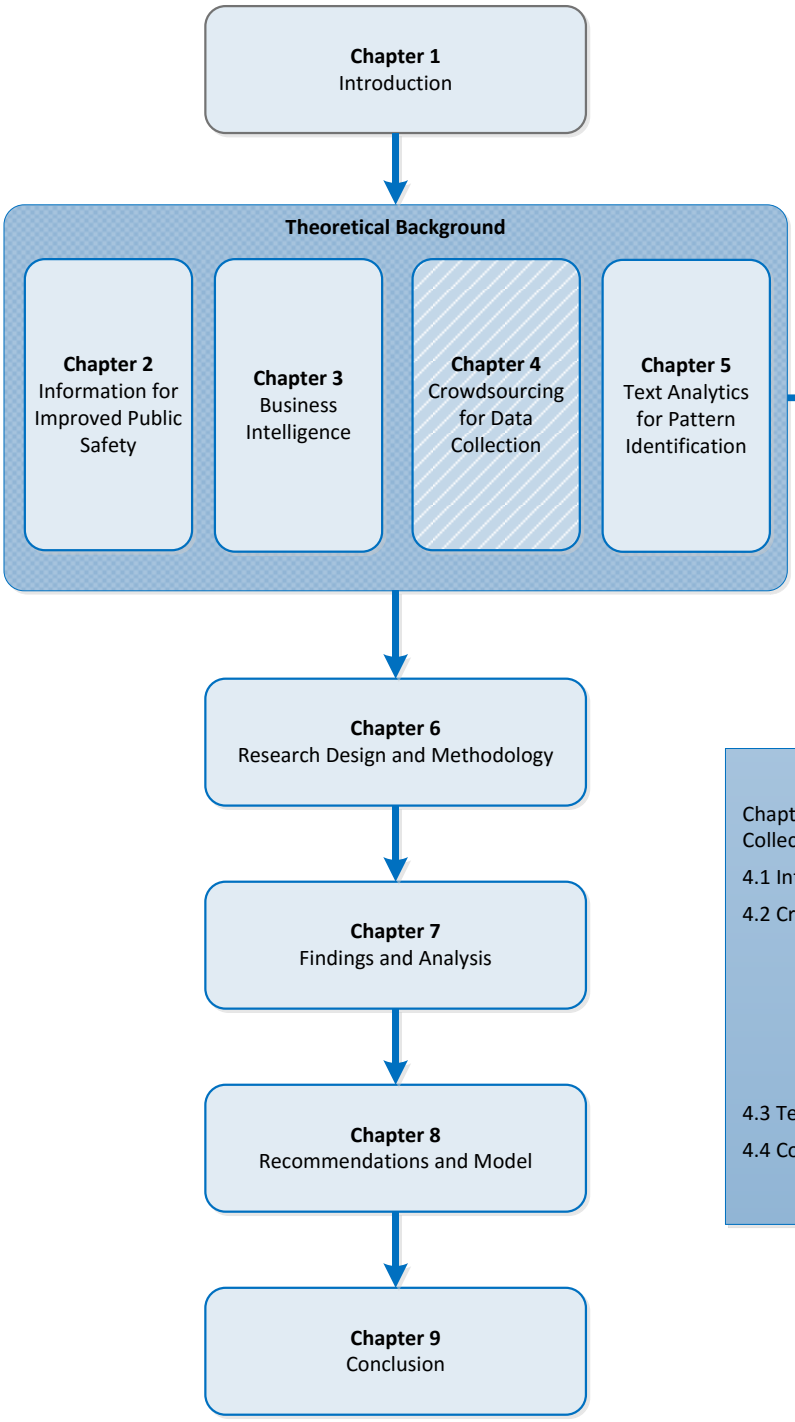
organisations. Though this list is by no means exhaustive, it does show some examples that are already available and accessible within the organisation. The business itself generates a large amount of unstructured text data, but the greater opportunity presented by big data and unstructured data is the access an organisation can gain to the public, customers, experts and generally, the crowd.

3.4 Conclusion

In order to develop an analysis strategy for the smart city, lessons can be learned from BI. The progression of BI is mainly centred on the input data available and how to make use of this data. This includes big data, crowdsourcing and other sources of unstructured data. There are tried and tested methods for modelling and reporting, as well as software from multiple vendors that can be employed to automate these functions for BI, but it is centred on structured quantitative data.

In order to make use of the newer data sources available in the form of qualitative, unstructured data, new methods of analysis will be required, thus pattern identification through text mining is suggested to be the appropriate way forward. In order to inform the analysis (or mining) process further, the data and its method of collection must also be understood. To this end the following chapter explores crowdsourcing as a method of data collection.

Chapter 4 – Crowdsourcing for Data Collection



Chapter 4 – Crowdsourcing for Data Collection

4.1 Introduction

4.2 Crowdsourcing

- 4.2.1 Types of Crowdsourcing
- 4.2.2 Data Generation by Crowdsourcing
- 4.2.3 Crowdsourcing and Business Intelligence

4.3 Text Analytics/Text Mining

4.4 Conclusion

4.1 Introduction

In order to develop an analysis process for text data, it is vital to understand the environment and data to be analysed. Chapter 2 explored the smart city concept in order to understand the requirements, goals and objectives of the information obtained from the analysis process. Chapter 3 aimed to understand the current situation and developments in intelligent decision making or Business Intelligence (BI). This chapter will consider the data which needs to be analysed. To this end, the source of the data (crowdsourcing) must be understood as well as the nature of the data that is obtained in this way. A link will be drawn between crowdsourcing and BI in order to complete the understanding of the entire environment for analysis.

4.2 Crowdsourcing

Since the term crowdsourcing was coined by Jeff Howe in 2006, a multitude of research has been published exploring the different aspects thereof (Yuen, King, & Leung, 2011). Howe (2006) compares the current popularity of crowdsourcing to the popularity of outsourcing in the early 2000s. Where companies were outsourcing labour to countries like China and India, they may now be obtaining it from anywhere and anyone so long as they are connected to a network such as the Internet or a large intranet.

Yuen et al. (2011) attribute the need for crowdsourcing to the fact that some tasks just cannot be computerised. Even with the advancements in modern computing, some tasks that are trivial for humans to do, such as image tagging, Natural Language Processing (NLP) and semantic inference, are extremely difficult or impossible for computers to do (Gao, Liu, Ooi, Wang, & Chen, 2013). Hiring staff to do these tasks on a permanent basis can then become costly. Thus, the attention of a variety of different organisations and businesses has been drawn towards crowdsourcing.

With Wikipedia exemplifying the success possible in harnessing networked users, specialist companies, such as eBay and MySpace, developed with the specific aim to take advantage of a networked world, prove that profitability is definitely possible (Howe, 2006). According to Howe (2006), this has in turn drawn the attention of older-style businesses that now see a new resource which can be leveraged for competitive advantage. Lebraty and Lobre-Lebraty (2013) explain that classic business management

methods have reached the peak of their usefulness and with the extreme environments that organisations are operating in today, efficiency can no longer be improved by using old methods. New opportunities for efficiency and potential customer networking, like crowdsourcing, have become vital (Lebraty & Lobre-Lebraty, 2013).

Crowdsourcing relies on technologies that enable interactive information sharing, collaboration and interoperability (Barbier, Zafarani, Gao, Fung, & Liu, 2012). Barbier et al. (2012) hold that the advancement of Web 2.0 and advent of social media is helping to collect data and spread problem solving across great geographical distances and a variety of cultures faster than ever before. Added to this is the convenience of being able to participate when and where appropriate for the individual while still contributing to the collective (Barbier et al., 2012). Lebraty and Lobre-Lebraty (2013) identify four factors of the crowdsourcing phenomenon:

- Calculation power (computers, smartphones, tablets, etc.)
- Digitized data and information
- A universal network protocol (the Internet, phone network)
- Individuals connected to each other

When these four elements are present, data elements circulate, are copied, changed and enriched (Lebraty & Lobre-Lebraty, 2013). To what end is this done and where can it be applied? The next section will discuss the different types and uses of crowdsourcing.

4.2.1 Types of Crowdsourcing

The crowdsourcing concept described in Howe's (2006) seminal article basically entails taking a specific task or problem and inviting the public or selected groups within the public (experts) to do it. Since then, various other ways of using crowdsourcing have been explored. A number of authors have attempted to classify crowdsourcing, but there is, as yet, no consensus among authors on exactly how to categorise or classify the different forms of crowdsourcing (Brussee, Rovers, Van Vliet, Swart, & Hekman, 2013).

Some authors, such as Kozinets, Hemetsberger, and Schau (2008), consider the characteristics of the intended participant group as a descriptive differentiator. In this view a 'crowd' actively contributes to solving a problem; a 'hive' is where some of the participants contribute more than others while still focusing on innovation; in a 'mob'

some will contribute more than others while focusing on lifestyle and special interests, and 'swarms' have participants that contribute and communicate freely (Kozinets et al., 2008). Other authors such as Oomen and Arroyo (2011) choose to classify crowdsourcing based on the results of the process proposing categories including: correction/transcription; classifications; co-curation; contextualisation; complementing collections, and crowdfunding. Lebraty and Lobre-lebraty (2013) discuss ten types of crowdsourcing according to the function and result of the crowdsourcing system as listed in Table 4.1.

Table 4.1: Forms of Crowdsourcing (adapted from Lebraty & Lobre-Lebraty, 2013)

Name	Description	Example
Crowdjobbing	Specific tasks are "outsourced" anonymous members of the crowd.	MechanicalTurk CloudCrowd Click Worker
Crowdwisdom	Questions are put to the crowd and a 'referendum style' group answer is deemed more accurate as the majority opinion is represented.	Voting online for pageants and game shows. Lulu ThreadLess
Crowdfunding	A project initiator receives funding from the crowd.	Kickstarter Kisskissbangbang Mymajorcompany
Crowdsourcing and forecasting	The crowd weighs in on decision-making. Can replace surveys and market research.	Qmarket Inkling
Crowdsourcing and innovation	Generation of new ideas or developments.	Innocentive InnovationExchange
Crowdsourcing and authenticity	Seeking to understand from the crowd current feelings and tastes regarding a product, brand or an organisation.	Eyeka Lionbridge
Crowdauditing	Analysis of business data is done by the crowd (related to the Open Data movement).	Open Data Paris Data Publica Extractive Industries Transparency Initiative
Crowdcontrol	Obtaining information from informers and giving information to potential victims in the crowd (i.e., outsourcing of surveillance and security activities).	Kaspersky Anti-cybercriminalité
Crowdcuration	Outsourcing the sorting and grouping of data according to specific subjects.	Wikipedia
Crowdcare	Allowing the crowd to assist people in need.	*ArrêtCardiaque* (Heart Attack)

Lebraty and Lobre-Lebraty's list of crowdsourcing forms as seen in Table 4.1 is quite comprehensive, but some of the types listed are quite similar in functioning and seem to be differentiated only by subject matter. It is the view of the researcher that these types are too specific and can rather be grouped according to whether they are used to generate data or to analyse it, under the headings data generation and data analysis as depicted in Figure 4.1 (note that crowdfunding can be part of either, but is closer to data generation as it generates funds).

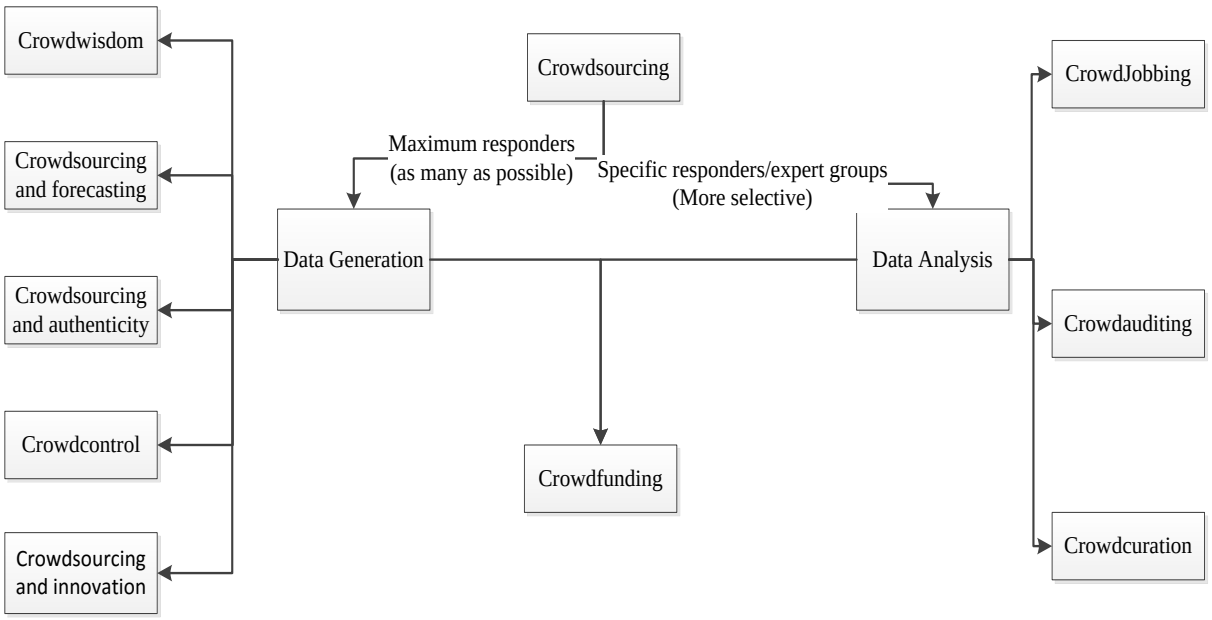


Figure 4.1: Grouping of Crowdsourcing Functions

4.2.2 Data Generation by Crowdsourcing

Ganti, Ye, and Lei (2011) explain that the Internet of Things (IoT) is leading to the emergence of mobile sensing as an important method for gaining data from the crowd. Everyday computing devices with integrated sensing capabilities are constantly connected to the Internet, generating a steady stream of data. Srivastava, Abdelzaher, and Szymanski (2012) discuss largescale networked data generation under the heading of sensing which they divide into three categories: participatory, opportunistic, and human-centric sensing. Bhana et al. (2013), however, feel it more appropriate to group these categories in terms of crowdsourcing under the headings of opportunistic crowdsourcing and participatory crowdsourcing as defined in Table 4.2. This

corroborates Muller et al. (2011) who trace the development from sensing applications to participatory sensing applications to the more state of the art crowdsourcing applications such as OurMaps that incorporates socialmedia tools in order to allow people (the crowd) to actively participate by posting and commenting online.

Table 4.2: Types of Crowdsourcing (Bhana et al., 2013)

Opportunistic crowdsourcing	**Participatory crowdsourcing**	
Crowds as targets of sensing	**Crowds as sensor operators**	**Crowds as data sources**
	Definition:	
(i) Sensing technologies are deployed to monitor individual or group behaviours, activities, and trends. (ii) Can be conducted with or without (acceptable if no private information about the user is collected) permission from the users.	(i) Users use the sensor device to collect data on their surroundings. (ii) Type of data collected is limited to data that can be quantifiably measured and does not require human interpretation.	**(i) Humans collect and disseminate data without the use of sensor devices.** **(ii) Data is usually provided based on human interpretation or background knowledge or experience.**
	Example:	
YouTube indicates how many people have watched a video by displaying the number of users who opened the video link.	Traffic congestion can be calculated by using users' GPS and speedometer, usually through a smart phone.	**Mugging, since it cannot usually be identified through a sensor device.**

With crowds as targets of sensing as well as the crowd as sensor operators, the data gathered comes in the form of sensor readings and aggregated counts, which results in numeric and quantitative data. This quantitative data may be somewhat useful for BI, but does not pose a new insight as it will slot neatly into traditional BI structures.

Participatory crowdsourcing can alternatively be used to obtain more informative qualitative data from the crowd. This concept is illustrated by Bhana et al. (2013) who describe a public safety smart city system which makes use of participatory crowdsourcing. The system allows the citizens of a specified location, the East London area in this case, to phone in and report public safety incidents via an automated IVR (Interactive Voice Response) system (Bhana et al., 2013). The system records the messages which are then transcribed into text files, ready to be analysed for the purpose of informing the decision making of the local municipality in terms of public safety resources. In this case, more informative data is obtained relatively cheaper by making use of ubiquitous equipment without needing purchase or installation. The drawback is

that this data does not fit neatly into current BI and database systems as it is unstructured, qualitative, and in a natural language format.

4.2.3 Crowdsourcing and Business Intelligence

Concerning BI, it has been established earlier in this chapter that obtaining information is usually the most expensive part of the BI process in terms of time and money. Added to this is the fact that businesses are looking for new sources of information as many of them are using their existing data sources fully. BI can thus 'outsource' the data gathering process to the crowd in order to reduce the expense of this task. Obtaining data from the crowd will also ensure more accuracy of the data (crowdwisdom), enable the business to do more accurate customer based forecasting (crowdsourcing and forecasting), and enable the organisation to better understand, communicate with, and broaden its customer base (crowdsourcing and authenticity). Therefore crowdsourcing is a highly appropriate way of gaining data into the organisation, but this concept is not without limitations.

Concerns in using crowdsourcing as data input for BI include (Schenk & Guittard, 2011):

- incorrect information submitted maliciously or unknowingly
- changes in opinion of participants may also skew data
- lack of knowledge or expertise on topics in question
- loss of management control
- closed questions may limit innovativeness in the answers
- ethics and ownership of data

Logically, these concerns can be mitigated as follows:

1. Open the call to as many respondents and responses as possible, which will cause the incorrect or manipulated input to be outliers and thus not skew the results.
2. Initiating the crowdsourcing responses by asking a strategically chosen question, which is simple enough for the majority of people to understand and answer, will allow for keeping management control by question selection and wording, and keeping the pool of participants as large as possible to ensure enough responses for accurate analysis.

3. Open-ended questions will allow for maximised creativity/innovation and descriptiveness in the answers, leading to maximum value from responses.
4. Making use of participatory crowdsourcing relies on respondents choosing to volunteer responses freely and knowingly. This can be incentivised to increase participation.

This is not an entirely new concept. Saxton, Oh, and Kishore (2013) refer to this crowdsourcing model as "Knowledge base building" as shown in Table 4.3. Saxton et al. (2013) explain that using the crowd to generate data allows for variety, collaboration and relationship building with the crowd. Additional business benefits mentioned include low costs to generate the data, as well as not having to enter into contractual agreements with the data providers as they are not employed and remuneration is up to the discretion of the individual business.

Table 4.3: Crowdsourcing Business Model (Saxton, Oh, & Kishore, 2013)

Business model	Services and/or products being outsourced	Role of community users	Level of collaboration	Compensation schemes	Escrow
Knowledge base building model	Encyclopaedia, building data, business trend watching, crisis (incident) information mapping, prediction(e.g., Emporis.com, Knol.Google.com)	Author of encyclopaedia, building data reporter, business trend spotter, crisis (incident) reporter, forecaster	High	Low ($0–$100)	Low

It becomes evident that there are correlations to be found between the concerns of BI and the characteristics of big data. When considering the shortfalls of traditional DW based BI and the characteristics of using participatory crowdsourcing, it becomes evident that great improvement is possible. The comparison is summarised in Table 4.4.

Table 4.4: Business Intelligence Shortfalls Addressed by Crowdsourcing

Traditional DW Approach	Crowdsourcing approach	BI shortfall addressed
Expensive equipment	Less equipment required (done online)	Cost
DW team	Analysis team	Flexibility, accessibility
Find right data	All data available (can ask specific question)	Flexibility, context, accessibility
Data is limited to exact figures	Data is rich and descriptive	Flexibility, Informative

As evident from Figure 4.1, crowdsourcing can also be used for the next step in BI, the analysis of data. Through crowdcuration and crowdauditing, the crowd (or targeted groups within it) can be tasked with (crowdjobbing) sorting, grouping and analysing data for use in strategic informed decision-making. The use of crowdsourcing for this purpose, however, is fraught with risk and complications for many businesses. Ethical and privacy concerns, size of the task, trust and difficulties in finding enough qualified and willing participants are listed as issues relating to analytical crowdsourcing tasks (Lebraty & Lobre-Lebraty, 2013). All of these concerns are in most cases too strategically important, complicated or costly to mitigate, and therefore make this option too uncertain and unstable for most businesses to implement as part of the BI process. The concept does show promise and is a worthy topic for future research, but for the time being, it is safer to keep analysis in-house, especially when already relying on crowdsourcing for data gathering.

Acknowledging that a great deal of BI has historically been focused on structured data, the DW and various database functions, the new directions and opportunities presented by crowdsourcing and particularly by qualitative crowdsourcing requires a different direction in information processing for use in strategic decision making. Considering that the key data to be exploited is unstructured, qualitative data, the next logical step after collection, the analysis, requires different methods.

The development of BI to a point where quantitative data is no longer enough has led it towards exploring qualitative text inputs. Technologies have evolved to support big data and specifically crowdsourcing, sensing and participatory crowdsourcing of qualitative data. This progression has drawn the concepts towards each other in a natural evolution, but they do not yet fit together. This is corroborated by Chen et al. (2012) who surveyed BI&A developments, use and current research to find that BI&A is moving towards the use of sensor based and unstructured content, applied in a variety of business spheres with developments towards topics including (big) data analytics and text analytics among others. The findings of Chen et al. (2012) are depicted in Figure 4.2. Thus, drawing from Figure 4.2, a link can be drawn from unstructured content (BI&A 2.0) to security and public safety to text analytics.

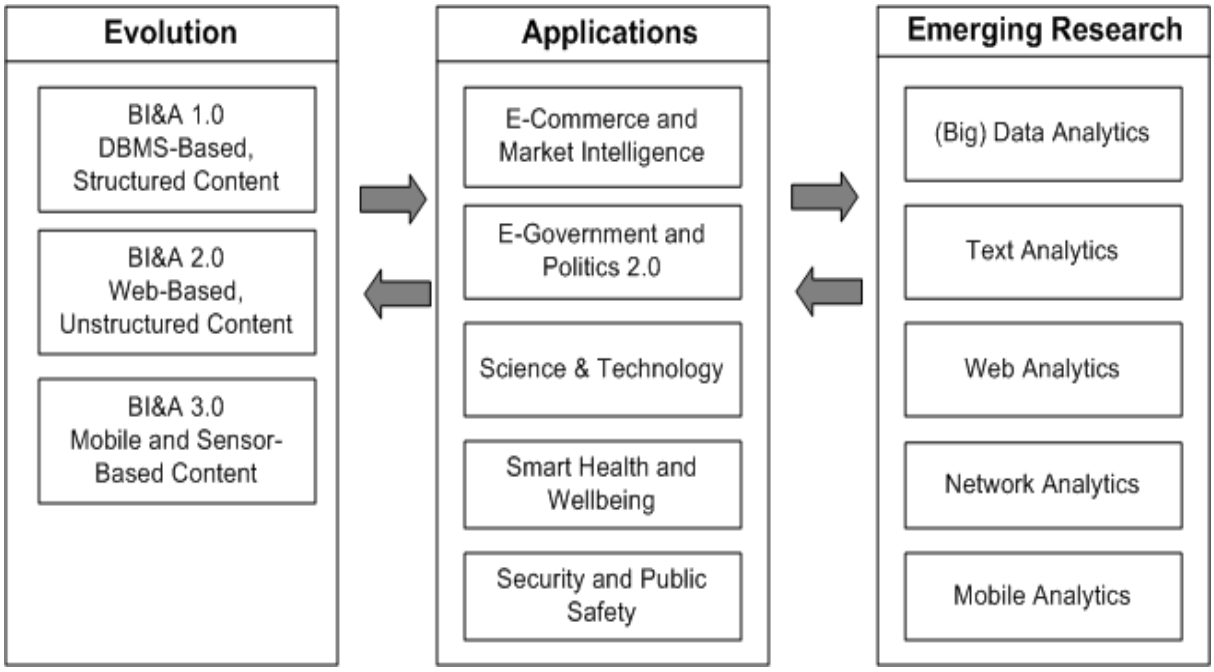


Figure 4.2: BI&A Overview: Evolution, Applications and Emerging Research (Chen et al., 2012)

4.3 Text Analytics/Text Mining

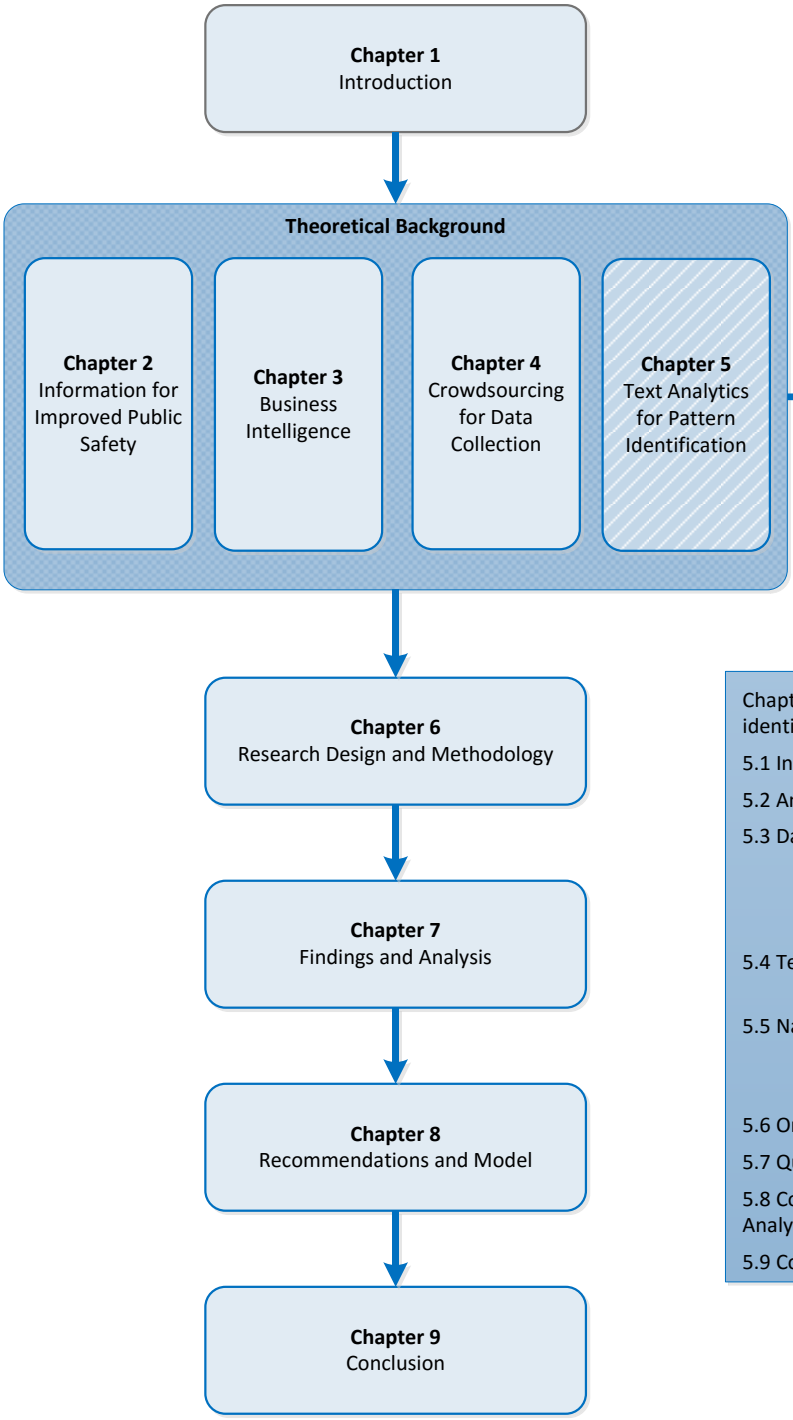
Maimon and Rokach (2010) define text mining as the discovery of new information by automatic analysis of textual resources. Text mining (TM), also referred to as knowledge discovery from text (KDT), is the name given to the process of identifying interesting patterns in large-scale text data for the purpose of discovering useful information (Maimon & Rokach, 2010). Maimon and Rokach (2010) further explain that text mining applies the same analytical functions as data mining, but also applies analytical functions with natural language and information retrieval techniques. It is explained that the text mining process is used initially to extract facts and events from the text sources, after which more traditional data mining and analysis techniques can be applied for further exploration (Maimon & Rokach, 2010).

4.4 Conclusion

Participatory crowdsourcing can be used to generate and gather informative qualitative text data to use as input data for BI. Considering recent developments in data analysis and use, especially the trending topics of big data, sensing and crowdsourcing, this chapter has identified unstructured text data as being the next logical resource that can be exploited for deeper insights into strategic decision making and customer communication. This is exactly the aim of the public safety smart city project. The project makes use of participatory crowdsourcing in order to obtain rich qualitative data. This data can be used to gain deeper insights in trends and patterns related to public safety, but BI techniques require the input of structured data and the data in question is unstructured.

Making use of unstructured data requires a shift away from the traditional ways of database use as implemented in the data warehouse. The data must be searched for patterns, trends and relationships which will reveal information which can be used for traditional BI functions such as forecasting, planning and general strategic decision making. The large amounts of unstructured text data that can be obtained via participatory crowdsourcing will have to be reduced or condensed and structured in order to be used for BI through the process of data mining or, more specifically, text mining. The following chapter, therefore, explores text analytics with specific focus on how it can be used for pattern identification.

Chapter 5 –Text Analytics for pattern identification



Chapter 5 – Text Analytics for pattern identification

5.1 Introduction

5.2 Analytics

5.3 Data Mining

5.3.1 The Cross-Industry Standard Process for Data Mining (CRISP-DM)

5.4 Text Mining

5.5 Natural Language Processing (NLP)

5.5.1 Semantic Analysis of Public Safety Reports:

5.6 Ontology

5.7 Qualitative Data Analysis (QDA)

5.8 Computer Assisted Qualitative Data Analysis Software

5.9 Conclusion

5.1 Introduction

Participatory crowdsourcing can be a source of rich unstructured qualitative data. This is a viable form of data for use to improve public safety in a smart city. Chapters 2 through 4 discussed applicable literature and theory in order to contextualise this study. First, by exploring the concept of a smart city and the information it would need in order to improve public safety within it. This was followed by a discussion on the method of obtaining the data (participatory crowdsourcing) and the resulting format of data obtained via this channel. The next step is how to get the data to the point where it can be used for improved and more proactive, better informed decision making in the smart city as per business intelligence strategies.

This chapter, therefore, attempts to answer the question of how to use analytics for the identification of patterns and trends in unstructured text data. This chapter illustrates a standardised process to follow when performing data mining (Crisp-DM), and then considers the implications of mining unstructured text data. To achieve this, the chapter traces a path from analytics back to the raw data, considering analytics, data mining, text mining, natural language processing and semantics. Qualitative data analysis techniques and automation thereof will also be explored.

5.2 Analytics

According to Delen and Demirkan (2013), "Analytics facilitates the realisation of business objectives through reporting of data to analyse trends, creating predictive models to foresee future problems and opportunities and analysing/optimising business processes to enhance organisational performance" (p. 361). The taxonomical view of analytics shows three main categories: descriptive, predictive and prescriptive. This is illustrated in Figure 5.1.

Descriptive analytics, sometimes also referred to as business reporting, makes use of data to understand current and past events. Doing so involves simple standard/ periodic reporting, ad hoc reporting or dynamic reporting such as OLAP (online analytical processing).

Predictive analytics applies mathematical techniques to data in order to discover explanatory and predictive patterns (trends, associations, etc.) representing relationships between data inputs and outputs. The process involves data mining, text mining, web/media mining and statistical time-series forecasting in order to accurately project future occurrences and the reasons for them occurring.

Prescriptive analytics aims to improve business performance by applying mathematical algorithms to data. These algorithms suggest sets of alternative directions to take in case specific sets of objectives, requirements and constraints occur. The algorithms may rely on data, expert knowledge or both. The outcomes of predictive analytics is thus either the best course of action to take in a given situation, or a set of rich information that will help a decision maker to choose the best course of action. These details are summarised in Figure 5.1 below.

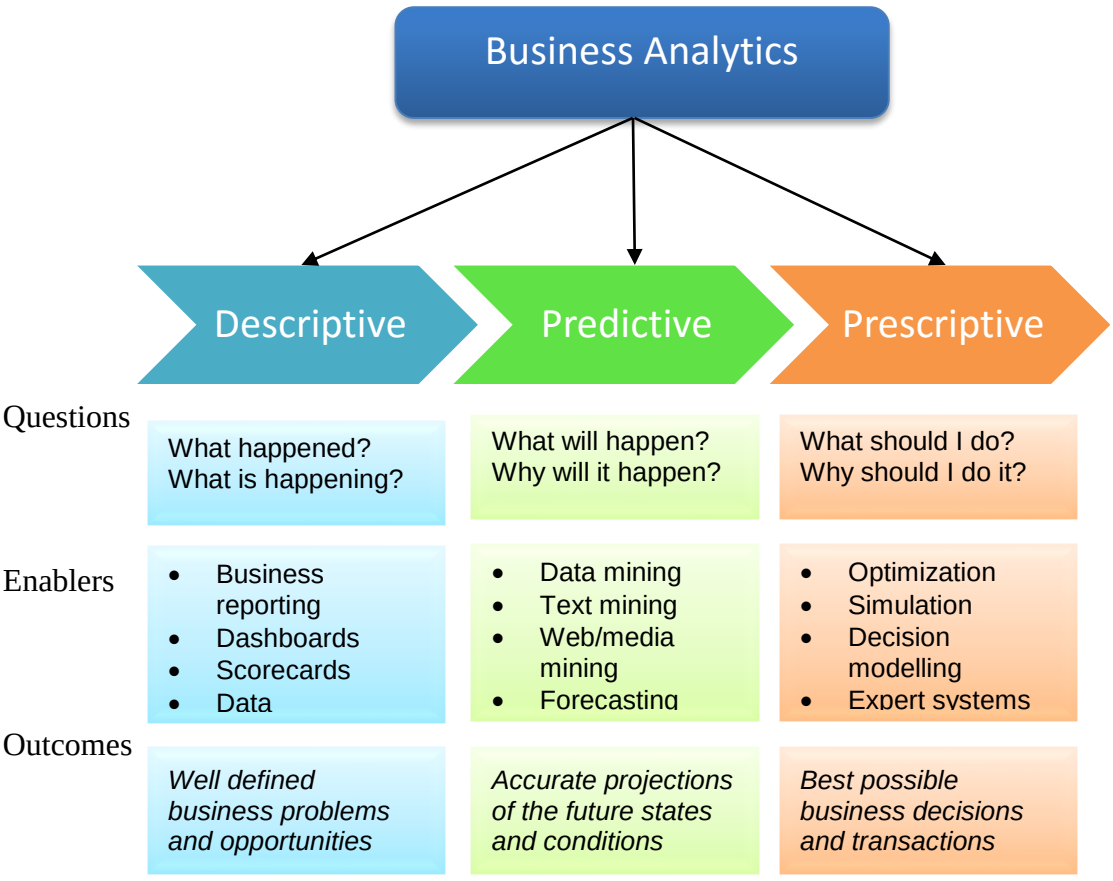


Figure 5.1: A Simple Taxonomy of Business Analytics (Delen & Demirkan, 2013)

In order to make public safety more proactive, the public safety reports must be put through the process of predictive analytics. The reports are obtained from participatory crowdsourcing in unstructured, natural language text format. As evident from Figure

5.1, data mining and text mining must be performed in order to do forecasting (web/media mining is not applicable in this study due to the nature of the data). The next two sections of this chapter will therefore explore data mining and text mining.

5.3 Data Mining

People interact with data and information on a daily basis, accessing information and creating or manipulating data often without even realising it; for some, in their daily work, or at least a part thereof, for others it may be as simple as checking the weather forecast on a smart phone application before deciding what to wear. This is nothing new. What is new, however, is the rate at which we are generating and recording data and the massive volumes of data that can be retained and stored. Focus has shifted from sifting out the most important bits of information to keep on limited storage space, to virtually hoarding every bit of data that can be found just in case it might be useful. This is the age of "big data". These vast amounts of data must be analysed in order to obtain useful information from it; some call this information retrieval (IR) or Knowledge Discovery and Data mining (KDD) or just Data Mining (DM).

DM is seen as the process of sifting through large databases for interesting patterns and relationships. The focus is usually on automated searching due to the size of modern data repositories and involves exploratory analysis and modelling of the data (Maimon & Rokach, 2010). Nisbet, Elder, and Miner (2009, p. 17) define knowledge discovery as "the entire process of data access, data exploration, modelling, model deployment, and model monitoring"; basically everything involved in transforming data into something usable.

Mitra, Pal, and Mitra (2002, p. 3) define KDD as "the overall process of knowledge discovery in databases", as depicted in Figure 5.2. These authors hold that DM is a specific step within this process of knowledge discovery, which involves the application of specific algorithms for pattern extraction from a set of data. In other words, the DM process includes only the actual modelling and statistical calculations and excludes the preparation of the data.

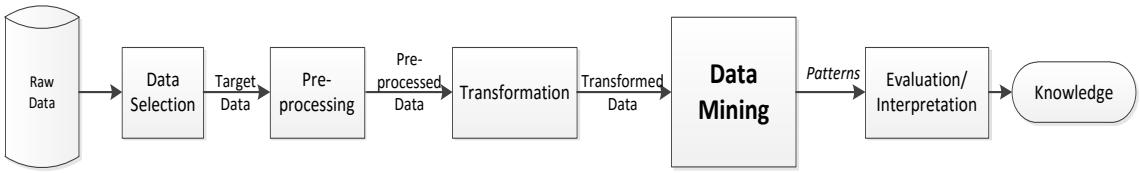


Figure 5.2: The KDD Process (Mitra et al., 2002)

The other steps in the KDD process then serve the purpose of ensuring that the knowledge obtained is useful (or applicable) by selecting the appropriate data, preparing it and interpreting it according to the purpose it is intended for (Mitra, Pal, & Mitra, 2002). Nisbet et al. (2009) prescribe the same process of knowledge discovery, but view the role of DM as a wider, more inclusive part of the process as shown in Figure 5.3.

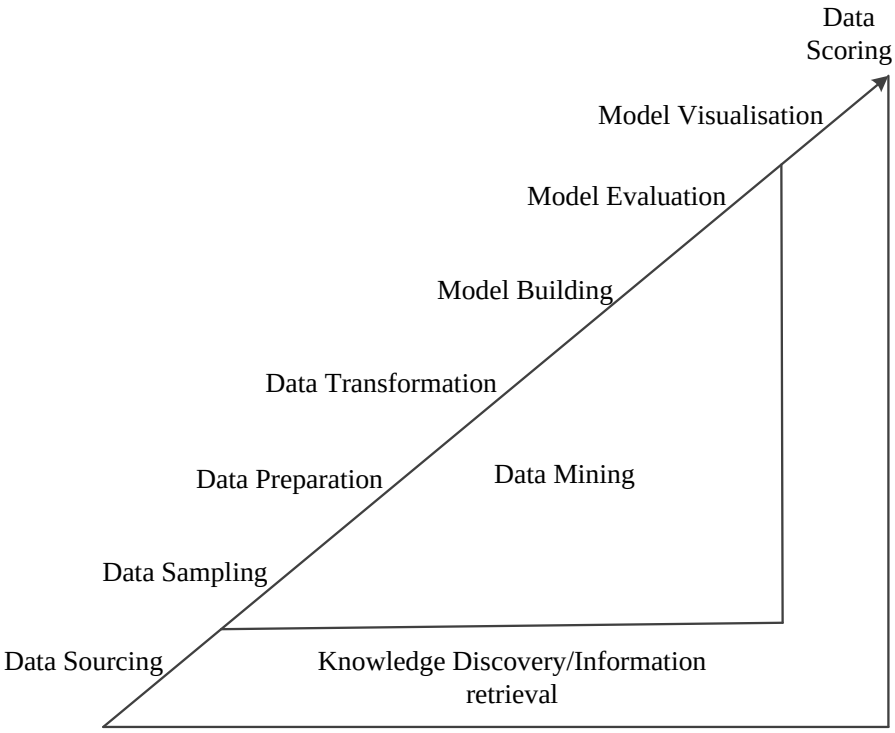


Figure 5.3: Data Mining and Knowledge Discovery (Nisbet et al., 2009)

There is thus a discrepancy as to what different authors believe should or should not be included in the DM process. This can be attributed to the variety of DM projects that have been undertaken. Projects can differ depending on the topic or area of the data, the format of the data, aims and goals of the project, and the tools and methods used. One of the differences evident between Figures 5.2 and 5.3 is the inclusion (or exclusion) of

selection/sampling, pre-processing and transformation steps. It is reasonable to assume that this is due to the variety of DM projects one could engage in, but there are also other debates concerning these specific areas. There is still debate as to the difference between data cleansing (also referred to as cleaning or scrubbing) and data quality (Maimon & Rokach, 2010). The significance of this in terms of this particular study is whether data scrubbing should be included in the analysis of the data or rather seen as part of quality assurance relating to sourcing of the data.

Data cleaning is a big issue in the data warehousing and database user communities, particularly when databases are merged, upgraded or transferred. Duplicated or missing values are regular issues which can cause problems for query resolution, with the source of the data often being a crucial factor (Maimon & Rokach, 2010). With regard to the public safety smart city project, extensive research has been done on ensuring good quality public safety reports are obtained via participatory crowdsourcing (Bhana et al., 2013). Most of the data cleaning issues created at the time of obtaining the data is beyond the scope of the actual analysis process as they can be avoided by adjusting the data gathering process. In the case of unstructured text which is coded, rather than structured database entries, missing information or variables will simply not be coded and be excluded from that particular query. The impact of this on any results will be negligible as one or two outliers will not skew a pattern if one is present. Transformation and formatting may or may not be required depending on the format of the data obtained and the type of analysis required.

The lack of standardisation in how DM should be performed resulted in analysts using various methods for their DM, leading to great variations in the findings of analysts working with the same data. This, in part, led to the development of a Cross-Industry Standard Process for Data Mining (CRISP-DM).

5.3.1 The Cross-Industry Standard Process for Data Mining (CRISP-DM)

The Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology is the result of European funded efforts to develop a standardised framework for data mining projects (Chapman et al., 2000). The motivations for the development of a standardised model included:

- the need for interoperable tools;
- to make simple data mining tasks easier and cheaper;
- to aid in project planning and management;
- to enable the replication of projects;
- to enable less experienced data miners to achieve the same results, and
- to encourage best practices for better results.

The development of CRISP-DM project was initiated in late 1996 by Daimlerchrysler, SPSS and NCR (Wirth & Hipp, 2000). With development and refinement input from over 300 organisations over a three year period, CRISP-DM 1.0 was published in 1999 (Chapman et al., 2000). The CRISP-DM is a non-proprietary, industry and technology neutral template for analysis with a business focus. As shown in Figure 5.4, CRISP-DM consists of six phases, which can then be divided into three to six tasks each (Wirth & Hipp, 2000).

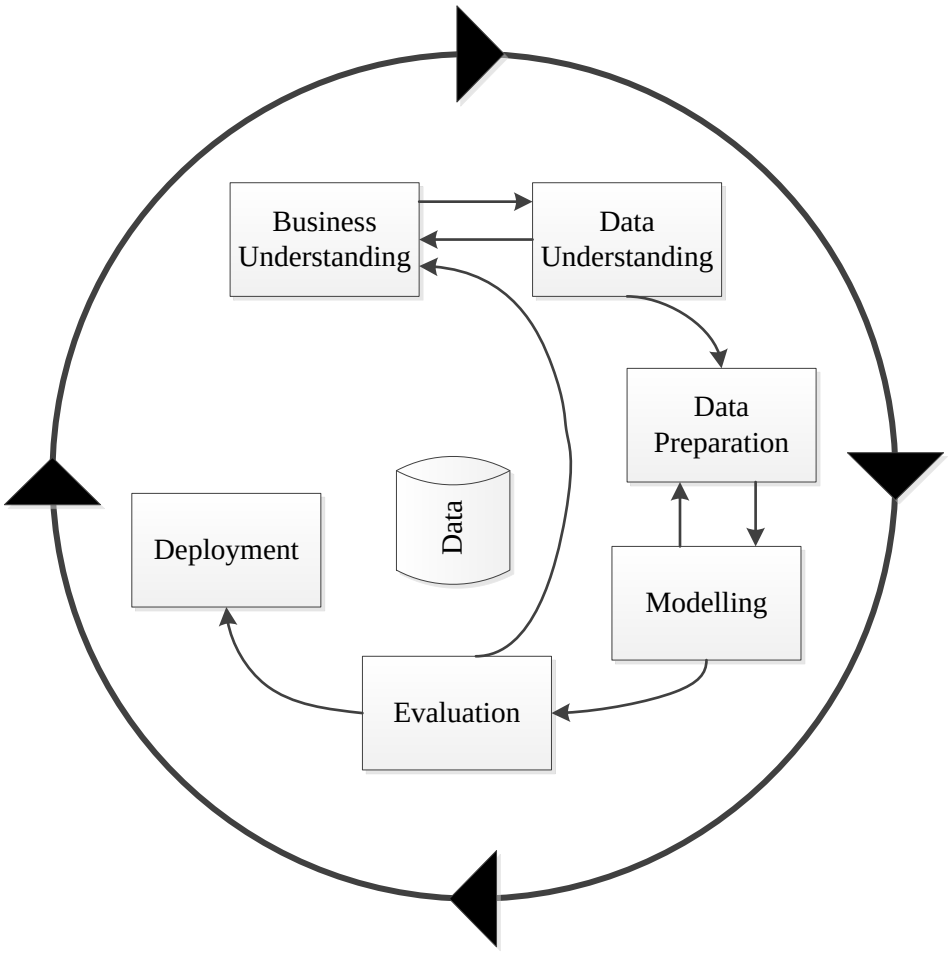


Figure 5.4: CRISP-DM (Wirth &Hipp, 2000)

The six phases of the CRISP-DM can be outlined as follows (Wirth & Hipp, 2000):

1. Business Understanding
   - Understanding the project's objectives and requirements from a business perspective.
2. Data Understanding
   - Collecting, exploring and describing the data while also considering quality issues.
3. Data Preparation
   - Construction of the final dataset. An iterative process which can include structuring, cleaning and formatting.
4. Modelling
   - Select and apply a modelling technique, build model and assess it.
5. Evaluation
   - Review steps taken to create model and ensure that the business objectives are properly achieved.
6. Deployment
   - Often done by the user; organising and presenting of the information for use.

Wirth and Hipp (2000) list the tasks associated with these phases as shown in Table 5.1 below.

Table 5.1: Overview of the CRISP-DM tasks (Wirth & Hipp, 2000)

Business Understanding	**Data Understanding**	**Data Preparation**	**Modelling**	**Evaluation**	**Deployment**
Determine business objectives Assess Situation Determine DM Goals	Collect initial data Describe data Explore data Verify data quality	Select data Clean data Construct data Integrate data Format data	Select modelling technique Generate test design Build model Assess model	Evaluate results Review Process Determine next steps	Plan deployment Plan monitoring and maintenance Review project

CRISP-DM thus gives a great deal of guidance to data miners in terms of a standardised process to follow and tasks that need to be performed for a successful DM project irrespective of what field or subject matter the project pertains to. In order to make the CRISP-DM applicable to a variety of data mining projects, it is designed to be generic and non-specific, leaving many questions and details to be figured out by the analyst. The overly generic process needs to be tailored to the environment, data available, and desired outcomes of the specific project. Thus, the CRISP-DM cannot be applied to the public safety smart city project.

DM is also traditionally applied to numeric or structured data within relational databases. The public safety smart city project receives unstructured natural language data input, thus a more specific method for analysis is required.

Increased interest in unstructured qualitative text data has led to increased focus on Text Mining (TM). Some call it a sub-category of DM, but other researchers suggest that TM is developing into a parallel entity to DM. For the public safety smart city project, DM alone will not be sufficient as the data obtained is in the form of natural language text. Thus, text mining options are explored in the next section of this chapter.

5.4 Text Mining

One of the earliest definitions of text mining states that it is "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources" (Hearst, 2003, p. 1). Miner et al., (2012) further explain: "Text mining and text analytics are broad umbrella terms describing a range of technologies for analysing and processing unstructured text data" (p. 30).

A more concise definition is hard to find as text mining is still a broad field spanning different areas and fields as depicted in Figure 5.5. What binds these various technologies together is the need to convert text to numbers in order to apply traditional Data Mining (DM) techniques to large text databases. What is shared by all techniques is that words or phrases, with particular meaning, must be coded (marked, flagged or tagged), counted and grouped. These groupings can then be counted in order to summarise and quantify data. In this way, analytical algorithms can be applied to large

quantities of text data by reducing or summarising the text to numerical values. There are, however, various techniques to choose from in order to do this.

Berry and Kogan (2010) group text mininng techniques into four main groups:

1. Keyword extraction
2. Classification and clustering
3. Anomaly and trend detection
4. Text streams

Miner et al. (2012), however, describe a further expanded seven main practice areas of text analytics, explaining that the breadth and disparity of contributing disciplines in conjunction with varying levels of maturity make it difficult even for experienced text mining professionals to further categorise text mining concisely and accurately. The seven practice areas of text analytics as listed by Miner et al. (2012) are:

1. Search and information retrieval (IR) – Storing and retrieving text documents, including keyword search and search engines.
2. Document clustering – Using data mining methods to group and categorise terms, phrases, paragraphs or whole documents.
3. Document classification – Using data mining methods to group and categorise terms, phrases, paragraphs or whole documents, based on models trained on labelled examples.
4. Web mining – Data and text mining on the Internet with special focus on the scale and networks.
5. Information extraction (IE) – Identifying and extracting relevant facts and relationships from unstructured text (structuring unstructured text data)
6. Natural language processing (NLP) – Low-level processing of language, for example parts of speech tagging.
7. Concept extraction – Grouping words and phrases according to their semantic meaning.

The seven areas exist at the major intersections of text mining with its six related fields as depicted in Figure 5.5 (Miner et al., 2012).

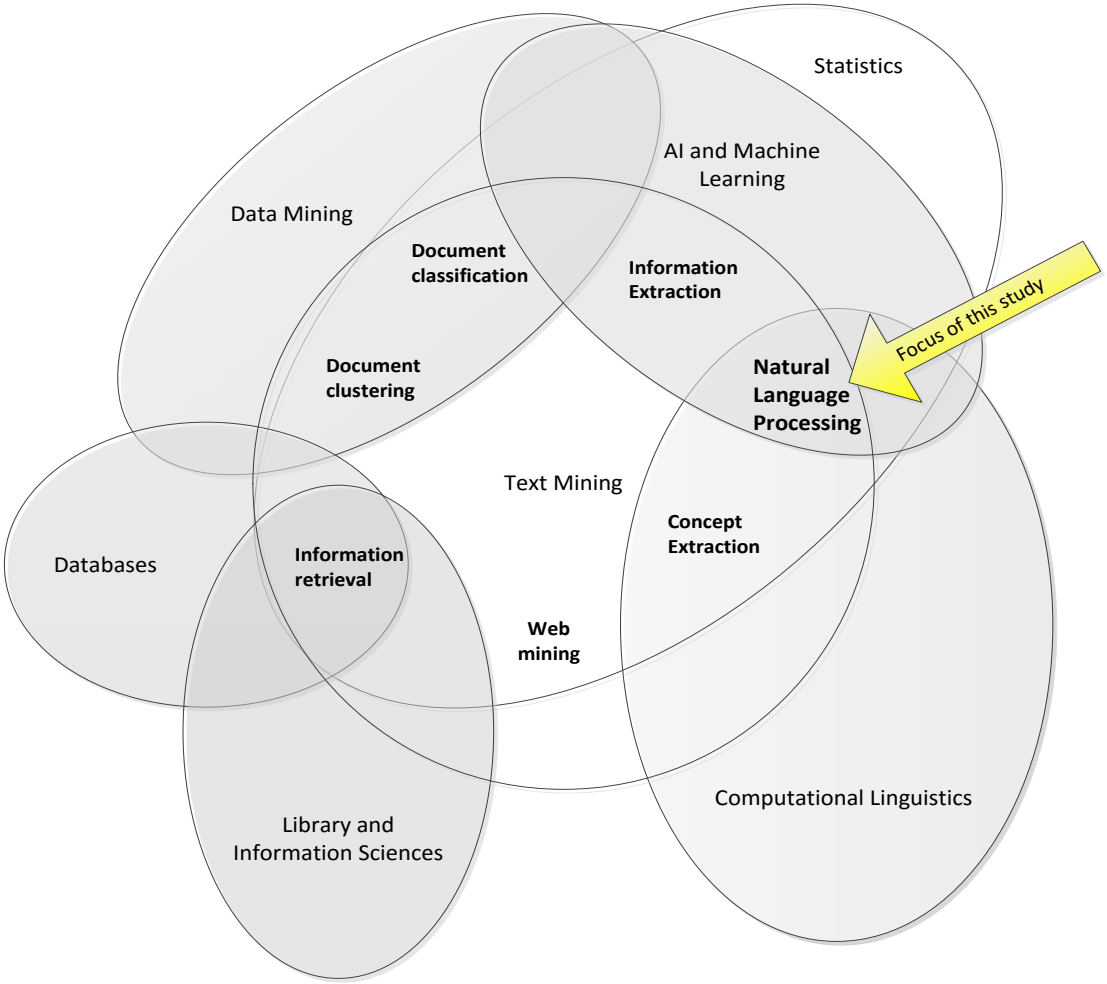


Figure 5.5: Venn diagram -Seven Practice Areas of Text Analytics (Miner et al., 2012)

Having established that text mining and text analytics span various techniques, technologies and strategies (or "practice areas"), one still has to choose the right one for the project. Miner et al. (2012) suggest that answering between two and four of the following five questions will lead one toward the correct option:

Table 5.2: TM Questions and the Characteristics they Address (Miner et al., 2012)

	Characteristic	Question
1	Granularity	Is the interest in results about individual words or at a higher level (sentences, paragraphs or documents)?
2	Focus	Is the interest in finding specific words and documents or characterising the entire set?
3	Available information	Are there predefined categories?
4	Syntax or semantics	Is the focus on the meaning of the text or the structure?
5	Web or traditional text	Are documents independent or connected via hyperlinks?

Miner et al. (2012) arrange these questions in a decision tree, as depicted in Figure 5.6, in order to help one choose the most appropriate TM technique for one's project. Following the tree with the public safety smart city project in mind, the public safety reports would fit into Natural Language Processing:

- Individual words such as a street name or name of a crime (such as "Southernwood" versus "Gonubie", or "burglary" versus "accident") can change the entire meaning of a report.
- Ultimately the meaning of the report is vital (what is the report actually about).
- The reports are structured in natural language which involves unique considerations.

Miner et al. (2012) describe NLP as tasks involving low-level language processing and understanding. Chopra et al. (2013) simply describe NLP as being everything a computer requires to understand and generate languages spoken by people. A more formal definition (essentially saying the same) states NLP as being "concerned with theories and techniques that address the problem of natural language communication with computers" (Lehnert & Ringle, 2014, p. 2). The two latter definitions explicitly state the involvement of computers in NLP; however, although Miner et al. (2012) do not explicitly include computers in this definition (See Figure 5.5), it is evident that these authors have placed NLP to include computational linguistics and/or artificial intelligence and machine learning.

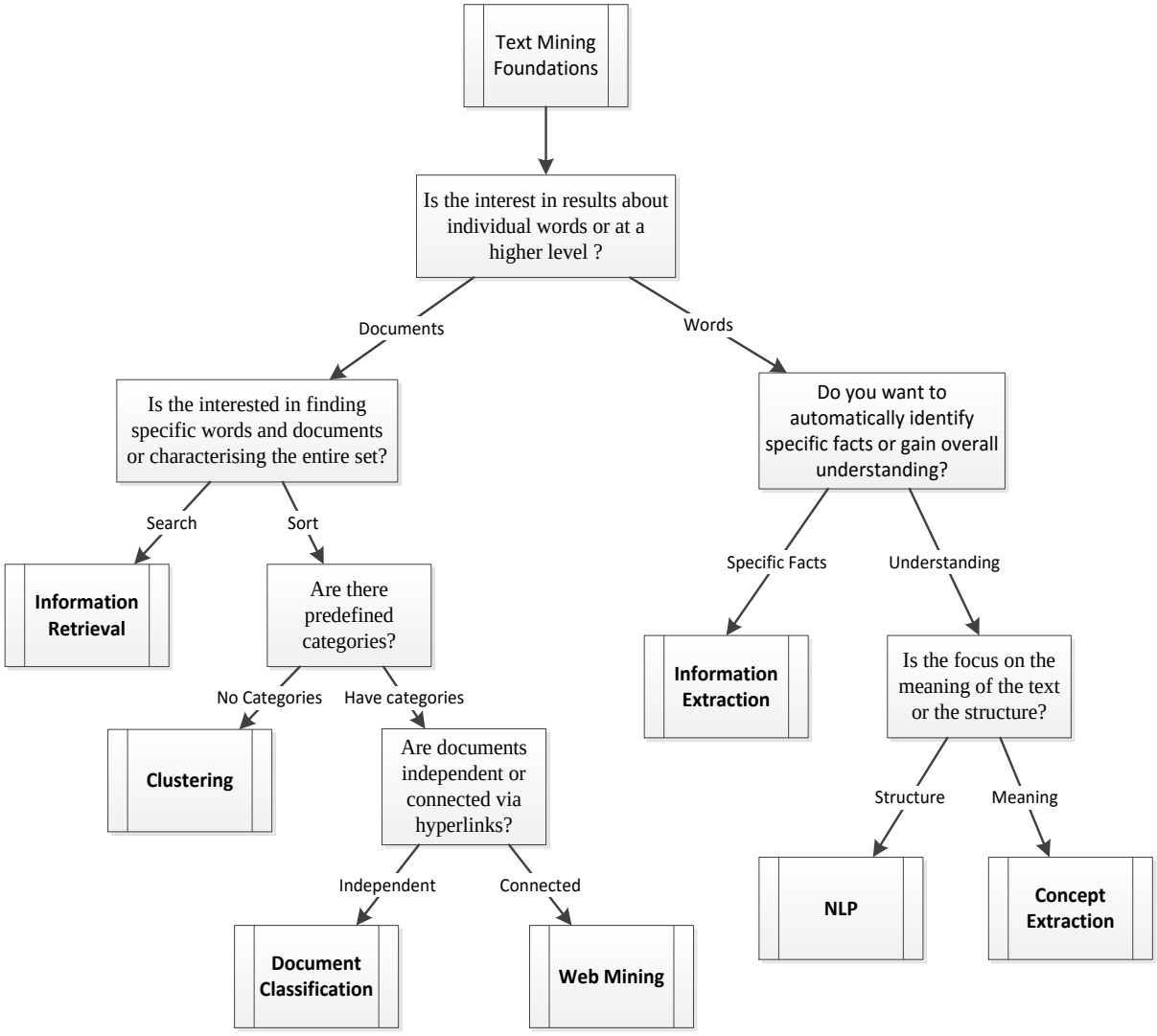


Figure 5.6: Decision Tree for Finding the Right Text Mining Practice Area (Miner et al., 2012)

5.5 Natural Language Processing (NLP)

Natural languages are those languages spoken by people, for example French, Mandarin and English, while computers use machine language. NLP encompasses everything computers need to understand, work with and create natural language. Hence, NLP is a subfield of both artificial intelligence and linguistics, with a focus on making computers understand human language in order for humans to use and interact with computers without having to learn machine language.

Mehrotra et al. (2013) practically defines NLP as "a field of computer science in which the computer is trained to "read" text to identify relevant data" (p._1234). Making use of this technology means that an analyst would not have to read and physically code

each individual word of natural language text. Using NLP, one will thus spend some initial time and effort on training the system, after which it should be able to run through vast quantities of text data and automatically identify and code the required sections or words for predictive modelling. This is in line with Miner et al. (2012) who show (in Figure 5.5) that NLP exists at the overlap between AI and machine learning, and computational linguistics. From this it can be said that NLP is a combination of content extraction and information extraction – not only finding specific words, but doing so based on the meaning associated with those words.

Chopra et al. (2013) loosely map areas of linguistic study to steps that form part of NLP:

Table 5.3: Linguistics and the Five Steps of NLP (Chopra et al., 2013)

Linguistics includes	The five steps of NLP
Sounds which refers to phonology	Discourse Integration
Word formation refers to morphology	Morphological and Lexical Analysis
Sentence structure refers to syntax	Syntactic Analysis
Meaning refers to semantics	Semantic Analysis
Understanding refers to pragmatics	Pragmatic Analysis

When analysing natural language public safety reports, it must be considered which of these areas are the most applicable. The smart city, in this case, aims to reduce public safety by obtaining information that will give insight into public safety occurrences. Therefore, phonology, morphology and syntax can be excluded as they will not provide insight into the public safety issues reported. Of the rest, semantics stands out as the area that this project should be concerned with. For extracting useful information from the reports, it is vital to understand what the respondent means when they are reporting an incident.

5.5.1 Semantic Analysis of Public Safety Reports:

The last 40 years has shown remarkable progress in automatic speech recognition, information retrieval, and statistical machine translation (Steedman, 2010). A full understanding of the concept and structure of human language, however, still seems to be beyond the reach of NLP technology (Steedman, 2010). Attempts at resolving this have primarily focused on grammar and conversation development through part of speech tagging and recognition. This still results in confusion with tenses and active

versus passive voice. An association can be made between two concepts, but the meaning is not properly transferred in different grammatical sentence structures.

Using a reliable, modern speech to text transcription system will prevent most problems related to phonology, morphology and syntax. Such issues are usually related to the collection of data and not necessarily the analysis thereof. Semantics (the inference of meaning from words), however, is still one of the greatest hurdles for NLP.
For example:

> "*Burglary at Sunnyridge on the 28th of September*."
> *"Break in at Trafford Road, Morningside at 2am on the 29th of September"*

The two reports above refer to the same type of event, yet a search for "burglary" would return only the first report and not the second. Different naming conventions, slang and colloquial meaning inference can change what a particular word or phrase means to different people, especially in places that are geographically far apart. Another example of this is the use of the word "izinyoka" in reference to illegal electrical cables and electricity theft. Although the word is derived from one of many local languages, it is known and understood by most South Africans, but would not likely be understood anywhere else in the world.

In the case of the public safety smart city project, the voice calls are in natural language text format. In order to summarise or quantify the reports for predictive modelling, the keywords must be identified and coded. Particular codes are then counted, allowing for calculations and statistics to be performed. In the preceding chapters the key elements that need to be coded are identified as being date, time, area, location and event. Additionally, metadata for time and date of the report can also be identified. Considering this, it becomes clear that grammatical understanding is not required and thus part of speech identification is also not needed. What is required is the identification of words or phrases that relate to the above mentioned categories.

The most successful NLP methods still use Supervised Learning from data labelled by humans, where precoded data is used as a sample which directs the system in how to automatically code the larger data set (Steedman, 2010). This is also called ontology development.

5.6 Ontology

Text mining is an effective means of finding useful information in large quantities of text data, but a high level of accuracy cannot be attained using conventional text mining technology (Jiang et al., 2013). Jiang et al. (2013) explain that this holds true as conventional text mining cannot effectively make use of the semantic information in the text. An ontology can, however, provide a theoretical basis and support for organising and representing semantic information, thus improving the accuracy of text mining projects. Domain specific ontologies enable text analysis tools to provide more accurate and relevant classification.

One of the most important building blocks in semantics is ontology. Ontology helps to solve many semantic problems and can help produce machine readable semantic models and semantically modelled data. An example of this is the Semantic Web where ontologies have been put to successful use by producing semantic aware solutions to existing web problems. Ontology helps machines understand natural language by producing abstract modelled representations of already defined sets of terms and concepts in natural language (Ahmed, Dandekar, & Majeed, 2012).

An ontology is an explicit, formal specification of a shared conceptualisation of a domain of interest (Gruber, 1994). An ontology thus is a structured, taxonomical hierarchy of various terms used for specific meanings within a certain group or community. It is dependent on knowledge of a particular domain as well as the semantic and colloquial jargon for that domain in a particular area. Domain specific ontologies enable text analysis tools to provide more accurate and relevant classification.

Ontologies are constructed and connected to each other in a decentralised manner to clearly express semantic contents and arrange semantic boundaries to discover required needed information (Jiang et al., 2013). Natural language based information is treated as the input to the ontology construction process, which then separates and structures text into categories and subcategories maintained in taxonomical hierarchy. The size of ontology varies depending on the amount of categories and groupings identified, as well as the complexity and desired outcomes of the related natural language text mining project. Ontologies can be created from raw data or by combining existing ontologies

into new ontologies. In order to develop such ontology from natural language text data, traditional qualitative data analysis methods must be used. The following section of this chapter will therefore explore the different types of Qualitative Data Analysis.

5.7 Qualitative Data Analysis (QDA)

Weitzman (1999) summarises the qualitative data analysis process as following a path from questions to conclusions. Some of the main steps, amongst others in this iterative process, are identifying questions, the development of a coding scheme, coding chunks of data, reducing the data, and entering the data into displays in order to draw conclusions (Weitzman, 1999). Though this may seem a simple process, these steps are in fact quite general and can vary widely as there are many different ways of doing each one. Weitzman (1999) explains that exactly how this is done depends on the nature of the project and the questions asked.

Namey, Guest, Thairu, and Johnson (2008) group aproaches to qualitative data analysis into either content analysis or thematic analysis. In content analysis the researcher evaluates the saliency and frequency of specific words or phrases within a particular document or piece of text data in order to find keywords or frequently repeated ideas. In addition to simple word counts, content analysis can be expanded to also include attributes of the keywords such as synonyms, location of the word within the text, and surrounding words or phrases. Content analysis techniques are known for their efficency and reliability. The use of appropriate software can allow for rapid scanning and tallying of keywords, and the use of "raw" data with minimal interpretation results in greater reliability. The main drawback to content analysis is that the context is usually not very broad which can limit the richness of the summarised data.

Thematic analysis includes more involved and nuanced approaches. Thematic analysis considers meaning when characterising both implicit and explicit ideas. Codes in this approach represent ideas and themes rather than a keyword, which are then used as summary markers for further analysis. Reliability is a greater concern with thematic analysis as the analyst is required to interpret the raw data in order to apply codes, but the interpretation may then vary between analysts. Strategies exist for more reliable interpretations but lengthen the entire process as they require additional time to select

and implement. The end result, however, is richer and more contextually qualified information.

Leech and Onwuegbuzie (2008) compiled a comprehensive compendium explaining the relationship between types of qualitative data analysis techniques and the source of the qualitative data to be analysed. For data obtained from documents (text), the following types of analysis techniques are listed:

Table 5.4: Analysis Techniques Applicable to Documents (Leech & Onwuegbuzie, 2008)

	Type of Analysis:	**Short Description of analysis:**
Documents:	Semiotics	Systematically analysing similarities and differences across cases; typically being used.
	Qualitative comparative analysis	Systematically analysing similarities and differences across cases, typically being used as a theory-building approach, allowing the analyst to make connections among previously built categories, as well as to test and to develop the categories further.
	Constant comparison analysis	Systematically reducing data to codes, then developing themes from the codes.
	Keywords-in-context	Identifying keywords and utilizing the surrounding words to understand the underlying meaning of the keyword.
	Word count	Counting the total number of words used or the number of times a particular word is used.
	Secondary data analysis	Analysing non-naturalistic data or artefacts that were derived from previous studies.
	Classical content analysis	Counting the number of codes.
	Text mining	Analysing naturally occurring text in order to discover and capture semantic information.

Leech and Onwegbuzie (2008) conclude that a more focused analysis using appropriate methods according to the above groupings will lead to deeper understanding. They further state that using more than one of these analysis types to triangulate results will

also increase the trustworthiness of findings and serve to reduce researcher (or analyst) bias. They are therefore not mutually exclusive.

In order to develop an ontology for the public safety smart city project in East London, a simple yet effective QDA technique such as content analysis is most applicable because this technique is simple yet accurate and efficient, making it easy to apply via an automated tool while accuracy is assured through its structured process.

The development of such an ontology is for the purpose of identifying and grouping various terms used for the same meaning. As there is thus no question to answer or hidden meaning to discover at this stage, findings would be most easily reproduced and verifiable if the technique used is as simple as possible while still ensuring accuracy. This step of using content analysis for ontology development would be followed by a more comprehensive process of coding and grouping, resulting in the previously unstructured data becoming structured for pattern identification and predictive modeling.

The above QDA tehniques are relevant and applicable, but with the rise of big data in recent years, it has become necessary to automate them when dealing with vast amounts of data. Considering the large quantities of data available, this process can become time consuming and tedious if done manually. Computer assisted qualitative data analysis software (CAQDAS) has thus been developed to help analyse qualitative data faster and more accurately.

5.8 Computer Assisted Qualitative Data Analysis Software

Due to the vast quantities of data available in the era of big data, and the drive towards ever shorter turnaround times, qualitative data analysis process must be automated. Statistical analysis software has been evolving over many years. Though the first of its kind was created in the 1960s, CAQDAS gained widespread recognition among researchers in the 1980s and 90s (Carvajal, 2002). Recently CAQDAS has become common among qualitative researchers; one could say it has become a necessity for most qualitative analysis (Rambaree, 2007).

The core functionality of these applications relies on coding, retrieving and recoding (Pruijt, 2012). CAQDAS thus does not change the way QDA is performed, but it facilitates the use of QDA methods and techniques. Besides the obvious benefits of speed and accuracy, Rambaree (2007) explains that the use of CAQDAS also brings additional rigour to qualitative analysis. Some other main advantages and disadvantages are listed in Table 5.5 below.

Table 5.5: Advantages and Disadvantages of a Dedicated CAQDAS

Advantages	Disadvantages
Being designed for this purpose, some are highly specialised	Expense - most offer a 30 day trial, but incur significant cost for the full version
Tedious tasks can be automated and sped up	Users need extensive training in order to take advantage of features
Usually able to process more than one kind of data (Text, video, graphic)	Can usually only use one language
Can generate graphs and models automatically	Not always available on all platforms (usually Windows first then others later)
Can facilitate group work by supporting multiple users	

Selecting a CAQDAS Package:

There are a number of CAQDAS packages available in the marketplace and via open source. Most of these packages can perform the basic QDA functions of tagging and coding, but do so with minor interface differences. To illustrate this, Table 5.6 compares three of the most popular CAQDAS packages (Gibbs, 2008; Saillard, 2011; Given, 2008).

The comparison shows that overall the three packages are quite similar. Basic functionality is available on all three platforms with differences becoming evident when considering price, platform and external integration with other applications. The commercial price on all three products is quite similar. With all three offering a greatly reduced price for students and educational institutions, MaxQDA and Nvivo are the cheaper options for scholars. MaxQDA and Nvivo both have free trial versions that expire after 30 days, allowing one to experience the program before deciding whether to purchase it, but the Atlas.ti offering stands out as there is no time limit on the trial product, only on the size of the project that one can use.

Ultimately, the three main CAQDAS packages on offer are similar in price and features. The minor variances will only make a difference to those few individual users who they are applicable to. For this study, the author has decided to make use of Nvivo as he personally finds it to be more intuitive and user friendly than the other two options. However, this is a personal preference rather than a scientific one.

Table 5.6: Comparison of Atlas.ti, MaxQDA and NvivoFeatures

	Atlas.ti	**MaxQDA**	**Nvivo**
Platform	Windows only	Windows and Mac	Windows and Mac
Trial	Free, no time limit, but project size limitations	30 Days free trial	30 Days free trial
Price	$2300 Commercial $670 Educational use	$2023 Commercial $119 Student	$2345 Commercial $120 Student
Input	Text, Audio, Video, Image	Text, Audio, Video, Image	Text, Audio, Video, Image, social media and web pages
Output	XML, Excel, HTML, SPSS	XML, Excel, HTML, SPSS	Word, Excel, HTML, SPSS, survey monkey, endnote
Project management	Yes	Yes	Yes
Coding	Yes	Yes	Yes
Analysis	Yes	Yes	Yes
Memo's	Yes	Yes	Yes Annotations, links
Visualisation tools	Cloud view, co-Occurrence table, network views	Maps, Matrix, relations, text comparison chart, cloud view	Charts, maps, models
Collaboration	Project merging, networking	Project merging, networking, multiple user roles and rights can be assigned	Nvivo Server, team work, collaboration
Drag and Drop	Yes	Yes	Yes
Linking	Internal/External, Google Earth	Internal/External Google Earth	Internal/External, web

It is important to note that while CAQDAS helps to make the analysis process easier, it cannot replace the analyst. A CAQDAS package is used to help manage and manipulate data. It can be used to perform queries and some can even model the data. However, it cannot describe the data nor can it understand, interpret or explain the results obtained, thus the analyst is still a necessity. As helpful as a CAQDAS package

may be, it is still only a tool for the analyst to use and not (for the time being) a replacement. The tool still needs an instructor.

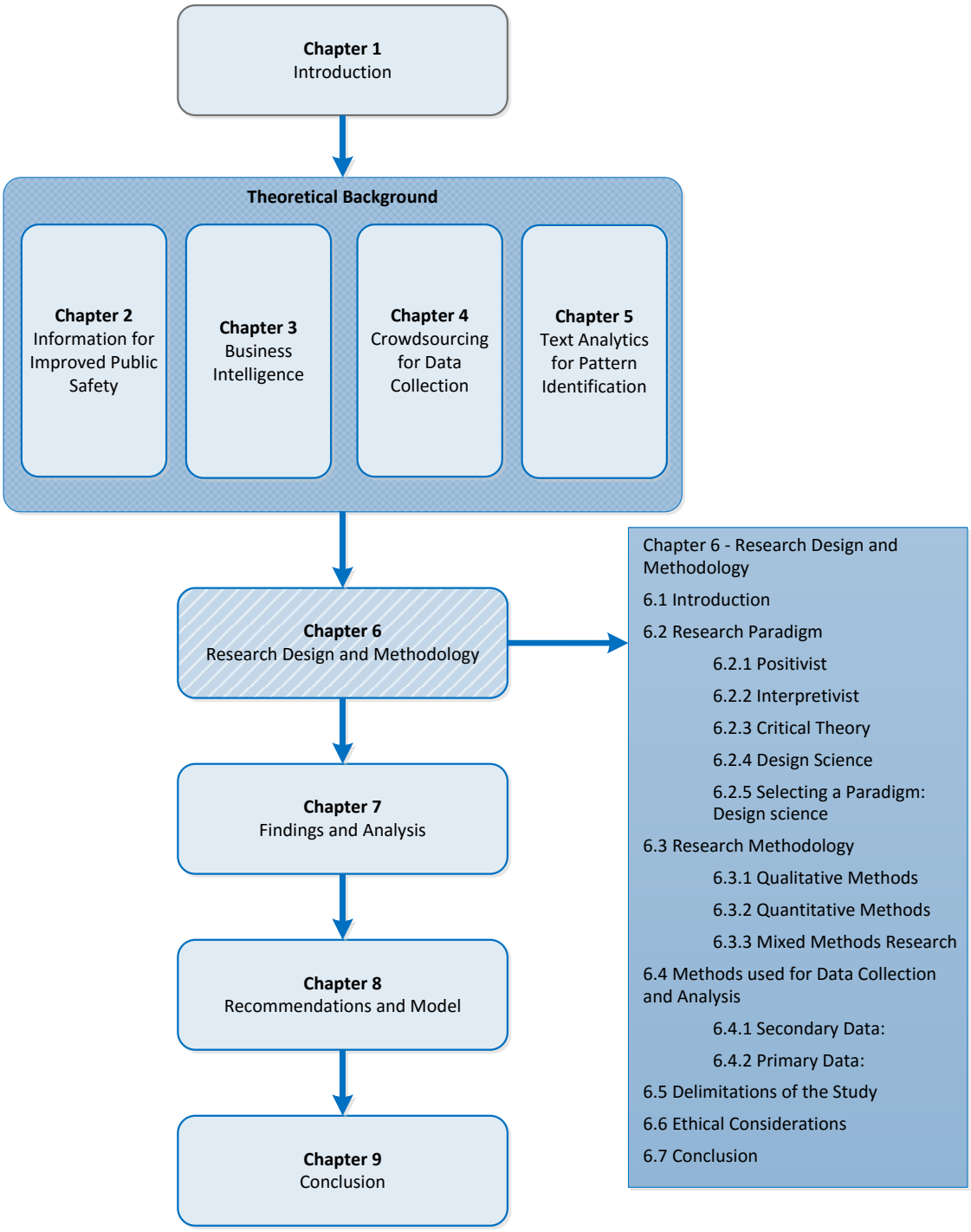
5.9 Conclusion

Data analytics can be used to understand, predict or direct by identifying patterns and trends within raw data. A widely accepted well-structured way of performing data mining is by following the CRISP-DM. In order to apply analytics to large quantities of unstructured text data, however, text mining is used to reduce or summarise the data into a more structured form that enables the use of data mining techniques. The process of text mining involves coding the text data by identifying words or phrases with similar meanings which can then be grouped and counted. This will result in sets of numeric data associated to certain concepts, which can then be modelled in order to reveal any trends or patterns that may be present.

As text mining spans many different subject areas, it is important to explore the data and the area in which the resulting information is to be used (in accordance with CRISP-DM) in order to select the best techniques for the text mining project. The focus area of the public safety smart city project is found to be NLP which involves both linguistics and machine learning (a culmination of information and concept extraction).

Within NLP the relevant issue regarding the public safety smart city project is semantic analysis (understanding of meaning) when extracting information. The proposed solution for this is the development of an applicable ontology. In order to retain usefulness while summarising and to compensate for semantic and colloquial references, an ontology of key terms and their associated meanings is required. The ontology is developed by performing a content analysis of the data using a CAQDAS to be used for machine learning to enable accurate analysis of larger amounts of data. The ontology can then be used to guide the automated coding and reduction of large text data sets such as a collection of public safety reports. The next chapter will describe how this study was conducted by presenting the research design and methodology.

Chapter 6 – Research Design and Methodology



6.1 Introduction

As stated in Chapter 1, the problem addressed by this study is **the analysis of natural language text data for improving public safety in a smart city**. Smarter informed decision making can help compensate for limited resources strained by increased urbanisation. The objective of this study is the development of a model which will guide the analysis process through the structuring of the unstructured text for pattern identification. More specifically, the study focuses on the public safety smart city project in East London. The method employed is for the analysis of public safety reports received from the citizens via participatory crowdsourcing to be used to inform the city's decision makers.

This chapter examines the research design and the methods used that are applied in this study. The methodology and research design selected for this study are informed by the problem statement and objective of this study as well as the data available. Hofstee (2006) explains that in research, there are often many possible routes that one can take in order to investigate a problem and find a solution. It is the responsibility of the researcher to select a good reliable method that fits the study and the limitations that may be faced in order to ensure credibility (Hofstee, 2006). According to De Vos, Strydom, Fouché, and Delport (2005), all research is conducted (or should be) within the boundaries of a particular research paradigm or philosophy. Considering the above, the rest of this chapter will discuss the research design including the paradigm of this study, the methods used for data analysis, and other applicable considerations.

6.2 Research Paradigm

Epistemologically, all research must have a sound basis in existing knowledge and theory in order to ensure that sound, rigorous research is done. Any research will have an underlying research paradigm that guides how the research should be conducted (Collis & Hussey, 2009). Olivier (2009) describes a research paradigm as an accepted model or pattern that guides all research. Additionally, the use of an accepted paradigm contributes to the credibility of a research project due to it being based on well-accepted methods (Olivier, 2009). There are several paradigms which can be followed, distinguishable by the philosophical assumptions on which they are based.

Oates (2006) recognises three philosophical paradigms in IT research, namely: positivism, interpretivism and critical research, while Collis and Hussey (2009) limit the discussion of paradigms to positivism and interpretivism. Vaishnavi and Kuechler (2008), however, motivate the inclusion of design science as an emerging research paradigm in this discipline.

6.2.1 Positivist

Positivist research functions according to two main assumptions: (1) the world is ordered and regular, (2) the world can be investigated objectively.

The belief here is that cause and effect are specific and measureable, independent of the researcher. Reality is concrete and unchanging and can thus be analysed through measurable attributes (Myers, 2009). The researcher is an impartial observer and facts are not altered by his personal values and beliefs.

This insures that positivist research can be replicated by multiple different researchers, with the same results. Conversely, the interpretivist paradigm is somewhat subjective and reliant on the social context of the researcher.

6.2.2 Interpretivist

This paradigm aims to understand IT as a practice constructed and developed by humans (Oates, 2006). Thus, interpretivism seeks to understand the social context of IT. Interpretivist research relies upon the view and activities of others in reference to how they understand their own reality (Collis & Hussey, 2009). Therefore, interpretivist research endeavours to understand human behaviour within a specific context.

Interpretivist research is conducted somewhat more subjectively, focusing on the meaning of a social phenomenon rather than its measurement. The researcher inevitably has an effect on the process and therefore this mustbe acknowledged (Oates, 2006).

Every individual has their own perception and interpretation of the world around them, transferred by their own language and shared meanings. These meanings can change over time and differ according to geographic location. Research following this paradigm

seeks to understand and explain phenomena, thus it has a strong preference towards qualitative data analysis. Interpretivist research is relatively subjective in comparison to positivism as it can be influenced by the researcher's beliefs, values and actions (Olivier, 2009). Critical theory, in contrast, is reliant upon economic, political and cultural influences and history.

6.2.3 Critical Theory

Critical theory research focuses on the oppositions, conflicts and contradictions in social reality (Myers, 2009). In this way it seeks to critique and change society rather than just to study or understand it (De Vos et al., 2005). Critical theory asserts that social reality is historically created and recreated by people in a consistent manner (Oates, 2006). This paradigm asserts that social reality has certain objective qualities which change experiences and the way of seeing the world; for example, political and economic systems.

Critical researchers reject the notion that people need to adapt to technology, but rather argue that people and society should shape the way technology is created.

As with interpretivist research, critical researchers acknowledge the influence their own values, beliefs and actions have on their research. However, critical researchers criticise interpretive research for failing to analyse the patterns of power and control that regulate views of reality (Oates, 2006). More recently, design science has been increasingly adopted as a research paradigm in IT research.

6.2.4 Design Science

Vaishnavi and Kuechler (2008) motivate the inclusion of design science as an emerging research paradigm in IT disciplines. Design science is fundamentally a problem-solving paradigm which ensures that knowledge and understanding of a problem domain are achieved through the building and application of an artefact (Hevner et al., 2004). This is depicted (below) in the design science research framework (Figure 6.1) developed by Hevner et al. (2004). The design science paradigm has also been referred to as the socio-technologist paradigm (Vaishnavi & Kuechler, 2008).

"*The design-science paradigm seeks to extend the boundaries of human and organizational capabilities by creating new and innovative artefacts*" (Childerhouse, Hermiz, Mason-Jones, Popp, & Towill, 2003).

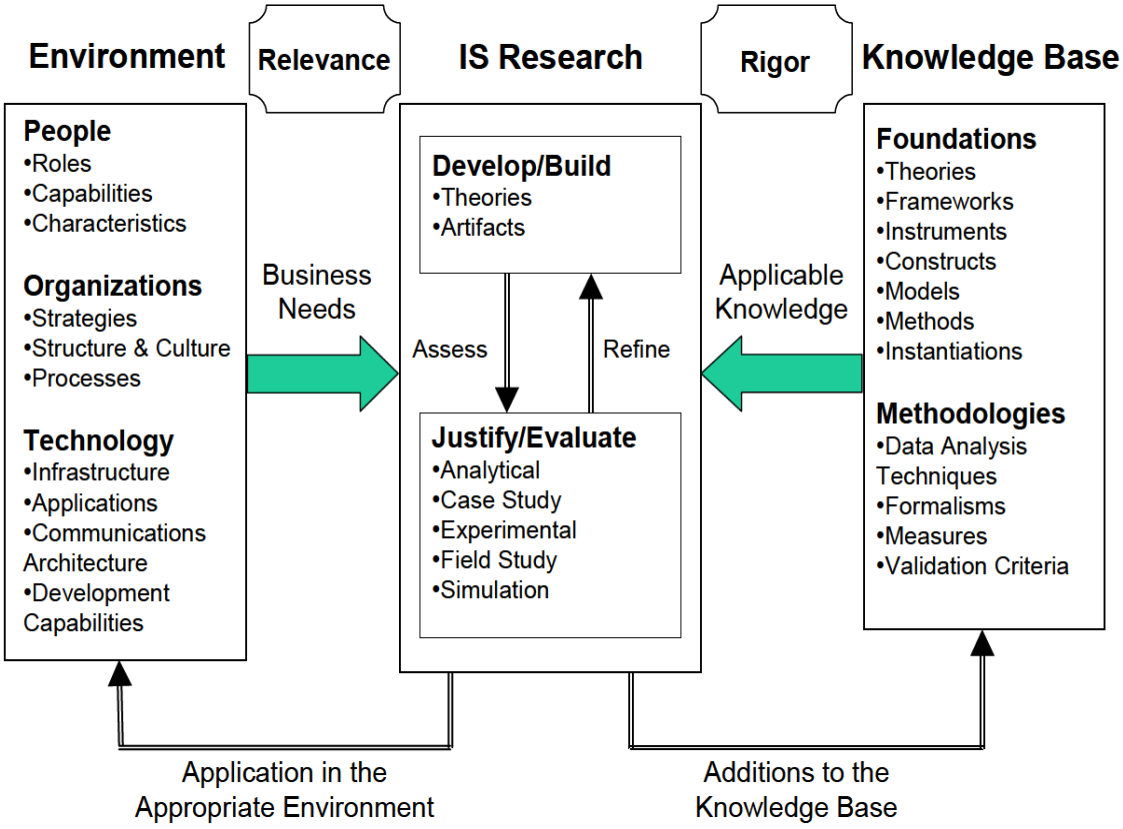


Figure 6.1: Information Systems Research Framework (Hevner et al., 2004)

Research in information systems incorporates people, organisations and technology (Hevner et al., 2004). In Figure 6.1 these are seen (along with their characteristics) on the far left, labelled Environment. This section ensures that the research problem or opportunity is relevant to the real world. The Knowledge Base section on the far right ensures rigor in the research is credible, and relevant previous IS research must be used to guide the researcher's study and methodology. The middle section of Figure 6.1 depicts the IS research process, showing the two main phases of which it is comprised: the development of an artefact or theory, and the testing or refinement of said artefact or theory.

The following section will compare the main paradigms in order to show why design science is the best choice for this particular study, followed by an explanation of how it is applied.

6.2.5 Selecting a Paradigm: Design science

Historically, the positivist approach was recognised as the norm for IT research, resulting in harsh judgement of the interpretivist and critical theory approaches (Oates, 2006). Interpretive research, however, has been adopted more frequently in recent years, with design science emerging recently as a dominant paradigm in many research areas. Critical theory is not as well-known and is much less often used in IT research than the other paradigms.

Before embarking on research, Collis and Hussey (2009) believe that the ontological, epistemological, axiological and methodological assumptions need to be considered in order to make an appropriate decision for the underlying philosophy of the research project. These are summarised in Table 6.1 below.

Table 6.1: Assumptions of the Main Paradigms (Vaishnavi & Kuechler, 2008)

Philosophical assumption	**Positivism**	**Interpretivism**	**Design Science**
Ontological assumption	A single reality, knowable, probabilistic	Multiple realities, socially constructed	Multiple, contextually situated alternative world-states, socio-technologically enabled
Epistemological assumption	Objective, dispassionate, detached observer of truth	Subjective (i.e. values and knowledge emerge from the researcher-participant interaction)	Knowing through making (objectively constrained construction with a context), iterative circumspection reveals meaning
Axiological assumption	Truth, universal and beautiful, prediction	Understanding, situated and description	Control, creation, progress (i.e. improvement), understanding
Methodological assumption	Observation, quantitative, statistical	Participation, qualitative, hermeneutical, dialectical	Developmental, measure artefactual impacts on the composite system

- The aim of this study is the development of a model that will guide the analysis of qualitative text data for use to improve public safety in a smart city.
- The data obtained via participatory crowdsourcing is in the form of natural spoken language, which by nature is subjective and varying and can change according to geographical location and perception of the caller.
- The analysis process must take into consideration the required information and its desired use in order to employ appropriate strategies and techniques.

- Results must be interpreted, although this should be replicable, the human interpretation means that a slight human variance should be expected.

Considering these points, it becomes clear that this study leans toward interpretivism rather than positivist theory. It is, however, more in line with design science. Then, considering that this research project aims to develop and refine an artefact (the model), applicable to a real-world problem, the design science paradigm stands out as the best fitting option.

Hevner et al. (2004) explain that design science aims to create a novel, useful artefact that is intended to be used to solve a real-world problem. This artefact must stand up to scrutiny, must be presented well, and be based on sound research. To ensure all the criteria are met, Hevner et al. (2004) give seven guidelines for design science research. An explanation of these guidelines and the manner in which this study will conform to them follows hereafter.

- **Guideline 1: Problem Relevance**

  From Chapter 1 of this dissertation, it is evident that there is a lack in context sensitive guidance for the analysis of unstructured text data (especially when in larger volumes). In order to address the scarcity of city resources caused by an increase in city population, this study focuses specifically on the analysis of public safety reports for the purpose of informing decisions pertaining to the improvement thereof. Thus this study addresses a relevant, real- world problem which can benefit a specific environment.

- **Guideline 2: Research Rigor**

  The model produced by this research is the result of logical conclusions that are drawn inductively from the review of sound secondary literature including peer reviewed articles, case studies, books and conference proceedings. The model is tested by its application to a working prototype and the related empirical data. The model is further refined using findings from the observation of the prototype. Input from experts and academics lends further rigor to this study via the conversational analysis of expert discussions which were also applied to the refinement of the artefact.

- **Guideline 3: Design as an Artefact**

  Design science states that the research must produce a viable artefact in the form of a construct, a model, a method, or an instantiation (Hevner et al., 2004). The aim of this study is to produce a model that will guide the analysis process of public safety crowdsourced data in text format. A model can be explained as an abstract representation of relationships and factors that can be used as a simple representation or example.

- **Guideline 4: Design Evaluation**

  Design evaluation is the process whereby the designed artefact is rigorously assessed to ensure utility, quality, and efficacy. The evaluation chosen must be based on the specific design artefact constructed by the researcher (Hevner et al., 2004). The model is evaluated by using it for analysing public safety reports obtained via the smart city prototype. The analysis process and application of the prototype produced observations that were used to refine the model. Expertise and opinion has been sought from experts at various stages of this study for guidance and evaluation of findings and conclusions which contributed to further refinement.

- **Guideline 5: Design as a Search Process**

  This guideline entails the description of the research process leading to the development of an effective solution to a problem. To ensure applicability to the problem domain, the research question was broken into three sub questions. The sub questions resulted in four chapters of extensive literature review which explored the setting and environment of the problem. The proposed artefact, constructed from the assessment of related secondary data, was thus developed considering the requirements and characteristics of the problem space. The testing of the artefact and analysis of conversations with relevant experts guided the development of the artefact and added to its refinement. Relevant comments and recommended modifications to the proposed artefact were taken into consideration. The model produced from this research is thus the culmination of an iterative process of literature review, as well as the review and analysis of primary data and experienced opinion.

- **Guideline 6: Research Contributions**

  All design science research must contribute to the specified area of the designed artefact. In design science research, at least one (or more) of three possible contributions must be provided. These contributions include the design artefact, foundations and methodology (Hevner et al., 2004). This research is focused on the development of a designed artefact. The development of an analysis model will allow the progression of the UFH public safety smart city project by ensuring that the reports obtained are analysed in a way that will ensure useful information is obtained for the improvement of safety and living conditions in the smart city.

- **Guideline 7: Communication of Research**

  The last guideline pertains to how the research is presented or communicated to others. Hevner et al. (2004) explain that the research must be communicated to technical and management audiences. The findings of this study will be published in academic journals and conferences and made available to the public for future research. The dissertation will be accessible through the library at the University of Fort Hare. IBM and Buffalo City Municipality (BCM) will also be provided with access to this study for future research and developmental purposes.

These seven guidelines are followed throughout this study, culminating in the development of a useful model. It must be noted that for the purposes of this study, these guidelines will not be followed in numerical order. Hevner et al. (2004) explain that all seven guidelines must be present for a design science study to be complete, but they advise against rigid, mandatory numerical implementation thereof as it is best left to the researcher to judge how, when and where to apply them. This study will apply these guidelines as depicted in Figure 6.2.

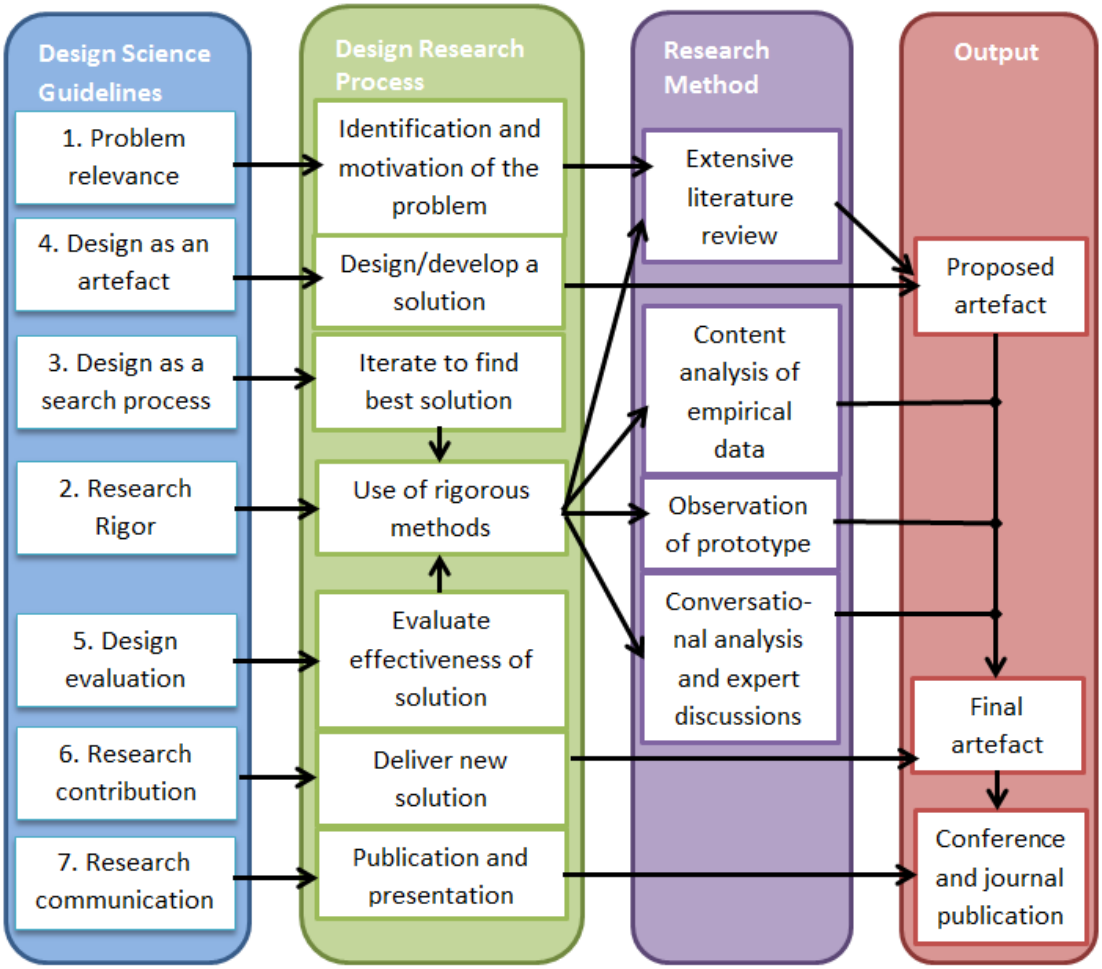


Figure 6.2: Application of Design Science Guidelines

Ensuring that these guidelines are followed throughout the research process will result in credible conclusions and productive, useful output. However, this is only true if appropriate methods are used. The following section will discuss the research methodology, methods chosen, and the reason for its implementation in this study.

6.3 Research Methodology

Collis and Hussey (2009) point out that the researcher needs to choose a methodology that reflects the philosophical assumptions of the chosen paradigm. A research methodology is an approach to the process of research and encompasses a body of methods.

The chief distinction to be made between research approaches is that between quantitative and qualitative research methods. Some studies, however, make use of a combination of qualitative and quantitative methods and this is called mixed methods.

6.3.1 Qualitative Methods

Qualitative research is commonly used in the social sciences context to understand or explain events, human behaviour or their causes. This research method focuses on the collection of qualitative data. Qualitative data is usually characterised as low in volume and high in detail (depth). Open-ended questions are used that result in data being presented in a rich, varying textual format. The most common methods used in qualitative research are case studies, observations, in-depth interviews, action research, physical experiences, theories, social interaction, surveys, Delphi technique and expert review.

6.3.2 Quantitative Methods

Research conducted quantitatively can usually be measured numerically by using closed-ended questions or measuring variables in a controlled environment. Quantitative data can be characterised as high in volume and low in detail. Quantitative research methods often include written surveys, questionnaires, observations, experiments, methods and theories. Presentation of data is usually in a statistical format (for example, graphs and tables) for statistical analysis (trends, average and totals, for example).

6.3.3 Mixed Methods Research

Mixed methods combine both qualitative research and quantitative research. As an alternative to quantitative and qualitative research, mixed methods advocates the use of whatever methodological tools are required to answer the research questions. This can be evident in the type of questions, the collection of data and the procedures and inferences used in data analysis (Teddlie & Tashakkori, 2009).

Although this study is predominantly qualitative, the qualitative data is processed into quantitative data which influenced the findings of the study. In order to achieve the goals set out by this study, a mixed methods approach has been selected as the most appropriate methodology to use.

6.4 Methods used for Data Collection and Analysis

Each research methodology is associated with a number of methods for data collection and analysis. Some of these methods are applicable in more than one methodology, but must be applicable to the problem at hand and its setting.

6.4.1 Secondary Data

Based upon the sub questions identified in Chapter 1, related literature is reviewed in order to understand the theoretical foundation and other empirical studies done in this area. The literature focused mainly on other smart city studies conducted in other locations in the world. Unfortunately, most of these are in first-world developed countries and not in developing countries like this study. Additionally, the literature reviewed also includes books and articles concerning crowdsourcing and techniques for the analysis of qualitative text data. This information is obtained from books, peer reviewed journals, websites, conference proceedings, frameworks and models, previous studies and other online publications. Frameworks and models relate to data analysis and case studies, and reports refer to smart cities and various crowdsourcing systems.

The literature was reviewed in order to inform a theoretical foundation which informed the development of a model for the analysis of qualitative text data. The model is based on inductive logic and was further refined and assessed using empirical data.

6.4.2 Primary Data

A prototype public safety smart city system has been developed by the University of Fort Hare (UFH) and International Business Machines Corporation (IBM) to harvest public safety reports via participatory crowdsourcing. The system has initially been targeted at the city of East London in the Eastern Cape province of South Africa. The city is part of the greater Buffalo City Metropolitan Municipality (BCMM), which in the 2011 census was recorded as having a population of 1.4 million people (City Population, 2014; Statistics South Africa, 2012). The initial focus for the project was chosen to be only the 761 996 citizens of East London itself (East London, 2013; ECSECC, 2012) due to time, financial and geographical constraints, and to have multiple reports by different citizens in close proximity to each other (concentrated area).

The citizens of East London were requested to volunteer public safety reports via a telephonic interactive voice response (IVR) system or a mobile website (or mobi site) which has been built specifically for the purpose of collecting these reports. Making use of these channels ensures that most East London citizens are able to participate.

The public safety reports are free- flowing responses (unstructured, natural language) obtained voluntarily from the public. The reports were transcribed into text and combined with text reports obtained from the mobi site.

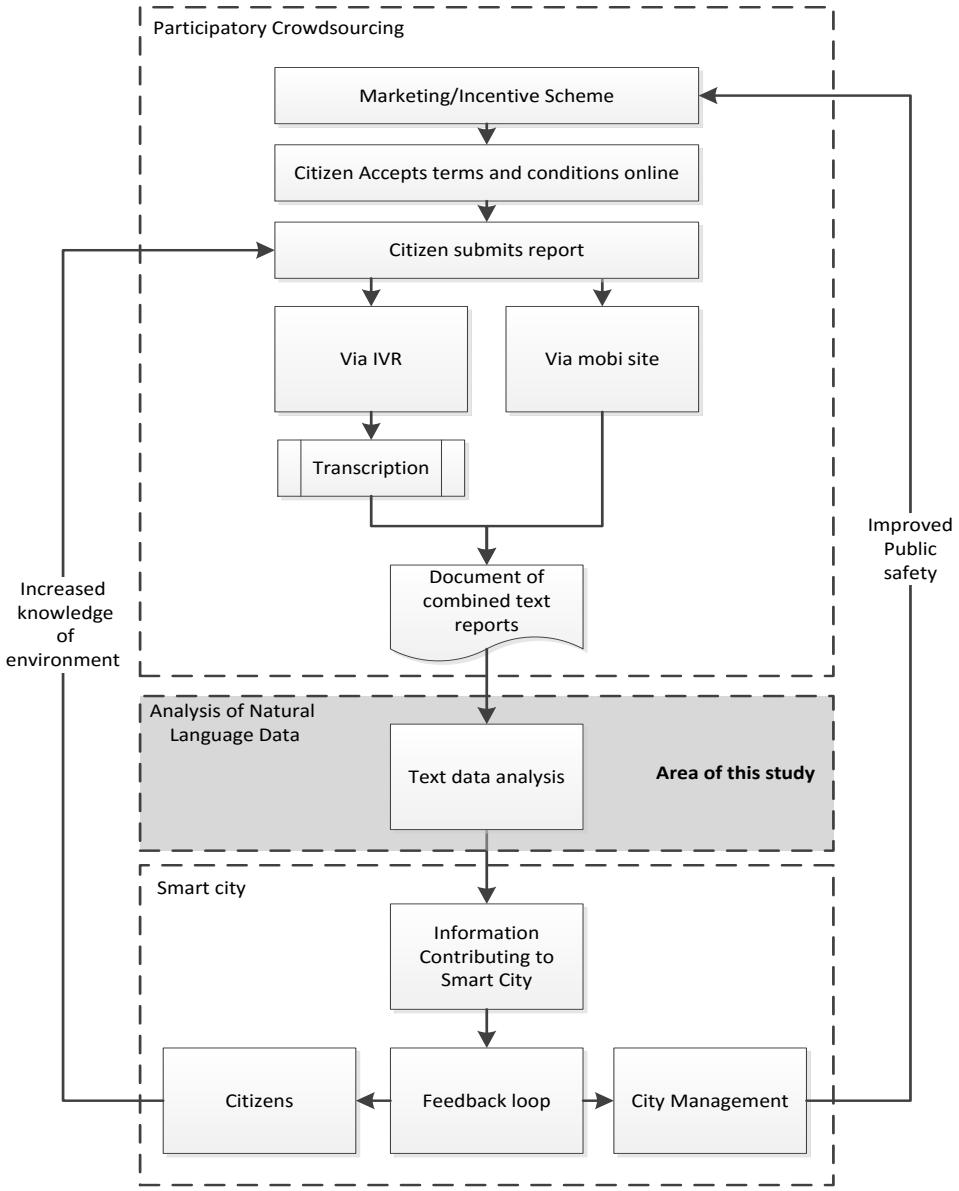


Figure 6.3: Flow of data from greater project to this study

The public safety reports obtained from the prototype make up the first set of empirical data used in this study. In order to assess the model produced by this study, the model was used with the prototype system by application to the empirical data obtained.

The analysis method which was applied to the citizen reports, according to the model, is the content analysis method. As described in Chapter 5, there are a number of qualitative data analysis methods that can be used to analyse text data. These include, but are not limited to, Semiotics, Qualitative comparative analysis, Constant comparison analysis and Keywords-in-context. Of the available options, content analysis has been identified as the best suited option for this particular project. This selection is based on the nature of the data as well as the goals of the analysis.

The public safety reports are relatively short and contain some specific identifiable information such as a description of the incident, place, date and time (Some examples can be seen in Appendix A). The main complication identified in Chapter 5 is that of semantic and synonym variations. The desired outcome of the greater analysis for the smart city is the quantification and modelling of public safety data as simply and quickly as possible for more proactive responder deployment and resource allocation. Content analysis is best suited to these factors as it is one of the simplest yet well-structured processes for reducing and quantifying qualitative data to a structured and quantifiable format.

In order to identify codes and coding hierarchies in the public safety reports data, this study will adhere to the guidelines as specified by Krippendorff (2004):

> "*A research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use*" (Krippendorff, 2004, p. 18).

By using specific rules of coding, large amounts of qualitative data is reduced and can be quantified in order to make it easier for analysts to read and identify trends and patterns. Following the content analysis technique, the data itself is used to identify the groups that need to be coded; the same data is then coded according to the groups. This is an iterative process and the groups (or code list) can be amended as more become evident during the analysis process.

The Nvivo software package was used as a tool to code the reports for performing the content analysis. There are a number of computer assisted qualitative data analysis software (CAQDAS) packages for researchers to choose from, available as open source and commercial applications. These applications are generally very similar, but have minor differences which are only relevant to certain specific cases (Lewins & Silver, 2009). The slight differences will not have any bearing on the findings of this study, and the choice of which one of the three came down to personal preference and ease of access. Thus, after experimenting with the trial versions of these three options, the author found a personal preference, based on user friendliness and intuitive controls, towards Nvivo 10. Additionally, a licenced version of the software is available at the institution where this research is carried out, thus allowing the researcher easy access to the full version and support if required.

Through the process of content analysis, Nvivo was used to develop an ontology as required by the model. The instances and frequency of codes in the reports are grouped and quantified using Nvivo in order to reduce the data to a structured form. The software served to organise, manage and perform queries on the data, but cannot interpret results and findings. The Nvivo queries will not in itself answer the research questions; ultimately it is up to the researcher to interpret the findings.

The assessment of the model with the prototype served to assess its functioning in a realistic environment. This process also resulted in the second set of empirical data as the researcher was able to make observations and thus record qualitative findings which served to aid in the refinement of the model. Observations of the processes and functioning of the prototype and proposed model informed the researcher's view on the real-world functioning of the model and added to its refinement. The most important of these findings are reported and discussed in Chapters 7 and 8.

A third source of empirical data was conversational analysis. Conversational analysis is a method of data collection through social interactions which include verbal and non-verbal cues (Goldkuhl, 2004). Regular, ongoing conversations with industry experts and academics in related fields (national and international) were held in order to help guide and inform this project throughout its progression.

These meetings included, but were not limited to:

1. Public safety smart city project group planning meetings, which included students and academic staff involved in various aspects of the project.
2. Research workshops:
    a. A four day research workshop attended by two visiting professors from Plymouth University in England, as well as two visiting professors from Nelson Mandela Metropolitan University (NMMU) and one from UFH.
    b. Two annual one- day workshops attended by lecturers from NMMU, Rhodes University, Walter Sisulu University (WSU) and UFH.
3. An IBM smart city presentation with BCMM departmental managers and international IBM representatives. Followed by a discussion session with the BCMM IT manager and local IBM representative.
4. Informal Conversations were engaged in with industry representatives from the UFH Department of Information Systems (IS) advisory board on three separate occasions at six month intervals.
5. Departmental Colloquia, held bi-monthly, involved progress reports and discussions with all academics from the IS department at UFH.
6. An artefact refinement question and answer session attended by all academics from the IS department at UFH as well as a visiting professor from NMMU.

The conversational analysis has been considered during the entire development of the analysis model. Much of the contributions and guidance obtained were applied to steering the research development and the main findings which applied to tempering the artefact and conclusions of this study are discussed in Chapter 8. The input from experts has been taken into consideration throughout this study and has helped to ensure the usefulness and applicability of the model produced.

Obtaining the right data and applying the appropriate methods ensure rigour and credibility of the study, as well as complying with the rest of the seven design science guidelines as depicted in Figure 6.3. Having outlined the data and methods included in this study, the following section will state what is excluded by presenting the delimitations of this study.

6.5 Delimitations of the Study

Delimitations of a study allow the researcher to create a parameter around the study by clearly stating what will be included and excluded in the research (Hofstee, 2006). The smart city qualitative data analysis (SCQDA) model developed in this study was tested in the context of public safety, but it would function in a similar way if applied to any other smart city subcategory. As principles of content analysis are applied to the data in order to form an applicable ontology, a change in the subject matter of the data would change the focus of the model to the topic at hand.

This study has made use of empirical data obtained through participatory crowdsourcing via an IVR and a mobi site. The use of participatory crowdsourcing entails humans acting as sensors by voluntarily reporting public safety issues in East London. This study will thus not be about the use of automated sensors, and the data will be restricted to an individual's willingness to contribute. The data obtained from the IVR system is transcribed into text format, matching the format of the mobi site data. This study will therefore not focus on analysing other formats of data such as graphics and video.

The data in question is obtained from the smart city public safety prototype run in East London by the University of Fort Hare. Although East London forms part of the greater Buffalo City Metro, the reports will be from only East London as the initial prototype area due to constraints such as time and cost and wanting a smaller geographically concentrated area. The reports will thus only be concerning public safety incidents that have occurred in and around East London as reported by citizens of East London.

The reporting format is in English and thus limits reports to English speakers with access to a mobile phone, telephone or mobile Internet. As this study is aimed at the analysis of the data, it will not include ensuring quality of data obtained, adoption, participation or ensuring the use of the information once analysed. Thus, the boundaries of this study have been emphasized through delimitations, which lead to the ethical considerations of the study.

6.6 Ethical Considerations

The researcher applied for ethical approval from the University of Fort Hare and its ethics committee prior to the commencement of this research. As the public safety reports used in this study had already been collected under ethical clearance (reference number: FLO011SCIL01), there was no further engagement of humans or animals, thus ethical clearance was granted. Punch (2006) holds that the responsible researcher must comply with academic integrity and honesty, while always respecting all people and animals involved. Of the range of considerations applicable to research, the following were found to be applicable to this study:

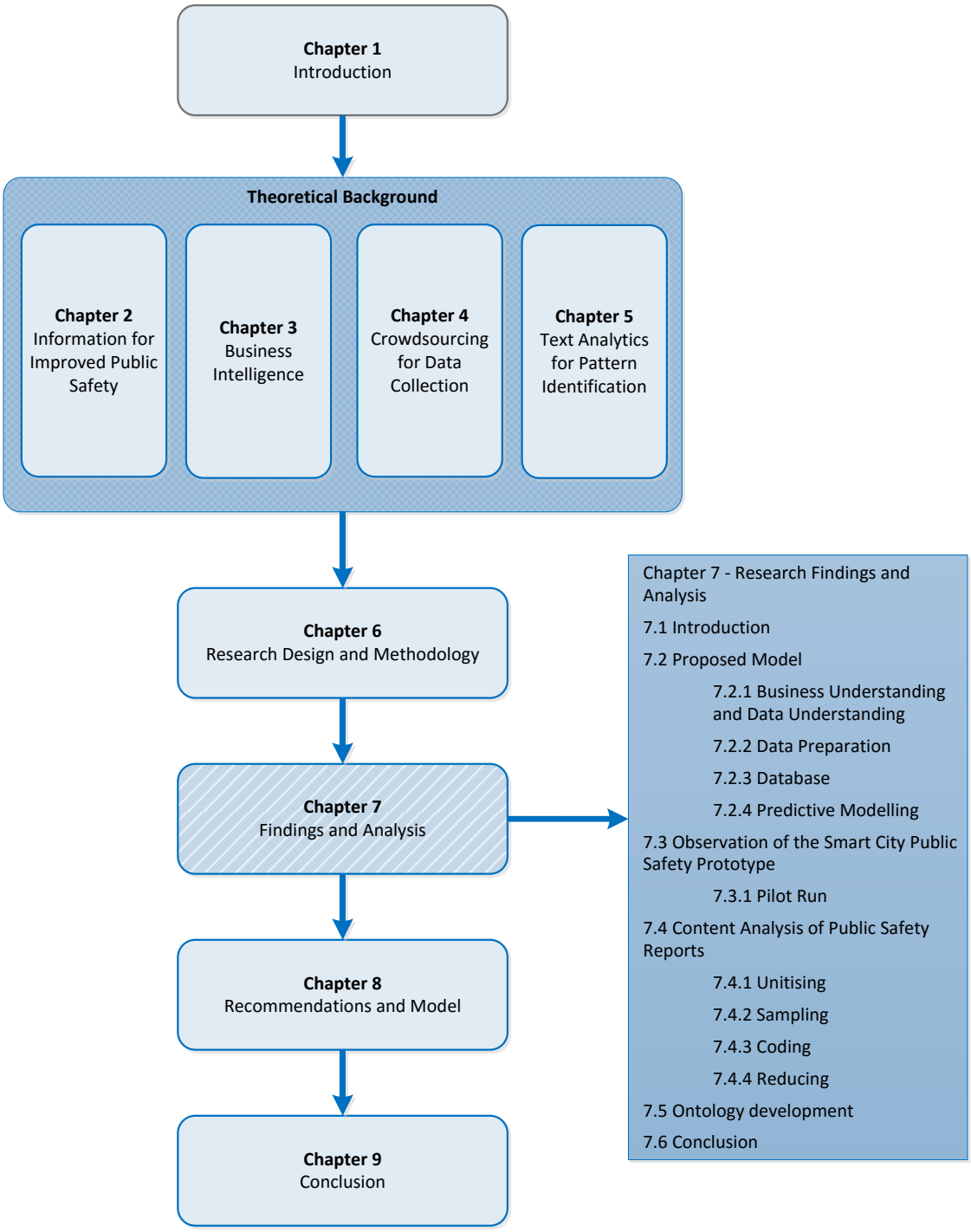
1. Anonymity – No personal identification is recorded or used in this study. Participants may choose to leave a contact number for inclusion in the incentives scheme, though this is optional. There is also no way to link this to any other personal information.
2. Privacy – Empirical data will only be used by the researcher and supervisor for its intended purpose.
3. Security –Data will be stored safely and securely at all times, accessible only by those authorised to do so.
4. Misuse of results –This ethical risk will be mitigated by ensuring information gathered from participants will only be used for the intended and expressed purpose of the research and in its original state, free of manipulation or adaptation.
5. Harm and risk – It is unlikely that harm could come to the participants if any personal information is provided. Nonetheless, all personal data will be removed from reports, as mentioned above.

6.7 Conclusion

This chapter provides a detailed description of the manner in which this research project was conducted. The chapter firstly discussed possible research paradigms including positivistic, interpretive, critical theory and design science. This was followed by a discussion on the most appropriate paradigm for this particular study, design science, and an explanation of how this study is structured in order to conform to the named paradigm and its seven guidelines. The research methods, qualitative, quantitative and mixed methods, were explained before motivating why a mixed methods research methodology was used for this study.

The next section discussed the data collection and analysis methods that included the greater structure of the public safety smart city project, identifying the area of this study within its context. The methods discussed included prototype observation, content analysis, and conversational analysis. The delimitations of the study followed the data analysis section to clarify the research area. The last section of this chapter listed the ethical issues considered applicable to this study. All discussions above allow this study to be conducted successfully, resulting in the generation of a viable business model for the analysis of qualitative text data for a public safety smart city project. As this chapter has described the research design and methods, the next chapter reports the findings and analysis process.

Chapter 7 – Research Findings and Analysis



7.1 Introduction

This chapter presents the findings of the study conducted in order to recommend the most suitable analysis process for qualitative text data used to improve public safety in a smart city within a developing nation.

As outlined in the preceding chapter, multiple data collection sources and methods have been used to compile the empirical data for testing and refining the proposed model. These include the observation of a prototype, content analysis performed on reports obtained from the prototype, and the conversational analysis of discussions with academics and other experts.

In this chapter, first a model is proposed based on findings from literature and built upon existing theoretical structure drawing from Chapters 2 through 5. Secondly, observations and findings are presented that are based on the use of the model. These findings have been used to refine the model into a final version which will be presented in the next chapter.

7.2 Proposed Model

The proposed model draws its initial foundation from the Cross-Industry Standard Process for Data Mining (CRISP-DM) as it is a tried and tested Data Mining (DM) methodology developed and used by industry leaders. Not all the elements, however, are used directly as they are in the CRISP-DM, but have been adapted and adjusted to suit the purpose of this study. Likewise, some parts of the CRISP-DM were not found to be applicable to text mining and were thus not seen to be applicable to this study.

The CRISP-DM also includes steps to evaluate and deploy the information obtained from the mining project. This does not necessarily form part of the analysis process, but can rather be viewed as what is done with the resulting information after the analysis process. It thus falls beyond the scope of this study, as the scope of this study is limited to analysis.

The resulting Smart City Qualitative Data Analysis Model (SCQDA) proposed by this study follows as Figure 7.1, with a description of the elements thereafter.

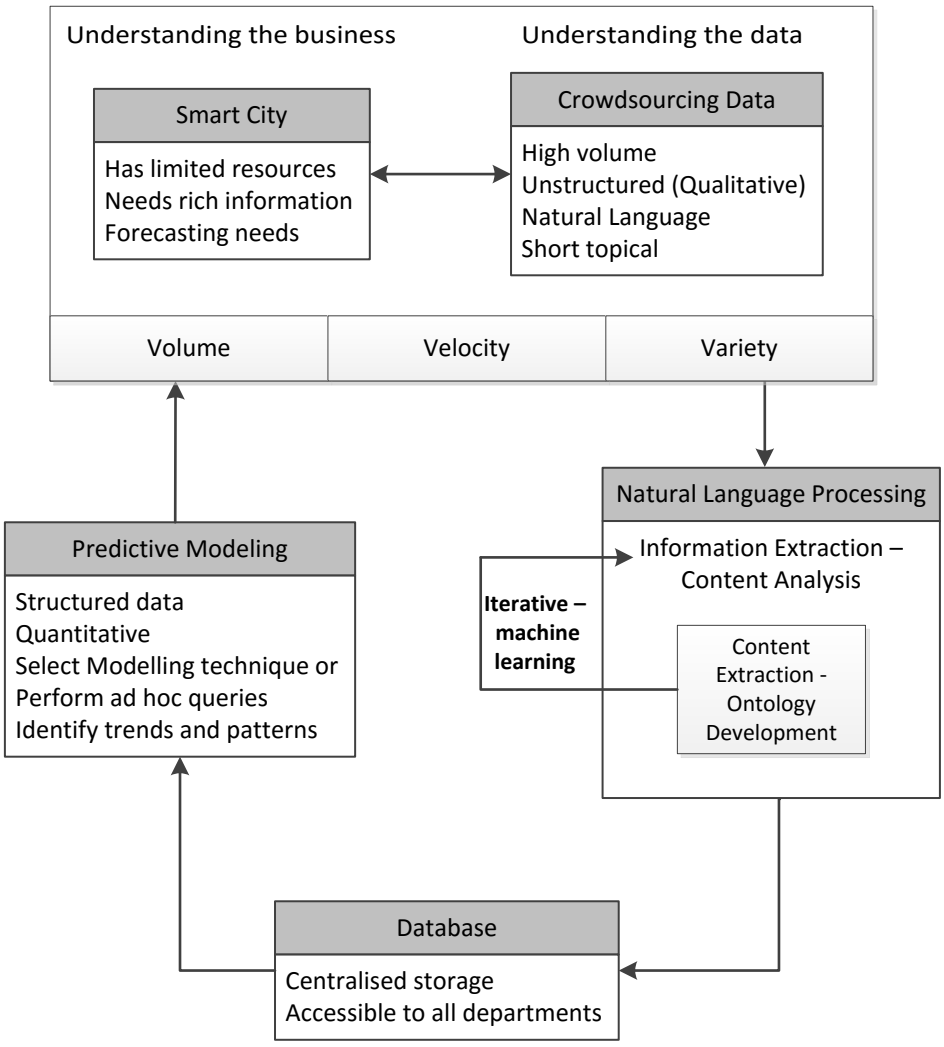


Figure 7.1: Proposed Smart City Qualitative Data Analysis Model (SCQDA)

The following few paragraphs explain the development of the SCQDA model in more detail with reference to the elements of the CRISP-DM and how they have been adapted.

7.2.1 Business Understanding and Data Understanding

Understanding the problem space, goals of the analysis and nature of the data available is vital for any analysis project to be successful. This can be seen as the planning for the Text Mining (TM) project. This is also an integral part of later steps and must be adjusted and updated throughout the lifecycle of the project. In this model the understanding of business and of data has been brought together within the greater framework of big data. To ensure the implications of big data are considered, the

characteristics of big data (volume, variety, and variety) must become part of the planning process, specifically when developing the understanding of the business and of the data.

A large part of this phase is deciding which computer assisted qualitative data analysis software (CAQDAS) to use. Literature has shown (Chapter 5) that there are multiple CAQDAS packages available which can help tag, code and manage unstructured data. Most of these packages are able to perform the required work, but they do vary slightly in the more minor characteristics. It is therefore left up to the individual or organisation to choose the right package to suit their needs.

7.2.2 Data Preparation

This step was adapted to suit Text Mining (TM), with the goal of structuring unstructured text data. This now becomes the biggest step in the TM process as the text needs to be prepared according to Natural Language Processing (NLP) methods for BI use. To enable querying of the data, it must become structured in order for the required information to be extracted. To achieve this end, the text must be tagged (or coded) for grouping and categorising (Concept extraction).

A key factor here is that the process of structuring the data must be largely automated to enable the use of large volumes of unstructured text data (big data). This requires a certain amount of machine learning as the software performing the tagging requires guidance in how and what to tag. To this end, a small amount of data (reports) is analysed manually as an example to the software in order to start forming a hierarchical ontology applicable to the subject area.

The ontology informs the AI in order for large scale content analysis (tagging and coding) to be performed automatically. Only the few outliers of the automated process should be analysed manually in order to assess whether it must be added to the ontology or if it can be disregarded or corrected as a one-time anomaly. This results in a largely automated functioning with a small amount of minor manual input.

7.2.3 Database

BI related principles regarding database use also arose from the literature. Most importantly was the use of a single unified source of information, easily accessible to all stakeholders with updates as close to real time as possible. A key factor in this model is the placement of the database entity. Instead of just scrubbing data, the preparation process herein structures the data prior to committing it to the database. Thus the information users can query the database directly in much the same way as they would have with a traditional structured database, without having to prepare the data prior to each individual and ad hoc query.

7.2.4 Predictive Modelling

The CAQDAS has structured, summarised and quantified the data. This model shows how the process can be automated to make gains in time and effort as required for good BI. Chapter 3 states that the greatest expense in time and cost is incurred when attaining and preparing the data.

As discussed in Chapter 5, a CAQDAS package can help manage and automate much of the work involved in analysis, but it cannot interpret and ascribe meaning to findings. The information still needs to be subject to human interpretation before it becomes meaningful. With the preparation step becoming the main analysis step, the data is structured and ready for reporting by the information users. The data can then be used for comparative modelling or queries. A dashboard can be used to display required data and figures. Structuring the unstructured data prior to database storage allows for greater flexibility in ad hoc reporting, but standardised reports can also be set up.

This also leaves room for further research to be conducted. The researcher recommends that research in the fields of computer science and statistics is conducted with the aim of exploring specific algorithms to be used in modelling the data (once structured and quantified) for use in forecasting public safety resource needs.

Making use of this model as a guide, the public safety reports harvested through a prototype participatory crowdsourcing system have been analysed in order to test and refine the model. A discussion of the observations from this process follows.

7.3 Observation of the Smart City Public Safety Prototype

In conjunction with IBM, a project team from the Department of Information Systems at the University of Fort Hare developed a participatory crowdsourcing system. This system allowed citizens to report public safety issues telephonically or via the use of a mobile enabled website (mobi site).

Conversational analysis led to the use of a speech based system rather than a purely web based option (such as just the mobi site) or purely text based option (such as a Short Message Service SMS). This is to ensure ease of use in an environment with low literacy rates and lack of technology infrastructure availability in the outlying regions of East London.

The ubiquity of the basic cellular phone in East London means that using a voice based system would allow all citizens to participate. Using the mobi site concurrently presented an additional option for those who preferred to submit typed messages.

The process of reporting a public safety incident was as follows (shown in Figure 7.2):

1. The respondent must first access a website presenting the terms and conditions of participation and choose whether or not to accept them. This only needs to be done once prior to submitting one's first report and not for subsequent reports.
2. Having accepted the terms and conditions, the respondent is then able to submit as many reports as they like, telephonically or via the mobi-site. If the respondent chooses not to accept the terms and conditions, they choose not to participate any further.

The communication process made use of an Interactive Voice Response (IVR) system in conjunction with a voicemail exchange. Calls were recorded, stored to the voicemail server, and transcribed into text. The flow of data through the entire system is depicted in Figure 7.1 below, highlighting the scope of this particular study.

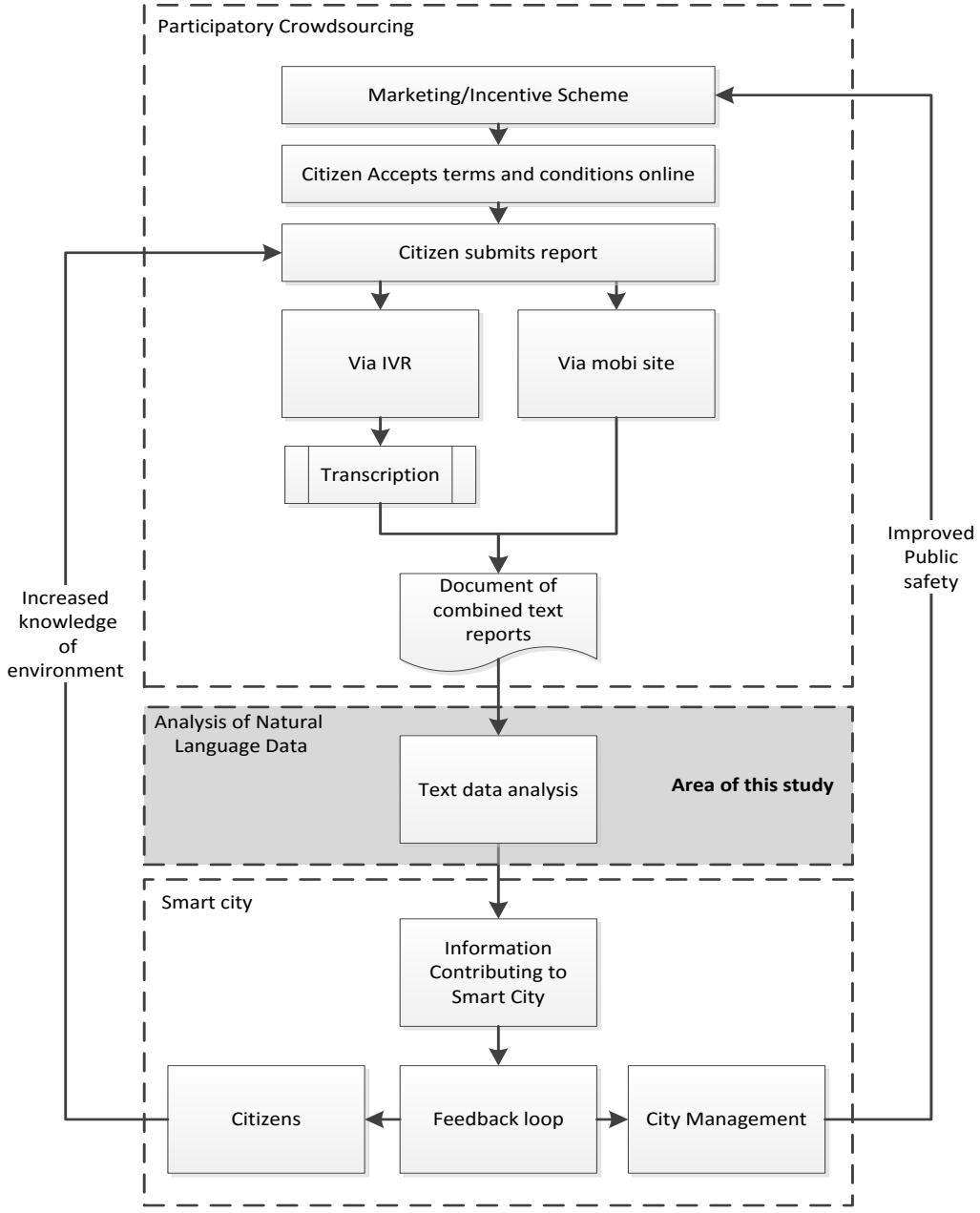


Figure 7.2: Flow of Data from Greater Project to This Study

The reports obtained from the prototype were analysed according to the proposed SCQDA model. A CAQDAS package was selected (in this case Nvivo 10 was selected as the researchers preference) and used to tag and code the reports according to the time, place and incident type described. A content analysis was performed on the reports which identified keywords that are used in East London to describe public safety incidents and situations. Reducing the code groups into hierarchical groups developed a hierarchical ontology to use for further analysis projects.

7.3.1 Pilot Study

Prior to opening the IVR system to the public, it was decided to conduct a trial run. The researcher and three colleagues phoned in and submitted a minimum of ten fictitious reports each; these fictitious reports were deleted prior to the inception of the prototype's actual running period and thus had no impact on the empirical data used in this study (Teddlie & Tashakkori, 2003). Olivier (2013) states that the results of a pilot study has little importance as the aim is to identify flaws in the process prior to actual use, rather than producing results.

The main contribution of the pilot study was in the observation that the interactive process and instructions given via the IVR could be deemed cumbersome by the caller. In light of this, the recordings used by the IVR system were adjusted in order to shorten instructions and lessen the amount of interaction required from the caller.

7.4 Content Analysis of Public Safety Reports

As discussed in Chapter 5, in order to overcome semantic problems in the process of automating the analysis of natural language text (especially within a developing country), ontology development was used. A hierarchical ontology merged synonyms and colloquialisms into structured "categories". This was done as a precursor to help automate the analysis process, but also makes up part of the analysis process itself as it is in line with the data reduction step of content analysis, which is required in order to identify patterns and trends within the data. For this purpose, the content analysis research technique, as specified by Krippendorff (2004), was used.

Although this qualitative research technique has often been applied to secondary data in order to develop the understanding of existing literature, in this study the technique has been applied to the empirical data in the form of the public safety reports.

Stemler (2001) explains that content analysis is "a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding" (p. 1). The following section analyses the various naming conventions used by East London citizens when reporting public safety issues and groups them into a hierarchical structure. The hierarchical grouping then forms an ontology which is used

for machine learning in order to automate the coding and analysis of public safety reports. To execute a reliable and valid content analysis, Krippendorff (2004) defines six steps: unitizing, sampling, coding, reducing, inferring, and narrating. The content analysis performed in this study is presented according to these categories hereafter.

7.4.1 Unitising

The content analysis was performed on the empirical data of this study: the public safety reports submitted to the smart city public safety prototype by the citizens of East London. During the live functioning of the prototype, a total of 663 viable reports were recorded.

7.4.2 Sampling

Sampling refers to drawing a manageable set of data from the greater whole when the entire population is too large to work with. In this study it has been emphasised that the developments in modern technology have enabled the use of big data through automating parts of the process. Thus, in this study the entire population of 663 reports were analysed.

7.4.3 Coding

Coding involves tagging words or phrases representing concepts within each of the reports. In this way the unstructured text is transformed into more structured analysable units. This was carried out using the Nvivo 10 qualitative data analysis software package. The transcribed calls and their associated metadata (date and time of the call) were transcribed in an Excel spreadsheet. This spreadsheet was then imported to Nvivo as a dataset.

The coding process started with manually highlighting the desired word(s) or phrase(s) which were then assigned to a particular node or code grouping. The Nvivo interface allows the user to do this with a simple “drag and drop” or by using a menu interface. Nodes can be created before commencement of coding and additional nodes can be created as the process continues. Thus all codes do not have to be identified before starting the process, but can be if available. This allows for concepts and ideas to emerge from the data itself (emergent coding).

Table 7.1: A Selection of the Public Safety Reports Obtained

Report verbatim	Date received	Time received
August 1 late afternoon, there is a bakkie with trailer and boat broken down on the last uphill before Kaysers Beach turnoff on outward bound side. The lane is partially blocked causing a hazard.	2014/08/11	15:49
In Southernwood on the 12th of August in the morning, I spotted 3 men jumping over the wall in my neighbour's yard with an ironing board and a mop.	2014/08/12	18:15
29/04/2014 a housebreaking occurred during the night at Fiddlewood Lane in Cambridge West. A vehicle was also broken into at Morrison Road.	2014/05/09	06:03
MVA on Mdantsane Access Road in East London at 10 a.m. on the 29th of September. EMS already on accident scene.	2014/09/29	11:14
Attempted vehicle theft prevented outside Falcon Ridge Complex - Harburn Road, Abbotsford tonight at 21.30 - Criminals fled when spotted by RED ALERT Patrol Guard Shannon Jasson during his patrol. The vehicle owner was visiting family and is happy to still have his vehicle (with only a broken window).	2014/09/25	14:01
On the 14th of August there were two burglaries in Coetzee Street.	2014/08/21	14:48

Table 7.1 displays just a few of the reports collected as an example to the reader.

Once some sections have been tagged, Nvivo can search the rest of the dataset for similar words based on what has already been coded. This further exemplifies the need for an ontology of locally used terms in order to automate the coding of larger datasets. The tagging process can be automated, but the software needs a starting point. It must be told what to look for and how to code it once found. A hierarchical ontology of terms used will give it this starting point and frame of reference.

For the coding of public safety reports, the literature has shown that three broad categories of data are required: time, place and issue (ie., The what, when and where of the incident).

7.4.3.1 Coding the Time Data

The date and time the report is submitted is captured automatically by the IVR and mobi site. These details are thus uniform and can be coded automatically in a chosen format, or even be assigned to the report as an attribute value. This coding thus delivered 663 tags each for the 663 reports.

However, the reports are not always submitted at the time the incidents occur. A number of reports are submitted at a later time, perhaps when more convenient for the respondent. In such cases the caller included a date and time in the report which does

not match the time recorded by the system. Coding this date and time of the issue was found to be much more complicated than anticipated as great variation will take place.

A listing of the 12 months of the year can easily be made mapping the name of each month to its number (January – 1, February – 2, etc.), but the format can vary beyond this. Variations in date format included the following examples (a random date is used):

- *1 February 2014*
- *February 1 2014*
- *first of February 2014*
- *February the first 2014*
- *1st of the second 2014*
- *1,2,2014*
- *2014,2,1*

Some reports omitted certain parts of the date, such as the year or even the month, where others made reference to a day such as *yesterday* or *last Saturday* instead of giving a specific date. The time references similarly varied in structure (*two thirty, half past two, thirty minutes past two*), but contained more references such as: *early, late, afternoon* and *lunch time*.

Recognising and understanding all the possible variations in date and time structure within natural language automatically may still be beyond the capability of most current technology. An easier alternative for this could be to control or guide the date and time aspect of the reporting in such a way that the details become recognisable to the software by enforcing a specific structure. As discussed in Chapter 3, when gathering data via crowdsourcing, many aspects of the quality and relevance of the data obtained can be steered in how the request is phrased and guided.

Considering that the current goal of the analysis is predicting future events and not the investigation of past events, time and date references may not have to be exact. The researcher suggests that a suitable structure be chosen and applied allowing automated coding to be done accordingly; mismatched or incomplete data can possibly be substituted or supplemented using the metadata date and time values.

It is further suggested that a list of references such as morning, midday, afternoon and evening be created. All time indicators falling within the specified time section can then be grouped together.

A suggested grouping structure could be formed as follows:

- Early Morning – 0:00 until 05:00
- Morning – 05:00 until 11:00
- Midday – 11:00 until 14:00
- Afternoon – 14:00 until 17:00
- Evening – 17:00 until 20:00
- Night – 20:00 until 0:00 (12 PM)

Categorising the above groups will allow the information user to draw reports that collate the public safety issues or crime categories according to the time of day they occur; or list all the public safety issues that occur within a specific time range such as *Evening* or *Night*. One could easily see, according to the amount of tags, which time category has the most occurrences of specific or general public safety issues. If there is more than one report per incident then a clearer picture can be established by linking reports.

7.4.3.2 Coding the Location Data

Coding the data regarding the area was separated into "suburb" and "street or landmark". The suburb was deemed the most important location related data and was coded into groups. The variation in the street or landmark data is very high and not always required and was therefore not coded individually. However, the data is retained and can be referred to as needed on a case to case basis.

As shown in Table 7.1, a total of 52 suburbs were named in the reports, culminating in 427 tags. This shows that a minimum of 236 reports had no suburb specified. Some reports were found to have made reference to more than one suburb. Automated coding based on a complete ontology will tag all named suburbs as individual tags.

Reports that contained no area reference are either incomplete or refer to general issues that are not limited to a specific suburb. The following excerpt shows a report that specifies no area information as it is a general warning of fraud which can occur anywhere:

"This issue occurred on the second of March, please be aware... there is a scam currently targeting pharmacies involving airtime… the con artist impersonates a doctor local to the pharmacy, he thoroughly researches and so answers any queries you may have. The impersonator takes thousands of Rands of airtime on his account and promises to call in and pay later, he never does and when you contact your local doctor he will think you have lost your mind. We have just been caught out as suckers. Please be aware all retail companies, he could change his strategy."

This respondent, however, should have reported the actual incident experienced, which would be associated with a particular suburb, rather than issuing a general warning, which can be inferred. Therefore it can be said that this respondent (and a number of others) did not have a technically correct understanding of exactly what details are wanted from the report, even though it was specified by the IVR and the mobi site.

Attempts to remedy problems such as this can take the form of adjustments to the IVR prompts or to the initial information that is disseminated to the public. One must also consider that the majority of respondents did not make this mistake as this type of report was in the minority. Therefore it would be reasonable to assume that long- term use of such a system could automatically solve such minor misunderstandings as users' experience levels increase.

The main finding for this section is that most suburbs are referred to using their actual name, with only two reported areas referred to using more than one name.

Table 7.2: Suburbs Reported

Number	Suburb	Tags
1	Haven Hills	3
2	Amalinda	29
3	Stirling	8
4	Gonubie	38
5	Southernwood	17
6	Beacon Bay	35
7	Nahoon	25
8	Collondale	2
9	Willow Park	2
10	Braelyn	9
11	Bonnie doon	1
12	Town	12
13	Abbotsford	9
14	Greenfields	10
15	Quigney	34
16	Vincent	27
17	Morningside	8
18	Duncan Village	16
19	Cambridge	35
20	Cambridge West	8
21	Stirling	8
22	Scenary park	2
23	Industrial area	2
24	Orange Grove	1
25	Berea	9
26	Dorchester Heights	3

Number	Suburb	Tags
27	Woodbrook	1
28	Bonza Bay	2
29	CBD	4
30	Sunnyridge	7
31	Summerpride	3
32	West bank	7
33	Rosemount	1
34	Arcadia	1
35	Selborne	1
36	Parkridge	1
37	Mdantsane	15
38	Wilsonia	6
39	North End	1
40	Egoli	2
41	Ziphunzana	5
42	Leeches Bay	1
43	Bunkers hill	1
44	Kaysers Beach	3
45	Kidds Beach	3
46	Fort Jackson	2
47	Stony Drift	1
48	Buffalo Flats	3
49	Oriental Plaza	1
50	Chiselhurst	1
51	Brookville	1
Total tags		427

Trying to make any sensible inferences is again very difficult due to the volume of the data, thus reiterating the need for further reduction. One can, however, already visualise the data in graphical format which lends insights to the comparison of how many reports were submitted per suburb. Figure 7.3 shows a graph of the 20 suburbs with the highest number of tags. It is also evident at this stage that some of these areas can be combined, but this was done as part of the reducing step discussed later.

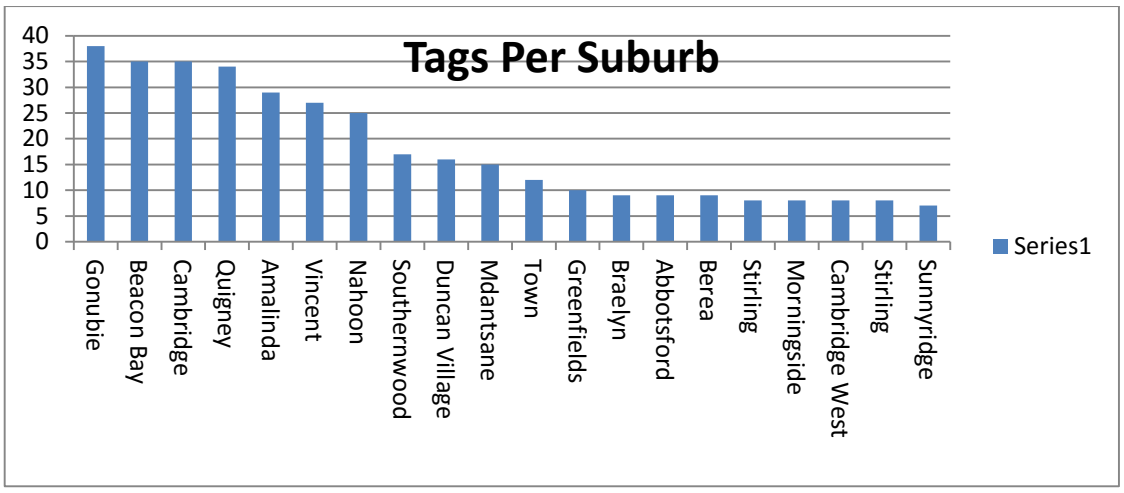


Figure 7.3: Tags Reported per Suburb

From Figure 7.3 it is evident that the suburb with the highest number of incidents reported is Gonubie with Beacon Bay and Cambridge being second and third respectively.

It is important to note that at this point, this does not mean that Gonubie has the highest rate of issues occurring in East London. It simply means that of the issues reported to the public safety smart city prototype, Gonubie had the most incidences reported. Once a fully functional system is put in place and adoption and participation rates are high, one will be able to generalise the findings of the system to the overall rate of issues. Thus another topic for future research is raised: how to ensure adoption and how to motivate participation.

7.4.3.3 Coding the Issue Data

To describe the "what" data, the researcher chose the word issue due to the fact that not all reports describe an event; some describe a situation or condition experienced.

Coding the issue, as displayed in Table 7.3, revealed 23 codes and a cumulative total of 723 tags within this category. The number of tags exceeds the total amount of reports as some reports describe more than one issue. Nvivo also allows one piece of text to be tagged in multiple nodes in case it qualifies as more than one. It can be noted that some of these issues are very similar and can be combined, but this was dealt with in the reducing step.

Table 7.3: Tags Per Problem Reported

Number	Problem	Tags
1	Potholes	45
2	Harassment	8
3	Reckless endangerment	27
4	Road hazard	58
5	Loitering	5
6	Motor Vehicle Accident (MVA)	104
7	Izinyoka (Electricity and cable theft)	19
8	Theft	91
9	Housebreaking	115
10	Vandalism	13
11	Mugging	61
12	High jacking	6
13	Fraud	25
14	Holdup	9
15	Assault/ Attack	21
16	Stabbing	8
17	Bomb threat	1
18	Suspicious activity	29
19	Fire	12
20	Illegal Drugs	6
21	Murder	9
22	Shoplifting	2
23	Unsafe conditions	49
Total Tags		723

Simply counting and comparing values gives immediate insight into which type of incidents are more common than others. Figure 7.4 emphasizes this by showing a visual representation of the incident tags ordered from highest to lowest.

It is no surprise that motor vehicle accidents (MVAs) are one of the most frequently reported issues. What is of interest is the high amount of theft occurring. Housebreakings are alarmingly high, being reported slightly more times than MVAs over the period that the prototype was running, with other general theft and mugging following just after.

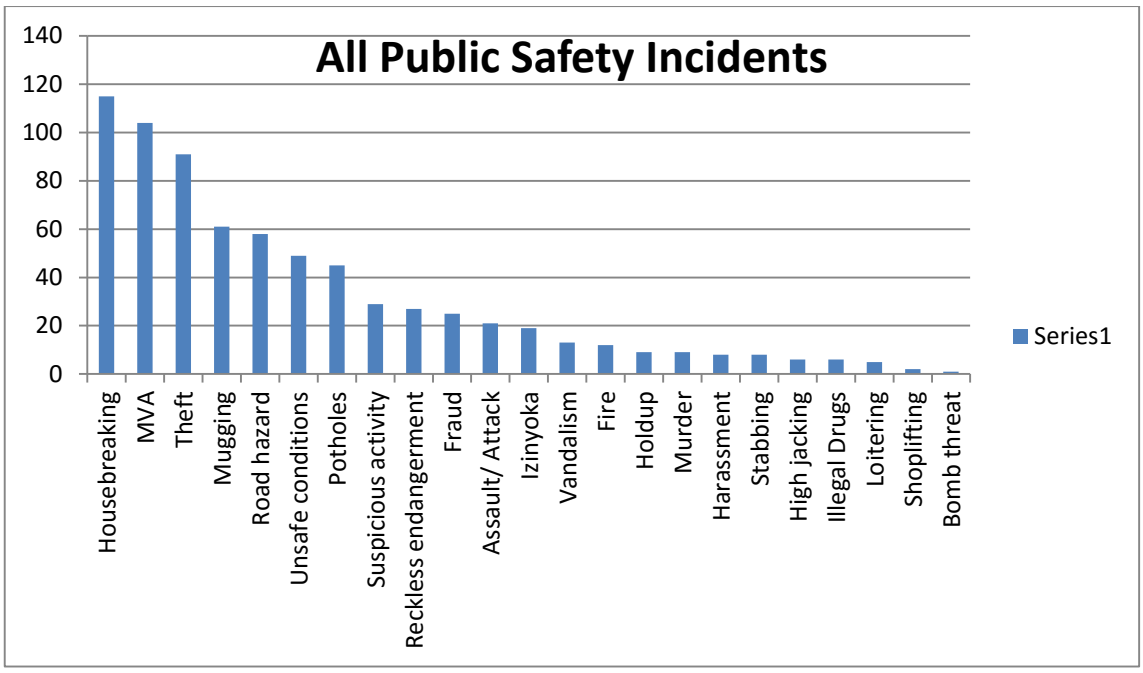


Figure 7.4: Public Safety Incidents by Amount of Tags

Some of the nodes coded, however, are already categories holding multiple variations of naming conventions and descriptions. For the purpose of developing the ontology, these must also be noted and are displayed in Table 7.4, thereafter the codes can be categorised further. This will be done in the reducing step.

7.4.4 Reducing

After the coding has been completed, the resulting data must be reduced or condensed in order for it to be easily interpreted.

Only two areas within East London were found to have multiple naming conventions used by respondents:

- The *CBD* (Central Business District) is also referred to as *town*,
- *Woodbrook Industrial* is referred to as *Industrial area* and *Woodbrook*

These can be combined under the headings CBD and Woodbrook respectively, with the other names forming part of the ontology. There are various ways to group these suburbs further, depending on the needs of the information users. As the purpose of this analysis is aimed at public safety, the researcher chose to group the reported suburbs according to the relevant police precincts in which they are located. These precincts are depicted in Figure 7.5 below.

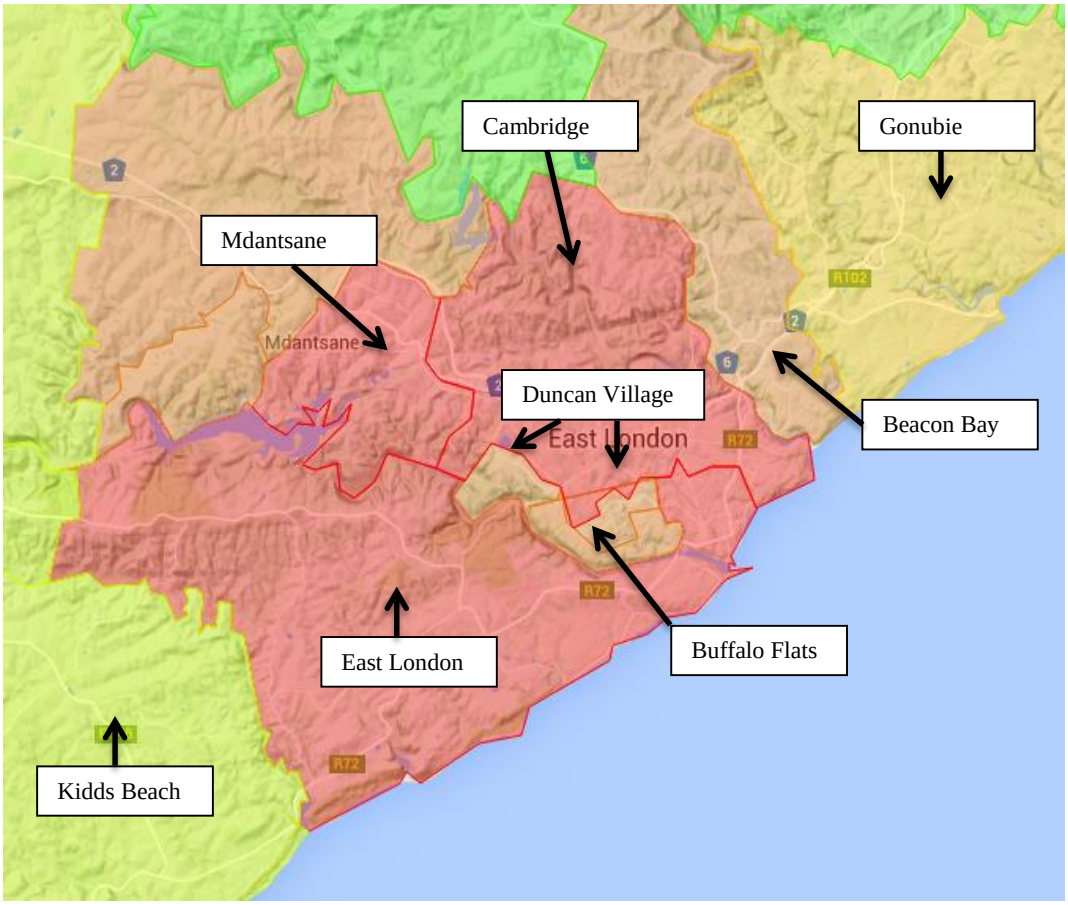


Figure 7.5: Police Precincts of East London (Crime stats simplified, 2015)

The relevant police precincts in East London are: Kidds Beach, East London, Buffalo flats, Gonubie, Mdantsane, Cambridge, Beacon Bay and Duncan village (Crime Stats Simplified, 2015). It must be noted that the precinct names are based on where the police stations are located and their geographically assigned areas, not on the formal East London map and its boundaries. Table 7.4 shows the reported suburbs grouped according to these precincts, with two alternative names and three subsections that were reported as suburbs.

Table 7.4: Grouping of Suburbs by Police Precinct

Police Precinct	Suburb	Alternative name/ sub section	Tags
Duncan Village	Duncan Village		16
	Scenary park		2
	Ziphunzana		5
Buffalo Flats	Buffalo Flats		3
	Parkridge		1
	Egoli		2

Gonubie	Gonubie		38

Beacon Bay	Beacon Bay		35
	Bonza Bay		2
	Ducats		
	Abbotsford		9

Cambridge	Cambridge	Cambridge West	43
	Chiselhurst		1
	Amalinda		29
	Nahoon	Bonnie doon	26
	Summerpride		3
	Vincent		27
	Berea		9
	Stirling		8
	Bunkers hill		1
	Selborne		1
	Haven Hills		3
	Morningside		8
	Dorchester Heights		3
	Wilsonia		6
	Stony Drift		1

East London	Arcadia		1
	Braelyn		9
	Baysville		
	Quigney		34
	CBD	Town	16
	Southernwood		17
	West bank		7
	Brookville		1
	Greenfields		10
	Sunnyridge		7
	Rosemount		1
	Collondale		2
	Willow Park		2
	Orange Grove		1
	Woodbrook	Industrial area	3
	North End		1
	Leeches Bay		1

Kidds Beach	Kidds Beach		3
	Kaysers Beach		3

Mdantsane	Mdantsane	Fort Jackson	17

As mentioned in the preceding coding section, the issues reported were logically grouped according to the type of issues reported. A listing of the issues tagged within each coding group is shown in Table 7.5.

It is evident that some of the groups contain much greater variations of naming conventions than others. There is also overlapping terminology which can be seen in cases such as *stabbing* which has been grouped as a standalone issue as well as within the group labelled *Assault/Attack*. The cause for this is reports referring to a stabbing incident as well as reports of an assault occurring, the execution of which included someone being stabbed. An automated system based on the ontology would tag the report containing both search words into both categories.

The duplication of words in multiple categories, however, was, also used as an aid in refining the groups further. An example of this is the finding that the majority of *reckless endangerment* reports are related to driving and motor vehicle accidents (*MVA*s), showing that these codes can be grouped together. Also, *potholes* were originally tagged into an individual grouping, but as the tagging continued, the need arose for a group including other road hazards. Thus, *potholes* became part of the *road hazard* grouping. Other duplications seen in Table 7.5 were eliminated as the reducing process continued

Table 7.5: Issue Grouped by Codes

What	
Potholes	Pothole(s)
Harassment	Harass, harassed
Reckless endangerment	Jumped the robot, drunk driver, out of control, nearly collided, accident, fire hazard, dangerous, exposed wire, drive fast, illegal, connections, didn't indicate, high speed, recklessly, through a red light, nearly hit, racing, throwing glass, throwing stones, dogs got away, without stopping, dangerous
Road hazard	Water drain, running water, pipe burst, flooded, manhole cover, uneven road surface, crossing, hump, robots, traffic lights, don't stop, tree fell, cattle walking, broken down, restricted visibility, restricted view, rubbish, roadwork, stray dog, disrupting traffic, protesters, burning tyres, drunk, sign missing
Loitering	Loitering, lingering, suspicious
MVA (Motor Vehicle Accident)	MVA, accident, pile up, car rolled, car crush, crash, pedestrian knocked, collided, bumped car, collision, car smashed, rammed, hit and run, run over, bumped into, knocked down, drove into
Izinyoka	Stealing copper, illegal connections, illegal electricity connections, illegal wiring, streaming wires, pinch electricity, stealing electricity, cable theft, electricity theft
Theft	Steal, stolen, took, pinched, broken into, breaking into,

	remote jamming, stole, taking, robbery, without paying, fleece, theft
Housebreaking	House breaking, break in, broken into, burglar, tsotsi, robber, forced entry, breaking and entering, gain entrance, intruder
Vandalism	Vandalised, broke, smashed, broken into, shattered, throwing stones
Mugging	Tsotsies, rob, robbed, mugged, grabbed, stolen, snatched, mug, took, take
High jacking	Highjack, high jacked
Fraud	Organised crime, scam, con, fraud, fleecing, posing, false, fraudulent, taken, swopping cards, ATM fraud, stolen
Holdup	Tied up, armed robbery, attack, robbed, breaking in, house breaking
Assault Attack	Stabbing, attacked, robbed, grabbed, beat, assaulted, raped, fighting, thieves, mob justice
Stabbing	Stabbing, stabbed
Bomb threat	Bomb threat
Suspicious behaviour	Looking into yard, looking over walls, suspect, suspicious, following, break-in, abandoned, loitering, watching, up to no good
Fire	Fire, smoke, burn, flames
Illegal Drugs	Drugs, marijuana
Murder	Dead, body, shot, strangled, stabbed, died
Shoplifting	Shoplifting, without paying, shop breaking, took, stole
Unsafe conditions	Garbage, abandoned, long grass, open electrical, exposed wires, slippery floors, dirty water, stagnant water, leak, burst pipe, stray dog, blocked drain, sewage

In order to further guide the grouping process, the researcher consulted current crime reporting conventions as published by the South African Police Service (SAPS). Making use of existing crime categories should serve to ease integration with existing systems and reporting structure. This resulted in the following categorisation of the groups.

7.4.4.1 Contact crimes

According to the SAPS, contact crimes are crimes that are inflicted via person to person contact. These crimes include murder, sexual crimes, attempted murder, assault with the intent to inflict grievous bodily harm, common assault, common robbery and robbery with aggravating circumstances (Department of Public Safety, 2014). Therefore, the code groups as listed in Table 7.6 were categorised as contact crimes.

Table 7.6: Contact Crimes

Category	Code Name	Tags: 116
Contact crimes	Assault/ Attack	21
	Murder	9
	Stabbing	8
	Mugging	61
	Holdup	9
	Harassment	8

7.4.4.2 Contact-related crimes

Crimes listed as contact-related crimes are crimes involving indirect contact with others. This is said to include arson and malicious injury to property. Due to this description, fire, vandalism and bomb threats were grouped into this category.

Table 7.7: Contact-Related Crimes

Category	Code Name	Tags: 26
Contact-related crimes	Fire	12
	Vandalism (loss of property)	13
	Bomb	1

7.4.4.3 Property-related crimes

The SAPS crime statistics list crimes such as burglary (residential and non-residential), theft of vehicles and stock theft as property related crimes (SAPS Services, 2014). This shows the category is based on theft of most forms of private property. Hence, the codes grouped into this category from the public safety reports are displayed in Table 7.8.

Table 7.8: Property Related Crimes

Category	Code Name	Tags: 237
Property-related crimes	Housebreaking	115
	Theft	91
	Highjacking	6
	Fraud	25

7.4.4.4 Other serious crimes

Theft not mentioned elsewhere, shoplifting and other commercial crime are listed as other serious crimes in the annual police crime figure reports. From this, as well as the content of the relevant reports, the following codes were grouped under the heading of other serious crimes.

Table 7.9: Other Serious Crimes

Category	Code Name	Tags: 27
Other serious crimes	Shoplifting	2
	Izinyoka	19
	Illegal Drugs	6

7.4.4.5 Traffic/Driving related

As the traffic department functions separately from the SAPS, it was decided to group together codes that relate to traffic and automobiles. This included reported MVAs as well as conditions or behaviour that could lead to MVAs. These are displayed in Table 7.10.

Table 7.10: Traffic/Driving Related

Category	Code Name	Tags: 234
Traffic/Automobile related	MVA	104
	Reckless endangerment	27
	Road hazard, Potholes	103

7.4.4.6 Potential incidents

Through the process of elimination, three codes remained unassigned. With traffic and crime reports eliminated, it became evident that the remaining codes pertained not to particular issues, but are closer to risk or potential for issues to arise. Thus, the final group was labelled *potential for crime*.

Table 7.11: Potential Incidents

Category	Code Name	Tags: 83
Potential incidents	Loitering,	5
	Unsafe conditions	49
	Suspicious activity	29

After creating the aforementioned groups and sub-groups, the hierarchical ontology could be created.

7.5 Ontology development:

Figure 7.7 depicts the ontology of suburbs in East London which can be used as a structure for automating the tagging of area related data for the structuring of unstructured public safety reports.

Figure 7.6 shows the ontology of public safety issues reported by East London citizens. Keywords made up of local naming conventions and descriptive references have been grouped for similar meaning. At this phase, words that were found to be repeated under different codes but within the same group were also reduced to only one most applicable instance.

These groupings can then be used for auto-coding the issues into code groups that combine the references describing the same kind of event. In other words, a search for these terms will return all the reports that should be grouped into the related category grouping. The code groupings are then grouped into crime and other groups which thus form a hierarchical structure.

The ontology can be used for the automation of the coding process and thereby structure the unstructured natural language text reports obtained via crowdsourcing. The tags can then be tallied and totalled at any or every level of the ontology depending on the information user's needs, thus enabling a multitude of reporting and visualisation options.

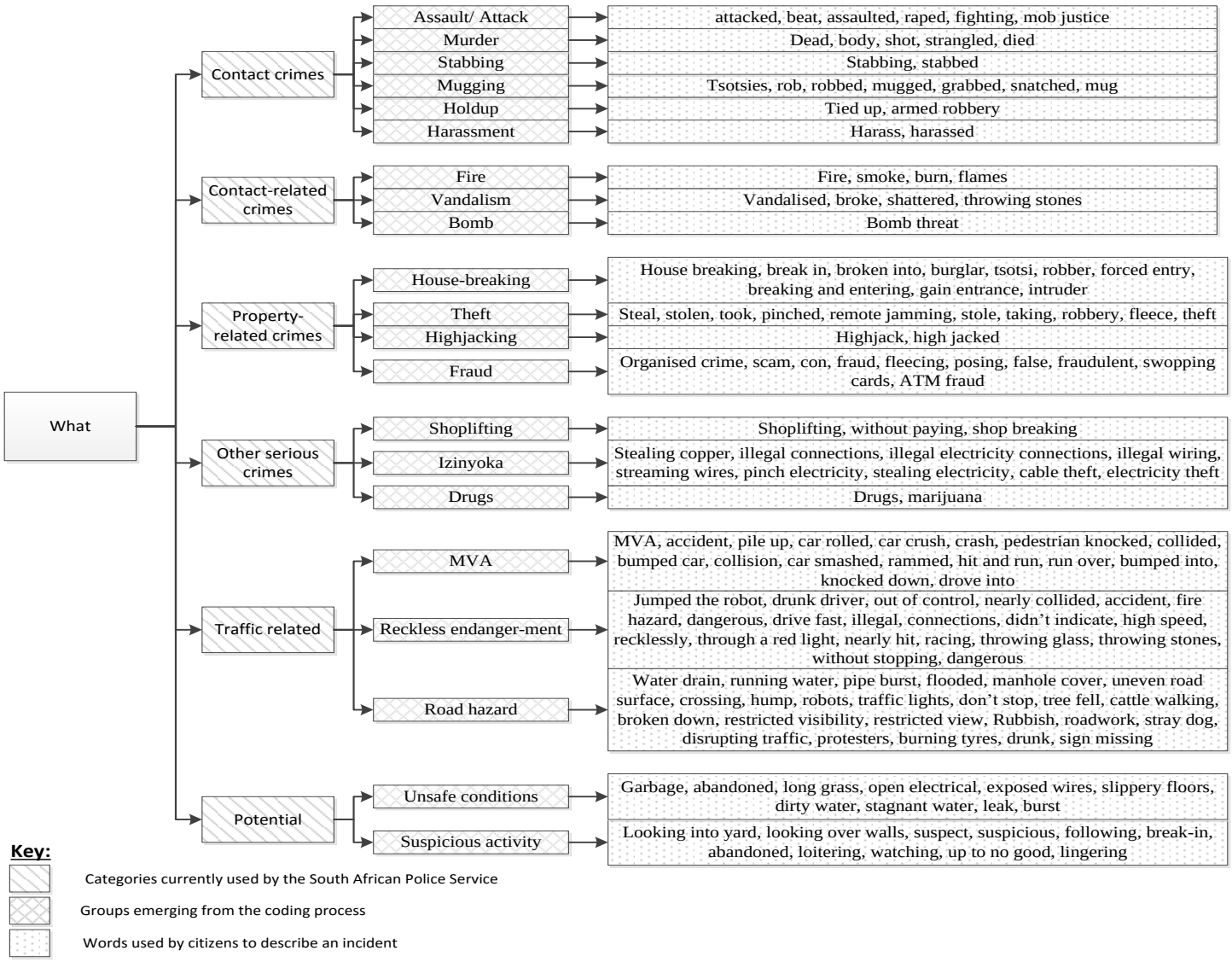


Figure 7.6: Ontology of Reported Issues

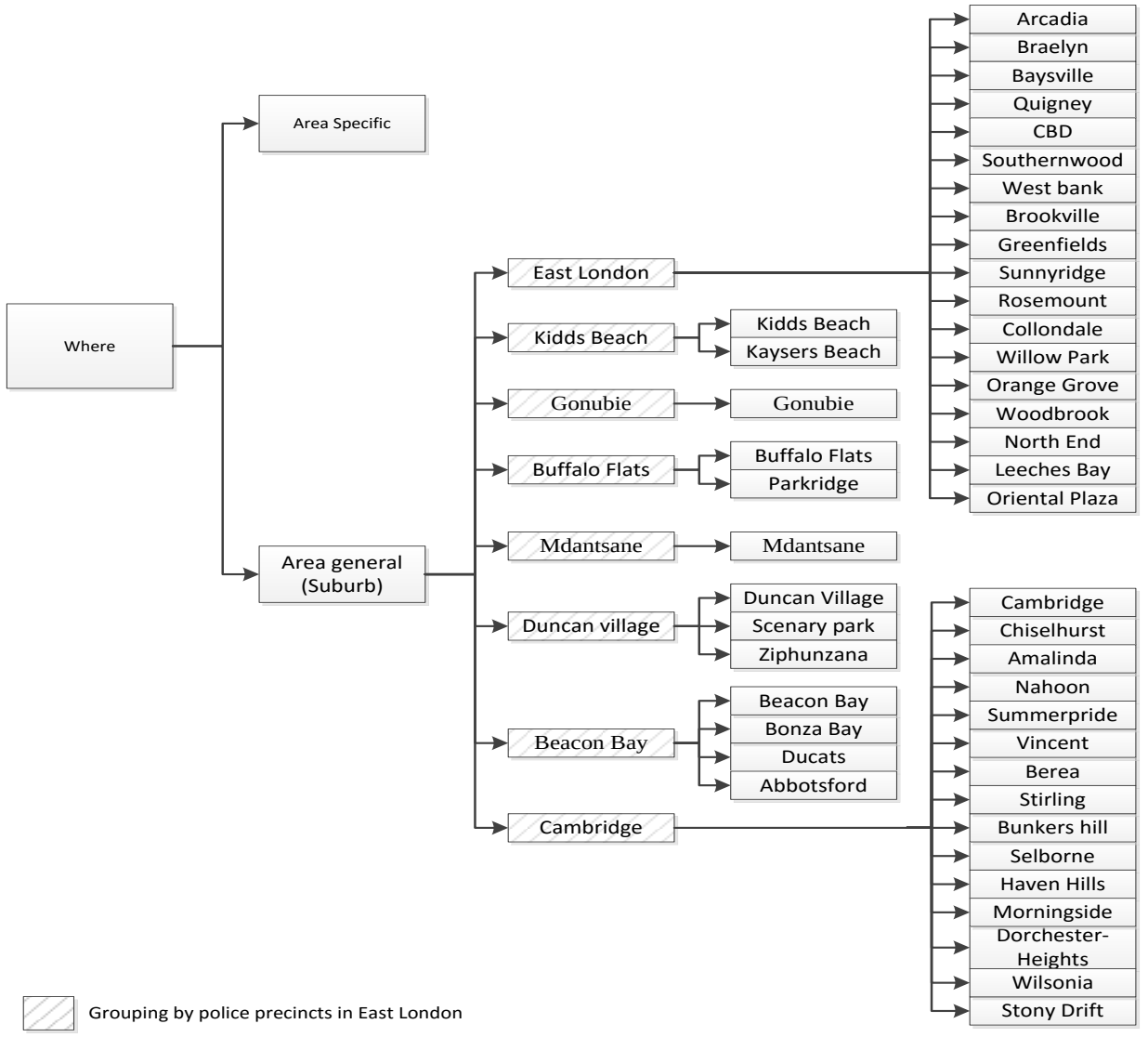


Figure 7.7: Ontology of Reported Suburbs

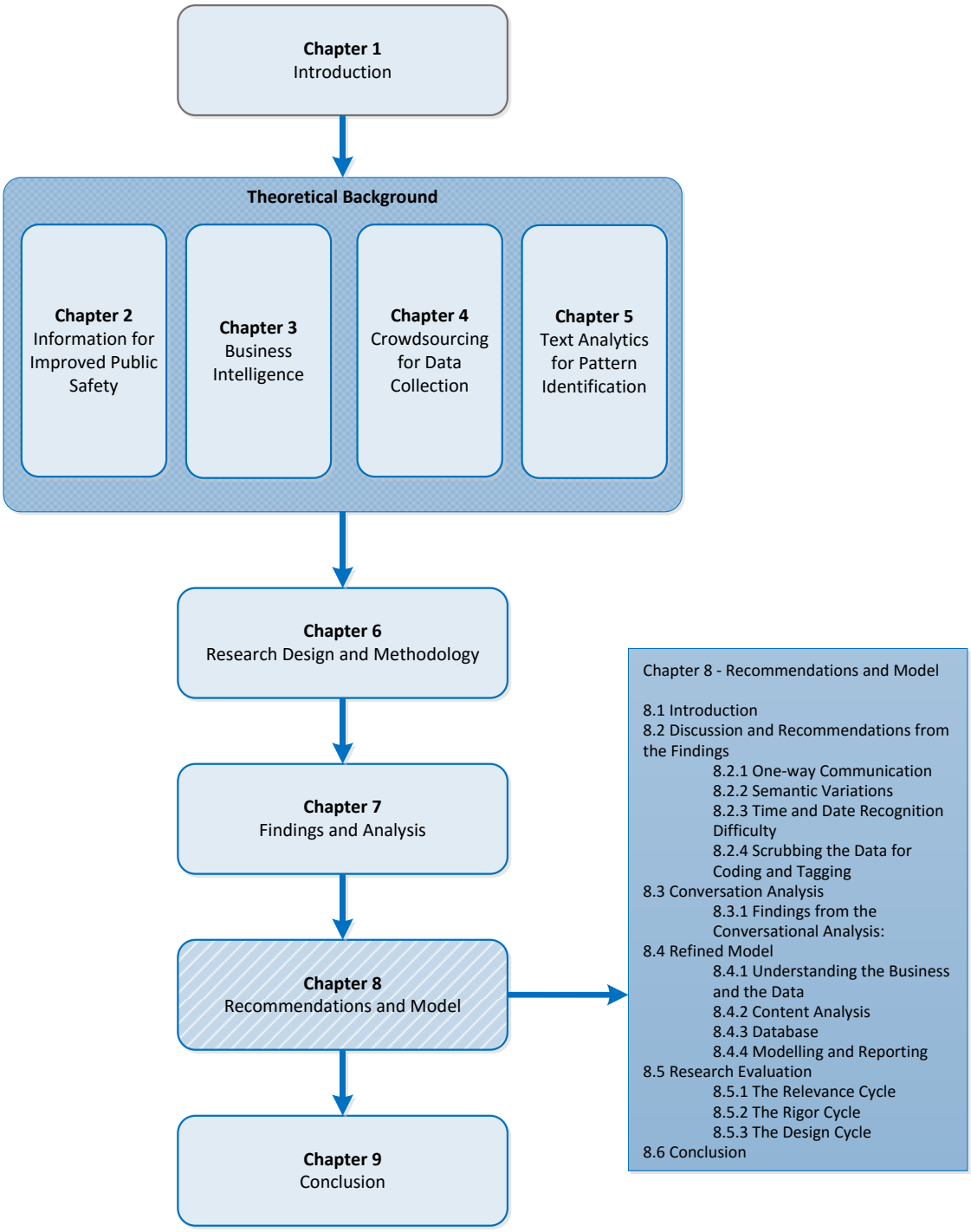
7.6 Conclusion

This chapter has presented a proposed model for the analysis of unstructured natural language text data: the SCQDA mode. This is followed by the description of a public safety smart city prototype which was used to gather the public safety reports discussed in this chapter.

Findings of a content analysis performed on the public safety reports were presented and discussed. The analysis of the reports served the purpose of testing the concepts of the SCQDA model. This led to the creation of hierarchical ontologies of the suburbs of East London as well as the types of issues reported by the citizens of East London, which will ultimately form part of the final model.

Further findings discussed in this chapter were from conversational analysis of discussions with experts and academics that are knowledgeable in the applicable fields. This analysis has led to further insights in confirming or refining the SCQDA model into its final form, which will be presented in the next chapter.

Chapter 8 – Recommendations and Model



8.1 Introduction

The preceding chapter presented the proposed model as well as observations taken from its application to a functioning prototype and empirical data. This chapter discusses the findings from the observation of the prototype, the content analysis of public safety reports, and findings from the conversational analysis of expert discussions. In order to verify and further refine the model produced by this study, these findings have been taken into consideration in order to refine the model, which is then presented with a brief description of the functioning of each of its elements. Finally, the model is followed by a short discussion explaining the evaluation of this study.

8.2 Discussion and Recommendations from the Findings

Using the prototype and performing a content analysis on the obtained reports illuminated the following major issues that affected the data analysis process.

8.2.1 One-way Communication

Making use of participatory crowdsourcing to have people volunteer reports anonymously is a "one way" communication process. If a report is not understood or is missing some information, one cannot ask the respondent to elucidate. Thus, for best results the report should be as complete as possible the first time. When high volumes of data is obtained, a few incomplete or incorrect reports should not affect results, as found in Chapter 3, but if there is a high volume of problematic reports, the results will be affected.

To ensure that the majority of data obtained is clear and usable, the question or request for reports must be as clear and informative as possible to ensure respondents are aware of what is required from them. This is not always an easy task, but adjustments can and should be made as signs of respondent uncertainty or confusion becomes evident from the reports. Hence the model shows a circular flow which is continuous. This correction would occur in the "understanding" phase.

8.2.2 Semantic Variations

As suggested in Chapter 5, there is a great variation in terminology used to name and describe public safety issues among respondents in the East London area. The citizens of developing nations often have great disparity in terms of educational level, experience and language use, which is also the case in East London. The reports, therefore, show large variations of ways in which similar incidents are described due to the use of slang, jargon, colloquialisms and the use of descriptive or abbreviated terminology.

This variety of semantics in the reporting greatly complicated the automation of structuring the unstructured text data. According to Miner et al. (2012), natural language processing (NLP) is a combination of machine learning and computational linguistics. Simply put, this means that the computer must learn in order to understand the language. Currently there are two ways in which to guide the computer to do so: by predefined structure, or by example. Using unstructured data means that the computer must learn by example, thus it requires a structured template that shows it what to tag and how to tag it. Therefore, a hierarchical ontology can be used by the computer to code by example.

8.2.3 Time and Date Recognition Difficulty

The time and date when an incident is reported can be captured automatically as metadata, but the incidents are not always reported as and when they occur (in real time), in which case a date and time may be referred to in the report. This means that the time of the incident occurrence and the time it is reported do not coincide and a time reference is therefore stated within the report. Ontologies could be developed for the issue and suburb related data, but doing this for time and date references within reports proved to be a much greater challenge. All the possible numeric variations combined with the various ways of structuring a date and time reference makes it virtually impossible to set parameters for automatic recognition of such data within a larger mass of text.

Making use of the automated date and time values found in the metadata of each report was much easier. These are uniform as they are automatically recorded by the IVR system. This allows for the values to be easily recognised as an attribute to the related

report. In order to make the reported date and time values easier to recognise and use automatically, the respondents could be guided with controls or instructions to report these values in a predefined way. This can be done by adjusting the request for reports or the IVR prompts.

8.2.4 Scrubbing the Data for Coding and Tagging

Working with the reports it is seen that individual words can be tagged or the entire report can be tagged, depending on the reporting and modelling needs. The impact of this is that data does not have to be scrubbed for joining words and stammers. Coding can be done making use of automated word count or word search functions or by simply selecting what is needed, thus there is no need to remove what is not needed prior to tagging as it is automatically excluded.

These points were all considered when refining the proposed model to the version seen in Figure 8.2. Other considerations for the models refinement and general steering of the study in its entirety came from the conversational analysis of discussions with experts which is discussed in the following section.

8.3 Conversation Analysis

Conversational analysis is a method of data collection through social interactions, usually including verbal and non-verbal cues (Goldkuhl, 2003). This was done throughout the duration of the study as depicted in Figure 8.1. Data relating to the topic and progression of this study was collected and discussed by having numerous meetings (formal and informal) with peers, academics and other industry experts in the related areas.

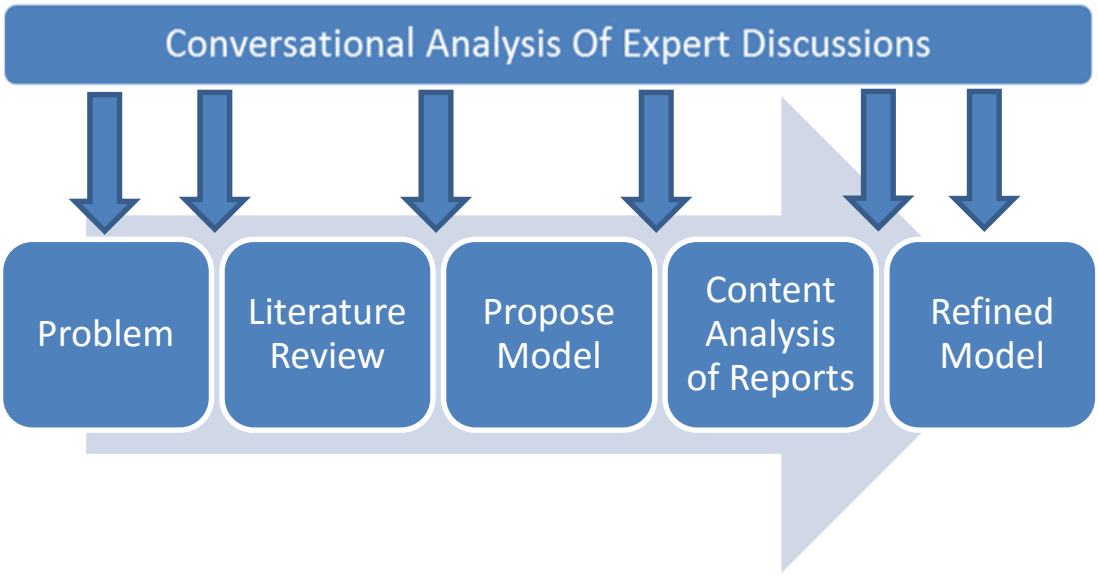


Figure 8.1: Conversational Analysis in This Study

8.3.1 Findings from the Conversational Analysis:

One of the first points raised by academics was a concern about the scope of this research project. It was noted that there are a number of points of interest involved in this study which could develop into large topics individually. The depth and width of this study was reassessed and discussed on a number of occasions and on an ongoing basis in order to avoid scope creep.

It was also found that there is a large need for ad hoc reporting. Every organisation requires certain standard reports to be used on a regular basis, but there is a great demand for unique reports per individual. Flexibility in generating reports is gaining importance as creativity and individuality becomes more prized in the workplace.

An industry expert raised the question of sampling when utilising big data. The discussion involved whether sampling is relevant and if so, what criteria would constitute a representative sample. The resulting conclusion from this discussion, and corroborated by literature, is that one of the implications of using big data and big data enabled technology is that sampling may no longer be required when one is able to use the entire population of data.

A practical topic of discussion concerned which CAQDAS package should be used for the execution of this study: should a best product be identified or can a convenient

option be identified. From existing literature and investigation it was found that there are a number of CAQDAS packages available. Most of these packages were found to share similar basic functionality and price range as discussed in Chapter 5. Most of these packages are able to perform the required work, but they do vary slightly in the more minor characteristics. It is therefore left up to the individual or organisation to choose the right package to suit their needs.

It was suggested that the label 'verbology' be considered and compared to the name ontology. Verbology makes reference to newly created words or vocabularies. This may be partially true in this case as some new words may surface, but the main purpose of the ontology in this study is the grouping and hierarchical structuring of words and phrases according to associated meaning rather than vocabulary development. It was therefore decided to use the term 'ontology'.

For the purpose of making the proposed model more generalizable for use in a greater variety of instances, it was suggested that allowance be made for more qualitative data input sources in addition to the current source (in this case crowdsourcing data). This was shown by illustrating the data as a stack of possible inputs, with crowdsourcing being the relevant one selected in this particular case.

Down the line, the system should compare reports to see if there are several reports referring to the same incident. Each report may shed light on a different aspect of the same incident, thus creating a clear picture by combining the data.

Lastly, it was mentioned that the model must show that there is a continuous flow of analysis. As more issues are reported, the system must keep cycling through the analysis process, adding the new reports and recalculating the outputs. This must be done in real time as and when reports are received without requiring any human intervention. The output can be checked manually if a problem occurs or on a regular basis. It was therefore decided to adjust the model to better emphasize this by increasing the weight of the arrows showing the circular and continuous flow of the model.

Taking these points into consideration, the model was refined into the version presented in the following section.

8.4 Refined Model

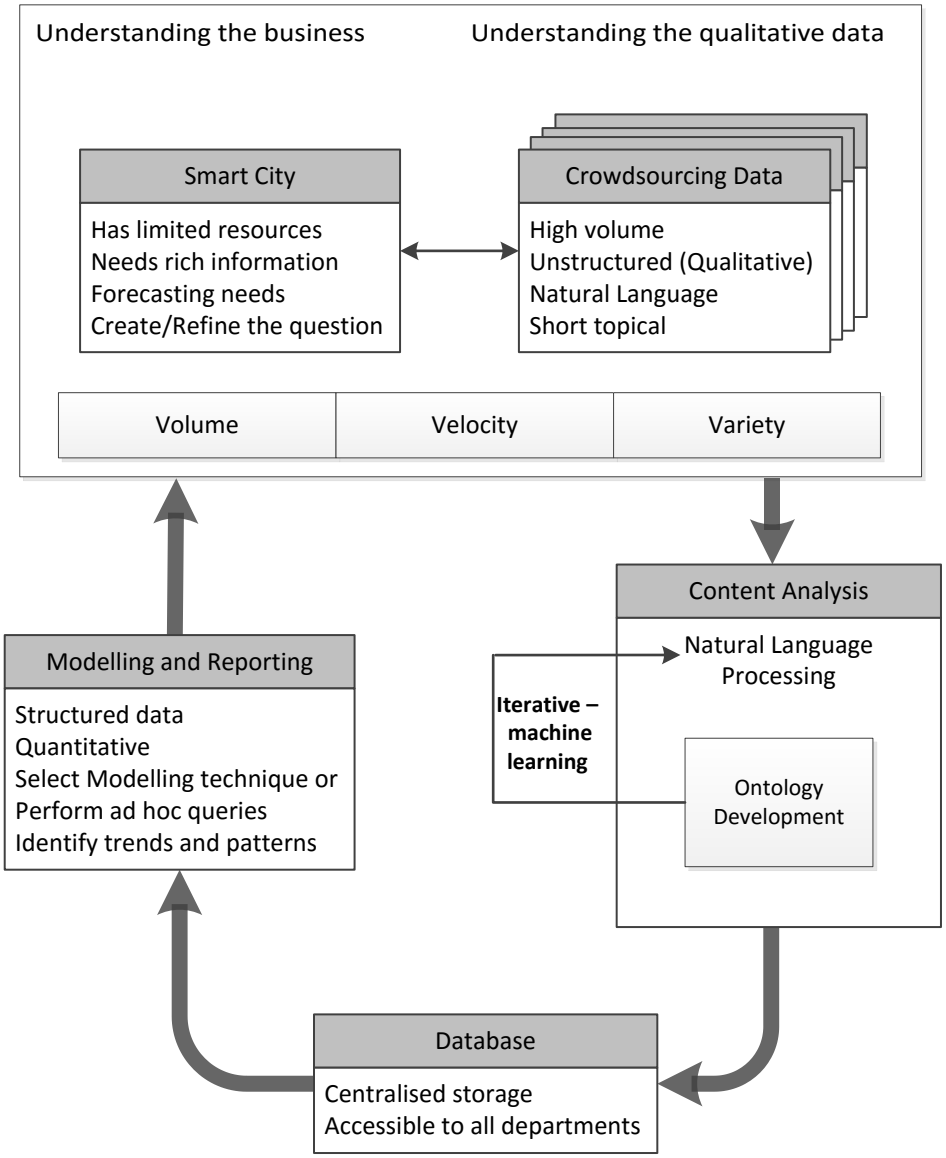
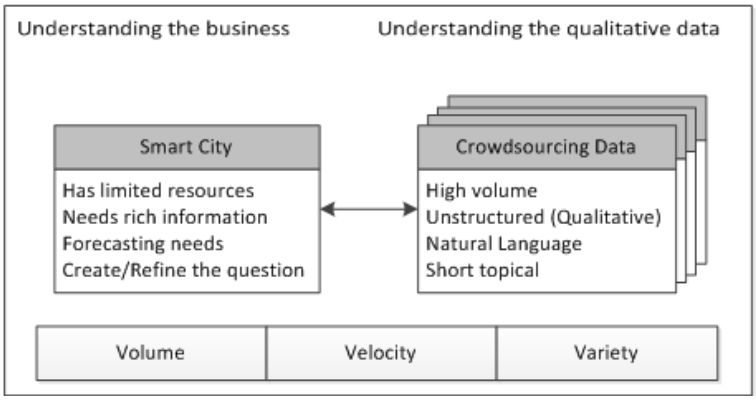


Figure 8.2: Refined Smart City Qualitative Data Analysis Model (SCQDA)

A review of relevant literature and theories led to the development of a proposed Public Safety Smart City Data Analysis Model (SCQDA). This model was tested by analysing reports obtained via a public safety smart city prototype. Observations from analysing the public safety reports as well as findings from conversational analysis of discussions with experts and academics were then used to refine the model. This section presents the refined model and the functioning of its elements.

8.4.1 Understanding the Business and the Data



In order to execute a successful text mining project, a good understanding of the business environment is needed. This includes the current situation and requirements of the solution. Together with the business understanding, one must also develop an understanding of the data that is to be used. The model makes provision for multiple data sources, but at this stage is focused on the crowdsourcing of qualitative data as the main input.

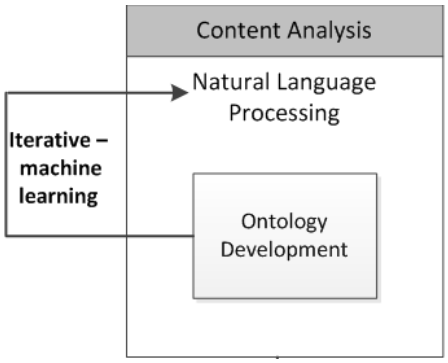
The understanding of business and of data has been brought together within the greater framework of big data. To ensure the implications of big data are considered, the applicable characteristics of big data (volume, velocity, and variety) must become part of the planning process and thus specifically when developing the understanding of the business and of the data.

As the SCQDA model has a continuous flow, this phase is not only the starting point, but the cycle flows back to this point during the system's functioning. Once the information output has been obtained, this phase is revisited with the new knowledge generated from the previous iteration of functioning. Thus, any experience or problems should be incorporated into the understanding and planning which gives rise to the opportunity to adjust the system as required. This should happen iteratively on an ongoing basis, regularly checking and adjusting for problems that become apparent or environmental changes that occur.

8.4.2 Content Analysis

The aim of this step is to give structure to the unstructured natural language text. The process of content analysis involves tagging words and concepts, and grouping the tags with similar meaning. This process can be automated by making use of a hierarchical

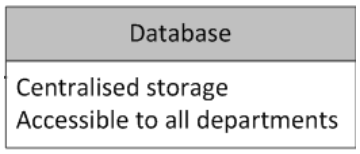
ontology that is developed in the same way making use of a CAQDAS package. The ontology can then be used for machine learning as it serves as a template and guide showing the system what to tag and how to tag it. The content analysis can then be automated enabling a fast and accurate structuring of high volumes of data.



The ontology is expected to develop and grow over time as some outliers may occur that would be added to the ontology as additional terminology is identified. Thus, there is an iterative cycle between the ontology development and the Natural Language Processing (NLP), even though both form part of the content analysis.

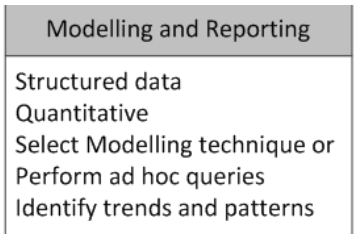
8.4.3 Database

Once the natural language data has been structured and quantified it should be stored in a centralised repository. All data users need to be provided with a single source of information (a singular, central database) via a dashboard of the organisation's (municipality's) own design. Changes, updates and access approvals should be controlled by someone who is appointed as administrator for controlled oversight. This database will be constantly updated as new reports are received and processed. Performing the content analysis prior to the storage allows for greater flexibility and least effort in the reporting step.



8.4.4 Modelling and Reporting

Although much of the analysis process can be automated, it is ultimately up to human interpretation to infer meaning from the data. The data is now structured and can be quantified at various levels of the hierarchical groupings. This allows for greater flexibility in ad hoc reports that can be generated. The data can now be modelled, visualised or queried. Calculations can be performed and various reports can be drawn depending on the specific information user's needs.



To be noted is that the model does not end here. An arrow from this entity back to the business and data understanding (planning) shows the continuous nature of this process. New reports can be received at any time which can change results that would need to be reassessed. Any lessons learned should also be taken into consideration as many aspects of the analysis process can then be readjusted or changed as required.

Having presented the contribution resulting from this research project (the model and its elements), the next section will outline the evaluation of the research.

8.5 Research Evaluation

The evaluation of this study, its findings and contribution is based on the foundation of the design science paradigm. Hevner (2007) describes design science research as the embodiment of three closely related cycles: the Relevance Cycle, the Rigor Cycle, and the Research Cycle. Hevner (2007) states that "The recognition of these three cycles in a research project clearly positions and differentiates design science from other research paradigms" (p. 88).

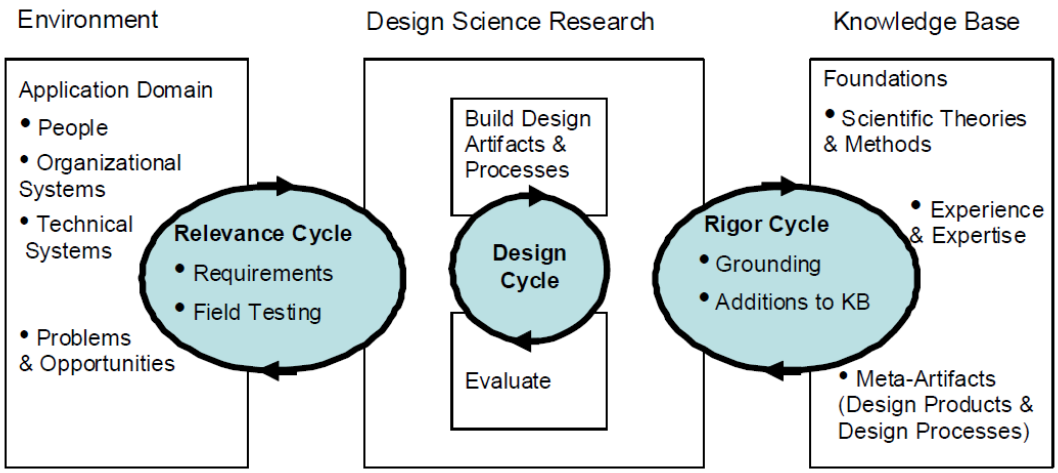


Figure 8.3: Design Science Research Cycles (Hevner, 2007)

8.5.1 The Relevance Cycle

The relevance cycle holds that good design science research is based on opportunities and problems in an actual application environment (Hevner, 2007). The environment thus provides requirements of the artefact and therefore the artefact must be applicable in the same environment.

The research question for this study is based on a real business problem: A lack of guidance and standardisation in the analysis of natural language text data, particularly for the use of public safety reports in a smart city. Thus requirements could be obtained from the contextual environment, which also contributed specifications for the environment in which the artefact could be tested. To this end, an extensive literature review has been conducted including journal articles, books, conference papers, white papers and case studies regarding the relevant topics of public safety in the smart city, crowdsourcing and qualitative data analysis. Thus, the research is relevant as it is based on a "business need" within a specific context and is designed to be usable in that specific business environment. This complies with the Relevance Cycle of design science.

8.5.2 The Rigor Cycle

According to the rigor cycle, design science research draws from the knowledge base of existing and appropriate theories and methods to ensure that artefacts produced can be defined as research contributions (Hevner, 2007).

A solution for the problem, a proposed artefact in the form of a model, was developed inductively based upon existing knowledge and theories in the form of secondary data which was critically reviewed. The model (the artefact) was initially developed making use of the widely accepted CRISP-DM as an initial frame of reference and foundation. Empirical data in the form of public safety reports were coded and grouped following the set guidelines of content analysis as specified by Krippendorff (2004) as an established method. Completing this report and the articles that will follow add back to the knowledge base, completing the Rigor Cycle as described by Hevner (2007).

8.5.3 The Design Cycle

Hevner (2007) states that in the design cycle an artefact is constructed, evaluated and refined based on the other two cycles.

In this study a model was developed based upon existing theory and sound literature from the existing knowledge base as presented in Chapter7. The model was then applied to a prototype based on the application environment to analyse empirical data in the form of public safety reports. Observations taken from the use of the model and the

prototype combined with findings from the conversational analysis of discussions with industry experts and academics were used to refine the proposed model in accordance with the Design Cycle into the final version presented in this chapter.

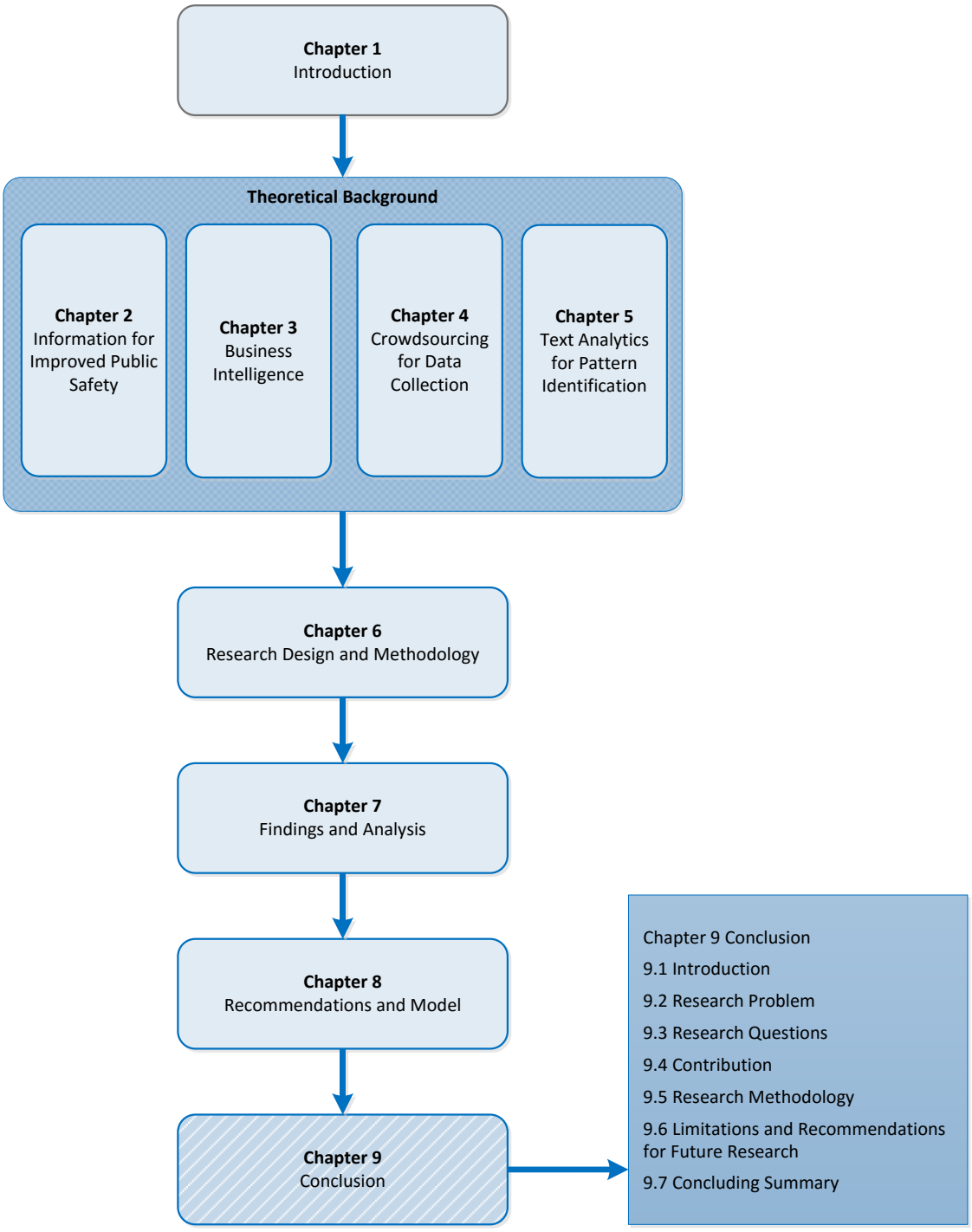
The validity of this study is thusly based on its compliance to the design science paradigm. This section has described how the study aligns with the three cycles of design science. For a more detailed description of how this study complies with the seven guidelines of design science, the reader may refer to Chapter 6 (Methodology chapter).

8.6 Conclusion

This chapter discussed recommendations based upon the findings reported in the preceding chapter. Additionally, the recommendations from the conversational analysis of expert discussions were also presented. These recommendations contributed to the refining of the model proposed in the previous chapter in order to develop the final model presented in this chapter.

After presenting the refined model and the functioning of its individual phases, a short explanation of the evaluation of this research followed. This evaluation is based on the three cycles that make up the design science paradigm which has been followed by this research. The aim of the study has thus been achieved. The next and final chapter will be a concluding summary of the research.

Chapter 9 – Conclusion



9.1 Introduction

This chapter aims to summarise and conclude this study. The research problem as introduced in Chapter 1 is revisited, followed by the research questions that this study set out to answer along with an explanation of how they were answered. Next the research contribution is summarised followed by the methodology which guided the achievement thereof. Finally, limitations of the study are discussed along with some suggestions for future research.

9.2 Research Problem

To mitigate the strain urbanisation has put onto limited resources, cities must become smarter. The smart city concept has been utilised in a number of developed first- world countries to improve public safety, but the focus has been on using automated sensors and digital cameras. In a developing nation this route is often too costly in terms of infrastructure and manpower.

Making use of the ubiquitous mobile phone, people can act as sensors by reporting public safety incidents they witness. The resultant data is a rich source of information, but is unstructured qualitative data in a natural language format. Other studies utilise quantitative data, therefore the analysis process used in developed nations is not applicable.

Looking to other organisations and business in general, it can be seen that most organisations utilise quantitative data in relational databases, but more recently there has been a drive to find ways to use qualitative data to better effect. Hence there is no direct example from business or other smart city projects on how to analyse large amounts of qualitative data. **Thus the research problem is that there is no specific method for the analysis of qualitative text reports for a smart city in a developing nation.**

9.3 Research Questions

The main research question asks: **How can natural language analytics be applied to qualitative public safety text data in a smart city project based in a developing country?** In order to answer this research question, the following sub-questions were addressed.

1. **What information is extracted from the data to enable a smart city to improve public safety?**

The first sub-question set a context for the data to be analysed by investigating the smart city concept with specific focus on public safety. It was shown that obtaining data relating to public safety incidents can contribute to improving public safety in a smart city. In many reported cases, the authorities chose to install or upgrade video surveillance systems or automated sensors in order to gather data. The data obtained is accurate and detailed, but video data also has negative points including: cost, location, time wastage and infrastructure.

Some of the important requirements of the post analysis data were also elucidated. This included the use of consolidated, singular, central storage points of information with multiple points of access provided to all authorised users from various levels and departments. Additionally, the analyst must remain conscious of the fact that different situations may require different models for this purpose. All the cases viewed show that time and location of the incident is important for analysis purposes, but geospatial information should include the greater area (or suburb) as well as a more specific targeted location such as an address, building or business for greatest impact.

2. **How can qualitative participatory crowdsourcing data be used for business intelligence?**

This sub-question aimed to explore the characteristics (and implications) of crowdsourcing data obtained via an IVR and mobi site. The concept of business intelligence was also explored in order to ensure usefulness of this data in a business context.

It was seen that participatory crowdsourcing can be used to generate and gather informative qualitative text data to use as input for BI. Considering recent developments in data analysis and use, especially the trending topics of big data, sensing and crowdsourcing, the literature identified unstructured text data as being the next logical resource that can be exploited for deeper insights into strategic decision making and customer communication. It was noted that there are tried and tested methods for modelling and reporting, as well as software from multiple vendors that can be

employed to automate these functions for BI, but it is centred on structured quantitative data and the data in question, and this research project, is unstructured.

It was concluded that the large amounts of unstructured text data that can be obtained via participatory crowdsourcing will have to be reduced or condensed and structured in order to be used for BI through the process of data mining or more specifically text mining. Thus pattern identification through text mining is suggested to be the appropriate way forward.

3. **How can text analytics be used to identify patterns and trends in unstructured text data?**

The literature showed that Data Analytics can be used to understand, predict or direct by identifying patterns and trends within raw data. In order to apply analytics to large quantities of unstructured text data, however, text mining is used to reduce or summarise the data into a more structured form that enables the use of data mining techniques. This will result in sets of numeric data associated with certain concepts, which can then be modelled in order to reveal any trends or patterns that may be present.

The focus area of the public safety smart city project was found to be Natural Language Processing (NLP), which involves both linguistics and machine learning. Within NLP the relevant issue regarding the public safety smart city project was found to be semantic analysis (understanding of meaning) when extracting information. In order to retain usefulness while summarising and to compensate for semantic and colloquial references, an ontology of key terms and their associated meanings is required. The ontology can be developed by performing a content analysis of the data to be used for machine learning to enable accurate analysis of larger amounts of data. The ontology would then be used to guide the automated coding and reduction of large text data sets such as a collection of public safety reports.

Considering the answers to the above sub-questions has led to the development of a model for the analysis of qualitative natural language public reports for the improvement of public safety in a smart city in a developing nation. The model served as the contribution of this study which is discussed in the following section.

9.4 Contribution

Increased urbanisation has put a strain on the limited resources in the urban environment leading to problems in categories such as traffic, health and public safety which all affect living standards. Thus new and innovative methods can help city officials use resources more effectively and efficiently. Smart cities can help in this regard, but it is a relatively new area of study with limited literature available.

Most smart city literature is currently focused on developed countries and therefore lack context and applicability to developing nations where challenges faced are often very different. Most previous smart city systems have also made use of expensive automated sensors which produce numeric data. The public safety smart city system focused on in this research is based in a developing country and utilises people as sensors via participatory crowdsourcing. This results in qualitative natural language data which cannot be analysed in the same way as the quantitative data obtained by automated sensors.

This study has culminated in the development, assessment and refinement of a model for the analysis of natural language public safety reports for a smart city in a developing nation. The model depicted in Figure 8.2 shows how unstructured natural language text data can be processed into structured quantifiable data which can be used for the identification of patterns and trends by modelling and reporting. This model is designed to be applied to qualitative public safety reports obtained through crowdsourcing for a smart city in a developing nation. It can, however, be generalised to a degree to include other qualitative data sources in addition to participatory crowdsourcing, or to be applicable to other smart city focus areas such as traffic and health.

Additionally, the research process has produced a hierarchical ontology of terms used to describe public safety incidents by the citizens of East London. This is a product of using the model so it can be developed from scratch at implementation. The ontology produced by this study, however, canbe used to fast-track implementation in an environment where similar terminology is used.

A journal paper presenting the findings has been written and submitted to IT Professional, which is a peer reviewed publication on the DHET approved list. This article is currently under review.

9.5 Research Methodology

Chapter 6 provides a detailed description of the manner in which this research project was conducted. The chapter firstly discussed possible research paradigms before justifying design science as being the most appropriate paradigm for this study due mainly to the need for a theoretically sound artefact to be produced as the solution to an industry specific problem. Quantitative methods were applied to qualitative data resulting in quantitative data linked to the qualitative data, showing why a mixed methods approach was used. The methods used included prototype observation, content analysis, and conversational analysis.

An initial model was developed through the inductive review of relevant literature. The proposed model was assessed by observation of the public safety smart city prototype and the application of the model to the 663 public safety reports obtained from the prototype. This process involved a content analysis of the public safety reports which gave the researcher insights into the model's practical application. These findings, in conjunction with findings from the conversational analysis of discussions with academic and industry experts, were used to refine the proposed model to its final form. This complies with the three cycles of design science ensuring viable research and contributions are achieved. Furthering this aim, the application of the seven design science guidelines to this study are explained in detail in Chapter 7.

9.6 Limitations and Recommendations for Future Research

In this study the Smart City Qualitative Data Analysis (SCQDA) model has been developed and tested in the context of public safety within the smart city. Public safety served as the test bed for this model, but the model is not limited to this focus area. The SCQDA model can be applied to other smart city focus areas when adapted and will work in the same way. A change in the subject matter of the data would change the focus of the model without changing its functioning, due to the content analysis principles used. This can be achieved by changing the request for data or finding different sources depending on the topic required for the alternate areas of focus.

A discrepancy in the date and time an incident occurs and the date and time that it is reported was found to be a complication to the analysis process. In this study, ontology development was shown to be an appropriate solution to resolve differences in naming conventions within public safety reports. However, standardising variations in the structure of date and time values within a report will be an easier remedy than ontology development for this specific category of data. Exactly how to achieve this and what structure would be most appropriate is still in question.

Research, possibly in the context of computer science or statistics disciplines, can be conducted into developing specific algorithms for the predictive analytics of the public safety reports once structured according to this study.

Other future research could directly consult with the information users at BCMM in order to establish exactly how they would like the feedback presented and what reports or modelling they require. A cost-benefit analysis may also be useful to BCMM officials in order to investigate the justification of expenditure on the implementation of a public safety smart city system.

In order for the analysis model developed by this study to be used, a steady flow of current public safety reports are required. As this study was focused on the analysis process, it did not explore the adoption of the system or the motivation of citizens to contribute the reports. This is another avenue which could be researched.

9.7 Concluding Summary

This final and concluding chapter has provided a summary of the research project and the conclusions reached. Initially the research problem is described followed by the research questions in order to show how the theoretical background informed this study. A summary of how each question was answered is presented, followed by a summary of the contribution made by this research. The research design and methodology section then briefly explained how this study was conducted and how the contribution was developed. Finally, the limitations of the study were discussed along with some recommendations for future research which can be conducted in related topics.

List of References

Ahmed, Z., Dandekar, T., & Majeed, S. (2012). Role of Ontology in NLP grammar construction for semantic based search implementation in product data management systems. *International Journal of Managment, IT and Engineering, 2*(2), 6-40.

Al-Hader, M., & Rodzi, A. (2009). The smart city infrastructure development & monitoring. *Theoretical and Empirical Researches in Urban Management, 2*(11), 87-94.

Allwinkle, S., & Cruickshank, P. (2011). Creating Smart-er Cities: an overview. *Journal of urban technology, 18*(2), 1-16.

Bakıcı, T., Almirall, E., & Wareham, J. (2012). A Smart City Initiative: the Case of Barcelona. *Journal of the Knowledge Economy, 4*(2), 1-14.

Barbier, G., Zafarani, R., Gao, H., Fung, G., & Liu, H. (2012). Maximizing benefits from crowdsourced data. *Computational and mathematical organization theory, 18*, 257-279.

Berry, M. W., & Kogan, J. (Eds.). (2010). *Text mining: applications and theory*. John Wiley & Sons.

Berthold, H., Rosch, P., Zoller, S., Wortmann, F., Carenini, A., Campbell, S., & Stohmaier, F. (2010). An architecture for ad-hoc and collaborative business intelligence. *Proceedings of the 2010 EDBT/ICDT Workshops* (pp. 13-18). ACM.

Bhana, B., Flowerday, S., & Satt, A. (2013). Using participatory crowdsourcing in South Africa to create a safer living environment. *International journal of distributed sensor networks,2013*, 1-13.

Boldon James. (2012). *Using classification to manage big data.* Retrieved May 23, 2014, from Boldon James: http://www.boldonjames.com/user-driven-classification/

Boldon James. (2014). *Empower your users, drive your business.* Retrieved May 02, 2014, from Boldon James: http://www.boldonjames.com

Boyd, D., & Crawford, K. (2011). Six provocations for big data. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society* (pp. 1-17). New York: SSRN.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: provocations for cultural, technical, and scholarly phenomonon. *Information, communication & society, 15*(5), 662-679.

Brandel, M. (2008). Crowdsourcing: Are you ready to ask the world for answers? *Computerworld, 42*, 24-26.

Brussee, R., Rovers, M., Van Vliet, H., Swart, D., & Hekman, E. (2013). Crowdsourcing: Classification, costs, benefits, and usage. In *Conference papers CARPE 2013*, 1-8.

Caragliu, A., Del Bo, C., & Nijkamp, P. (2011). Smart cities in Europe. *Journal of urban technology*, *18*(2), 65-82.

Carvajal, D. (2002). The artisan's tools. Critical issues when teaching and learning CAQDAS. In *Forum: qualitative social research, 3*(2)*Art*. 14.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 *Step-by-step data mining guide.*

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. *MIS Quarterly, 36*, 1165-1188.

Chiang, R. H., Goes, P., & Stohr, E. A. (2012). Business Intelligence and Analytics Education, and Program Development: A Unique Opportunity for the Information Systems Discipline. *ACM Transactions on Management Information Systems, 3*(3), *Art*. 12.

Childerhouse, P., Hermiz, R., Mason-Jones, R., Popp, A., & Towill, D. (2003). Information flow in automotive supply chains - identifying and learning to overcome barriers to change. *Industrial Management and Data Systems, 103*(7), 491-502.

Chopra, A., Prashar, A., & Sain, C. (2013). Natural language processing. *International journal of technology enhancements and emerging engineering research, 1*(4), 131-134.

Cisco. (2006). *County government capitalizes on network to improve public safety and quality of life.* San Jose: Cisco Systems.

Cisco. (2010). *Police department increases situational awareness.* San Jose: Cisco.

City Population. (2014, July 17). Retrieved from City population: http://www.citypopulation.de/php/southafrica-buffalocity.php

Collis, J., & Hussey, R. (2009). *Business research: A practical guide for undergraduate and postgraduate students.* London: Palgrave Macmillan.

Cozens, P., & Grieve, S. (2011). Investigating crime precipitators and the 'environmental backcloth'of the nigth time economy: an environmental criminology perspective from an Australian capital city. *York Deviancy Conference* (pp. 24-27). Perth: University of Western Australia.

Crime Stats Simplified. (2015). Retrieved from Crime stats SA: http://www.crimestatssa.com/

De Vos, A., Strydom, H., Fouche, C., & Delport, C. (2005). *Research at Grass Roots for the Social Sciences and Human Service Professions* (Third ed.). Pretoria: Van Schaik.

Delen, D., & Demirkan, H. (2013). Data, information and analytics as services. *Decision Support Systems, 55*, 350-363.

Department of Public Safety. (2014). Retrieved December 17, 2014, from Buffalo city metro municipality: http://www.buffalocity.gov.za

Dinu, B., & Iovan, S. (2014). Harnessing big data volumes. *Fiability & Durability, 1*, 250-257.

Dirks, S., Gurdgiev, C., & Keeling, M. (2010). Smarter cities for smarter growth: How cities can optimize their systems for the talent-based economy. *IBM Institute for Business Value, May 2010*

Dodgson, M., & Gann, D. (2011). Technological Innovation and Complex Systems in Cities. *Journal of Urban Technology, 18*(3), 101-113.

East London. (2013). *East London, Eastern Cape, South Africa*. Retrieved June 30, 2013, from http://www.eastlondon.org.za/

ECSECC. (2012). *Eastern Cape Development Indicators.* Retrieved January 20, 2013, from http://www.ecsecc.org/files/library/documents/EasternCape_withDMs.pdf

Exner, J. P., Zeile, P., & Streich, B. (2011). Urban Monitoring Laboratory: new benefits and potential for urban planning through the use of urban sensing, Geo-and Mobile-Web. In*Real corp, 2011*, 1087-1096.

Fan, W., & Bifet, A. (2013). Mining big data: current status, and forecast to the future. *ACM SIGKDD explorations newsletter, 14*(2), 1-5.

Ganti, R. K., Ye, F., & Lei, H. (2011). Mobile crowdsensing: current state and future challenges. *Communications Magazine, 49*(11), 32-39.

Gao, J., Liu, X., Ooi, B. C., Wang, H., & Chen, G. (2013). An online cost sensitive decision-making method in crowdsourcing systems. *Proceedings of the 2013 international conference on Management of data* (pp. 217-228). ACM.

Gibbs, G. R. (2008). *Analysing qualitative data.* London: Sage.

Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler-Milanovic, N., & Meijers, E. (2007). Smart cities-Ranking of European medium-sized cities. *Vienna University of Technology*.

Given, L. M. (2008). *The SAGE Encyclopedia of Qualitative Research Methods.* Thousand Oaks: Sage.

Goldkuhl, G. (2003). Conversational Analysis as a Theoretical Foundation for Language Action Approaches? *8th International Working Conference on the Language Action Perspective,* (pp. 1-14). Tilburg.

Goldkuhl, G. (2004). Design theories in information systems-a need for multi-grounding. *Journal of Information Technology Theory and Application (JITTA), 6*(2), 59-72.

Golfarelli, M., Rizzi, S., & Cella, I. (2004). Beyond data warehousing: what's next in business intelligence? *Proceedings of the 7th ACM international workshop on Data warehousing and OLAP* (pp. 1-6). ACM.

Gruber, T. (1994). Towards principles for the design of ontologies used for knowledge sharing. *International Journal of human and computer studies, 43*, 907-928.

Gutierrez, S. (2012). *Social and legal aspects related to citizens empowering.* Aalborg: SafeCity.

Hawking, P. (2012). Business Intelligence Excellence: A Company's Journey to Business Intelligence Maturity. *Intelligence*, 2008.

Hearst, M. (2003). What is text mining?*SIMS, UC Berkeley: Association for computational linguistics.*

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian journal of information systems*, *19*(2), 87-92.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004, March). Design science in information systems research. *MIS Quarterly*, *28*(1), 75-105.

Hofstee, E. (2006). *Constructing a good dissertation: a practical guide to finishing a master's, MBA, or PhDon schedule.* Johannesburg: EPE.

Hollands, R. G. (2008). Will the Real Smart City Please Stand Up? City: Analysis of Urban Trends, Culture, Theory, Policy, Action. *City: Analysis of Urban Trends, Culture, Theory, Policy, Action, 12*(3), 303-320.

Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, 1-5.

IBM. (2011). *Memphis PD: Keeping ahead of criminals by finding the "hot spots".* New York: IBM Corporation.

IBM. (2012). *Report m153 - IBM SPSS ROI case study: city of Lancaster.* Nucleus Research.

Jiang, L., Zhang, H.-b., Yang, X., & Xie, N. (2013). Research on Semantic Text Mining Based on Domain Ontology IFIP Advances in Information and Communication

Technology. *Computer and Computing Technologies in Agriculture VI*, 336-343.

Kaiserswerth, M. (2010). *Creating a smarter planet, one collaboration at a time.* Zurich: IBM Research.

Kalakota, R. (2011, November 11). *Big Data Infographic and Gartner 2012 Top 10 Strategic Tech Trends.* Retrieved March 05, 2014, from Business Analytics 3.0 (blog): www.practicalanalytics.wordpress.com/2011/11/11/big-data-infographic-and-gartner-2012-top-10-strategic-tech-trends

Komninos, N., Pallot, M., & Schaffers, H. (2013). Special issue on smart cities and the future internet in Europe. *Journal of the Knowledge Economy*, *4*(2), 119-134.

Kozinets, R. V., Hemetsberger, A., & Schau, H. J. (2008). The wisdom of consumer crowds collective innovation in the age of networked marketing. *Journal of Macromarketing*, *28*(4), 339-354.

Krippendorff, A. (2004). *Content analysis: an introduction to its methodology* (2nd ed.). Sage Publications, Inc.

Langseth, J., & Vivatrat, N. (2003). Why proactive business intelligence is a hallmark of the real-time enterprise: Outward bound. *Intelligent Enterprise, 5*(18), 34-41.

Lebraty, J.-F., & Lobre-Lebraty, K. (2013). *Crowdsourcing: one step beyond.* London: ISTE Ltd.

Leech, N. L., & Onwuegbuzie, A. J. (2008). Qualitative Data Analysis: A Compendium of Techniques and a Framework for Selection for School Psychology Research and Beyond. *School Psychology Quarterly*, *23*(4), 587–604.

Lehnert, W. G., & Ringle, M. H. (2014). *Strategies for natural language processing.* Psychology Press.

Lewins, A., & Silver, C. (2009). Choosing a CAQDAS package. *QUIC - qualitative innovations in CAQDAS,* 6.

Lönnqvist, A., & Pirttimäki, V. (2006). The Measurement of Business Intelligence. *Information Systems Management, 23*(1), 32-40.

Maimon, O., & Rokach, L. (Eds.). (2010). *Data mining and knowledge discovery handbook* (2nd ed.). New York: Springer.

McAfee, A., & Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data. *The management revolution. Harvard Bus Rev*, *90*(10), 61-67.

Mehrotra, A., Dellon, E. S., Schoen, R. E., Saul, M., Bishehsari, F., Farmer, C., & Harkema, H. (2012). Applying a natural language processing tool to electronic health records to assess performance on colonoscopy quality measures. *Gastrointestinal endoscopy*, *75*(6), 1233-1239.

Miner, G., Delen, D., Elder, J., Fast, A., Hill, T., & Nisbet, R. (2012). The seven practice Areas of text analytics. In *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications* (pp. 29-41). Elsevier.

Mitra, S., Pal, S. K., & Mitra, P. (2002). Data mining in soft computing framework: a survey. *IEEE Transactions on neural networks*, *13*(1), 3-14.

Muller, M., Hupfer, S., Levy, S., Gruen, D., Sempere, A., & Priedhorsky, R. (2011). Circles of crowdsourcing: The social organization of participatory sensing. *MobileHCI*.

Myers, M. D. (2009). *Qualitative Research in Business & Management.* London: Sage Publications.

Namey, E., Guest, G., Thairu, L., & Johnson, L. (2008). Data reduction techniques for large qualitative data sets. *Handbook for team-based qualitative research*, 137-161.

Nam, T., & Pardo, T. A. (2011). Conceptualising Smart City with Dimensions of Technology, People, and Institutions. *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times* (pp. 282-291). New York: ACM.

Negash, S. (2004). Business Intelligence. *Communications of the association for information systems*, *13*(15), 177-195.

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of statistical analysis & data mining applictions.* Elsevier.

Oates, B. J. (2006). *Researching Information Systems and Computing.* London: Sage Publications.

Olivier, M. S. (2009). *Information technology research.* Pretoria: Van Schaik.

Oomen, J., & Aroyo, L. (2011, June). Crowdsourcing in the cultural heritage domain: opportunities and challenges. In *Proceedings of the 5th International Conference on Communities and Technologies* (pp. 138-149). ACM.

Pruijt, H. (2012). InterviewStreamliner, a minimalist, free, open source, relational approach to computer-assisted qualitative data analysis software. *Social science computer review*.

Punch, K. F. (2006). *Developing Effective Research Proposals.* London: Sage Publications.

Rambaree, K. (2007). Bringing rigour in qualitative social research: the use of a CAQDAS. *Uom research journal, 13A,* 1-16.

Russom, P. (2011). Big data analytics. *TDWI best practices report, Fourth quarter,* 1-36.

Saillard, E. (2011). Systematic versus interpretive analysis with two CAQDAS packages: NVivo and MAXQDA. *Forum: qualitative social research, 12*(1).

SAPS Services. (2014). Retrieved December 18, 2014, from South African Police Service: http://www.saps.gov.za/

Saxton, G. D., Oh, O., & Kishore, R. (2013). Rules of crowdsourcing: Models, issues, and systems of control. *Information Systems Management, 30*(1), 2-20.

Schaffers, H., Komninos, N., Pallot, M., Trousse, B., Nilsson, M., & Oliveira, A. (2011).Smart Cities and the Future Internet: Towards Cooperation Frameworks for Open Innovation.*Future Internet Assembly, 6656*(1), 431-446.

Schenk, E., & Guittard, C. (2011). Towards a characterisation of crowdsourcing practices. *Journal of innovation economics & management*, 93-107.

Smart Cities background paper. (2013). UK Government, Department for Business, Innovation and Skills. London: Ove Arup & partners Ltd. Retrieved August 13, 2014, from http://www.gov.uk/bis

Smart cities international case studies: global innovators.(2013). UK Government, Department for business innovation and skills. London: A. Retrieved August 13, 2014, from http://www.gov.uk/bis

Srivastava, M., Abdelzaher, T., & Szymanski, B. (2012). Human-centric sensing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 370*(1958), 176-197.

Srivastava, P., & Hopwood, N. (2009). A practical iterative framework for qualitative data analysis. *International journal of qualitative methods, 8*(1), 76-84.

Statistics South Africa. (2012, July 17). Retrieved from Statistics South Africa: http://beta2.statssa.gov.za/?page_id=1021&id=buffalo-city-municipality

Steedman, M. (2010). Some Important Problems in Natural Language Processing. *Informatics Hamming Seminar*, (pp. 1-44). Edinburgh.

Stemler, S. E. (2001). An Overview of Content Analysis. *Practical assessment, research & evaluation, 7*(17) 137-146.

Tait. (2012). *Case study: São Paulo Military Police, Brazil.* Brazil: Tait Limited.

Teddlie, C., & Tashakkori, A. (2003). *Handbook of mixed methods in social and behavioral research.* Thousand Oaks: Sage.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research.* Washington: Sage.

Thomsen, E. (2003). BI's promised land. *Intelligent enterprise San Mateo, 6, 20-25*.

Tien, J. M. (2013). Big data: unleashing information. *Journal of Systems Science and Systems Engineering, 22*(2), 127-151.

Troester, M. (2012). *Big Data Meets Big Data Analytics.* USA: SAS Institute Inc.

Vaishnavi, V., & Kuechler, W. (2008). *Design Science Research Methods and Patterns: Innovating Information and Communication Technology*. Florida: Auerbach Publications.

Vossen, G. (2014). Big data as the new enabler in business and other intelligence. *Vietnam J Comput Sci, 1*, 3-14.

Washburn, D., Sindhu, U., Balaouras, S., Dines, R. A., Hayes, N., & Nelson, L. E. (2009). Helping CIOs understand smart city initiatives. *Growth, 17*, 1-15.

Watson, H. J., & Wixom, B. H. (2007). The current state of business intelligence. *Computer, 40(9)*, 96-99.

Weitzman, E. A. (1999). Analyzing Qualitative Data with Computer Software. *Health services research, 34*(5 Pt 2), 1241.

Williamson, K. (2002). *Research methods for students, academics and professionals: Information management and systems.* Elsevier.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, (pp. 29-39).

Yuen, M.C., King, I., & Leung, K.-S. (2011). A survey of crowdsourcing systems. *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)* (pp. 766-773). IEEE.