Anomaly Detection Beyond the Research Setting

An exploration of the use of statistics and machine learning to detect cyber attacks

Gunnar Daníel Sæmundsson

Anomaly Detection Beyond the Research Setting

An exploration of the use of statistics and machine learning to detect cyber attacks

by

Gunnar Daníel Sæmundsson

in partial fulfilment of the requirements for the degree of

Master of Science
in Systems Engineering, Policy Analysis and Management,
Information Architecture track,

at the Delft University of Technology,
to be defended publicly on Tuesday October 6, 2015 at 14:00.

Graduation committee:

Committee chair	Prof. Michel J.G. van Eeten (Professor section EoC)	TU Delft
First supervisor	Hadi Asghari, M.Sc. (Assistant Prof. section EoC)	TU Delft
Second supervisor	Dr. Wolter Pieters (Assistant Prof. section ICT)	TU Delft
Daily supervisor	Dr. Dina Hadžiosmanović (Postdoc section ICT)	TU Delft
Ext. supervisor	Irfaan Santoe, M.Sc. (Mgr. Cyber Risk Services)	Deloitte

Public version

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft Delft University of Technology

Deloitte.

Acknowledgements

This thesis marks the final step on my path to obtain a M.Sc. degree in Systems Engineering, Policy Analysis and Management (SEPAM) programme at Delft University of Technology. This graduation project *"Anomaly Detection Beyond the Research Setting"* was conducted in collaboration with Deloitte Cyber Risk Services and two other organisations that will remain anonymous due to confidentiality reasons.

I am very grateful of all the good, inspiring and supportive people who have been with me on this journey and helped me in so many ways.

First of all, I would like to thank my daily supervisor Dina Hadžiosmanović for her excellent guidance, brainstorming sessions, care for quality and personal commitment to see this project through to the end. It was good to have her as a supervisor as I could count on her to point me in the right direction with honest and well-reasoned feedback.

Secondly, want to extend my thanks to the rest of my graduation committee, Michel van Eeten and Hadi Asghari. Their research experience and way of seeing the big picture was valuable for this study, and their comments and suggestions influenced the course of this research for the better. Furthermore, the milestone meetings and discussions I had with the committee were very rewarding on their own.

I also want to acknowledge Jan van den Berg for inspiring me to pursue a project on data analytics and cyber security, and for his help during my first steps of this project.

This work was conducted as a part of my graduation internship at Deloitte Cyber Risk Services where I was surrounded by good people. I want to thank Robert for his detailed feedback and in particular his support during the uncertain project definition phase. Moreover, Irfaan for his supervision and involvement with both the academic and practical side of this project. Furthermore, special thanks to Sergio and Lourens for their help with connecting to organisations involved in this topic in practice, and to my fellow interns and other co-workers for their camaraderie. The people of the anonymous case study organisation also have my gratitude. The time spent with them was not only essential for this work, but also inspiring and fun.

The support and kindness of my family and friends during the project was invaluable to me, and I wish to thank Óli for his help and advice.

Very special thanks to my girlfriend Guðbjörg for all her kindness, encouragement and help during this project.

Gunnar Daníel Sæmundsson
Rotterdam, September 2015

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Governments and businesses around the world operate complex and interconnected information systems and networks. In recent years organisations have made their networks increasingly open to the outside world. For instance, many organisations allow partner organisations to access parts of their inner network, customers to directly interact with a company's databases in e-commerce transactions, or allow employees to access their private networks from home. This increase in external access has made today's networks more susceptible to attacks [2].

In the past, cyber attacks were generally perceived to be the work of the stereotypical lone hacker. In recent times they're increasingly observed to be the work of disgruntled employees, hacktivists, vandals and script kiddies, organised crime, terrorist organisations, and state actors [3, 4].

These developments, along with increasingly sophisticated methods to attack being more accessible, are considered some of the reasons why cyber attacks are on the rise [2]. In recent times there have been several highly publicised attacks, e.g. when an unknown group of hackers reportedly stole $300 million from banks [5], published private pictures of celebrities after hacking cloud services [6] and stole sensitive personal data on over 22 million U.S. government employees [7].

In order to detect cyber attacks organisations often monitor their networks, systems and applications for suspicious activity or known attack patterns. To do so they often deploy solutions like an intrusion detection system (IDS), intrusion prevention system (IPS) [8], and/or a security information and event management (SIEM) system [9].

Most of the methods and systems currently in use for detecting cyber attacks are rule-based, i.e. pre-existing knowledge is used to define monitoring rules. For example, a rule might generate an alert if the system observes 5 failed login attempts for a given user within a period of 30 minutes. With a well-defined set of rules, the main advantages of such systems is reliable detection and low false alarm rate. However, the main disadvantage of such systems is that without the relevant monitoring rules they can not detect unknown or novel attacks [2, 10].

Furthermore, newly discovered vulnerabilities often do not immediately become public knowledge. Hackers and computer security researchers that discover such vulnera-

bilities can sell the information on the black market. The buyer might then use the information to develop and execute new attacks [11], i.e. vulnerabilities that are known to some are unknown to others.

Anomaly detection refers to finding patterns or instances in data that do not conform to what is normal and expected, i.e. anomalies are rare and different from the norm. Anomalous patterns found in organisational system and application data can reveal attacks against an organisation without the need for pre-existing rules for each attack [12]. Anomaly detection for the detection of cyber attacks has been extensively researched by academia since it was originally proposed in 1987 [13]. However, few such systems have been successfully implemented in an operational environment for improving cyber security [10, 14]. One of the reasons for limited operational success is that technologies failed to provide the environment needed to do anomaly detection. For a long time it was not economically feasible to store and retain the needed (and vast) amounts of data and build the computational capabilities needed. With the emergence of new tools and technologies that are capable of handling and analysing large amounts of data (e.g. distributed technologies like Hadoop and Mapreduce) it is becoming increasingly feasible to do anomaly detection on a large scale [15].

In this work we approach this problem from an organisational perspective and try to identify promising practices (e.g. regarding data choices, output expectations, performance requirements) for deploying anomaly detection for cyber security purposes in a business environment. While anomaly detection is used in other problem domains (e.g. detecting anomalies in medical imagery and damage in industrial machinery [12]) we use the term 'anomaly detection' to refer to its application in cyber security.

1.1 Overview

In chapter 2 we describe the research methodology, including objectives, approach and research questions that we aim to answer in this work. Next, in chapter 3 we explore research topics related to this research. The topics include the current state of research efforts and challenges of anomaly detection, the execution anomaly detection projects within organisations, and usability of cyber security tools. Thereafter, in chapter 4 we construct and describe an anomaly detection methodology based on scientific literature. In chapter 5 we do case studies at business organisations that use anomaly detection as a tool for detecting cyber attacks. Furthermore, we link the case studies back to the theoretical methodology and compare. In chapter 6 we discuss important topics that arose from this study. Finally, we conclude the thesis in chapter 7 with conclusions and reflections.

Chapter 2

Research Methodology

In this chapter we describe the research methodology. First, we identify and explain the knowledge gap and problem statement. Second, we outline the aim of the research, high-level objectives, and the means to achieve them. Third, the research questions, and the academic and practical relevance of this research. Last, we describe our approach of answering the research questions.

2.1 Knowledge gap and problem statement

2.1.1 Complex projects in business organisations

Deploying an anomaly detection system in an organisation is a technological project. Generally, these type of projects are often complex and commonly exceed the time and budget allocated [16].

On one hand, such projects have to deal with technological complexity that is influenced by several factors. For instance, the size of the system, the number of tasks it has to perform, and the number of sub-systems [17]. More importantly, the extent of the interaction and dependency between sub-systems, e.g. bi-directional communication, and a fault in one sub-system easily spreads throughout the entire system [18]. Lastly, technological innovation increases complexity as the implications of using that technology may not be well understood, resulting in unique sub-systems that are often challenging to integrate with other sub-systems [17, 18].

On the other hand, technological projects face social and organisational complexities. Developing the interaction between different parts of the system is a shared task performed by fragmented groups of professionals (e.g. highly specialised but different types of engineers) [16]. Due to technological complexity, individual tasks often require certain skill-sets to perform [19]. The interaction between the different specialists may prove difficult (e.g. frequent misunderstandings and disagreements), hence adding to the project complexity [16].

Deploying innovative technology requires high-tech professional workers that have different motivations than managers and other types of workers [20]. The high-tech pro-

fessionals are highly educated, seek autonomy, and have strong ties to their technical speciality through external communities of professional peers. Moreover, they take pride in the quality of the technology, are motivated by challenging work, and are engaged in the innovation process and the realm of ideas. Furthermore, they may desire to disseminate knowledge that contributes to their field, or to push for the organisation to adopt standards set by the external communities of professional peers. Unsurprisingly, these characteristics may clash with the characteristics of managers that are more concerned with the health of the business than the technology itself [20].

Another essential point about complex projects is the complexity added by having multiple objectives and conflicting goals when working with many stakeholders (e.g. managers, clients, project team, business owner, public bodies) [21]. In fact, Engwall [22] defines environmental factors like sudden opposition of stakeholders as one of the three types of project management failures. The other two are general deficiencies in project management (e.g. poor planning or coordination) and the problems with aligning the project to its goals (e.g. they are vague, unclear, or constantly changing) [22]. As a result, a project manager of complex technological projects must be sensitive to organisational politics and to the (real or imagined) concerns of stakeholder groups resulting from the changes and disruptions brought by the project [23].

As mentioned before, the organisational deployment of anomaly detection will likely include these general complexities of technological projects. In this subsection we have introduced issues that are common in technological projects. In the following subsection we present issues that are more specific to anomaly detection.

2.1.2 Anomaly detection in business organisations

Anomaly detection has promising applications for cyber security. Mainly, for detecting unknown (or undefined) attacks that traditional rule-based detection can not do [24,25]. However, there are many practical challenges that apply to this problem in particular. For example, a high rate of false positives, difficulties with modelling normal activity, and costly evaluation [10,14]. In addition, organisational issues play an important role in the preparation, execution, and success of anomaly detection projects. Gaining access to data and necessary experts, and privacy and legal concerns are organisational issues that commonly affect anomaly detection projects [10,12,26,27,28,29]. Furthermore, the combination of a high false positive rates and alerts that are generally less interpretable than those of rule-based systems make usability especially challenging [10,14,30].

Some research exists (e.g. [10, 14]) where researchers provide a set of guidelines, i.e. best practices, that are designed to improve the approach of researchers or others that want to deploy operationally sound and effective anomaly detection techniques for detecting cyber attacks. These are general guidelines on how to avoid common pitfalls and address potential problems when applying anomaly detection. However, the guidelines are too general and abstract to be directly applied to specific types of organisations or for addressing specific cyber security tasks in business environments. Therefore, it is a challenging task for an organisation to create its own pathway for deploying a usable anomaly detection approach.

It is likely that many researchers do not have a clear insight into how different business organisations approach anomaly detection in practice and what issues may play an important role that environment. This is partly due to the fact that research on the topic is sparse and this sort of insight is not made readily available by business organisations. Therefore, it may be difficult for researchers to provide clear guidelines for business organisations interested in deploying anomaly detection tools.

As mentioned before anomaly detection has been extensively researched for almost three decades while it has seen relatively little operational success. We argue that we have to better understand this gap in order to move research and practice towards more operational and usable anomaly detection.

Problem statement:

> **There are significant discrepancies between the application of anomaly detection in research and business organisations, resulting in low utility of anomaly detection solutions in industry.**

2.2 Objectives

The main objective of this research project is to contribute to the research area of anomaly detection by investigating the gap, or discrepancies, between anomaly detection in practice and research. We argue for importance of further exploring the best practices and practical challenges that face anomaly detection in the organisational environment. With a better understanding of the gap we aim to provide recommendations for business organisations that want to deploy anomaly detection.

In order to reach our goals this research has three sub-objectives. First, we propose a a theoretical methodology for producing usable anomaly detection approaches for business organisations based on the guidelines discussed in academic research (chapter 3). We gather best practices covering different issues that we find are often overlooked or underestimated in research and practice. Some of these issues apply more specifically to anomaly detection (e.g. costly evaluation, variability of data) while others apply to data mining projects in general (e.g. working with problem owner, access to data or experts). Furthermore, we focus on two usability issues that are seldom the main focus of studies on anomaly detection, the amount of false alarms and actionable alerts. Second, we gain insight into how business organisations deploy anomaly detection in practice. We do a case study comprising of interviews with organisations that do anomaly detection. In the interviews we broadly cover the content of the theoretical methodology and any other issues that play an important role in the specific organisational context, e.g. requirements, expectations and business case. From the interviews we identify the main success factors and challenges that these organisations have encountered. Moreover, we reconstruct and understand their methodology for deploying anomaly detection. Third, we aim to understand the similarities and differences between anomaly detection in practice and in theory by analysing and comparing different methodologies.

2.3 Research questions

Main research question:

What are the core discrepancies between theoretical guidelines and operational approaches when using anomaly detection in business organisations?

Research questions:

RQ 1 - What is the state of the art of research on anomaly detection for detecting cyber attacks?

In this research question we explore literature on anomaly detection in business organisations to determine the existing recommended practices and guidelines. We focus on the technical and organisational challenges of anomaly detection and the usability of the alerts generated by an anomaly detector. This literature study has two goals. First, to define a set of statements (propositions) to use as building blocks when constructing a theoretical methodology for deploying operational anomaly detection approaches. Second, to design interviews that cover the broad set of issues related to making anomaly detection operational.

RQ 2 - What methodology for anomaly detection does academic research propose for business organisations?

With the propositions derived from the recommended practices and guidelines from academic research (**RQ 1**) we construct a theoretical methodology for deploying operational anomaly detection. More specifically, we tailor a generic methodology for data mining projects to the guidelines for deploying operational anomaly detection from academic research. The main purpose of this research question is to create the foundation for comparing theory with practice. This theoretical methodology is general in the sense that it is not tailored to specific types of attacks, data, tools, or organisations. However this will enable us to compare theory with a diverse set of practical approaches.

RQ 3 - How do business organisations approach anomaly detection?

In this research question we attempt to understand and reconstruct the implicit or explicit approaches followed by business organisations doing anomaly detection. We do that in a series of case studies where we use the interview questions defined in **RQ 1** to explore how they have approached this problem in practice. Furthermore, we aim to identify success factors and challenges of anomaly detection from the perspective of the business organisation. In addition, we aim to highlight stakeholder complexities that are likely to occur in complex technological projects as discussed in subsection 2.1.1.

RQ 4 - How does a theoretical methodology compare to practical approaches in business organisations?

In this research question we analyse the differences (and similarities) between anomaly detection in research and industry. More specifically, we compare the theoretical methodology (**RQ 2**) with the approach of business organisations (**RQ 3**).

Firstly, we evaluate the *soundness* of the theoretical methodology. For instance, we identify and understand which propositions are most relevant and irrelevant in practice. Secondly, we look at the *completeness* of the theoretical methodology in practice. That is to see whether we have identified significant technical, organisational or usability issues or concerns that are not covered by the theoretical methodology.

2.4 Relevance

Scientific relevance. A large body of research on anomaly detection exist. However, a small subset of it focuses on ways to make these techniques operational in industry. In this work we will add to this research interesting by testing the applicability of these theoretical propositions in practice. These propositions have, to our knowledge, not been explicitly tested in practice. Furthermore, in an effort to help bridge the gap between theory and practice we perform an analysis of current anomaly detection practices in business organisations.

Management relevance. From a practical perspective this research is relevant as it aims to identify good practices for organisations deploying anomaly detection. The outcome of this work can help managers and experts to define or follow anomaly detection approaches that focus on usable and operational results, and avoid common pitfalls on the way.

2.5 Research approach

This research follows an *exploratory* approach. As is common in exploratory research, this work makes use of propositions based on a literature review [31]. The propositions are the main findings of the literature review, forming building blocks of the theoretical methodology for doing anomaly detection. These are provisional tools for advancing the research with the aim of leading to the discovery of new insights or facts. Thereafter evidence is collected that may or may not support the propositions [31].

In this study we collect evidence in case studies at two business organisations that are doing anomaly detection in practice, a financial institution (FI) and an internet service provider (ISP). By means of interviews and observations of practical work in such organisations we explore how they approach the problem of making anomaly detection operational. In summary, we explore how well the propositions from academic research hold in their intended environment.

The following paragraphs outline the research approach followed for the major phases in this work.

Literature review (chapter 3). We study literature on the state of the art of anomaly detection for detecting cyber attacks. More specifically, we study on both the technical and organisational aspects of such projects. In addition, we explore the usability criteria of alerts produced by such tools as they eventually have to be verified by experts.

Propositions and a theoretical methodology (chapter 4). We extract statements, or propositions, e.g. in the form of best practices, challenges, guidelines, from the literature review and construct a theoretical methodology for anomaly detection in practice. More specifically we tailor CRISP-DM [1], a well-known generic framework for data mining projects, to anomaly detection. This is a provisional 'ideal' methodology for doing anomaly detection that will guide us in the process of conducting this research and discovering new insight.

Exploration of practices in business organisations (chapter 5). We do case studies of business organisations that are actively pursuing or deploying anomaly detection. We interview various stakeholders related to those efforts, e.g. business managers, project managers, data scientists, data engineers, and security experts. We select interviewees from different roles that reflect the broad range of topics from the literature survey, i.e. the purpose is to look at anomaly detection from different perspectives, technical, organisational and usability.

The interviews will be conducted in face to face meetings. These will be semi-structured interviews, structured around the theoretical methodology while remaining open, allowing interviewees explore and discuss different views. Firstly, they will broadly cover the contents of the literature review and theoretical methodology. E.g. success criteria, modelling and maintaining normal activity, evaluation, organisational support, usability of alerts. Secondly, we ask open questions that enable the exploration of ideas and topics from the perspective of the interviewee. These questions will emphasise how the organisation approaches anomaly detection, what has been successful and what has been unsuccessful.

Analysis (chapters 5 & 6). Following the interviews we start the analysis phase where we compare theory and practice. Starting the analysis phase, we first revisit the theoretical methodology and link the data from the case studies back to the propositions. We identify the main differences and similarities between the two by looking at whether the data from the case studies support, fail to support, or provide an alternative perspective on the proposition(s).

Moreover, we investigate some of the major discrepancies between practice and theory. We elaborate on the main issues that come up during the case studies but are not identified in the theoretical methodology, i.e. the additional insight from the case studies. These may be issues important for operational deployment that researchers could further explore, or incorporate into their research on anomaly detection.

2.6 Thesis structure

Figure 2.1 is a visual presentation of the research methodology, research questions, their relation to the case studies and the individual chapters of this work.

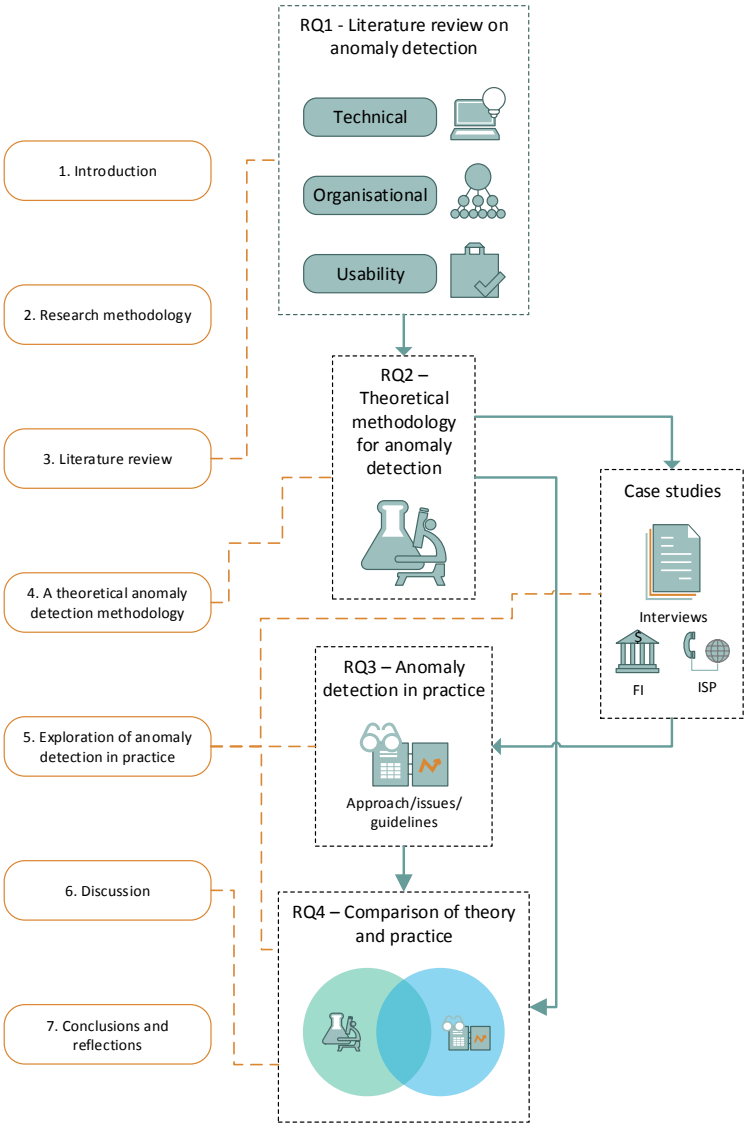

Figure 2.1: Structure of the thesis

Chapter 3

Literature review

In this chapter we answer **RQ 1** by exploring research topics related to anomaly detection in practice. First, we discuss the state of the art of using anomaly detection to find cyber attacks. There we explore what advantages anomaly detection can hold for business organisations. Moreover, we research the potential reasons behind the relatively limited operational success of anomaly detection. Second, we explore organisational issues that can affect the execution of anomaly detection projects. Third, we look into the topic of usability of alerts produced by cyber attack detection systems. With the number of false alarms and interpretability of alerts being some of the main challenges of anomaly detection [10, 14] we consider these as key usability issues for organisational success.

3.1 The state of the art of anomaly detection for cyber attacks

Anomaly detection is an active area of research, and has been so for almost three decades.

In 1988, a year after Denning [13] first proposed it, Lunt [32] presents a prototype host-based intrusion detection system that uses both rule- and anomaly-based methods. The system learns the normal behaviour of the user over time and reports anomalous behaviour. The researcher argues that using both methods the system benefits from the strengths of both approaches while overcoming their weaknesses (e.g. false alarms and inability to detect unknown attacks) [32]. Ever since the research area on using anomaly detection for detecting both *network-* and *host-based* attacks has been active [12, 24].

One of the main advantage of anomaly detection is the ability to detect unknown (or undefined) attacks, which rule-based methods are unable to do [24, 25]. In addition, knowledge gained from the detecting anomalies can be used for developing and refining detection rules [25, 33, 34], possibly reducing the time spent and cost of the process. This is for example done by using rule extraction techniques to generate comprehensible rules [35] from the activity identified as anomalous, i.e. translate anomalies into understandable rules.

Labelled training data for both normal and malicious activity is often unavailable

in practice as it is often technically challenging or prohibitively expensive to obtain [12]. Therefore, methods that do not rely on labelled data (*unsupervised*) are the most widely applicable compared to the other types of machine learning (i.e. *supervised* and *semi-supervised*) [12]. However, the unsupervised methods assume that normal network activity is far more frequent than anomalous, which is not always valid [24]. Many semi-supervised methods can be applied under the same assumption. That is, as long as the resulting model is robust to the anomalous activity present in the data [12]. In such cases the unlabelled data used while it is assumed to contain only instances of normal activity [36].

Research remains active and recent publications make promising claims. For instance, in [37] the authors propose a real-time method to detect both encrypted and unknown attacks in high-level network data. In the first phase they find time slots where the network traffic changes, within that period identify outlier traffic patterns, and send to human expert validation. In case of possible Botnet attacks the system goes into another phase where it attempts to identify communication between bots and botmasters. As future work, the authors intend to test their method on simulated datasets. In [38] the author attempts to detect attacks in an organisation's private network. Using real-life data, they employ algorithms that splits network entities into two groups, clients and services. Next, usage profiles are constructed for each group and how they interact with each other. Anomalous activity then triggers an alarm, for instance when a client uses an uncommon service that does not fit his normal usage profile. This results in an anomaly detection that generates a few hundred anomalies per day. While there are claims that a human administrator can easily inspect this amount of anomalies in a day [38], the usability of this approach is questionable as the author does not provide support for this claim. In both [39] and [40] authors claim their anomaly detector can detect network attacks like distributed denial-of-service (DDoS) attacks, port- & network scans, and spreading worms. The methods are applied on high-level network traffic data without any prior knowledge, labels, or rules. Both find clusters of patterns within the traffic data to detect previously unknown attacks. Both claim to have a high detection rate and a low false positive rate using simulated datasets.

Summary

In this section we took a broad look at the research developments and promising applications of anomaly detection for cyber attacks. We now highlight a few conclusions. Firstly, anomaly detection has the potential of detecting attacks without the prior knowledge needed for constructing conventional detection rules. Moreover, the knowledge gained from using these methods can be used for defining and refining conventional detection rules. Secondly, unsupervised (and semi-supervised) machine learning techniques seem to be the most applicable since labelled training data is often unavailable and hard to obtain.

Characteristic	Description

Potential benefits	Detection of new attacks [24, 25]. Refinement of detection rules [25, 33, 34].
Types of learning in anomaly detection	Unsupervised is widely applicable since labelled data is often unavailable [12]. Semi-supervised can be used on unlabelled data, as long as anomalies are rare enough and the model is robust to the anomalies [12, 36]. Supervised requires a dataset with accurate labels of both normal and anomalous activity [12].

Table 3.1: The state of the art of anomaly detection (3.1)

3.2 Technical challenges in anomaly detection for cyber attacks

The machine learning methods that make the foundation of anomaly detection have been successfully implemented in several problem domains (e.g. spam detection and recommendation systems). Even though this topic has been studied in a large body, these techniques have not become widespread and operational in business organisations [10]. Two articles [10, 14] that examine the reasons for this gap between academic research and implementation form the foundation of this research.

Assumptions. Gates and Taylor [14] identify several common assumptions that are found in literature on anomaly detection and question whether these assumptions hold for the operational network environment that the anomaly detectors are being developed. The authors define three categories of assumptions about the problem domain, training data, and operational usability.

The assumptions about the **problem domain** are that attacks are rare and distinguishable from normal behaviour, and that anomalous activity is malicious. There are several examples available where these assumptions do not hold [14]. Firstly, attacks are not anomalous if attackers deliberately and successfully conceal their activity within the variable nature of network- or host activity [41, 42]. This was observed in a recent case where an unknown group of hackers reportedly stole $300 million from banks after months of lurking in their systems and observing, eventually mimicking the behaviour of employees who's accounts the attackers had compromised [5]. Secondly, attacks are not always rare as is shown in [43] where 8 million scanning attempts took place in one day on a single organisation's network, two times greater than the number of normal incoming connections for that day. Thirdly, the assumption that most anomalies are malicious does not hold in network environments, as is shown in studies [44, 45] where few to none of the detected anomalies represent malicious activity.

The assumptions about the **training data** are that attack-free training data is available, simulated data is representative, and that network traffic is static. The authors

discuss examples where these assumptions do not hold. Firstly, research on the topic often requires attack-free data, but as discussed before real-life data is likely to contain a large number of attacks and the process of making it attack-free is a cumbersome process [14]. Secondly, simulated data is often used due to availability and privacy reasons. The most popular datasets available for intrusion detection research are old, lack variability, and not representative of current real-life network traffic conditions [14]. Maxion and Tan [46] demonstrate the effects of variability in data has on performance by tuning parameters for generating simulated datasets. While maintaining a 100% detection rate, the s range from being close to none up to a rate of 100% in increasingly variable datasets. Thirdly, network traffic has been shown to be highly variable in both the short term and the long run, most research acknowledge this fact while failing to address issues related to updating the anomaly detectors [14].

The assumptions about the **operational usability** are that false alarm rates above 1% are acceptable, that there is consensus on what is considered malicious activity, and that experts can interpret anomalies. There are several examples available of where these assumptions do not hold. Firstly, even with small percentages of false alarms, the anomaly detection system can quickly become unusable, for instance if the amount of data means that 1% false alarm rate generates thousands of alerts each day [14]. Hadžiosmanović et al. [30] evaluate the usability of algorithms used in anomaly-based intrusion detection system research by looking at thresholds for false positives: 10 false positives per day represents the number false alarms for a user to maintain trust in the system, and 1 false positive per minute represents the maximum number of alarms that a human can verify for traditional intrusion detection systems. When confronted with highly variable data, high detection rate was not possible without unacceptably high levels of false positives. Secondly, research is often missing a discussion about what is defined as malicious activity and treat all anomalies as interesting and potential attacks. However, this definition is not universal but has more to do with the security policies and priorities of each organisation, e.g. scanning activity might be treated as an attack in one organisation but merely a nuisance to the other [14]. Thirdly, unlike rule-based systems anomaly detectors do not report information on what types of attacks have been detected, hence it is likely that more time is required to verify alerts from such systems [30]. The assumption that experts have the time, interest and ability to verify the anomalies is questionable since the same experts are likely busy enough with dealing with known threats [14].

Unique challenges. Sommer and Paxson [10] discuss unique challenges related to the use of machine learning techniques in anomaly detection for cyber attacks and divide them into five categories: challenges of outlier detection, the high cost of error for this problem domain, the semantic gap, diversity of network traffic, and the difficulties with evaluation.

If the goal of the anomaly detector is to discover unknown attacks it has to be able to do meaningful **outlier detection** which detects activities that deviate from the norm. To do so normal activity has to be modelled which is hard for the high dimensional

and variable data produced in today's organisational environment. Anomaly detection for credit card transactions have been shown to work in practice, but that problem resides in better-defined and lower dimensional data than the detection of cyber attacks. Furthermore, machine learning generally performs better when classifying into known groups, like in the case of spam detection where a message is either spam or not, and large number of examples of both types are available [10].

Another challenge is the **high cost of errors** for this problem domain. False positives are costly since verifying alarms from anomaly detectors takes a longer time than for traditional systems [30] and normally has to be done by expensive experts, e.g. system administrators and analysts. Moreover, false negatives are attacks that went unnoticed and can cause serious harm. Other applications of machine learning have lower cost of errors: in product recommendation systems an error means lost opportunities rather than significant harm, text recognition is cheap to verify, and in the case of spam detection false positives are expensive while false negatives are not and the detection tools tuned accordingly [10].

The **semantic gap** refers to the challenge of transforming detected anomalies into actionable alerts for the end user. Ultimately the goal of the anomaly detector is to detect attacks as opposed to abnormal activity. This challenge relates to the issue of different definitions of malicious activity between organisations, i.e. different security policies, and what kind of conclusions it is possible to make given the data that are put into the model. Thus, anomaly detectors should also be evaluated based on their ability to produce alerts that have meaning to the organisational environment that they operate in [10].

As discussed before, the **diversity of network traffic** (and application and system activity) in operational environments is often much higher than people expect. This leads to unrealistic expectations of what anomaly detection tools can achieve in real-life environments. In order to overcome this variability, anomaly detectors often use highly aggregated data, e.g. by calculating sums of traffic volume, or by counting the number of connections in an hour. In these cases, it is possible that other more simple and non-machine learning approaches, like establishing simple thresholds, can perform just as well [10].

Lastly, Sommer and Paxson [10] discuss the **difficulties in evaluation**. Most issues they mention have already been discussed, e.g. the problem with the availability of real-life data, the problems with simulated data, and the time-consuming process of validating the detected anomalies. Additionally, they mention the adversarial setting of this problem domain. A simple example is that while retail customers are not likely to intentionally mislead product recommendation systems, attackers are more likely to try to evade detection. As in the case of the bank heist mentioned earlier, evasion is a real problem [5]. However, the authors conclude that it does not have the same impact as the other challenges mentioned [10].

Summary

In this section we explored the challenges of operationalising anomaly detection. We looked at the challenges from two different perspectives. Firstly, they lie in assumptions that are commonly made in anomaly detection research. However, these assumptions may not hold for anomaly detection an operational environment. Secondly, in unique problems of using machine learning techniques for detecting cyber attacks.

The challenges discussed cover a wide range of topics (see Table 3.2 below). For instance, that not all attacks are anomalies, and anomalies are not necessarily malicious. Moreover, real-life network and system activity is much more irregular and unpredictable than most people expect. A high variability in the short term and on the long run leads to difficulties in defining normal behaviour needed for detecting anomalies. Furthermore, the high cost of both false positives and false negatives make it hard to tune the model. Evaluation is also costly as it can take a long time to verify each alert produced by anomaly detectors. Moreover, the evaluation normally has to be done by expensive and busy experts.

Research on anomaly detection sometimes overlooks these issues important for operational success. For example, how to cope with constantly changing notion of normal activity, or whether the alerts produced are manageable and useful for the experts that verify them.

Challenge area	**Description**
General	Attacks are not necessarily anomalous and/or rare [5, 10, 14, 41, 42, 43].
	Anomalies are not necessarily malicious or interesting [10, 14, 44, 45].
	Machine learning performs better when classifying into known groups with labelled data [10].
	The high cost of both false positives and false negatives [10, 14].
	To overcome variability, data is often aggregated to a point where simpler methods work just as good [10].
Data	Attack-free data or 'normal activity' is often unavailable [10, 14].
	High dimensionality and variability makes it hard to model normal activity [10, 14].
	Results on simulated data are not a good indicator of real-life results [10, 14, 46].
	Underestimation of variability of data leads to unrealistic expectations of anomaly detection [10].
Usability	With a large amount of data, a small false positives rate can result in an unusable anomaly detector [10, 14, 30].
	Challenging to transform detected anomalies into actionable alerts [10, 14, 30].
Evaluation	No universal definition of malicious activity. Definition dictated by security policies and priorities [10, 14].
	False positives are costly. Verifying takes a long time and performed by expensive and busy experts [10, 14, 30].

	Ability to produce alerts that have meaning to the organisational environment that they operate in is important. Evasion tactics of attackers can pose a problem in evaluating effectiveness [5, 10].

Table 3.2: The challenges in anomaly detection (3.2)

3.3 Organisational challenges in anomaly detection projects

Anomaly detection projects within business organisations are not only a technical challenge. An important part of such projects is to understand, from a business perspective, the objectives, requirements, and issues or considerations that might affect the project. Thereafter, translating this understanding into the project definition and plan [1].

Business issues are seldom the main focus of research on anomaly detection, or data mining in general. However, such issues are wide-ranging and play an important role in a business environment [26, 47, 48]. In the *Cross-Industry Standard Process for Data Mining* (CRISP-DM) reference model these organisational issues are embodied in first phase of the model, the *Business understanding* phase. There the project is defined from the perspective of the problem owner. Moreover, the outcomes of the Business understanding phase have an impact on the whole project life-cycle, the implementation, evaluation, and eventually the deployment of data mining projects [1].

In the following paragraphs we explore organisational issues that can affect anomaly detection projects. First, we look at people and organisational issues encountered in anomaly detection and data mining projects in general. Second, we identify issues that apply in particular to anomaly detection. Some of the other organisational issues were mentioned in section 3.2. For instance, the availability and ability of experts to evaluate the anomaly detection results and accuracy requirements that are impossible to achieve given the variability of the data.

Success factors. Nemati and Barko [47], and Hilbert [48] identify and test success factors for organisational data mining projects. In both papers, the authors hypothesised several success factors based on literature survey. Both test their hypotheses by surveying organisations involved in data mining and propose a set of significant success factors based on the results. The significant success factors are a diverse collection of organisational issues that influence data mining projects.

Some of the factors have to do with data and technology. For instance, the quality of data [47, 48]. Moreover, the presence of information technology (IT) that enables the integration of data mining results into the organisation's work flow [47, 48]. Other success factors apply to the people executing the projects, like the level of employee data mining expertise [47, 48], technological expertise [47, 48], and an outsourcing strategy for data mining projects [47]. Success factors related to management were also found

to be significant. Firstly, the commitment and understanding towards such projects from top level management, e.g. through guarantees like a financial budget. Secondly, effective change management for integrating the outcome of the project into the business work flow [48]. Lastly, successful projects were generally associated with relatively little dependency of resources, limited scope, and short time frame [47].

Common issues. Weiss [26] discusses organisational and human issues that often affect data mining in business organisations. Those who execute data mining projects usually require substantial support from the problem owner organisation, e.g. access to the necessary data, documentation, and experts with relevant domain knowledge. The author states that such support is often lacking, including in projects where all parties belong to the same organisations. He mentions several possible causes for this lack of support. For instance, experts may not be willing to cooperate if they feel that sharing their unique knowledge decreases their own job security or power within the organisation. Furthermore, budgeting constraints and lack of time can hinder necessary expert support. To overcome these issues, it is important to assess the organisational support before starting data mining projects. Another issue that can affect data mining in organisations is the level of commitment from the problem owners. As data mining is a new tool for many organisations it is common that the problem owners do not immediately recognise the need for the new techniques. Instead, it is common that data mining experts actively seek out problems to solve, educate problem owners and sell ideas. As a result, the experts often dive into projects without full commitment, hoping to gain it with quick and promising results [26].

Organisations and anomaly detection. We explored literature on anomaly detection projects in organisations and identified three issues that apply to anomaly detection projects in particular.

First, anomaly detectors with a clear objective and clearly defined targets generally perform more accurately than those without [10]. The objective and purpose in anomaly detection projects are generally defined by the problem owner, a business organisation [1]. Hence, it is important that a clear problem definition and objectives for the project are defined by (or in collaboration with) the problem owner. However, that can be a challenge as problem owners often do not have a clearly defined problem to solve [26]. For example, when an organisation is interested in using anomaly detection on a particular dataset to find 'unknown attacks' without a clear idea for success criteria for the project.

Second, anomaly detectors are evaluated based on their ability to detect malicious activity. However, the definition of 'malicious' depends on the often subjective and organisation-specific security policy and priorities [14]. Moreover, such company policies are often written in vague legal language rather than in clear technical terms [10]. For example, a military organisation might consider all peer-to-peer activity as malicious while a university might consider it normal activity [14]. In addition, the university's policy might allow file sharing activity on its network as long as no inappropriate content is shared and the activity does not negatively affect the performance of the network [10].

Thus, the same technique produces a good detection rate for the military organisation while it would produce a high false alarm rate and complicated evaluation process for the university.

Third, privacy concerns and legal issues often affect anomaly detection research projects. For example, organisations often can not provide the researchers with access to relevant data [27] or can only allow researchers test their models without direct access to the organisation's data [28]. Moreover, the use of privacy preserving technologies is sometimes required [12] or organisations can only give the researcher to access to anonymised data [10]. These concerns can make the process more cumbersome. In addition, privacy preserving- and anonymisation techniques often remove important information that could have been used for the anomaly detection model [29]. For example, when anonymising IP addresses that can be used to determine geographical locations and internet service providers.

Summary

In this section we discussed the organisational aspect of anomaly detection (and data mining) projects. A good understanding of the problem from a business perspective and translating this understanding into concrete parts of the project definition is crucial. Other success factors include data quality, organisational support and integration of results into business processes. Moreover, smaller anomaly detection projects (scope, resource dependency, time) tend to be more successful. In addition, we outline human and organisational issues that can hinder the progress of anomaly detection projects. For instance, common problems with access to necessary data, documentation, and experts with relevant domain knowledge. Lastly, we explore how issues such as privacy and legal concerns, and security policies written in legal language can affect anomaly detection.

Issue	**Description**
Problem owner	The objectives, requirements, issues and constraints are important [1,10, 26,47,48].
	Impacts the whole life-cycle of anomaly detection projects, e.g. preparation, execution, evaluation and deployment [1,26,47,48].
	Collaborate with the problem owner to work out a clear problem definition and project objectives [1].
Success factors	Data quality and technological expertise [47,48].
	Data mining expertise or a strategy to obtain the expertise [47,48].
	Organisational support and integration of anomaly detection outcome into business processes [48].
	Little dependency on resources, limited scope and short time frame [47].
Common barriers	Lack of access to necessary data and experts [26,27,28].
	Lack of commitment from problem owner [26].
	Problem owners not recognising the benefits of anomaly detection [26].
	The definition of malicious activity varies and is often described in vague language [10,14].

	Privacy and legal concerns can affect the researcher's way of working and limit access to data [10, 12, 27, 28, 29].

Table 3.3: Organisational challenges in anomaly detection projects (3.3)

3.4 Usability in anomaly detection

The evaluation of practical usefulness of systems can be divided into two categories: utility and usability. Utility is the question of whether a system has the functionality to do what is required. Usability applies to every part of the system where there is human interaction and is the question of how well the users of the system can utilise that functionality [49].

Usability considers several different aspects of the user experience interacting with the system. This includes how easy it is to learn and remember how to use the system, how efficient the system is to use for experienced users, the system error rate and ability to recover from them, and how pleasant the user finds using the system [49].

Nielson [50] states that there are two basic ways, that work well in practice, of evaluating usability of user interfaces. First, *informally* using rules of thumbs, and the experience of usability experts. For example, through expert-based evaluation where a range of usability rules and heuristics (i.e. design principles) or the knowledge of usability experts are used to evaluate the system or guide its design [51]. Second, *empirically* testing the usability with real users. For example, by conducting user studies where a sample of the typical end-user participates in experiments to test a system's usability [52].

Nurse et al. [52] recommend combining both methods for the development of cyber security tools. First, use expert-based usability knowledge to guide the system design from early stages. Second, conduct user studies in later phases to identify overlooked usability problems and confirm previous design choices.

Usability is an important concern for cyber security tools. In the past cyber security tools have tended to have poor usability due to a confusing user interface [53], high workload, and increasing level of complexity. Poor usability can lead to inadequate use, which limits the effectiveness of the tool [52]. Furthermore, literature on the topic states that by using usability heuristics tailored towards cyber security tools, more severe usability problems are found than when using general usability heuristics [54, 55, 56, 57].

In the previous sections we discussed several issues related to usability. For instance, highly variable data can result in false positive rates that make the solution unusable for the problem owner [30, 46]. Moreover, anomaly detection alerts are harder to verify than alerts generated by rule-based system because alerts from anomaly-based systems generally provide the end-user with less useful information than in rule-based systems [10, 14, 30].

Usability of cyber security tools can mean many different things, e.g. quality of documentation [54], and supporting and facilitating the division work between different

users of the tool [55]. However, we will focus on usability of alerts in this research. More specifically, a workable amount of alerts and alerts that are interpretable and actionable. We focus on these issues because we have seen (see section 3.2) that there are practical challenges for anomaly detection that tend to increase error rates [10]. Adding to the challenge of high error rates, alerts from anomaly detectors generally take longer to verify than alerts from rule-based systems [30]. In the following paragraphs we take a closer look these two topics.

Actionable alarms. Zhou, Blustein and Zincir-Heywood [54], Jaferian et al. [55], Ibrahim et al. [56], Patil, Bhutkar and Tarapore [57] propose usability heuristics and metrics for a wide range of cyber security tools (intrusion detection systems [54, 57], personal internet security tools [56], and a wide range of related cyber security tools for protection, detection, or user management [55, 57]). While the studies propose a wide set of guidelines, we focus on those related to producing actionable alerts. Compared to other types of tools, the display of information and information navigation plays a greater role in the usability of cyber security tools [54]. Flexible representation of information is also seen as important so that alerts can be catered to different stakeholders through a variety of means such as reporting or visualisation [55]. For example, this is important for an organisation that provides managers with weekly high-level reports and security officers with detailed alerts and visualisation of real-time network activity. This implies that the usability criteria of the output of anomaly detectors varies depending on different stakeholders. In addition, the researchers discuss the importance of using simple and consistent vocabulary for describing alerts and severity levels. Furthermore, the tool should help the users to evaluate the alerts through the use of colours, tables, charts and other visuals [56,57]. Moreover, alerts should avoid overwhelming users with information, instead, make it simple to retrieve additional information [56, 58].

Manageable number of alarms. Even with a good detection rate, false alarms can cause the end-users of anomaly detectors to spend hours each day to investigate and dismiss false alarms. As a result, users may quickly lose trust in the system [59].

Axelsson [60] states that in an operational setting the ability to suppress false alarms is more important to an intrusion detection system's performance than a high attack detection rate. Moreover, he argues that a rate of 1 false alarm for every 100.000 events (normal and malicious) in the system is an acceptable false alarm rate for effective detection systems.

Lippmann [59] discusses evaluation of both anomaly- and rule-based tools for detecting cyber attacks. A common primary success-metric in intrusion detection research is the attack detection rate (true positives). The author argues that in an operational setting, this metric is insufficient without also looking at false alarm rates. Together these two metrics allow for evaluating the workload required to verify the alerts. Moreover, the author argues that 10 false alarms per day is an acceptable number of false alarms [59].

Summary

In this section we take a closer look at two important usability issues of anomaly detection in practice: a manageable amount of false positives, and alerts that convey meaning to the end-user (i.e. actionable alerts). These issues are important as we have seen in section 3.2 looking at unique challenges of anomaly detection for in an operational environment. According to papers discussing actionable alerts we see that flexible presentation (e.g. visualisation, reports) of information and facilitation further investigation are important factors. To achieve a manageable number of alerts we found that a very low rate of false alarms is needed to maintain users' trust in the results. In addition, we see researchers stating that a low false positive rate is of greater importance than a high true positive rate. Lastly, we identified best practices for working with usability: In early phases, use relevant expert-based usability knowledge to guide the process of deploying anomaly detectors. In later phases, verify usability in practice with end users and problem owners.

Issue	**Description**
Usable design	In early stages, guide design with expert-based usability knowledge (heuristics) [50, 51, 52]. In later stages, verify design decisions with end-users (testing, surveys, etc.) [50, 51, 52].
Actionable alerts	Different stakeholders have different criteria for actionable alerts/output [54, 55]. Alerts should be clear and simple, while helping the user retrieve additional information [56, 57, 58].
Manageable alerts	Low false alarm rate is more important than a high detection rate for usability in an operational setting [59, 60]. Acceptable false alarm rates: 1 false alarm for every 100.000 events and 10 false alarms every day [59, 60].

Table 3.4: Usability in cyber security (3.4)

Chapter 4

A theoretical anomaly detection methodology

In this chapter we answer **RQ 2**. We construct and describe a theoretical methodology for doing anomaly detection in practice. The building blocks of this methodology are based on the literature review (chapter 3). More specifically, the building blocks are made up of several propositions, statements about the nature of anomaly detection in practice. Propositions are commonly used in qualitative case studies, and similar to hypotheses, they are an educated guess to the possible outcome of the study [61]. The underlying structure of the theoretical methodology is based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) [1] reference model (Figure 4.1), a well-known framework for data mining projects within business organisations.

The purpose of the theoretical methodology, and propositions is to guide this work in the process of discovering new insights. In a later stage, these new insights come from case studies, exploring and analysing how anomaly detection is deployed in practice.

In the following section we introduce CRISP-DM. In the second section we discuss our approach for constructing this methodology. Next, we construct and present the theoretical methodology in section 4.3. Lastly, we outline the conceptual framework of the thesis in section 4.4.

4.1 Introduction to CRISP-DM

CRISP-DM was conceived in 1996 when the data mining market was still in its infancy, young and immature. Relatively few organisations where practising data mining and its authors wanted to develop a standard approach of doing data mining for two main reasons. First, to capture what different organisations have learned so that every new adopter did not have to go through the same learning process of trial and error. Second, to demonstrate to their prospective clients that data mining was mature enough for them to adopt these techniques into their business processes. The result is a freely available model built from practical experience and generic, i.e. it is not tailored to a specific application, tool, nor type of organisation [1].

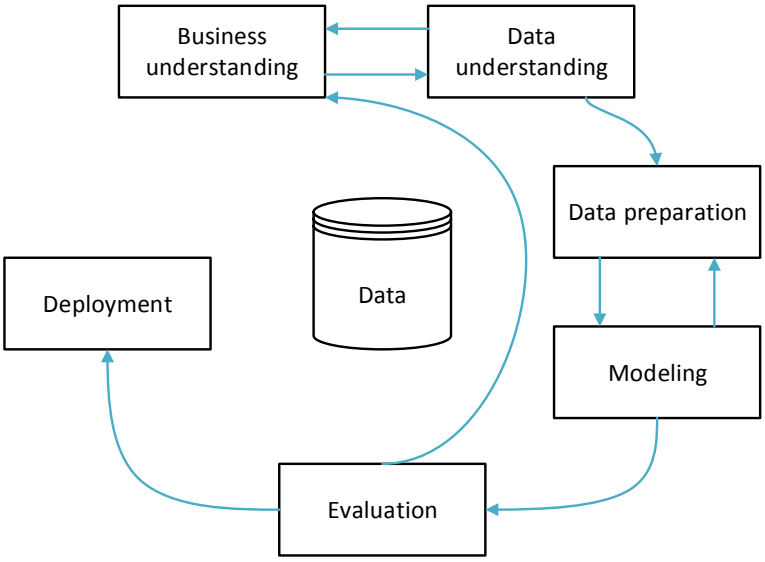

Figure 4.1: The CRISP-DM reference model [1]

The six phases of CRISP-DM CRISP-DM consists of six phases of the life-cycle of data mining projects: From the initial project definition, to the eventual deployment and plan for maintenance. While the model is depicted in an (ideal) sequential order, the designers of CRISP-DM acknowledge that in a practical setting tasks and phases will likely be performed in different order [1].

A brief description of the six phases is as follows: Firstly, the **business understanding** phase where business requirements and objectives are translated into a data mining definition and project. Secondly, the **data understanding** phase where data is collected, preliminary exploration is performed, and potential quality problems are identified. Thirdly, the **data preparation** phase which is often performed iteratively and comprises all activities related to preparing the raw data and constructing the final dataset. Fourthly, the **modelling** phase where modelling techniques are selected, calibrated, applied, and evaluated from a data mining perspective. Fifthly, the **evaluation** phase the model is thoroughly evaluated on how well it achieves the business objective, and which business issues have not been sufficiently considered, e.g. some of the criteria may have changed during and/or as a result of the data mining. Lastly, **deployment**, where ready models are put into use by the customer / problem owner. In CRISP-DM, deployment can mean different things, e.g. a simple report or a deploying a live and repeatable data mining tool across the business organisation [1].

In Table 4.1 we provide an overview of the generic tasks of each phase.

Business understanding	**Data understanding**	**Data preparation**	**Modelling**	**Evaluation**	**Deployment**
Business objectives	Collect initial data	Select data	Select modelling technique	Evaluate results	Plan deployment
Assess situation	Describe data	Clean data	Generate test design	Review process	Plan monitoring and maintenance
Data mining goals	Explore data	Construct data	Build model	Determine next steps	Produce final report
Project plan	Verify data quality	Integrate data	Assess model		Review project
		Format data			

Table 4.1: Phases and generic tasks of CRISP-DM [1]

Guideline for mapping the generic model to specific As mentioned earlier, the authors of CRISP-DM provide practitioners with guidelines and step-by-step instructions on how to tailor it to a specific organisational context. The guideline is as follows: First, analyse the specific data mining context in four different dimensions: application domain, problem type, technical aspect, and tools & techniques. Next, remove and add details depending on their applicability to your context. Thereafter, specialise the generic contests (e.g. tasks, outputs) according to the characteristics of your context. Last, if applicable rename some contents for clarity and to convey meaning [1].

4.2 Approach for constructing a theoretical methodology

In this work we have studied literature on anomaly detection. The literature comprises technical, organisational, and usability issues that present a challenge when deploying such techniques in business organisations. In addition, we have identified several success factors and best practices from the literature.

The literature study covers topics beyond the statistical/machine learning parts of anomaly detection in practice, e.g. topics related organisational and usability. To ensure that the theoretical methodology reflects the span of the literature review, an existing framework was chosen to serve as the foundation for the theoretical methodology.

We choose CRISP-DM for three main reasons. First, because it is widely used in practice and in academic publications. Second, it is comprehensive. It covers more

phases of the life-cycle of data mining projects (e.g. business understanding and deployment) compared similar models [62]. This makes it a good match since this work emphasises those same phases. Third, CRISP-DM was created with the intention that it is tailored to specific situations and scenarios, and the authors even provide general guidelines for doing so.

4.2.1 Propositions

In qualitative research, the use of *propositions* [31] can be equated to *working hypotheses* of quantitative research [61]. As the data collected in this work will mainly come from qualitative interviews we term propositions to describe the building blocks of the theoretical methodology.

The propositions are a provisional set of statements supported by literature. During the case study data is collected that is then linked back to the propositions during the analysis phase. New insight comes from analysis of the data collected, e.g. statements that support or contradict the propositions, or provide additional perspectives.

Defining specific propositions are useful in qualitative case studies for three main reasons. First, they increase the likelihood a researcher can define a limited scope for the case study. Second, as a result of the limited scope, specific propositions make the study more feasible to complete Third, revisiting the propositions makes the analysis phase more reasonable in scope and helps the researcher prioritising which data to analyse [31].

In this work, the theoretical methodology comprises a set of propositions based on the literature study (chapter 3) and the generic CRISP-DM. The specific propositions of this work come in the form best practices, recommendations, guidelines, success factors, common challenges or pitfalls from scientific literature that are relevant for business organisations deploying anomaly detection.

4.2.2 Mapping theory to a generic methodology

The main purpose of the theoretical methodology in this work is to guide the process of conducting this research and discovering new insight. Furthermore, it provides structure and helps extract the main ideas from the literature review. It is general, i.e. it is not tailored to specific types of attacks, data, tools, or organisations. Consequently, we do not follow the steps for tailoring CRISP-DM to a full extent as they are presented in section 4.1. Instead, we take an approach that we deem more suitable for this work.

CRISP-DM consists of phases and generic tasks (Table 4.1) that provide a framework guiding the process of extracting key issues from the literature review. We divide this framework into three pairs ('bins') of phases: First, *Business understanding* and *Data understanding*. Second, *Data preparation* and *Modelling*. Third, *Evaluation* and *Deployment*. Pairing reduces unnecessary complexity as some phases are iterative (e.g. the second bin) and/or the same parts of the literature review are most relevant (e.g. usability (section 3.4) for the third bin).

We define the propositions by examining the generic phases and tasks of each bin and identify relevant issues from literature. For instance, the first two phases include

the tasks of assessing the organisational situation and collecting initial data. Since we have identified issues related to those tasks in the literature, we will define a proposition around them. Moreover, we define an equal number of proposition for each bin so as to have a balanced focus on the different parts of the life-cycle of anomaly detection projects.

4.3 A theoretical anomaly detection methodology for business organisations

In this section we present the theoretical methodology that guides the process of this research. It is based on the framework of a generic data mining methodology and tailored to anomaly detection in the form of several propositions based on scientific literature. The propositions are statements on practices and issues that are relevant for business organisations deploying anomaly detection according to scientific literature.

4.3.1 Phases 1-2: Business understanding and Data understanding

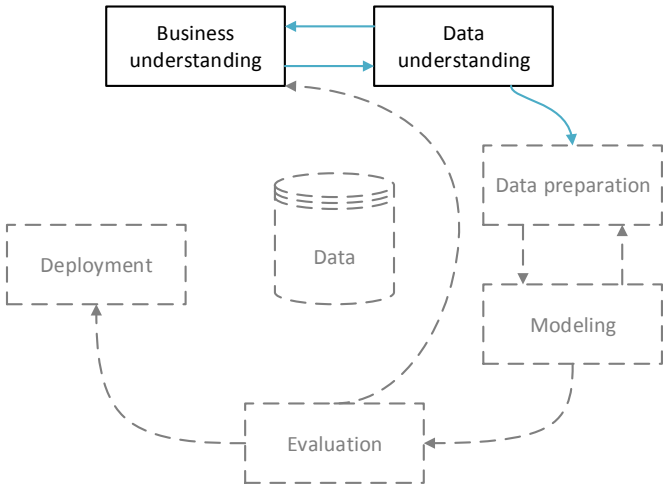

Figure 4.2: Phases 1-2: Business understanding and Data understanding

P1 Business organisations are motivated to deploy anomaly detection for detecting new/unknown attacks and for automatic refinement of detection rules.

Literature: [24, 25, 33, 34].

Rationale: *Anomaly detection has the potential to detect attacks without the prior information required for conventional detection rules. Furthermore, the process of defining and refining detection rules for conventional rule-based systems is costly*

in terms of labour and time. The knowledge gained from anomaly detection can make that process more effective and efficient.

P2 It is important that the problem owner defines clear objectives, requirements, constraints, and success criteria for the anomaly detector in early phases.

Literature: [1, 10, 26, 47, 48]

Rationale*: These criteria impact the whole life-cycle of anomaly detection projects, e.g. preparation, execution, evaluation and deployment. Moreover, when clearly defined the detector is more likely to achieve tolerable false positive rates.*

P3 Anomaly detectors should have a narrow scope, e.g. around a specific attack or malicious activity. A small scope is a success factor for reduced false positive rate and increased likelihood of project success.

Literature: [10, 47]

Rationale*: With a narrow scope it is feasible to effectively argue why anomaly detection is a good tool and how different attributes in the data relate to detecting the target activity.*

P4 It is important to take inventory of technical and organisational resources in the early stages. E.g. ensure that data is available and of good quality, arrange access to domain experts in advance and verify that the project is sufficiently supported within the organisation.

Literature: [10, 12, 26, 27, 28, 29, 47, 48]

Rationale*: A common challenge in anomaly detection is getting the necessary access to data (e.g. due to technical and privacy reasons) and documentation about the data. Furthermore, anomaly detection requires substantial support from within the organisation in the form of management support and access to experts with the domain knowledge to explain the data and evaluate the results.*

P5 It is essential to use data that is representative of the traffic of an organisation's networks and systems, e.g. real-life operational data. It is difficult to simulate realistic activity. The results of anomaly detectors on simulated data are not a good indicator of operational results.

Literature: [10, 14, 46]

Rationale*: Simulated data is often used due to availability and legal/privacy reasons. The most of the popular and freely available (simulated) datasets for intrusion detection research lack the variability of real-life activity in organisational system/network data. False positive rates of anomaly detectors can go from close to 0 to a 100% just by increasing the variability in the simulated data.*

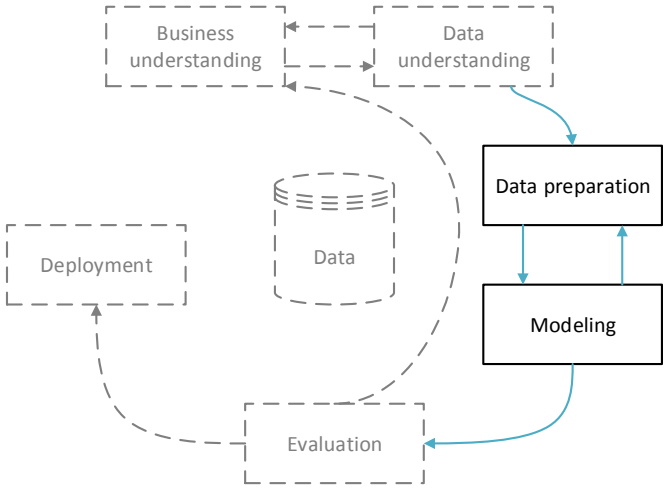

Figure 4.3: Phases 3-4: Data preparation and Modelling

4.3.2 Phases 3-4: Data preparation and Modelling

P6 It is important to reduce the variability and dimensionality of the data. For instance, by filtering and aggregating it as much as possible around the scope of the anomaly detector.

Literature: [10, 14, 46]

Rationale: *The high variability and dimensionality of organisational system- and network activity is the cause of anomaly detectors generating an unusable amount of false positives.*

P7 It is important to keep in mind that attacks are not necessarily anomalous and/or rare. Furthermore, anomalies are not necessarily malicious or interesting. The goal is to find specific malicious activity, not statistical anomalies.

Literature: [5, 10, 14, 41, 42, 43, 44, 45]

Rationale: *In practice anomaly detection often generates high rates of false positives. The main data mining success criteria for anomaly detection is to find malicious attacks or security-related events, i.e. the end goal is not to find statistical anomalies (rare and different from normal activity). Advanced attackers may try to evade detection by deliberately mimicking normal behaviour. Attacks are not necessarily rare (e.g. examples of more scans than normal connections). Some studies find many anomalies, but none that are malicious or interesting.*

P8 When selecting an algorithm or tool, it is essential to consider the interpretability of its output. It is challenging to transform detected anomalies into actionable alerts.

Literature: [10, 14, 30]

Rationale*: Generally, it is harder to verify the output of anomaly-based systems than of rule-based ones since the alerts from the former provide less useful information than the latter. Evaluate how easy it is for a human to understand an anomaly detector's decision whether an event/pattern is normal or anomalous.*

P9 Clearly argue why anomaly detection is suitable for a particular problem or for detecting certain activities, and how the features in the data (selected or extracted) are significant for detecting the activity.

Literature: [10, 14]

Rationale*: Anomaly detection is not necessarily a suitable or efficient solution for every detection problem. For example, to overcome high variability, anomaly detectors sometimes have to use data that is so highly aggregated that simple (rule-based) thresholds work equally good.*

P10 While supervised machine methods cannot be used on unlabelled data, they can be used to post-process the alerts to reduce false positives and/or provide additional information that accelerates verification of alerts.

Literature: [10, 12, 14, 36]

Rationale*: Machine learning performs better when classifying into known groups with labelled data from each group (supervised learning). While these methods are infeasible for detecting anomalies in operational data, they can be used to facilitate the inspection of anomalies, e.g. by extracting interpretable rules for why an event is an considered an anomaly, or asses how similar a new anomaly is to previous false positives.*

4.3.3 Phases 5-6: Evaluation and Deployment

P11 A low false positive rate is essential for operational usability of anomaly detectors. As a result, organisations should prioritise a low false positive rate over a high detection rate.

Literature: [59, 60]

Rationale*: False positives are costly because evaluating alerts generated by an anomaly detector takes a long time and is performed by expensive and busy experts.*

P12 It is important to achieve a very low rate of false positives when working with large amounts of data. Even a small false positives rate (>1%) can result in an unusable anomaly detector.

Literature: [10, 14, 30]

Rationale*: For example, with enough data a relatively good false alarm rate of 1% can result in thousands of alerts. Examples of acceptable false alarm rates from literature: 1 false alarm for every 100.000 events and 10 false alarms every day.*

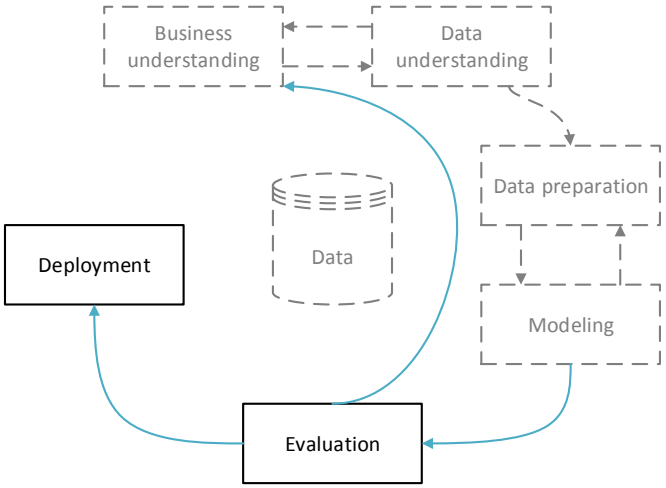

Figure 4.4: Phase 5-6: Evaluation and Deployment

P13 It is essential that an anomaly detector generates actionable and interpretable alerts. Evaluating alerts is time consuming and is done by expensive and busy experts.

Literature: [10, 14, 30]

Rationale: *An alert is interpretable when its message is clear, simple, and provides an indication of the severity of the alert. Moreover, it should enable the user to retrieve additional information in an easy way. Furthermore, different stakeholders have different criteria for actionable alerts/output. Therefore, the system should be flexible in presentation of results, e.g. weekly reports for managers, visualisations for analysts.*

P14 It is important to address the issue of updating operational anomaly detectors as the notion of normal traffic tends to change over time.

Literature: [14]

Rationale: *Operational anomaly detectors have to be updated as organisational system and network data is highly variable, changes frequently in both the short and long term. Hence, knowing how and when to update the anomaly detectors is important.*

P15 Organisations should recognise the value understanding what makes an alert either a true- and false positive. This knowledge can be used to improve detectors or construct detection rules.

Literature: [10]

Rationale: *An organisation should have a good process for handling alerts, learn from them, use them to tune the detectors or construct simple detection rules.*

An examination of true positives can produce valuable knowledge for new anomaly detectors or detection rules. False positives should be used to improve the anomaly detector. Furthermore, a collection of data labelled as true and false positives can be used as training data for future supervised anomaly detectors.

4.4 Conceptual framework

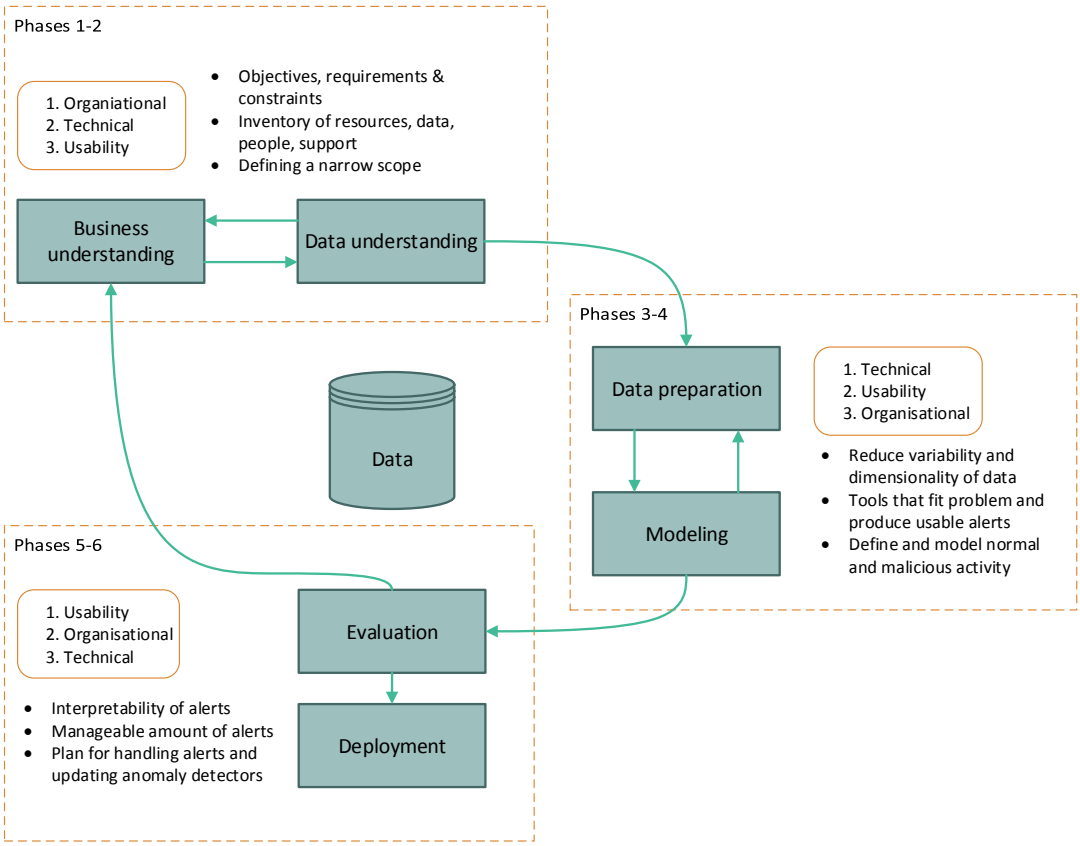

Figure 4.5: The conceptual framework of this research

The conceptual framework [63] in Figure 4.5 is a summary of the theoretical methodology. It illustrates the relation between the three bins of phases of the methodology, the importance of different topics of the literature survey, and propositions.

For case studies, a conceptual framework serves several purposes. First, it helps explain the scope of the study. Second, describes the relationships between theory, logic, and experience. Third, helps the researcher organising the contents of the study into 'intellectual bins' [64].

The 'bins' of this work consists of three pairs of phases of CRISP-DM. To further

explain the relationship with literature we indicate which part[1] of the literature study is most relevant for what pair of phases. For example, there are many organisational issues that may arise and need to be addressed during early phases of anomaly detection projects, e.g. ensuring access to data and experts. In later phases, evaluation and deployment, the literature survey suggests that usability issues are most important, e.g. that a low false alarm rate is vital for operational success. Furthermore, we list bullet points with summaries of the propositions from section 4.3 for each bin.

4.5 Summary

In this chapter we have answered **RQ 2** by constructing a theoretical methodology (section 4.3) based on CRISP-DM framework. The framework helps us build a methodology that covers the whole life-cycle of organisational anomaly detection project, just as CRISP-DM is designed to do for data mining. We divide the phases of the framework into three 'bins' each holds five key propositions from the literature.

The theoretical methodology is revisited in the next chapter. Firstly, it guides the case study of practices in business organisations. Secondly, one of the core findings of this work is the comparison of the results from the case studies with the theoretical methodology.

[1] technical (3.2), organisational (3.3), or usability (3.4)

Chapter 5

Exploration of anomaly detection in practice

In this chapter we describe the case studies performed as a part of this research and answer **RQ 4**. Firstly, we provide a general description of the organisations that participated in the case study. Secondly, we describe the methodology of the case studies, conducting interviews, collection and analysis of empirical data. Thirdly, we link the empirical data back to the propositions from section 4.3 and examine how well they are supported in practice. Fourthly, we identify issues missed or underestimated by the theoretical methodology. Lastly, we discuss some of the stakeholder complexities observed during the case studies.

5.1 Introduction to the case study

In this section we describe the business organisations that participated in the case studies. We provide a general description of their background and motivation for doing anomaly detection. In addition, we describe the case study process for each organisation as the two organisations had a different degree of involvement in this work.

5.1.1 A financial institution (FI)

The 'cyber security team' of a financial institution based in the Europe (henceforth referred to under the alias **FI**) are in the early phases of building a threat management system that includes the use anomaly detection to detect cyber attacks. The main motivator of the project stems from the fact that the cyber security team frequently receives indicators of compromise (IoCs) from various internal and external sources, e.g. law enforcement agencies and computer security companies.

IoCs usually comprise a list of domain names or IP-addresses that are considered malicious. There are several reasons why it is challenging to follow up and investigate these IoCs in large organisations. First, the format of an IoC varies between sources, e.g. threat intelligence feeds, e-mails, phone calls. Second, in order to investigate they need to

match the IoC with internal data that is geographically and organisationally distributed. Third, data collection policies may differ between locations, e.g. some collect and store data for months, while others only store samples of the data for a short time. Fourth, IoC notifications are dynamic, meaning that a list may be updated while investigation is at various stages (e.g. pending, ongoing, completed) at different locations.

The threat management system is being built to overcome these challenges and has a few broad objectives: to collect, structure, and correlate IoCs and internal network data, use statistical/machine learning tools for anomaly detection on the network data, and provide visualisation of network activity. In brief, match IoCs with actual data, do anomaly detection, and visualisation.

This project and the success criteria for anomaly detection are not defined around detecting a specific type of attack. Instead, the team is interested in deploying a number of anomaly detection models to the platform to detect different threats.

The involvement of FI is the more extensive one of the two case studies. Eight formal and individual interviews were conducted as well as several informal discussions.

5.1.2 Internet Service Provider (NET)

NET (alias) is a ISP for educational institutions based in Europe. Essentially, their goal is to protect two networks: their internal (office) environment, and the external network where they provide for their customers with internet services.

For ISPs it is important to know what is happening on their networks, and for NET, anomaly detection is one of the tools that helps them distinguish between good and bad traffic. Throughout the years several anomaly detection tools have been deployed at NET with different levels of success.

For their internal network they are interested in using anomaly detection to find compromised hosts on the network, e.g. user infected via a spear-phising e-mail. As for the external, customer network DDoS attacks are the biggest threats to NET's infrastructure and customers overwhelmed with traffic experience a less safe and less reliable network.

The case study at NET was executed in a single interview with two experts involved in anomaly detection for both their internal and external network.

5.2 Methodology for exploring anomaly detection in practice

5.2.1 Case studies

In this study, case studies are the means to explore and understand anomaly detection in practice. Moreover, the propositions and theoretical methodology (chapter 4) based on the literature review (chapter 3) guide the work performed in the case studies.

Yin [31] states that a case study should be considered when four conditions are met. First, when the case study revolves around answering 'how' and 'why' questions.

Second, when the researcher in the case study can not manipulate the behaviour of the participants. Third, when the researcher wants to study contextual conditions (business organisations) that he/she believes are relevant to the subject being studied (anomaly detection). Fourth, when the boundary between the context and subject are not clear [31].

In this work all four conditions are met. We look at how business organisations have approached anomaly detection in the past. Moreover, one of the main motivation for this research is the gap between research and practice. Hence, understanding the context of doing anomaly detection in a business organisation is a key part of this research.

More specifically, the purpose of the case study is to understand the implicit or explicit approaches business organisations follow, their experiences and learning from doing anomaly detection (**RQ 3**). Ultimately, this understanding is then used to compare anomaly detection in practice and theory.

5.2.2 Collecting empirical data

During the case studies we gather empirical data by doing interviews with professionals and managers from organisations that have deployed anomaly detection tools to detect cyber attacks. Considering the exploratory nature of this research the interviews are conducted in a semi-structured way.

Semi-structured interviews are guided by pre-defined topics that must be covered during the interview. The purpose of the interviews is to learn and understand anomaly detection from the perspective of professionals of business organisations. The style of the interviews is conversational, questions are standardised and the interviewee may use 'probes' to steer the interview to a certain topic [65].

In this work, all questions are open in the sense that the respondents are not restricted to predefined answers. Most questions are general, enabling the interviewees to discuss their experiences from their own point of view.

Anomaly detection projects within organisations will likely involve people of different roles and backgrounds. With the aim of emphasising topics that fit the interviewees' roles, several interview questions focus specifically on one of the main topics[1] of the literature review that we divide into the following categories: The *data*-, *management*-, and *security*-domains.

For each interviewee we assign them to one or more categories. For instance, a person who is working with creating anomaly detection models, collecting data, or building an anomaly detection platform is asked the data domain questions. A respondent who is a manager of the team, project manager, product owner, or communicates with outside stakeholders is asked the management domain questions. Lastly, people who are security experts, work with alerts, do incident investigations are asked the security domain questions.

Furthermore, Figure 5.1 shows how these domain specific questions relate to the different phases of theoretical methodology. Each set of domain questions directly

[1] technical (3.2), organisational (3.3), or usability (3.4)

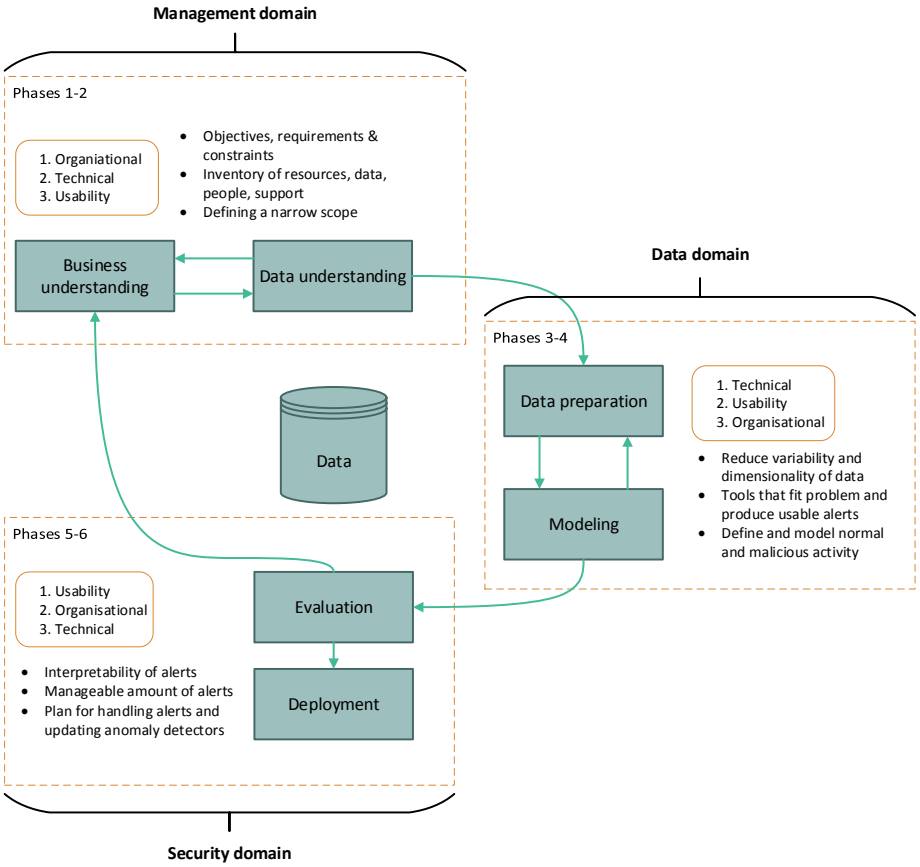

Figure 5.1: Conceptual framework and domain experts

contributes to one of the three bins of phases in the methodology. For instance, the security-domain questions emphasise usability issues, and the questions belonging to the data-domain are mainly focused on the issues related to defining and modelling normal and malicious activity from the data.

See interview questions in Appendix A.

5.2.3 Interviewee selection

At the financial institution (FI) the interviewees were selected based on the informal discussions with the project team during the start of the case study. The goal was to identify at least more than one stakeholder for each domain (data, management, security). In total there were eight interviewees that participated in the formal interviews and few others that took part in informal discussions.

As for the ISP (NET), we approached the organisation where we described the research, main topics covered, and purpose of the interviews. The outcome was a case study executed in a single interview with two professionals, one involved in anomaly de-

Organisation	Date	Interviewee	Role
FI	2015-07-27	A	Data engineer/scientist
FI	2015-07-28	B	Security intel. officer
FI	2015-07-28	C	Manager cyber security dpt.
FI	2015-07-29	D	Security intel. officer
FI	2015-07-29	E	Security intel. officer
FI	2015-08-04	F	Product owner / Security expert
FI	2015-08-11	G	Data engineer
FI	2015-08-11	H	Data scientist
NET	2015-08-21	I	Security expert / Manager
NET	2015-08-21	J	Security expert / Manager

Table 5.1: List of interviewees

tection of their 'outside' network, and the other involved in the security of their internal network. From the initial discussions it was deemed appropriate to ask them from all three domain categories.

In Table 5.1 we list the interviewees, dates of the interviews, and the roles of the respondents related to the anomaly detection.

5.2.4 Interview process

The interviews were recorded on both audio recordings and written notes. Having recordings of the interviews proved valuable for summarising the interviews afterwards as there is less risk of interviewer mixing in his own interpretation of the interview.

5.2.5 Analysing empirical data

As expected, the output of the interviews is a collection of answers, unstructured text, where each response can cover various issues, e.g. most interviewees mentioned more than one lessons learnt from working with anomaly detection. In addition, notes from informal discussion that were collected during the course of the case studies were also analysed.

The interviews were analysed using a qualitative data analysis tool, QDA Miner lite. This tool enabled us to code and organise the data collected during the case studies. Statements related to propositions were marked as such, and frequently mentioned issues were assigned their own category. Furthermore, each coded statement included meta-information about the source of the statement, e.g. the role of the interviewee, organisation and interview question, which simplified the process of looking at contrast between groups and/or organisations.

5.3 Interview findings and propositions

In this section we present an overview of the core findings of the case studies by linking the data from the case studies to the theoretical methodology (propositions) defined in section 4.3. In chapter 6 we further discuss important topics, the propositions and additional insights gained from the data collected during the case studies.

For each proposition we give an indication of whether the data collected during the case studies support (+), strongly support (+ +), contradict (−), or strongly contradict (− −) the propositions, or whether the interviews gave an alternative perspective ($\star$) or no answer/data (n.a.) was provided.

5.3.1 Phases 1-2: Business understanding and Data understanding

Proposition	Description	FI	NET
P1	Business organisations are motivated to deploy anomaly detection for detecting new/unknown attacks and for automatic refinement of detection rules.	+ / $\star$	+ / $\star$
P2	It is important that the problem owner defines clear objectives, requirements, constraints, and success criteria for the anomaly detector in early phases.	− / $\star$	− −
P3	Anomaly detectors should have a narrow scope, e.g. around a specific attack or malicious activity. A small scope is a success factor for reduced false positive rate and increased likelihood of project success.	+	+ +
P4	It is important to take inventory of technical and organisational resources in the early stages. E.g. ensure that data is available and of good quality, arrange access to domain experts in advance and verify that the project is sufficiently supported within the organisation.	+ +	$\star$
P5	It is essential to use data that is representative of the traffic of an organisation's networks and systems, e.g. real-life operational data. It is difficult to simulate realistic activity. The results of anomaly detectors on simulated data are not a good indicator of operational results.	+ +	+

Table 5.2: Propositions and case study (Business- and Data understanding)

P1 Business organisations are motivated to deploy anomaly detection for detecting new/unknown attacks and for automatic refinement of detection rules.
Partially supported

From the case studies we see that organisations are motivated by the detection capabilities that anomaly detection promises (detection of unknown/undefined attacks).

However, there was no explicit mentioning of a motivation being to use anomaly detection to make the refinement of detection rules more efficient. While not a motivating factor, both organisations have (or intend to) create detection rules based on the output of anomaly detector, e.g. to use in real-time detection systems.

An important motivation factor for security experts of both organisations is that anomaly detection will increase their understanding of what happens on their networks. Furthermore, both FI and NET both mention their organisational situation. As a financial organisation, FI states it is a likely target of attackers and welcomes any additional tool for detecting cyber attacks. As an ISP, it is important for NET to provide their customers with a safe and reliable network.

P2 It is important that the problem owner defines clear objectives, requirements, constraints, and success criteria for the anomaly detector in early phases.
Not supported

Anomaly detection is a new tool for most organisations, most techniques and information comes from academic research, the anomaly detection market is not mature, and there are few commercial solutions available. From the case studies we understand that it may not be realistic to expect problem owners to set clear and specific criteria for the anomaly detectors, in particular in an early stage. An example of specific criteria would be requiring a 0.1% false positive rate before starting the data preparation and modelling.

At FI, general high-level criteria, objectives and requirements were defined in an early stage. At NET, the deployment of anomaly detection does not happen within a clearly defined project, rather as a part of their day-to-day operations.

We see that business organisations want the empowering capabilities that anomaly detection tools promise. In addition, they want to use the tools to increase their understanding of what happens on their networks and systems. For both organisations, these criteria are general in the begin, based on broad goals, that will evolve into specific ones as deployment progresses. In brief, the case study organisations know what capabilities they want, but the problem and solutions are not known well enough to be able to define specific criteria in early stages.

P3 Anomaly detectors should have a narrow scope, e.g. around a specific attack or malicious activity. A small scope is a success factor for reduced false positive rate and increased likelihood of project success.
Supported

The case study at NET revealed a strong support for having a narrow scope when doing anomaly detection. The only way to detect anomalies is when there is a definition of what is normal. In reality this definition varies greatly between different parts of large organisational networks. For example, with NetFlow data, the only way they have

been able to establish a reliable model of normal activity (baseline) is by individually analysing small segments of the network activity (e.g. combining specific ports, protocols, and clients). As a result, the anomaly detectors they have in place look at small subsets of their network activity at a time, trying to find specific anomalies of specific activities (e.g. DDoS attacks).

The data scientists of FI are looking at research papers that revolve around detecting specific type of activity (e.g. scans, anomalous paths in network), which we consider as evidence supporting the proposition. However, they did not discuss real-life experiences that directly support the proposition, so whether they are defining a narrow enough scope remains to be seen. For instance, there was no mention of defining small subsets of activity like at NET.

> **P4** It is important to take inventory of technical and organisational resources in the early stages. E.g. ensure that data is available and of good quality, arrange access to domain experts in advance and verify that the project is sufficiently supported within the organisation.
> **Supported**

From the case studies, the most critical part of the proposition has to do with ensuring and getting access to data, i.e. without data the anomaly detector cannot detect anything. Thus, it is essential to know where a project stands in terms of getting data from within the organisation.

Interestingly, this was not an issue for NET, while it was often mentioned as the biggest challenge for the anomaly detection project at FI.

NET collects sampled NetFlow (1/100) from their core routers, with plans to switch to unsampled flows (collect all flows) in the coming months. From the case study this seemed like a simple task while they saw some challenges ahead regarding their capacity to store and analyse unsampled flows. Obviously, it is important for them to ensure that they have access to data, but in their case it was not a challenge. This is likely due to the fact that the people who are responsible for deploying the anomaly detectors also 'own' the data.

For FI, this was a technical and an organisational challenge that some of the interviewees felt they had underestimated. First, the data is 'owned' by other departments that may need convincing to start collecting and delivering the data. Second, these departments may not have the technological expertise or tools to collect and deliver data. Third, legal and privacy laws on a local, national and international level make it difficult to collect some data. Fourth, once the data is delivered it does not come with any quality assurance, and monitoring quality is a challenge on its own. Fifth, it is challenging to be context-aware in such a large setting, with data from so many sources.

While getting data is a challenge, the team at FI has already secured data from many departments. Doing so have had to approach and 'convince' different different departments to start collecting the data. Furthermore, they support these departments with advice on hardware, best practices, or with funding for building the capabilities to

collect the data that is needed. Nevertheless, they had to exclude some data sources they were interested in using because system generating it were not built to export and deliver data.

As for organisational support and access to experts, both organisations seem to agree with the proposition. Support from within the organisation, and upper management, is neither static nor everlasting. Instead, it is dynamic and has to be maintained throughout the life-cycle of an anomaly detection project/deployment, e.g. by delivering valuable results. From literature and case studies it is apparent that anomaly detection is unlikely to immediately produce usable results, thus it is important that there is understanding that such projects/deployments need long-term support. Moreover, people outside the project team sometimes make decisions that directly affect how the team can work with the anomaly detector, e.g. by restricting access to data. As we observed at FI, the project has had good support and understanding within the organisation and from upper management. However, they are aware of the fact it can change. To maintain support and understanding they involve managers and relevant stakeholders, and make the effort of increasing their understanding of the project and its requirements. For instance, it is important that there is an understanding of why there is a need for the data scientists of having a seemingly extraordinary access to live data.

P5 It is essential to use data that is representative of the traffic of an organisation's networks and systems, e.g. real-life operational data. It is difficult to simulate realistic activity. The results of anomaly detectors on simulated data are not a good indicator of operational results.
Supported

Both organisations exclusively use operational data to build and test their anomaly detectors, i.e. models are not developed on simulated data. From working on fraud detection the team at FI their experience in fraud detection the experts at FI have learnt that simulated data cannot be used build a model that will distinguish between normal and anomalous activity.

In one discussion a data expert at FI mentioned the possibility that when moving the anomaly detection system to an operational phase (with more users allowed access to the system) may introduce stricter limitations to access of data. Restricting access to data may present new challenges for developers and data scientists, e.g. if models have to be built on simulated data, but are tested on live data in a way that is inaccessible to people.

5.3.2 Phases 3-4: Data preparation and Modelling

Proposition	Description	FI	NET
P6	It is important to reduce the variability and dimensionality of the data. For instance, by filtering and aggregating it as much as possible around the scope of the anomaly detector.	−	+ +
P7	It is important to keep in mind that attacks are not necessarily anomalous and/or rare. Furthermore, anomalies are not necessarily malicious or interesting. The goal is to find specific malicious activity, not statistical anomalies.	+	+ +
P8	When selecting an algorithm or tool, it is essential to consider the interpretability of its output. It is challenging to transform detected anomalies into actionable alerts.	n.a.	+ +
P9	Clearly argue why anomaly detection is suitable for a particular problem or for detecting certain activities, and how the features in the data (selected or extracted) are significant for detecting the activity.	+ / −	− / ⋆
P10	While supervised machine methods cannot be used on unlabelled data, they can be used to post-process the alerts to reduce false positives and/or provide additional information that accelerates verification of alerts.	+ +	+ / −

Table 5.3: Propositions and case study (Data preparation and Modelling)

P6 It is important to reduce the variability and dimensionality of the data. For instance, by filtering and aggregating it as much as possible around the scope of the anomaly detector.
Partially supported

The literature is clear about the problems associated with the variability of organisational data. While both organisations agree that a narrow scope (**P3**) is the right approach, the case studies reveal interesting differences between the two organisations.

As discussed before, NET's approach for anomaly detection is to have specific attacks in mind and only use narrowly scoped subsets of the data each time. Moreover, it was clear that they have not been successful with anomaly detectors that apply to their whole network, nor when applied to individual customers.

FI is currently working on anomaly detectors with the goal to detect specific type of malicious activity, e.g. scans and worm infections. The tools and techniques used come from academic literature, for instance in the form of specific features that can be built from NetFlow that may indicate scans. However, they currently do not apply anomaly detection to specific subsets of data (e.g. ports, protocols).

The differences in approach for the two organisations may be explained with the fact that they are not looking for the same type of attacks. For NET, the main goal is to

detect DDoS attacks where they know of specific vulnerable ports/protocols they choose to focus on. At FI, they are looking for activity within their network that indicates advanced (and unknown) attackers or targeted attacks. It is perhaps premature to expect a clearly defined subset of data with them still being in an exploratory stage with the goal of detecting 'unknown' attacks.

P7 It is important to keep in mind that attacks are not necessarily anomalous and/or rare. Furthermore, anomalies are not necessarily malicious or interesting. The goal is to find specific malicious activity, not statistical anomalies.
Supported

The case studies indicate that both organisations are using statistical models or machine learning to find specific malicious activities, and neither assume that anomalies automatically represent attacks or other interesting behaviour.

At FI, the security experts state that because they are in an early phase with anomaly detection, they are interested in all anomalies. In their view, anomalies indicate strange activity that can help them increase their understanding of what happens on their network. During this phase they are tolerant to higher rates of false positives as any anomaly in and of itself is interesting. However, this phase is temporary and they eventually expect a manageable amount of false positive.

The contrast between early and current stages of NET's anomaly detection efforts are a good example of how the opinion of organisations on this matter may change. One of the first 'anomaly detection' models that NET deployed was a simple model that generated an alert if IP addresses were responsible for a large number of flows within the network. While this solution seemed like exactly what they needed at the time, NET quickly realised that there were many different, non-malicious, non-interesting, reasons why an IP address generates an anomalous numbers of flows. For years they tried different solutions, commercial ones and models from literature, but it proved difficult to deploy general models that generate reliable output. Now they only look at (and have only had success with) very specific scenarios, e.g. a single port that is relevant for parts of DDoS attacks.

P8 When selecting an algorithm or tool, it is essential to consider the interpretability of its output. It is challenging to transform detected anomalies into actionable alerts.
Supported

The case studies, at FI revealed that the interviewees, and especially the security experts have a clear idea of what they would define as an interpretable and actionable alert. However, possibly due to being in the early phases, the interpretability of alerts is not a big part of the selection criteria for anomaly detection tools/algorithms. The tools they

are currently looking at do help with interpretation, e.g. in the form of an anomaly score that helps with prioritising which alerts to first evaluate.

NET have a long experience of alerting their customers when NET detects anomalies indicating a DDoS attack. From the interviews we find two main examples of this challenge. Firstly, one of NET's tool built internally produces alerts that they do not consider has good usability. More specifically, the alerts only say that for a certain period of time 'something strange was happening', and which protocol was involved. As a results, the person receiving the alert has to manually collect a great amount of additional data needed to properly evaluate the alert. Secondly, it is challenging to communicate alerts to external customers. It is difficult to make these alerts good for every stakeholder. They have to be careful with the amount of text, i.e. too much text and it may not get read. Furthermore, while the security experts at NET may be able to interpret summaries of NetFlow records, these summaries may not say much to the recipients of the alerts. For both these challenges they constantly try to make improvements to increase the usability of the alerts generated.

P9 Clearly argue why anomaly detection is suitable for a particular problem or for detecting certain activities, and how the features in the data (selected or extracted) are significant for detecting the activity.
Not supported

On a general level, both organisations are motivated to deploy anomaly detection for its promising capability of detecting unknown/novel attacks. In that sense they can argue why anomaly detection is suitable for a particular problem, e.g. FI will not detect advanced targeted attacks by matching internal data with threat intelligence feeds of known attackers, however, statistical anomalies on the internal network traffic may reveal such attacks.

On a detailed level, these arguments come from academic literature, studies where anomaly detection is applied to a particular problem. Both FI and NET use academic research to guide their efforts or use commercial tools offered as a solution to their problems.

From the case studies we conclude that it is not realistic to expect organisations to have a clear and detailed argument for tool and data choices. For both organisations, in the early phases these clear and specific arguments come from external sources (research papers, commercial tools). Moreover, these efforts are driven by cyber security teams that will go through a phase of trial and error on their path of deploying these tools. In essence, both organisations trying to solve a business problem with innovative tools rather than doing scientific research.

However, as we see from the case study at NET this capability can grow as their maturity level and experience with anomaly detection increases.

P10 While supervised machine methods cannot be used on unlabelled data, they can be used to post-process the alerts to reduce false positives and/or provide additional information that accelerates verification of alerts.
Supported

Both FI and NET are solely using unsupervised machine learning methods. One interviewee at FI discussed how using unsupervised machine learning and real-life makes it challenging to model normal traffic as they do not know if they only have normal traffic in their datasets. The only way to overcome that would be to use supervised methods on their own real-life datasets that has been manually checked for normal and anomalous behaviour. However, the large scale of their datasets make that infeasible. As a result they look for academic research that propose models that work on unlabelled datasets.

As for post-processing the security experts at FI discussed the usefulness of applying automatic classification on the alerts. Firstly, this classification could be used to group alerts into categories of similar alerts. Secondly, it could be used to send the alerts to decrease false positives, presumably by comparing with previous false positives. While they are not actively developing this functionality, they do see its value. Post-processing of alerts may become a necessity as they plan to incorporate a greater number of data sources into the system, possibly resulting in too many false positives for a small team to handle.

5.3.3 Phases 5-6: Evaluation and Deployment

Proposition	Description	FI	NET
P11	A low false positive rate is essential for operational usability of anomaly detectors. As a result, organisations should prioritise a low false positive rate over a high detection rate.	+ / −	+ / ⋆
P12	It is important to achieve a very low rate of false positives when working with large amounts of data. Even a small false positives rate (>1%) can result in an unusable anomaly detector.	+ +	+ +
P13	It is essential that an anomaly detector generates actionable and interpretable alerts. Evaluating alerts is time consuming and is done by expensive and busy experts.	+ +	+ +
P14	It is important to address the issue of updating operational anomaly detectors as the notion of normal traffic tends to change over time.	n.a.	+ / −
P15	Organisations should recognise the value understanding what makes an alert either a true- and false positive. This knowledge can be used to improve detectors or construct detection rules.	+ +	+ ⋆

Table 5.4: Propositions and case study (Evaluation and Deployment)

P11 A low false positive rate is essential for operational usability of anomaly detectors. As a result, organisations should prioritise a low false positive rate over a high detection rate.
Partially supported

From the literature study we found that the high false alarm rates are a barrier for deploying a usable anomaly detector within business organisations. Looking at this issue in the case studies showed that the two organisations had a more relaxed view on false positives.

At FI, most acknowledge that there will be a large number of false positives, particularly in the beginning of deployment. The vision is that when an anomaly detector is deployed an improvement and learning process begins. Alerts generated that prove to be false positives will further the experts' understanding of the network. This understanding is then used to explain why an event is a false positive and should not be detected again. More specifically, this process of tuning the models, and understanding the false positives brings valuable knowledge. However, this process should eventually lead to a model that is usable. As a result, false alarms are not a great concern for them in the early phases.

Anomaly detectors that require a lot of time and resources to tune will eventually be abandoned as we saw at NET. From their experience, having a good (combination

of) tools to analyse the alerts and find out what is happening is essential. For example, if these tools can reduce the time it takes to investigate some alerts from 2 hours to 2-3 minutes, like in the case of NET, the cost of a false alarm is reduced.

In summary, the experts do agree with the fact that high rates of false positives is a problem. However, there are many ways to reduce the impact of them, e.g. having the tools to quickly investigate. Moreover, the insights a false alert brings can be of value for the security experts as it enhances their understanding of what type of (strange) activity takes place on their networks. More specifically, for the security experts it is a priority to better know their networks.

P12 It is important to achieve a very low rate of false positives when working with large amounts of data. Even a small false positives rate (>1%) can result in an unusable anomaly detector.
Supported

Reflecting their experience with both anomaly-based and rule-based detection systems the interviewees recognise that a manageable number of false positives is key to the operational success of detection tools.

As discussed before, the experts understand that a recently deployed anomaly detector is likely to produce a high rate of false positives in the beginning. However, they expect that after a reasonable period of tuning the model the false positive rate will reach usable levels. The expectation of such an improvement process is in line with their experience with rule-based detection systems, e.g. working with firewall rules.

This tuning process is not always successful as is clear from the case study at NET. They have abandoned models embedded in commercial solutions and models coming from academic literature due to difficulties with tuning. For example, they tried to tune one model from academic literature (Holt-Winters) for more than a year before abandoning it due to difficulties with getting reliable results. Furthermore, the problems with tuning models of some commercial systems is that it is only enabled to a limited degree, e.g. a few parameters can be changed but the underlying model is hidden.

While both organisations fully agree with the proposition it is challenging for them to reduce the false positive rate of an anomaly detector in operation.

P13 It is essential that an anomaly detector generates actionable and interpretable alerts. Evaluating alerts is time consuming and is done by expensive and busy experts.
Supported

Based on the case studies we find that interpretable and actionable alerts are essential for the operational success of anomaly detectors.

At FI the team and especially the security experts have clear usability criteria for alerts generated by anomaly detectors. From their experience they know that it takes

a long time to investigate alerts, data has to be collected from various sources before it becomes clear whether they have a true or false positive. The more intepretable and actionable it is, the less time it takes to investigate. Simply put, they want the alerts to automatically answer as many 'w'-questions (who, what, where, when, why) as possible. For instance, it should be clear why an alert is being raised, what entities (e.g. IP addresses) are involved, and some indication of severity. While these usability criteria are there from the perspective of the security experts, they are not used as 'requirements' for the work of the data scientists building the anomaly detectors. At the moment their main concern is developing working models that detect relevant, interesting (and/or malicious) anomalies.

The same goes for NET who both send alerts to customers and receive alerts that they have to investigate. Actionable alerts, that indicate IP addresses and provide good means to start investigating (e.g. a hyperlink to visualisation of the event) are important for reducing investigation time. As for the alerts they send to customers they have had to deal with challenges of producing interpretable alerts. For example, they sometimes send alerts identifying potential victims of DDoS attacks, with large amounts of traffic coming through them. However, some of their customers (instinctively) block the IP address because most of NET's alerts identify the attackers that need to be blocked.

Both organisations seem to follow a similar way of working with usability. They clearly understand its importance, but usability is not of high priority early on, at least not in a way that directly impacts the construction of anomaly detection tools/models. When they reach a stage of sending alerts, they start to emphasise and put more effort into making the alerts usable. For instance, NET are continuously working to improve the interpretability of their alerting, and FI intend to learn that by looking at what is frequently requested information with an alert.

P14 It is important to address the issue of updating operational anomaly detectors as the notion of normal traffic tends to change over time.
Not supported

The case studies did not reveal any decisive statements of support for this proposition. The team at FI have not reached the stage where these issues are addressed. As for NET, without access to the data and tools at the time it is difficult to say whether the long time spent on tuning models was only due to the poor performance of the algorithms or whether what should be considered as normal was changing during the same time. Hence, this proposition is not supported as neither organisation has consciously dealt with this problem.

More importantly, it is unclear whether the organisations would be able to determine whether false positive rates are caused by a model that needs to be updated because 'normal' is changing, or by a model that performs poorly or does not fit the problem. Moreover, before deployment it may prove difficult to decide a good frequency for updating an anomaly detector.

P15 Organisations should recognise the value understanding what makes an alert either a true- and false positive. This knowledge can be used to improve detectors or construct detection rules.
Supported

The case organisations recognise the value of a careful examination and understanding of both true and false alarms.

The team at FI intend to research detected anomalies that are either malicious or interesting and turn them (where possible) into rules. Furthermore, they want to understand why an alert is a false positive, e.g. by looking whether the model is overemphasising some features in the data, or whether the model is missing key contextual data.

In a sense, NET has used false positives to develop detection rules. To elaborate, NET is using a system that uses algorithms to generate rules to mitigate attacks. While the rules that the system generates are reportedly 'clever', they are often too specific. In some cases, these very specific rules would fail to detect even a slightly different variation of the same attack. As a results, NET develops more general rules based on the specific ones.

5.4 Incompleteness of the theoretical methodology

During the case studies we conducted semi-structured interviews with open questions. The questions purposely covered the broad contents of the theoretical methodology and allowed the freedom of discussing other issues relevant for business organisations. In this section we outline some of the issues observed from practice that were either understated or overlooked in the the propositions. In short, these are issues missing from the methodology or more important in practice. With regard to **RQ 4** this section looks at the *completeness* of the theoretical methodology while in section 5.3 we examined its *soundness*.

In the following subsections we list (in no particular order) and summarise some of these issues. Furthermore, we present relevant (and anonymous) quotes from interviews that capture the issues discussed. Subsequently, we elaborate on them in chapter 6.

Data governance

> *"It is hard to get the data, everything is hard about it, technically and organisationally speaking."*

In one proposition (**P4**) of the theoretical methodology we mention the importance of securing access to data and other resources from within the organisation. Considering the insights from the case studies it can be concluded that the theoretical methodology understates this issue.

Even when a security department is allowed to use the data, getting the data is not as simple as showing up and asking for it. It can still be an immense technical and

organisational challenge to actually retrieve the data. For instance, the data may be owned by other departments that have neither the interest nor capability to collect and deliver network logs.

Once the data is collected it is has to be made accessible to the data scientist. Furthermore, the raw data is not always enough, metadata describing the data is also needed, e.g. a network/asset model that describes the topology of a particular network segment, and helps identify the actors on the network. Again, this information may belong to other departments that have to agree to deliver this meta-information.

There are many other challenges related to data, including the storage of large amounts of data, managing performance simultaneously for data scientists and security experts, building a system that is compliant to an organisation's security and privacy policy, annotating the data with contextual information, and dealing with poor data quality.

In summary, the theoretical methodology does not emphasise the issues of getting and working with data with the same weight as these issues appear in the case studies. The problems with data may arise at different times from different directions, but all directly affect the anomaly detector.

Laws and regulations

> *"There are laws in some countries that prohibit logs from leaving the country."*

In the theoretical methodology, the only challenge related to laws and regulation has to do with whether it is allowed to use live data (**P5**), i.e. simulated data is often used for legal and privacy reason.

In practice, these issues have a wide impact and can introduce complexities to an anomaly detection project. Firstly, everything from departmental policies to international law has to be considered. Secondly, data collection and transfer is subject to many laws. For instance, at FI we saw that banking secrecy laws of one country dictates whether a certain data source is available, and data protection laws in another country forbids logs ever leaving the country. Thirdly, laws regarding privacy are not always clear, and subject to interpretation, meaning lawyers can interpret the same law differently. Fourthly, a deployed and operational anomaly detector has to be in compliance with company regulations. Reaching the point of compliance can be costly in terms of time spent doing the necessary documentation and analysis. Furthermore, as observed at FI, requirements and constraints for these types of projects often come from people working with risk management, compliance or legal departments.

Understanding the network

> *"We do basic monitoring, we know our baseline, and we know our network. Without this knowledge, and without tools for doing and checking your baseline you cannot effectively do anomaly detection."*

The propositions of the theoretical methodology are focused on the end goal, deploying an operational and usable anomaly detector. The methodology is about defining a clear scope and requirements, apply statistics or machine learning techniques, generate usable alerts, and deploy the technology in the organisation.

From the case studies we see that the respondents, and particularly the security experts, want to use these tools to better understand what is happening on their networks and systems, i.e. gain situational awareness. Considering the detailed problem definition and narrow scope the theoretical methodology ask of organisations we argue that it assumes this situational awareness is already at hand.

However, we see that the tools that support business organisations in developing a good understanding of the networks are being deployed simultaneously as those that allow them to do anomaly detection, e.g. Hadoop and Mapreduce frameworks, visualisation tools. Furthermore, a good understanding of the network and tools that support exploring the data helps with investigating of alerts.

In brief, the theoretical methodology does not sufficiently address the need of security experts, or those who verify alerts, to have a good understanding of what goes on in the organisation's networks and systems.

Finding and retaining talent

> *"In the current labour market it is difficult to get the right people for a project like ours."*

The theoretical methodology implicitly assumes that the business organisation already has all the experts needed to deploy anomaly detection. In reality, finding and retaining talent is a significant challenge that affects anomaly detection projects in various ways.

For instance, we saw that the case study organisations need three main types of experts: data scientists, security experts, and data engineer (Hadoop specialists, infrastructure developers). Reportedly, these are all in great demand and in short supply, making them difficult to find and retain in the current labour market.

To get the necessary expertise they have had to partially rely on external employees. Naturally, this is a temporary arrangement and in the end of the project the external employees leave and take their knowledge with them. Moreover, the organisations have invited students to deploy anomaly detection models, e.g. as a part of their graduation. Similarly, the students are likely to leave with valuable knowledge, leaving the internal employees with the difficult task of tuning the deployed models.

As for the internal employees, they gain valuable experience from working on these 'cutting edge' projects, e.g. building a threat management system and deploying anomaly detection. This increases their value on the labour market which may encourage them to leave for other opportunities. Furthermore, it may be hard to provide talent with an environment where they can enjoy their work. For example, providing good flexibility and freedom while working with such sensitive information.

The issues with talent can affect various parts of anomaly detection projects, like deciding what kind of tools to use.

The open source community and the commercial market

> *"Some models seemed promising, with many academic papers about them, but in practice they did not work."*

Another important issue missed by the theoretical methodology is the decision whether to go an open source route, a commercial one, or combining the two. This is not an easy decision, both types have their strengths, weaknesses, and accompanying requirements.

Open source tools, including the use of scientific research, are often the only choice available as the vendors of commercial solutions have not been offering many solutions in recent years. However, these tools introduce a high level uncertainty and dependency on highly skilled people into anomaly detection projects. Moreover, they often lack functionalities that are required in a corporate environment that have to be built, e.g. authentication and user management.

As for commercial tools, though the market has not offered many solutions, a few vendors and commercial anomaly detection tools have become increasingly mature and sophisticated in the recent years. The main drawback of the more sophisticated tools is their high price. However, they come with less uncertainty than the open source tools, they are likely to come with a range of functionalities organisations require, and have less demand on talent.

5.5 Stakeholder complexities

> *"Some things in this project are much harder to achieve than we thought. Those things have more to do with people."*

In subsection 2.1.1 we introduced some of the general complexities of technological projects in business organisations. For example, deploying innovative technology often calls for the involvement of different types of highly specialised experts. On top of technological complexity, managing the coordination and communication between the varying groups also increases project complexity.

Interacting with the various experts involved in anomaly detection in practice, we observed these stakeholder issues (complexities and tensions). We divide our observations into two main categories: the project teams, and innovating in a business organisation. Furthermore, we extract and present relevant quotes from the case studies in the same way as we did in the previous section.

Teams of security experts and data scientists

> *"It is difficult to manage a group like that, everybody has their own understanding, own way of working."*

At its core anomaly detection requires the bridging of two professional domains, statistic/data science (e.g. building and tuning models), and cyber security (e.g. defining goals,

evaluation and verifying alerts). From what we have seen in practice, people generally fall into one or the other of these categories, i.e. people are seldom experts in both. Thus, both data scientists and security experts need to be involved to take care of their parts of the system, but they also need to communicate and work together. As discussed in subsection 2.1.1, the stakeholder complexities come from these interactions between different sub-systems and specialists.

It is a challenge to introduce and integrate data science, anomaly-based detection, into a team of people that are used to rule-based detection/way of working.

On one hand, the team has to know when to use what type of tool, which is difficult when the experts are not familiar with the anomaly-based tool. More specifically, when people do not have good knowledge of statistics/machine learning, they do not have a good understanding of the capabilities and limitations of anomaly detection. As a result, they are less likely to identify problems and formulate them in a way that fits anomaly detection.

> *"When you start to see what anomaly detection can do, you start to formulate your questions in terms of them being handled with anomaly detection."*

On the other hand, the data scientists are most often making anomaly detection models based on scientific literature that they intend to deploy within the organisation. However, the security experts have a better understanding of the threats, network activity, and other contextual information related to the organisation. It is difficult for the security experts to know how much they should try to integrate this context and understanding into the process of making an anomaly detector in a useful way. In their view, it may be better to keep this contextual information separate from the data scientist so as to not distract him in his work.

> *"Maybe they should be more separate so that you are not distracting the data scientists with such information."*

More importantly, the two groups have to work together, agree on an approach, and divide responsibilities. To explain this we provide three relevant examples from the case studies. First, we see that the data scientists are focusing more on making advanced models that detect specific threats while the security experts are more interested in situational awareness, smaller (modular) models, and monitoring capabilities. Second, in neither case study was there a clear agreement on how to handle false positives, who should handle them and when (see more detailed discussion in section 6.3). Thirdly, how these two groups want to approach the problem is reinforced by their external professional peers, e.g. by sharing ideas with security experts at other financial institutions.

At the same time, we see the project and project management is more focused on the end goal and deploying the various sub-systems. There is less focus on 'designing' the interaction between the people belonging to different parts of the system.

In summary, the two types of specialists are needed to work on different parts of an operational anomaly detection system. However, they are also faced with the challenge

of deciding how to work together, i.e. the different types of experts need to find out how to bridge their domains.

Innovating in business organisations

> *"Similar to 'building a rocket in mid-air' but in something thick, you feel pressure from everywhere, you have to work fast, but around every corner there is a 'but' and many stakeholders intervening in the project."*

For both organisations, anomaly detection is a new way of detecting security incidents. They see the operational success of statistics and machine learning techniques in the field of business intelligence, customer intelligence, and fraud detection. However, when deploying anomaly detection, organisations are essentially doing research projects as there are few tools and models readily available, and little practical information available and shared across communities. For that reason organisations take it upon themselves to develop and build (large parts of) the anomaly detection system.

As mentioned in subsection 2.1.1, the deployment of innovative technology can clash with the way business organisations operate in projects and the interviewees mention several examples of how these affect them.

> *"I love prototyping. Try something, if it does not work, try something else. I think this way of working would be good in this project, but it is very hard to get this approach into this organisation."*

Firstly, there is friction between the exploratory and experimental way of working and standard business practice. As this is innovative technology these projects may need a long exploratory phase where the team 'dives' into the technology and tries different solution in an effort to understand the effects of different decisions. In other words, to better understand the technology the experts need room to explore and play around with it. However, his way of working is difficult to incorporate into how an organisational way of running projects, with a set budget, time constraints, and results to deliver.

> *"A lesson learned would be to externally develop the capabilities of the system, then fit it into the organisation, or change the organisation to fit the system."*

Secondly, it is challenging to make the system fit to the organisation, i.e. comply with its standards. During deployment are many internal dependencies to work with, needs for approvals, discussions and meetings (elaborated on in section 6.1). Reaching this point of compliance is costly in terms of time spent on documentation, analysis and presentations in order to get approval. In other words, costing time the experts would much rather spend on building the system.

> *"You sometimes get the feeling that these rules and guidelines have not evolved as fast as technology, but this is something we have to do."*

In summary, these are not issues that are confined to anomaly detection or innovative technology. However, they are do affect the case study organisations, and deploying a relatively unknown and cutting edge technology only exacerbates these challenges.

5.6 Summary

In this chapter we have answered **RQ 4**. Firstly, by testing the theoretical methodology by comparing it with how business organisations approach and experience anomaly detection (section 5.3, summarised in Table 5.5). Secondly, by identifying key issues with anomaly detection that are important in practice but overlooked or underestimated in the theoretical methodology (section 5.4). Thirdly, describing how general stakeholder complexities of doing technological projects surfaced in the case study organisations (section 5.5).

On the soundness of the theoretical methodology, we observe that the majority of propositions were supported by the experts of the case study organisations, i.e. organisations where following the propositions to some degree, or were aware of (and agreed with) the important issues from the literature. The ones that were not supported seemed to expect a mature experience of anomaly detection or certain way of working and thinking that did not fit the case study organisations, e.g. early definition of requirements or clear arguments for their selection of methods.

On the *completeness* of the theoretical methodology, it did miss a few crucial issues important to the case study organisations. Namely, the selection of open source or commercial tools, retention of talent in today's labour market, gaining situational awareness, and how laws and regulations can add complexities to anomaly detection projects.

Proposition	**Description**	**Support**
Business understanding & Data understanding		
P1	Business organisations are motivated to deploy anomaly detection for detecting new/unknown attacks and for automatic refinement of detection rules.	Partial
P2	It is important that the problem owner defines clear objectives, requirements, constraints, and success criteria for the anomaly detector in early phases.	No
P3	Anomaly detectors should have a narrow scope, e.g. around a specific attack or malicious activity. A small scope is a success factor for reduced false positive rate and increased likelihood of project success.	Yes

P4	It is important to take inventory of technical and organisational resources in the early stages. E.g. ensure that data is available and of good quality, arrange access to domain experts in advance and verify that the project is sufficiently supported within the organisation.	Yes
P5	It is essential to use data that is representative of the traffic of an organisation's networks and systems, e.g. real-life operational data. It is difficult to simulate realistic activity. The results of anomaly detectors on simulated data are not a good indicator of operational results.	Yes
Data preparation & Modelling		
P6	It is important to reduce the variability and dimensionality of the data. For instance, by filtering and aggregating it as much as possible around the scope of the anomaly detector.	Partial
P7	It is important to keep in mind that attacks are not necessarily anomalous and/or rare. Furthermore, anomalies are not necessarily malicious or interesting. The goal is to find specific malicious activity, not statistical anomalies.	Yes
P8	When selecting an algorithm or tool, it is essential to consider the interpretability of its output. It is challenging to transform detected anomalies into actionable alerts.	Yes
P9	Clearly argue why anomaly detection is suitable for a particular problem or for detecting certain activities, and how the features in the data (selected or extracted) are significant for detecting the activity.	No
P10	While supervised machine methods cannot be used on unlabelled data, they can be used to post-process the alerts to reduce false positives and/or provide additional information that accelerates verification of alerts.	Yes
Evaluation & Deployment		
P11	A low false positive rate is essential for operational usability of anomaly detectors. As a result, organisations should prioritise a low false positive rate over a high detection rate.	Partial
P12	It is important to achieve a very low rate of false positives when working with large amounts of data. Even a small false positives rate (>1%) can result in an unusable anomaly detector.	Yes
P13	It is essential that an anomaly detector generates actionable and interpretable alerts. Evaluating alerts is time consuming and is done by expensive and busy experts.	Yes

P14	It is important to address the issue of updating operational anomaly detectors as the notion of normal traffic tends to change over time.	No
P15	Organisations should recognise the value understanding what makes an alert either a true- and false positive. This knowledge can be used to improve detectors or construct detection rules.	Yes

Table 5.5: Summary of the case study findings on the propositions

Chapter 6

Discussion

To answer the main research question, **"What are the core discrepancies between theoretical guidelines and operational practices when using anomaly detection in business organisations?"**, we elaborate on key topics from the literature review and case studies that exemplify the discrepancies between anomaly detection in practice and in theory. Firstly, in section 6.1 we take a close look at an unfortunate combination of organisational challenges that can negatively affect anomaly detection projects. Secondly, in section 6.2 discuss how the challenge of defining and modelling normal and anomalous activity looks in practice. Thirdly, in section 6.3 we use insights from the case studies to expand on our understanding of working with alerts and addressing false positives. Fourthly, in section 6.4 we describe the challenge of choosing between open source and commercial solutions, and the factors influence that decision.

For each topic we provide our opinion on the matter and come up with recommendations for practice. Future research directions and recommendations for researchers will be discussed in section 7.4.

6.1 Gathering resources and support

The topic of gathering the necessary resources and support was explored to some extent in section 3.3. During the case studies we saw that these issues are critical. For example, in one of the case studies all eight respondents said that getting data from within the organisation was the biggest (or one of the biggest) challenge they have encountered. In the following paragraphs we elaborate on these issues in light of what we observed in the case studies (summarised in Figure 6.1 and Figure 6.2).

Data. In one case, where the company charter allows the cyber security team to use or access any data it needs, getting it is not as simple as showing up and asking for it. In their view the hardest part of the project is getting the data, "everything is hard about it", from both a technological and organisational perspective. Without the data the system cannot detect anything and the interviewees mention several challenges they have faced surrounding data.

First, the network logs are not centrally owned or managed, instead various local network/IT departments have the data. These departments do not necessarily understand the usefulness of the data. They might use it for operational purposes, e.g. store it for 2 weeks or collect only samples of the data. Often it takes some convincing to get them to collect and deliver the data needed for the system.

Second, capabilities like generating NetFlow do not necessarily exist in all these departments. In some cases, the project team has to support them with advice on specific devices for collecting the data, best practices, or funding for building the capabilities to collect and deliver the data. This process can take months.

Third, legal and privacy laws on a local, national and international level make it difficult to collect some data. Commonly, it takes a long time to get all the necessary approvals (e.g. from legal departments) and some laws even prohibit logs leaving the country of origin.

Fourth, once the data is delivered it does not come with any quality assurance. This can cause a problem for the anomaly detection if the models are being built with faulty data. As a simple example, if no data was collected in the last two Tuesdays, the model may produce a lot of alerts next Tuesday as it does not expect any traffic. Currently there are no tools or processes in place to address this, but if someone suspects that something is wrong it is investigated. The project team have recognised that the need for processes and tools for monitoring and dealing with data quality.

Fifth, it is challenging to be context-aware in such a large setting, with data from so many sources. Currently, the system is not collecting and providing these contextual information, meta-information about the data in an organised way. This makes it difficult for the data scientists and anyone that needs access meta-information to explain the data, e.g. for verifying alerts.

Sixth, as the amount of data grows the scale of the infrastructure has to grow as well. The data currently in the system is a fraction of what the team intends to have in the system. It is difficult to know whether the decisions that make the system run today will not impede future growth.

Figure 6.1: Summary of main challenges with getting data

Project team. From the case studies we observed that within the project team the system is highly dependent on people, the talent running the system. In one case study organisation, once operational, the threat management system (including anomaly detection) needs considerable human resources from within the security team working on the system. There is need for talented data scientists, experienced security people, and data engineers (e.g. Hadoop specialists). From the interviews we observed some uncertainty whether the small team will be able to handle investigating all interesting alerts (both

from anomaly detection and threat intelligence matching) once the system contains all the data they plan to collect and ingest.

Organisational support and dependencies. Within the organisation we see that the project is dependent on several different departments and stakeholders. The interviewees discussed a few of them: Firstly, there is a critical dependency on local IT- and network departments for getting new data sources into the system. Secondly, people from risk management, data protection, legal, and compliance departments have to be actively involved in the project. This is mostly because the data (NetFlow, DNS, Proxy) used for the system are privacy sensitive. The project team has constant open discussion with these parties, incorporate their criteria (requirements/constraints) into the project. Moreover, these parties have to be consulted before using certain data or expanding the scope of the system to incorporate more data sources. Thirdly, the (continued) support and funding from top level management. This type of project is not necessarily high on the agenda of all members of senior management that may have other priorities in mind. The priority of this type of project depends on many factors like the result it generates and the perceived importance of detecting all cyber attacks.

Environment. Outside the organisation one case study project is mainly dependent on three factors.

Firstly, the system is highly dependant on various (privacy, data protection, banking) rules policies and laws on different levels (local, national, international) that affect how organisations can work with data. For example, some countries have banking secrecy laws that affect whether and how it is allowed to extract certain data. Moreover, others have laws prohibiting the logs from leaving the country. Currently the system is collecting data from FI's country of origin and even though the country's privacy law is not always clear these issues were sorted surprisingly fast. However, the team expects this will be a difficult challenge for other countries.

Secondly, the system is dependent on the outside anomaly detection 'market' consisting of academic research on the topic, open source communities and tools, vendors of commercial solutions, other organisations pursuing anomaly detection, and freely available or commercial models.

Currently, doing an anomaly detection project is essentially a research project as they have to rely on academic papers and research. One challenge is that there is limited amount of (open source or commercial) readily available tools and models for doing anomaly detection. However, the project team has observed the outside world becoming more mature in the two years of working on the project, both vendors of commercial software and other organisations pursuing similar projects.

Thirdly, the labour market. Finding and retaining talent for the system is difficult. In the current state of building the system the project is highly dependant on external employees. These external employees will leave, taking knowledge with them. In addition, internal employees working on the project increase their value on the job market and may leave for other opportunities. Furthermore, the expertise needed for the build-

ing, maintaining and operating the system are all hard to find in the current labour market.

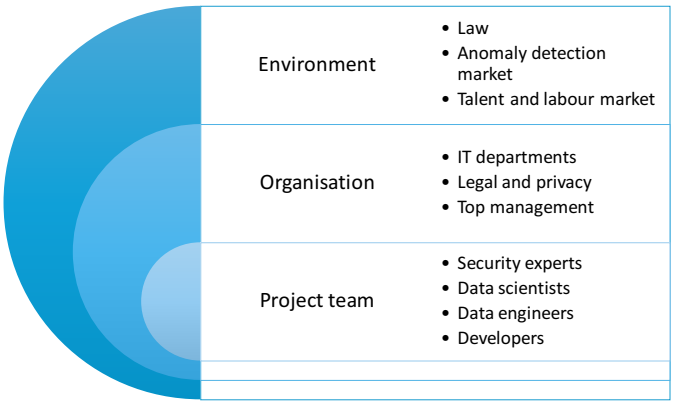

Figure 6.2: Summary of main organisational dependencies

Discussion

While none of these issues are unique to anomaly detection we have observed all of them directly influence such efforts. Most of the issues are hard to solve and come from outside the core team involved in deploying the tools. In essence, the combination of challenges is unfortunate for the potential success of these tools in a practical setting.

Recommendations for practice. Many of these issues are ingrained the culture of business organisations, so unless the organisation is able to change they will likely remain. A recommendation for practice is to try to identify all these types of issues before starting a project or in an early phase, and then make a decision whether any of these issues are likely to obstruct the project. In some cases it may be advisable to start small and generate quick wins that may give the project the additional support needed to clear some obstacles.

6.2 Defining and modelling normal and anomalous

One of the fundamental statement from the literature study about the challenges of anomaly detection is that it is hard to define normal with operational data, and that non-malicious and non-interesting anomalies are common. In the following paragraphs we discuss how the case study organisations looked at the topic.

Defining and modelling normal A common success criteria in anomaly detection is to detect malicious activity. From the discussion we understand that the distinction between malicious and anomalous can be difficult. More specifically, an anomaly in and of itself is not automatically considered malicious. Before making that distinction

organisations have to define what an anomaly is and the only way to do that is to define what is normal traffic.

From the case studies we observe that the definition of normal in statistics is not the same as in cyber security (Figure 6.3). From a data science perspective, 'normal' is when an event looks like most others, and has many neighbours around. From an organisational cyber security perspective, the definition is more complex. Normal is when the organisation is not being attacked, and there are no threats of interruptions by internal or external parties, i.e. everybody and everything is running for business goals.

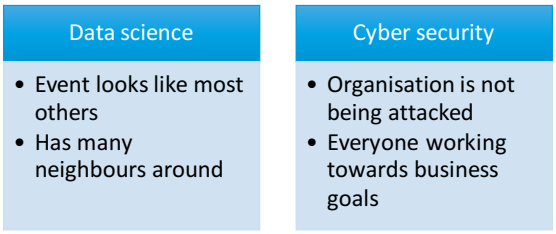

Figure 6.3: 'Normal' according to data science and cyber security

We also observed support for the proposition on using real operational data (**P5**) is the only way to build an anomaly detector. From working on fraud detection one case organisation has learnt that it is not possible to define normal unless they are using live data, i.e. simulated data cannot be used to model and distinguish normal and anomalous.

It is hard to define normal traffic because organisations do not know if they have normal traffic, i.e. live operational data may or may not contain malicious activity. Furthermore, there is no one model or formula for defining normal, it is dependant on each scenario. Therefore the data scientists have to create models based on the observations at hand, and may never know if the model only contains normal behaviour. The only way to overcome this challenge would be to use supervised machine learning on a training dataset that has been manually checked for normal and malicious behaviour, but on real-life scale that is not feasible. Therefore, the data scientists look to academic papers where researchers use models for detecting anomalies in unlabelled dataset, i.e. they are trying to find known models that are capable in distinguishing outliers in such datasets.

We observed some real life challenges that organisations have had with defining normal. For instance, two similar customers of an ISP (e.g. universities in the same country) do not share a common definition of normal traffic patterns. Moreover, tools based on available research designed to find statistical anomalies in traffic data simply fail to find patterns in the data.

The limited success we observed only happened in cases where the scope of the detector is small. For instance, it was possible for them to define normal traffic, and subsequently anomalies, for UDP (User Datagram Protocol) traffic on for the DNS (Domain Name Server) port (port 53). In their experience, such narrowly defined subsets of traffic have a relatively stable baseline, enabling the use of anomaly detection.

Defining and modelling anomalies. As mentioned before, anomaly detection is about detecting malicious activity, and similarly to the problems with the definition of 'normal', 'anomaly' in data science and cyber security is not the same (6.4). From a data science perspective, an event is anomalous when it does not look like the others, there is sparsity in the neighbourhood. From an organisational cyber security perspective, the anomalies that they are after are not statistical anomalies as such, instead malicious activity, i.e. an attack on the organisation. In addition, organisations are interested in finding non-malicious but security related anomalies, e.g. data leakage or misconfiguration.

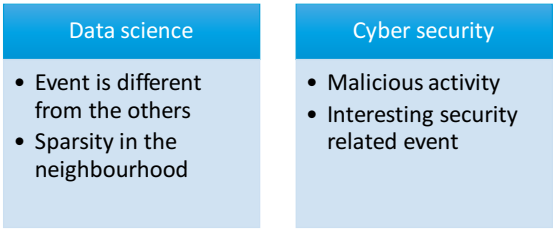

Figure 6.4: 'Anomaly' according to data science and cyber security

The case studies produced both promising and alarming insight into the issue of detecting anomalies in practice.

One one hand, it is important to keep in mind that interesting anomalies are not only detected using advanced statistical models. The case studies revealed a variety of different ways anomalies have been detected when building a detection platform or working with data. First, working with data on other parts of the system (e.g. data ingestion or IoC matching). On several occasions the developers handling the data have stumbled upon some strange events and patterns in the data. Some have been investigated but none turned out to be security related anomalies but rather issues with data quality. Second, doing basic visualisation and overview of the data can reveal interesting fluctuations that the team wants to investigate. Through visualisation they have seen interesting fluctuations in the data that were investigated. Third, using basic statistics, looking at extreme values in the network traffic logs one data scientist did detect (quick) port scans by internal employees. From investigating they found out that the scan was intentional, but not malicious. Fourth, by doing a time-series analysis they detected and investigated some of the outliers. These investigations lead to the discovery of a malfunctioning router that was causing some of the anomalies. To summarise, working with the data in new ways can reveal interesting anomalies and insights that bring value to the organisation.

On the other hand, some experts discussed the negative experience they have had with detecting anomalies. Firstly, it is difficult to define a narrowly scoped problem that will generate relevant anomalies. For instance, in an early phase one case study organisation was enthusiastic about a simple anomaly detector, detect IP addresses that are responsible for an outlier amount of flows. While this seemed like the detector they wanted, they quickly realised that there are many reasons for something to have many

flows. Secondly, the organisation deployed a system that once an attack is detected, it uses algorithms to generate rules to mitigate future attacks. As it turned out, the rules it generated were too specific to detect anything remotely different to the original attack. Thirdly, they have had to abandon models based on promising research; for instance, after struggling to tune them for over a year.

Discussion

Based on the findings of the case studies and literature review it can be argued that the problem of defining and modelling normal and anomalous is difficult to overcome. Unlabelled data used to build anomaly detectors may well contain the attacks the goal is to find. Promising ideas and models do not work as intended, and promising efforts get abandoned. In fact, unless with a very narrow scope, we have yet to see an operational anomaly detector that is producing satisfactory results.

The main promise of anomaly detection is to detect unknown or undefined attacks. Having to define a very narrow scope seems to beat that purpose of using the technique in the first place. Organisations want to deploy models that can detect all kinds of anomalies, including attack patterns no one has thought of. In other words, anomaly detection does not seem to live up to expectations.

Recommendations for practice. For organisations it is important to manage expectations of what anomaly detection can deliver. From looking at both theory and practice they will likely end up with poor results unless they are ready to accept and define a narrow scope for the anomaly detector.

6.3 Working with alerts and false positives

The literature and theoretical methodology is clear on the importance of usability, high rates of false positives combined with the cost of verifying alarms that may quickly render an anomaly detector unusable in a practical setting. From studying literature we find that most research on anomaly detection is focused on achieving a certain detection rate while issues related to usability see less attention. During the case study it was obviously an important issue and in the following paragraphs we describe our findings.

Working with alerts. From the literature we understand that it is costly (time consuming and expensive) to evaluate alerts from anomaly detectors. This was supported in case studies were discussion with experts that do investigations of alerts enable us to better understand and explain the process.

Investigations can take considerable time (hours) and requires substantial manual labour of answering various 'w'-questions (who, what, where, when, why). Firstly, the alert is examined. Based on the experience and intuition of the security experts a decision is made whether to follow up with an investigation. For example, there are cases were a single alert is not a cause for concern, but with 10 alerts in a row an investigation is

started. Second, more intelligence is collected in order to create context, e.g. identify the machine(s) generating the traffic, which are receiving it, and whether this activity might be part of something bigger. The experts use various sources for building this contextual information, e.g. different tools, users, processes and contacts within the organisation. Third, when sufficient contextual information has been collected it is possible to deliver actionable information to the people responsible for handling the event, e.g. local security team or customer that is affected.

Figure 6.5: The basic process of working with an alert

Usability of alerts. The case studies also helped us gather understanding of what usability criteria apply to alerts of anomaly detectors. Those that have experience with investigating cyber security related alerts can generally express clear requirements and expectations for the alerts that an anomaly detector generates. In cases where there are no alerts being generated by anomaly detectors, these criteria come from experience with rule-based detection systems, i.e. people have the same expectations for rule-based and anomaly-based systems.

For instance, when an anomaly detection model raises an alert it is important to explain why the alert is being raised, i.e. the logic of the detector should be interpretable for humans. Furthermore, there should be an indication of the severity or importance of the alert, e.g. how certain the model is about it's decision or an indication of how anomalous an event is compared to normal. At minimum, the alert should enable the investigators to collect intelligence. More specifically, by making is possible to dig into and explore the data and events connected to the alert. Moreover, as it takes a long time to investigate alerts it should automatically provide as much contextual information and directly answer as many 'w'-questions as possible. This information should not only be limited to data available for models, but also use additional data (e.g. perform 'Whois' lookup). Other usability criteria include automatically sending actionable alerts to the parties responsible for handling them.

Addressing false positives. Most acknowledge the fact that anomaly detectors produce many false positives, particularly in the beginning. From the case studies we observed two main types of false positives.

On one hand, when the anomaly detector is not sensitive to the context surrounding the anomaly event. Similar to when a smoke detector goes off when someone is overcooking their food. The alert is justified but the detector is missing key contextual information that makes it a false positive. The security experts expect an improvement

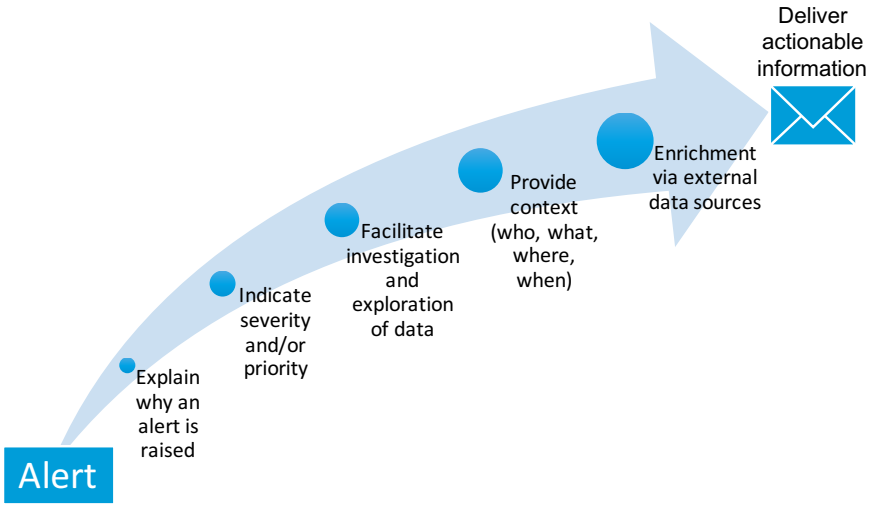

Figure 6.6: Usability criteria for alerts

process where they can specify reasons/rules why an alert is a false positive and should not be produce any more false positives under the same conditions (i.e. white-listing).

On the other hand, when the model is not good enough, producing too many alerts. Similar to when a smoke detector goes off when a match is lit. For this type of false positive the sentiment is that it is the role of the data scientists to fine-tune or fix the models, for instance by re-evaluating decisions on models, feature selection and extraction, weights and thresholds.

However, the boundary between the two types of false positives (Figure 6.7) is unclear. As a result, it is unclear who will be responsible for reducing false positive some scenarios. Surprisingly, the security experts are not particularly worried about false

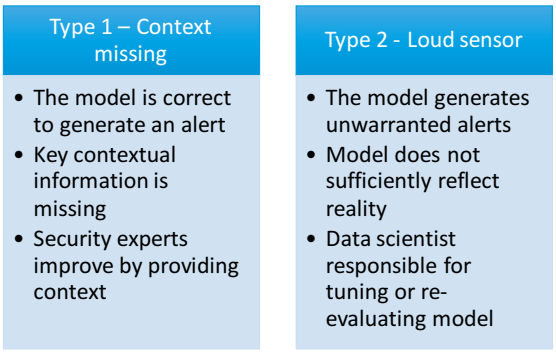

Figure 6.7: The two types of false positives from the case studies

positives but welcome them as a part of increasing the understanding of what goes on in the network. Their expectation is that the false positives will be reduced to a man-

ageable amount in a learning process consisting of an increased understanding of both the network and models (partly due to false positives) that is used to improve the false positive rate.

On the contrary, we have observed examples where this learning process does not live up to expectations. For example, a student research project where an anomaly detector based on promising research results was deployed in a case study organisation. Once deployed, the maintenance of the anomaly detector was in the hands of the security team. The detector did not produce reliable results, and after spending over a year trying to tune the model, it was abandoned. Likewise, tuning commercial solutions has its own unique challenges. Software vendors may hide (or 'black box') the underlying models and limit tuning of the detector to few parameters. These restrictions may make it impossible to properly tune the model to an organisation's specific situation.

Discussion

Producing interpretable and actionable alerts. It is challenging to meet the different usability criteria (Figure 6.6), especially if the organisation is building its own anomaly detection platform. This is a challenge both for the data scientists and the developers of the platform that should be considered from early stages.

On one hand the data scientist has to take interpretability into account when selecting algorithms and selecting or extracting features. In other words, it requires the data scientists to consider whether a model/algorithm produces a human interpretable and relevant explanation of why it labels an event as anomalous. This may be simple to do if the problem has a very narrow scope and low dimensionality, e.g. finding anomalies in packet rates on a single port/protocol. This may also be a very complex task if the anomaly detector has a broad scope, e.g. finding anomalies for all flows of the organisation's network. Moreover, certain machine learning algorithms transform (or need to use transformed) data in order to work properly. For example, many algorithms normalise data so that different features of the data are in the same scale [66], like transforming both the number of packets and bytes sent on a scale between 0 and 100. If the algorithm decides what is normal and anomalous based on transformed features the interpretability suffers unless it is simple to translate the decision back to the original data.

On the other hand, it takes a serious development effort to build all the capabilities needed to fulfil the usability criteria of alerts. The more usable an alert is (Figure 6.6) the more tasks of the verification process (Figure 6.5) are automatically delivered with the alert. It is difficult to automate these tasks, as they require skilled manual work, come from many different sources of information, and in many cases require human interaction. Furthermore, the alerting mechanism must be flexible enough so that the output of different anomaly detectors can be processed, e.g. some identify a single anomalous event while others produce a single alert for multiple events at a time (suspicious communication patterns).

These are not easy tasks, but if organisations that want interpretable and actionable alerts they need to consider these issues well before an algorithm starts distinguishing

between normal and anomalous. Specific criteria for interpretable and actionable alerts act as constraints for the data scientists when choosing tools, and act as requirements for the developers of the detection system. If possible, these criteria are important to define early, for example, so that resources are not wasted on tools that can not possibly produce acceptable alerts.

Addressing the three types of false positives. It is important to reflect on the distinction of different types of false positives from the case studies. Organisations understand that there are different causes of false positives that call for different reactions.

Looking at theory and practice, and assuming that the data used is correct[1], we define three main types of false positives. First, false positives caused by the lack of key contextual information can be dealt with by white-listing events such as scheduled back-ups that are statistically anomalous but normal activity given the context. Second, false positives caused by a poorly performing model, generating too many unwarranted alarms should be fixed by tuning or redesigning the anomaly detector. Third, from theory we know that the notion of normal activity is variable in both the short term and in the long run. In order to address this type of false positives the anomaly detector has to be periodically updated so that the model reflects current activity.

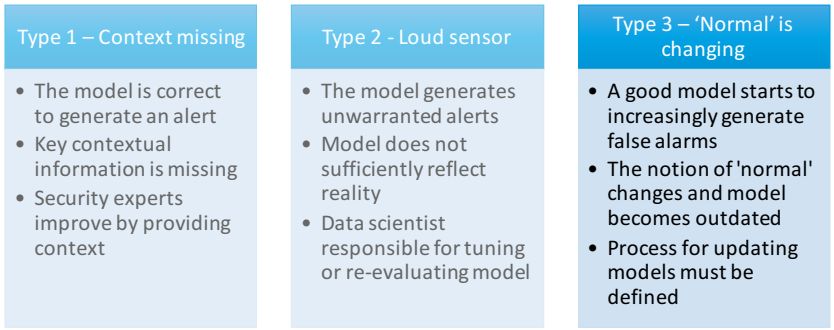

Figure 6.8: The third type of false positive from the literature review

As discussed in a previous paragraph, the boundary between different types of false positives is unclear. If it unclear what is the main cause of false positives at a given time, it is difficult to decide how to address the problem. Furthermore, it is not an easy task to address any of the different types. For the first type, the contextual information has to be collected, stored and translated into rules that white-list certain scenarios. As for the second type, the process of building detector may need to be repeated. Lastly for the third type, both the data scientists and security experts must define a process of updating models based on how frequently 'normal' changes for that particular situation.

[1]Incorrect or missing data may cause false positives if the anomaly detector. For instance if it includes the error in its model of normal activity, or when erroneous data looks like an anomalous event.

Many ways to achieve usability. The case studies confirm that false positives are a problem. How serious a problem depends on more factors than the amount of false positive, and interpretability of alerts (see examples in Figure 6.9).

Figure 6.9: Factors that compensate high false positive rates

Surprisingly, the case study organisations were more relaxed about false positive rates compared to the academic research. For one thing, the case studies did not reveal any hard limits of acceptable number of false positives before a detector is considered unusable. Instead, the definition of acceptable number of false positives depends on many different things. For instance, organisations in the early phases may welcome false positives as a mean to increase their understanding of what happens on the network. Moreover, organisations may have improved the efficiency of their investigation process, reducing the cost of false alarms.

In essence, false positives are a problem because alerts are costly to verify. As we observed in the case studies, improving the tools used to investigate and verify alerts can have an immense impact. For example, by developing a good combination of basic monitoring, baselining (knowing what activity is normal), and predefined visualisations (important ports, protocols), one organisation reduced the investigation time for some alerts from taking them hours, to a few minutes.

Specific false positive rates do not automatically make a solution unusable and there are many ways of reducing the problem and cost of false positives. From literature and practice we see that high rates can be compensated with interpretable alerts, good investigation tools, methods of white-listing certain false positives, post-processing of alerts, and more (Figure 6.9).

Recommendations for practice. For business organisations it is important that they define usability for a particular problem and how they intend to achieve it. The process of making a solution usable should not start after an anomaly detector has been built or purchased. Fortunately, there are many ways to achieve usability other than having a very low false positive rate.

6.4 Open source vs commercial tools

The decision on whether to develop the anomaly detector (system) *yourself* using open source technologies or purchase commercial solutions on the market (or deciding on a balance between the two) was prominent in the case studies. This was possibly the most pressing issue not included in the theoretical methodology (section 4.3). In the following paragraphs we discuss the practical implications both types of tools.

Open source. On one hand, organisations can use freely available, open source tools to build the anomaly detector.

Developing an anomaly detection system (or large parts of it) internally is challenging and makes the project unpredictable. Obviously it is challenging for a small team of developers to build such a system if they do not have strong development capabilities with similar technology to begin with. Furthermore, it is difficult to find and retain talented developers on the labour market. While external employees may help address that problem the organisation may not want to lose valuable knowledge and learning when the project is over and external employees leave for other projects. Moreover, end-users of such a system are used to working with systems with functionalities and interfaces that would require development resources and capabilities that way exceed those of the project organisation. For example, having a simple (web-search like) interface for searching the data, or be able to seamlessly explore events from different perspectives.

On a related note, it is challenging to implement many of the functionalities and controls required by the business organisation into a system that is built with and around open source components. For example, the poor authorisation mechanisms of the Hadoop framework which is central in the system.

When one case study organisation started its project there was little supply of commercial solutions that could do anomaly detection for cyber security purposes. As a result, the project team is doing anomaly detection in an 'open source' way, getting its models and inspirations from academic research. Still there are few ready models available to purchase, some that are available are highly patented or not shared between communities, others are incorporated into commercial software products.

As discussed in section 6.2, another case study organisation has already tried implementing models based on available promising research. All things considered, their experience has been negative as models based on promising research fail to live up to expectations. For example, they have had students deploy models as a part of their studies. Once in operation they have proved too difficult and time-consuming to tune for reliable results.

Commercial. On the other hand, the organisation can buy commercial solutions on the market.

From the case studies, we identify two main advantages of buying solutions on the market. Firstly, commercial solutions make projects more predictable. Developing the system, its different features and functionalities is unpredictable. Open source tools are not necessarily designed with organisational concerns in mind, e.g. authorisation and access management. Secondly, the project is less dependent on the technical talent of its people and the users of the system. Commercial solutions are often designed with broader types of users in mind, i.e. people with less technical skills.

In the last two years, one case study organisation is starting to see more promising commercial products available (albeit expensive) and are looking into how whether such solutions could replace some of the open source components of their system. More specifically, commercial solutions for both the matching of threat intelligence feeds, managing and storing large amounts of data, and doing statistical anomaly detection.

As discussed in section 6.3 the other case organisation shared some negative experience with commercial solutions that are often sold as the solution 'to all your problems'. Firstly, inflexibility with tuning commercial systems as the underlying models and algorithms are often hidden by the software vendors, only allowing users to tune a few parameters or simple thresholds. In many cases they have disabled some of these functionalities as the limited tuning did not allow them to properly adapt the model to their scenario.

Looking back some respondents do believe that doing this type of project with more emphasis on commercial solutions is a better way to go.

Discussion

Whether to build the system internally or purchase commercial solutions is a critical decision for organisations doing anomaly projects.

Both options are expensive, have their advantages and their disadvantages. Building a system requires strong development capabilities, and talent that is difficult to find and retain. Furthermore, the project is more unpredictable, e.g. since many functionalities required by the organisational environment are not included in open source solutions. Buying the solution on the market often comes with a hefty price tag, promises that fail to live up to expectations, and less flexibility to adapt the solution to specific scenarios.

Weighing the advantages and disadvantages the discussion we find ourselves in favour of commercial solutions.

We assume that most organisations are after the capabilities that the techniques promise, as opposed to wanting to have the ability to build such a system. Building a good anomaly detection platform that includes data ingestion, data storage, machine learning capabilities, visualisation, and alerting requires substantial development effort. In addition, there are many functionalities required by organisational rules (privacy, authorisation, authentication, data segregation) that have to built around the open source tools. From what we see in the case studies, going the 'open source route' is

unpredictable, and requires talent that is hard to find and retain in today's labour market.

In brief, defining the models are research projects and building the infrastructure is an immense development effort. Unless an organisation is a research driven software development company, they are likely to find themselves way out of their comfort zone.

With commercial solutions organisations saves them from the potential problems that come with building such a platform, although they are sacrificing the flexibility of that open source solution provide.

Recommendation for practice. Business organisations starting this journey must decide whether they want to have the capabilities anomaly detection promises or have the ability to build a detection platform. Building a platform is an immense development effort unless the anomaly detection will consist of a small group of people looking for strange patterns in a relatively small amount of data. Commercial solutions are not perfect, but a more realistic option for most organisations.

Chapter 7

Conclusions and reflections

In this research we have explored the topic of anomaly detection, focusing on the discrepancies between practice in academia and industry. When compared to promising research results, anomaly has been relatively unsuccessful when deployed in a business environment. It is important to better understand this topic as these techniques have the potential to detect harmful and sophisticated attacks.

We answer the main research question, **"What are the core discrepancies between theoretical guidelines and operational practices when using anomaly detection in business organisations?"**, in chapter 6 where we discuss in detail four key issues that show the contrast between doing anomaly detection in practice and in research.

First, we identify an unfortunate combination of organisational challenges that can affect the success of anomaly detection projects. Second, we explain how the technical challenge of defining and modelling normal and anomalous activity affects practice. Third, we identify the main discrepancies between alerts and false positives in research and industry. Fourth, how the decision of choosing between open source and commercial solutions is an important factor for anomaly detection in business organisations.

On our path of answering the main research question we answered the four subquestions that we outline in the following paragraph.

We first answered the research question **"What is the state of the art of research on anomaly detection for detecting cyber attacks?"** (**RQ 1**) in an extensive literature review (chapter 3) on research topics related to deploying usable anomaly detection in practice.

Here we found out that anomaly detection has the promise and potential to detect attacks without the prior knowledge needed for constructing conventional detection rules. The process of building rules is costly and rules constrained by the experience and imagination of the experts that construct them.

However, there are significant technical challenges of doing anomaly detection in business organisations, i.e. assumptions often made in research do not hold in reality and there is an unfortunate combination of unique problems with using machine learning to

detect cyber attacks.

Furthermore, the literature also identifies substantial organisational challenges that anomaly detection faces in practice. With these techniques being new to most organisations it is difficult to define a problem and gather the necessary resources.

Lastly, usability has to be considered from early phases and can be achieved with a combination of two things. First, by having a very low false positive rate. Second, by producing actionable and interpretable alerts.

In chapter 4 we answered the next research question **"What methodology for anomaly detection does academic research propose for business organisations?"** (**RQ 2**) by using the results of **RQ 1** to construct a theoretical methodology for deploying operational anomaly detection (section 4.3).

We used CRISP-DM [1], a well known framework for organisational data mining projects as the underlying structure for the methodology. The framework covers all phases of the life-cycle of anomaly detection projects we expect organisation will have to undergo before deployment.

The methodology consists of 15 propositions, statements from the academic literature on how organisations should approach the problem. The propositions are evenly divided between the different phases of CRISP-DM and address the most important technical, organisational, and usability issues we have identified (see summary in Figure 4.5).

Then we aimed to understand what anomaly detection looks like in practice by answering the research question, **"How do business organisations approach anomaly detection?"** (**RQ 2**). In chapter 5 we describe the case studies where we answered this research question. More specifically, the results of the case study are input for the testing of the theoretical methodology in section 5.3, and in the discussion on the main discrepancies between theory and practice in chapter 6.

Conducting a series of interviews (Appendix A) we collected information from two different organisations at different stages of deploying anomaly detection.

In brief, we learnt that business organisations approach anomaly detection differently than in academic research, and many important issues from practice have little impact on academic research. For example, our observations suggest that they are motivated by the promising capabilities of the techniques, i.e. detect unknown attacks, and its use as a tool to better understand what goes on in the organisation's networks and systems. Consequently, they do not approach this problem in a scientific way with clearly defined problem to solve using statistics or machine learning, it is a capability that they want to have. For another example, people issues play a big role in business organisations. For instance, the choice of tools (e.g. open source or commercial) has implications on what kind of talent is needed within the team. Getting the right talent then depends on the labour market, and as our interviews suggest, it is hard to find and retain people with the right skills (security, data mining, analytics infrastructure).

Finally, we connect our understanding of theory and practice when answering the last research question, **"How does the theoretical methodology compare to practical approaches in business organisations?" RQ 4**, in section 5.3 where we test the

theoretical methodology (**RQ 2**) by comparing them with approaches observed from practice (**RQ 3**). Furthermore, we identified important issues left out in the theoretical methodology (section 5.4) and discussed the stakeholder complexities of doing anomaly detection projects (section 5.5).

We find that overall the theoretical methodology is *sound*, i.e. most propositions were supported in practice. Some are strongly supported, like the challenge of false positives and importance of producing interpretable alerts for organisational success. For others, the comparison gave new insights, like the fact that one of the main motivators for using anomaly detection is to provide a better understanding of the activity on their networks. The ones that were not supported seemed to expect a mature experience of anomaly detection or certain way of working and thinking that did not fit the case study organisations. For instance, it is unrealistic to expect early on a precise definition of requirements or clear arguments for selection of algorithms or tools.

On the soundness of the theoretical methodology, we observe that the majority of propositions were supported by the experts of the case study organisations, i.e. organisations where following the propositions to some degree, or were aware of (and agreed with) the important issues from the literature. The ones that were not supported seemed to expect a mature experience of anomaly detection or certain way of working and thinking that did not fit the case study organisations, e.g. early definition of requirements or clear arguments for their selection of methods.

On the *completeness* of the theoretical methodology, it did miss a few crucial issues important to the case study organisations. Namely, the selection of open source or commercial tools, retention of talent in today's labour market, gaining situational awareness, and how laws and regulations can add complexities to anomaly detection projects.

In this chapter we have answered **RQ 4**. Firstly, by testing the theoretical methodology by comparing it with how business organisations approach and experience anomaly detection (section 5.3, summarised in Table 5.5). Secondly, by identifying key issues with anomaly detection that are important in practice but overlooked or underestimated in the theoretical methodology (section 5.4). Thirdly, describing how general stakeholder complexities of doing technological projects surfaced in the case study organisations (section 5.5).

7.1 Reflection on research approach

In this section we reflect on the developments and choices made during the course of this research. First, we reflect on overall type of research approach chosen. Second, we discuss the long process of studying literature and its eventual use. Third, look at the case studies conducted. Fourth, we reflect on the analysis phase of this work.

Exploratory approach

The choice of approach enabled us to discover new insights and facts about the research problem. Doing an exploratory research approach proved a good choice for two main

reasons. First, we headed into uncharted territory as this exact research topic is relatively small, and few researchers have looked at this topic from the different perspectives of this work (technical, organisational, usability). Second, we did not have a clearly defined practical problem to solve, instead we wanted to gain a better understanding of the discrepancies between theory and practice.

Literature and propositions

The choice of literature evolved in a process of defining the objectives of this research.

Furthermore, this process significantly impacted the direction of the research. More specifically, two articles by Sommer [10] and Gates [14] on the discrepancies between anomaly detection in theory and practice set us on this path. From there we expanded the literature review around the ideas they presented and finalised the research design.

In this work we used *propositions*, that are commonplace in qualitative research, to guide us in the research and case studies. Moreover, they enabled us to condense the literature review into 15 statements that could be tested in practice.

Case studies

The case studies were obviously essential for this work as the purpose of this work was to study and compare both theory and practice.

A key decision when designing the case studies was to collect data/investigate what business organisations have experienced with anomaly detection, i.e. capture the learning of the case study organisations. Initially, the idea was that in the case study the researcher would test the theoretical methodology by following it in a business environment. Looking back we consider it unlikely that this initial idea would have brought as deep insights as we got from the experience and learning of people who have worked with anomaly detection for years.

As a way to collect empirical data, we conducted semi-structured interviews. The main advantage of that choice was that the freedom given to interviewees allowed for responses that were outside the scope of the literature study. The main disadvantage was that the interviews generate large amounts of unstructured data that was challenging to analyse and quantify.

Analysis

Following the interviews we were confronted by vast amounts of unstructured interview responses. We overcame that problem by using a qualitative data analysis tool to code the interviews in an organised way, extract the key statements and determine the key topics.

Again we find that the use of propositions was a good research decision. Linking the empirical data back to the propositions enabled us to compare theory and practice (**RQ 4**) in a structured way.

7.2 Reflection on findings

In this study we have discovered many new insights, both from theory, practice, and the comparison of the two. Exploratory case studies often have no clear, single set of outcomes [31] and the same goes for this work. In this section, we reflect on the outcome of this work, what we find as the most interesting findings, and the implications of this study.

Key findings

With many findings, insights, and no single clear outcome, we reflect on our key findings that contribute to the academic debate.

The case study organisations want to deploy anomaly detectors that monitor their networks/systems and generate alerts for anomalous activity. For these efforts to be successful many different things have to go right, apart from finding strange patterns in network/system activity.

For both case study organisations their main source of information and inspiration is academic literature. This is because the commercial market does not currently offer many solutions, although it has started to catch up and offer promising anomaly detection tools in the last years. Moreover, business organisations deploying these methods may be reluctant to openly share information and best practices.

Due to privacy reasons, or other difficulties with gaining access to operational data academic research often uses simulated datasets. Furthermore, these anomaly detectors are in many cases only evaluated based on their detection rate, not operational deployability, or usability of alerts. As we know from the literature review, results from simulated data are unlikely to indicate real-world results, and other metrics than detection rate are important.

From the case studies we also see that the main challenges revolve more or less about issues like getting the necessary data, ensuring data quality, managing alerts, tuning operational models, developing alerting capabilities, or making the system fit to the organisation.

It is apparent that there is a mismatch between what organisations want to achieve and the resources that are available to them. In other words, finding and building an algorithm that detects cyber attacks in a lab environment is a small step towards solving the problem at hand.

It is important that organisations not only focus on the exciting capabilities of anomaly detection driving their efforts. They must also understand the challenges they will likely encounter and effort required to build an operational and usable anomaly detector.

With this in mind it is also important to keep in mind that these challenges can be addressed and overcome in a variety of ways. For instance, high false positives rates can be addressed in many ways including reducing the scope of the problem, generating interpretable alerts, having good investigation tools, or by good post-processing of alerts.

Implications of study

This study is a step on the path of developing a methodology for deploying operational anomaly detection, i.e. not solving a practical problem, but providing a better understanding of a complex research problem. The theoretical methodology of section 4.3 is built around a generic framework for data mining in a practical setting. Moreover, we have tested propositions, statements specific to anomaly detection, in a case study of two business organisations (section 5.3). It served as a tool for achieving the objective of this research, to investigate and understand the core discrepancies between theory and practice. The findings of this work can be used to iteratively improve, adapt and test the methodology in practice.

7.3 Limitations

Every research has its limitations and we will outline the main limitations of this research in the following paragraphs.

Firstly, the case study may be considered small, interviews with ten experts from two business organisations. While we have discovered new insights it is difficult to generalise based on the results. For instance, the two organisations may share similarities (e.g. culture, resources, talent) that affect their experience with the subject.

Secondly, we only used one type of empirical data, qualitative evidence from semi-structured interviews. More specifically, we are not working with 'hard data' to support our claims, but the opinions and experience of people who have worked on anomaly detection.

Thirdly, this research is overall on a general level, i.e. not tailored to specific types of organisations (e.g. size, industry), tasks (e.g. clustering, visualisation), or resources (e.g. data sources, human expertise). Therefore, it may be a challenging task for organisations to apply good practices we have identified in this research.

7.4 Future research

In chapter 6 we proposed several recommendations for organisations practising anomaly detection. In this section we propose recommendations for researchers. A work like this can not thoroughly explore all directions and topics that present themselves. These topics remain as research interest for future research projects. Considering these topics, limitations of this work, and main findings we propose the following directions to be explored further in future research.

A practical anomaly detection methodology. In this work the theoretical methodology's main use was to guide the research and analysis of empirical evidence. Furthermore, data from the case studies was used to test the theoretical methodology, and give new insights into anomaly detection in practice. However, the methodology was not put to the test in practice, e.g. by following it when addressing a cyber security problem.

In addition, two case studies are not sufficient to generalise for anomaly detection in general. Future research topic is to use the results of this work to iteratively refine and test the methodology in practice and in different anomaly detection scenarios with the goal of constructing a practical anomaly detection methodology.

Working with alerts and false positives. Further research on usability is needed. The problem of maintaining an anomaly detector and dealing with different types false positives presents a future research challenge. In this work we have found that false positive rates are and will remain a challenge. Addressing the problem only by decreasing the rate by reducing scope may not fit with project where the goal is to detect unknown attacks, and it is challenging to produce interpretable alerts. From the literature these two were identified as the ways to address usability, but the case studies revealed more ways of achieving usability (see section 6.3).

First, research needs to focus on ways to manage and handle false positives, e.g. how to incorporate white-listing of events that are anomalies but should not generate alerts. Second, research should identify good practices of monitoring and diagnosing performance of anomaly detectors so that practitioners can better respond when performance is poor. Third, research must clearly explain any assumptions made regarding the updating of the anomaly detector.

In summary, there is need for research specifically on the usability of anomaly detectors and methods of handling alerts and false alarms.

Gathering data and support. Research is needed on the best practices with gathering and working within this type of data across business organisations, especially large multinational ones that face even greater challenge of working with the different rules and regulations of different countries.

Furthermore, it is important to further explore these insights to better understand the organisational side of doing anomaly detection. The organisational issues from the literature review mostly come from research on data mining in general. However, as discussed in section 6.1, many organisational challenges can affect anomaly detection projects. For example, getting data that is distributively generated and governed throughout an organisation, setting up a project team and responsibilities, getting support, and working with different legislation.

Open source vs commercial tools. From the case studies we saw that organisations look to academic research for models to deploy. We observed the difficulties of implementing the tools described in research, e.g. in many papers an algorithm is described or a certain technique used but key information is missing. To help lower the cost of the 'open source route' researchers should increasingly share code used in their work, or the very least provide clear information about the tools used, parameters set, and other criteria that make it easier to replicate the approach.

In summary, a future research direction is to explore, combine and test freely available tools for their potential for deployment, and operational success in business organisations

Appendix A

Interviews

This appendix contains the formal interviews conducted in this work. For each interview we outline the questions asked and provide summaries of the response.

Interview questions

General questions from theoretical methodology

1. What makes your organisation interested in using anomaly detection?
2. What kind of threats or attacks are you trying to address?
3. What do you consider a successful deployment of anomaly detection in this project?

Domain specific

Data	Management	Security
D4. What data is used in this project? *Who supplies it and who is responsible for quality and explaining the data?*	**M4**. How are criteria such as objectives, requirements and constraints for the anomaly detection defined and updated?	**S4**. How do you evaluate whether an anomaly is interesting or malicious (an attack or threat)?
D5. How do you define (or model) normal activity?	**M5**. Do any legal or privacy concerns directly affect this project? If yes, how?	**S5**. What action do you take once you have an interesting or malicious anomalies?
D6. What kind of anomalies are trying to detect?	**M6**. Are there any parties that this project is highly dependent on?	**S6**. What makes the alerts (or the output of the anomaly detector) usable for you?
D7. How will false alarms generated by the anomaly detector be addressed?	**M7**. What amount of resources (time and people) will work on anomaly detection once operational?	**S7**. How will false alarms generated by the anomaly detector be addressed?

Open/exploratory questions

8. What have been some 'lessons learned' in this project?
9. What has been successful in this project?
10. What are some upcoming challenges for the project?

Interview transcripts

Confidential

Bibliography

[1] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. CRISP-DM 1.0 Step-by-step data mining guide. 2000.

[2] Animesh Patcha and Jung-Min Park. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, 51(12):3448–3470, August 2007.

[3] Bert-Jaap Koops. The internet and its opportunities for cybercrime. *Transnational Criminology Manual, M. Herzog-Evans, ed*, 1:735–754, 2010.

[4] National Cyber Security Centre - Ministry of Security and Justice. Cyber Security Assessment Netherlands 2014. Technical report, 2014.

[5] David E. Sanger and Nicole Perlroth. Bank Hackers Steal Millions via Malware, 2015.

[6] Adrianne Jeffries. 'Celebgate' attack leaks nude photos of celebrities, 2014.

[7] Patricia Zengerle and Megan Cassella. Millions more Americans hit by government personnel data hack, 2015.

[8] Mohammad A Faysel and Syed S Haque. Towards cyber defense: research in intrusion detection and intrusion prevention systems. *IJCSNS International Journal of Computer Science and Network Security*, 10(7):316–325, 2010.

[9] European Union Agency for Network and Information Security. ENISA Threat Landscape 2014, 2015.

[10] Robin Sommer and Vern Paxson. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. *Security and Privacy (SP), 2010 IEEE Symposium on*, 2010.

[11] Charlie Miller. The legitimate vulnerability market: Inside the secretive world of 0-day exploit sales. In *In Sixth Workshop on the Economics of Information Security*. Citeseer, 2007.

[12] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15, 2009.

[13] DE Denning. An Intrusion-Detection Model. *IEEE Transactions on Software Engineering*, SE-13(2):222–232, February 1987.

[14] Carrie Gates and Carol Taylor. Challenging the anomaly detection paradigm: a provocative discussion. In *Proceedings of the 2006 workshop on New security paradigms*, pages 21–29. ACM, 2006.

[15] Cloud Security Alliance. Big Data Analytics for Security Intelligence. Technical report, 2013.

[16] V W Veeneman. The strategic management of large technological projects. *TBM, Delft, NL (chapters 1, 2 and 3)*, page 13, 2004.

[17] Charles Perrow. *Normal accidents: Living with high risk technologies*. Princeton University Press, 2011.

[18] John P van Gigch. *System Design Modeling and Metamodeling*. Springer Science & Business Media, 1991.

[19] David Baccarini. The concept of project complexity—a review. *International Journal of Project Management*, 14(4):201–204, 1996.

[20] Susan Resnick-West and Mary Ann Von Glinow. Beyond the clash: Managing high technology professionals. *Managing complexity in high technology organizations*, pages 237–254, 1990.

[21] Terry M Williams. The need for new paradigms for complex projects. *International journal of project management*, 17(5):269–273, 1999.

[22] Mats Engwall. The Futile Dream for the Perfect Goal. 2002.

[23] Jeffrey K Pinto. Understanding the role of politics in successful project management. *International Journal of Project Management*, 18(2):85–91, 2000.

[24] Monowar H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita. Network Anomaly Detection: Methods, Systems and Tools. *IEEE Communications Surveys & Tutorials*, 16(1):303–336, January 2014.

[25] Shelly Xiaonan Wu and Wolfgang Banzhaf. The use of computational intelligence in intrusion detection systems: A review. *Applied Soft Computing*, 10(1):1–35, January 2010.

[26] Gary M Weiss. Data Mining in the Real World: Experiences, Challenges, and Recommendations. In *DMIN*, pages 124–130, 2009.

[27] Wenke Lee, SJ Stolfo, and KW Mok. Mining in a data-flow environment: Experience in network intrusion detection. *... on Knowledge discovery and data mining*, 1999.

[28] Christopher Kruegel and Giovanni Vigna. Anomaly detection of web-based attacks. In *Proceedings of the 10th ACM conference on Computer and communications security*, pages 251–261. ACM, 2003.

[29] Kevin S Killourhy and Roy A Maxion. Toward realistic and artifact-free insider-threat data. In *Computer Security Applications Conference, 2007. ACSAC 2007. Twenty-Third Annual*, pages 87–96. IEEE, 2007.

[30] Dina Hadžiosmanović, Lorenzo Simionato, Damiano Bolzoni, Emmanuele Zambon, and Sandro Etalle. N-gram against the machine: On the feasibility of the n-gram network analysis for binary protocols. In *Research in Attacks, Intrusions, and Defenses*, pages 354–373. Springer, 2012.

[31] Robert K Yin. Case study research design and methods third edition. *Applied social research methods series*, 5, 2003.

[32] Teresa F Lunt and R Jagannathan. A Prototype Real-time Intrusion-detection Expert System. In *Proceedings of the 1988 IEEE Conference on Security and Privacy*, SP'88, pages 59–66, Washington, DC, USA, 1988. IEEE Computer Society.

[33] Wenke Lee, Salvatore J. SJ Stolfo, and KW Mok. A data mining framework for building intrusion detection models. In *Proceedings of the 1999 IEEE Symposium on Security and Privacy (Cat. No.99CB36344)*, pages 120–132. IEEE Comput. Soc, 1999.

[34] S Axelsson. Intrusion detection systems: A survey and taxonomy. *Technical report*, 99, 2000.

[35] A. Hofmann, C. Schmitz, and B. Sick. Rule extraction from neural networks for intrusion detection in computer networks. *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme - System Security and Assurance (Cat. No.03CH37483)*, 2, 2003.

[36] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. Semi-supervised learning. 2006.

[37] Payam Vahdani Amoli and Timo Hamalainen. A real time unsupervised NIDS for detecting unknown and encrypted network attacks in high speed network. In *2013 IEEE International Workshop on Measurements & Networking (M&N)*, pages 149–154. IEEE, October 2013.

[38] R. Vaarandi. Detecting anomalous network traffic in organizational private networks. In *2013 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, pages 285–292. IEEE, February 2013.

[39] Pedro Casas, Johan Mazel, and Philippe Owezarski. Unsupervised network intrusion detection systems: Detecting the unknown without knowledge. *Computer Communications*, 35(7):772–783, 2012.

[40] Monowar H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita. An effective unsupervised network anomaly detection method. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, pages 533–539. ACM, 2012.

[41] Mark Handley, Vern Paxson, and Christian Kreibich. Network Intrusion Detection: Evasion, Traffic Normalization, and End-to-End Protocol Semantics. In *USENIX Security Symposium*, pages 115–131, 2001.

[42] Kymie M C Tan, Kevin S Killourhy, and Roy A Maxion. Undermining an anomaly-based intrusion detection system using common exploits. In *Recent Advances in Intrusion Detection*, pages 54–73. Springer, 2002.

[43] Ruoming Pang, Vinod Yegneswaran, Paul Barford, Vern Paxson, and Larry Peterson. Characteristics of internet background radiation. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pages 27–40. ACM, 2004.

[44] Matthew V Mahoney and Philip K Chan. Learning rules for anomaly detection of hostile network traffic. 2003.

[45] Roy A Maxion and Frank E Feather. A case study of ethernet anomalies in a distributed computing environment. *Reliability, IEEE Transactions on*, 39(4):433–443, 1990.

[46] Roy A Maxion and K M C Tan. Benchmarking anomaly-based detection systems. In *Proceedings of the 2002 International Conference on Dependable Systems and Networks*, pages 623–630, 2000.

[47] Hamid R Nemati and Christopher D Barko. Key factors for achieving organizational data-mining success. *Industrial Management & Data Systems*, 103(4):282–292, 2003.

[48] Andreas Hilbert. Critical Success Factors for Data Mining Projects. In *Data Analysis and Decision Support*, pages 231–240. Springer, 2005.

[49] Jakob Nielsen. *Usability engineering*. Academic Press, Boston, 1993.

[50] Jakob Nielsen. Usability inspection methods. In *Conference companion on Human factors in computing systems*, pages 413–414. ACM, 1994.

[51] Stephanie Rosenbaum. Usability evaluations versus usability testing: When and why? *Professional Communication, IEEE Transactions on*, 32(4):210–216, 1989.

[52] Jason R C Nurse, Sadie Creese, Michael Goldsmith, and Koen Lamberts. Guidelines for usable cybersecurity: Past and present. In *Cyberspace Safety and Security (CSS), 2011 Third International Workshop on*, pages 21–26. IEEE, 2011.

[53] Alma Whitten and J Doug Tygar. Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0. In *Usenix Security*, volume 1999, 1999.

[54] Andrew T Zhou, James Blustein, and Nur Zincir-Heywood. Improving intrusion detection systems through heuristic evaluation. In *Electrical and Computer Engineering, 2004. Canadian Conference on*, volume 3, pages 1641–1644. IEEE, 2004.

[55] Pooya Jaferian, Kirstie Hawkey, Andreas Sotirakopoulos, Maria Velez-Rojas, and Konstantin Beznosov. Heuristics for evaluating IT security management tools. *Human–Computer Interaction*, 29(4):311–350, 2014.

[56] Tarik Ibrahim, Steven Furnell, Maria Papadaki, and Nathan Clarke. Assessing the Usability of Personal Internet Security Tools. In *Proceedings of the 8th European Conference on Information Warfare and Security*, pages 102–111. Academic Conferences Limited, 2009.

[57] Tulsidas Patil, Ganesh Bhutkar, and Noshir Tarapore. Usability evaluation using specialized heuristics with qualitative indicators for intrusion detection system. In *Advances in Intelligent Systems and Computing*, volume 176 AISC, pages 317–328. Springer, 2012.

[58] Tarik Ibrahim, Steven M Furnell, Maria Papadaki, and Nathan L Clarke. Assessing the usability of end-user security software. In *Trust, Privacy and Security in Digital Business*, pages 177–189. Springer, 2010.

[59] Richard Lippmann, Joshua W Haines, David J Fried, Jonathan Korba, and Kumar Das. The 1999 DARPA off-line intrusion detection evaluation. *Computer networks*, 34(4):579–595, 2000.

[60] Stefan Axelsson. The base-rate fallacy and its implications for the difficulty of intrusion detection. In *Proceedings of the 6th ACM Conference on Computer and Communications Security*, pages 1–7. ACM, 1999.

[61] Patricia M Shields and Hassan Tajalli. Intermediate theory: The missing link in successful student scholarship. *Journal of Public Affairs Education*, pages 313–334, 2006.

[62] Ana Isabel Rojão Lourenço Azevedo. KDD, SEMMA and CRISP-DM: a parallel overview. 2008.

[63] Pamela Baxter and Susan Jack. Qualitative case study methodology: Study design and implementation for novice researchers. *The qualitative report*, 13(4):544–559, 2008.

[64] Matthew B Miles and A Michael Huberman. *Qualitative data analysis: An expanded sourcebook*. Sage, 1994.

[65] Margaret C Harrell and Melissa A Bradley. Data collection methods. Semi-structured interviews and focus groups. Technical report, 2009.

[66] Antti Juvonen. Intrusion detection applications using knowledge discovery and data mining. 2014.