CASE STUDY PROTOCOL - WHY AND HOW TO ADAPT THE CRISP-DM DATA MINING PROCESS? A CASE STUDY IN THE FINANCIAL SERVICES DOMAIN

Case Study Protocol

Protocol version	Version released	Content or update summary	Update date	Update done by
1 Version 0.1.	16/10/2019	First draft version of protocol to guide the research process	15/11/2019	
2 Version 0.2	07/01/2020	Interview Instrument update, Research Questions Update, other updates based on improved scoping of study	20/12/2019	
3 Version 0.3	02/04/2020	General update to reflect final documentation of the study	20/03/2020	

1 Preamble

The purpose of this case study protocol is to document all key aspects of the given case study research, including its design and execution process. The coding scheme is an integral part of this protocol and is available on FigShare alongside the latest version of the protocol. The concurrent publication with reporting of this case study findings is intended for submission to IEEE BigData 2020 conference. This case study protocol has been constructed based on the best practices as outlined in [1, 2].

2 Research Project and Method - General Overview

2.1 Background

Given research project is a continuation of the previous studies conducted by authors on the topic of data mining methodologies application in financial services domain . In particular, by conducting SLR-based study, we have identified 3 key adaptations scenarios which focused on tackling technology-centric (scalability), business-centric (actionability), and human-centric (mitigating discrimination) aspects when applying data mining methodologies in financial services industry settings. Further, it was concluded that the primary purpose of adaptations had been associated with the fact that given aspects were not addressed adequately in existing reference data mining methodologies (e.g. CRISP-DM), thus constituting gaps. To understand their nature and more importantly understand the respective solutions and remedies applied or potentially propose new ones, the gaps require further investigation. Therefore, the overarching objective of this case study is an in-depth investigation of the most common reference data mining methodology (CRISP-DM) gaps when applied for data mining, data science projects in the banking industry settings. The study outcome to achieve is to identify perceived gaps when applying data mining methodologies

applications in the financial services domain, their implications and how practitioners address them. The research questions are formulated as follows:

- RQ1 - What CRISP-DM gaps do practitioners perceive in the financial services industry?
- RQ2 - Why do practitioners perceive these gaps, i.e. what is the impact of these gaps?
- RQ 3 - How they adapt CRISP-DM to address these gaps?

The case study adopts interpretative research paradigm as per [5], investigating and providing an understanding of the particular phenomenon (in our case, data mining methodologies applications and associated process model gaps) through interpretations of data mining projects participants. This research approach is underpinned by the exploration of documentary evidence, i.e. documentation of all projects covered in this study.

2.2 Key principles

In the research process, we intend to rely on qualitative data. The research process is structured as a flexible design study, e.g. [7], where we expect to adjust key parameters of the study during the research process. Due to this, there is an increased need for triangulation in research process due to heavy reliance on qualitative data, e.g. [8]. Referring to four common triangulation approaches such as data (source), observer, methodological, and theory, we will employ triangulation for data sources. In the context of this study that implies that we will be using two data sources - data mining/data science projects documentation and semi-structured interviews with projects participants.
Further, we have adhered to the recommendation on the characteristics of the exemplary case study as set out by [3, 9], and in particular, by [10] for case studies in software engineering domain. The key aspects include clearly formulated research questions, planned and consistent data collection (with triangulation), analysis and synthesis, well-documented research process (with the chain of evidence), adequate exploration and reporting.

2.3 Case Study Research Design

Case study methodology Referring to [3], we adopt a single, embedded case study approach motivated by the following reasons:

- we perform the case study in the context of one, single financial organisations, and we investigate data mining practices within one department. Further, we study usage of one data mining methodology - CRISP-DM. Thus, we by design adopt single case study approach
- however, we investigate data mining practices over long period of time (ca 2 years) and across number of various data mining projects. Further, we examine each project in-depth with support of two data sources - documentation and interviews, analysing and reporting on complex details in each project. Therefore, based on [1], embedded case study design with each data mining project serving as a separate unit of analysis become appropriate.

Selection of research organisation and case studies With research questions defined and selected methodology, we proceeded with the selection of the relevant organisation and associated most suitable projects portfolio. Concurrently, based on guidelines of [1] we took into account:

- accessibility aspects (relevant data required should be easily accessible including documentation as well as domain experts), as well as
- findings potential (projects with maximum potential to derive data to answers research questions).

Combining domain criteria (financial services industry), scope (data science projects), and accessibility, findings potential considerations, we have postulated the following organisational criteria:

- organisation has been systematically engaged with data mining over the last 3 years. That means that data mining adoptions should have been performed in all aspects covering: (1) specialised function creation, (2) infrastructure (platforms, tools) investments, (3) formalised data science process development, (4) governance model
- further, maturity has been achieved in data-driven decision making
- organisation has accumulated a significant portfolio of completed data mining projects.

Based on these considerations, we have selected universal financial institution which is one of the largest universal financial institutions in the Northern Europe and is present across different geographic markets. The organisation is one of the early adopters of data analytics and data mining for decision-making. Further, we have selected the largest centralised function in the organisation (center of excellence) directly responsible for the execution of data mining/data science projects.
For case selection, we have considered [12] approach, which defines four variants of case studies selection:

1. extreme/deviant - for capturing unusual cases (extremely problematic or extremely good)
2. maximum variation - to capture information about the significance of various circumstances
3. critical - to achieve information permitting logical deductions
4. paradigmatic - to develop a metaphor or establish a school for the domain that the case concerns.

In our research, we directed our case selection strategy towards maximum variation and critical dimensions paradigms to observer data mining process realisation in various circumstances. Also, in order to achieve maximum possible information richness to identify common patterns which persist across the projects (both in the form of gaps in CRISP-DM as well as their solutions).
For the data mining projects itself, we have formulated the following selection criteria:

- projects should cover various distinct purposes and business problems and as a result, also provide distinctly different solutions (e.g. automated payment classification systems, strategic data-driven complex clustering solutions, business process mining, etc.)
- projects should cover different geographic markets
- project should cover different clients' universes (private or corporate) or alternatively support various functions across the bank
- projects have been conducted relatively recently, advisable if unit has applied formal data mining process methodology
- projects should be completed and encompass all data mining process phases.

Data Collection The data collection itself is guided by three key principles (for example, defined in [11]):

- usage of multiple sources of data - we use two sources of data in the form of projects documentation and semi-structured interviews
- creation of case study database - projects documentation is logged; interviews are recorded, fully transcribed, signed-off, logged into QDA Miner Lite qualitative analysis software package, and coded based on developed coding instrument
- data validation and maintenance of a chain of evidence - we validate project documentation with interviews while data database ensures traceability (validity aspects are in detail discussed in section 7)

Quality Assurance Quality Assurance routines are implemented throughout the design and execution process to maintain the validity and reliability of the study. Based on [1]:

- draft case study design has been peer-reviewed
- case study has been piloted to confirm the case study design

- case study progress has been regularly reviewed and evaluated with supervisors to ensure that deviations towards expected process and results (if any) do not result in case study weakness or are addressed to prevent it
- research process is well documented, the chain of evidence is constantly maintained - protocol and research instruments versions (based on iterations) are logged along with raw data.

2.4 Case Study Research Process

Based on [1], there are five key case study research process steps to which we also adhere. Cross-referencing to both case study protocol and respective research paper, they are:

- Phase 1 - Case study design - outlined in this protocol, section *2.2 Key principles*
- Phase 2 - Preparation for data collection - outlined in this protocol, chapter *3 Procedures* and chapter *4 Research instruments*
- Phase 3 - Collecting evidence - outlined in this protocol, chapter *6 Execution, Results and Reporting* as well as in research paper itself
- Phase 4 - Analysis of collected data - outlined in this protocol, chapter *6 Execution, Results and Reporting* as well as in research paper itself
- Phase 5 - Reporting - outlined in this protocol, chapter *6 Execution, Results and Reporting* as well as in research paper itself.

3 Procedures

Data Collection Techniques We have selected first degree (direct methods) and third-degree (independent analysis of artefacts) data collection techniques as per [13]. Further, direct methods for this case study are interviews (eg. [7]) while artefacts analysis refers to a collection of projects documentation. The main steps of research process are planned as follows: (1) Step 1 - *projects Documentation Collection and Initial Analysis Phase*, (2) Step 2 - *Interviews (incl. Pilot interviews) Phase*, (3) Step 3 - *Information Synthesis*. Justification for each of the phases is as follows.

Initially, projects documentation is to be gathered from participating institution to get acquainted with the context of projects and perform initial analysis. Documentation will contain all internally recorded, written information on the projects. We also note that the set of documents has not been created for the purposes of the study nor to serve the purpose to provide a comprehensive account of the context of the respective data mining projects. Therefore, we expect natural deviation between the quality and completeness of documentation as it is not fully standardised and projects were carried out by different teams.

After the analysis of documentation, interviews are to be conducted for exploratory and explanatory purposes. Exploratory objectives are : (1) to provide additional in-depth context on data mining project by projects participants, (2) to ascertain completeness and correctness of documentation, and (3) to eliminate deviations across projects documentation (obtain missing factual data as recommended by [14]), and make case studies collected evidence comparable. The explanatory objective is (4) to gain in-depth participants perspective, insights on the conducted data mining process, and (5) to elicit data mining process deviations in comparison to benchmark standard data mining methodology CRISP-DM.

The key benefit of the proposed combined data collection approach is improved research validity (see, for example [15, 1]). In particular, conclusions drawn from the combination of documents analysis and interviews as two data sources will be more robust. Also, as mentioned, interviews provide observer triangulation (eg. [13]) where not only documentary but also participants' perspective is accounted for. At the same time, based on the flexible case study design approach, data collection sub-phases and elements could be adapted as the study progresses.

Project name	Geography /Coverage	Clients' Universe	Solution	Outcome
Project 1 - Product propensity model	1 geography, 1 product	Private customer	Business Delivery - Propensity models available for consumption by campaign management engine	Customers score
Project 2 - Retail customers Microsegmentation model	3 countries	Private customer	Business Delivery - Clustering algorithm with integrated, automated vizualization dashboards	Customers assigned to microsegments
Project 3 - Product propensity model	3 geographies, 1 product	Private customer	Business Delivery - Propensity models available for consumption by campaign management engine	Customers score
Project 4 - Lending process mining	3 geographies, mortgage lending business process	Private customer	POC - Comprehensive business process diagnostics and improvements	Process KPIs dashboard
Project 5 - Private Payments Classification	3 geographies	Private customer	Model Rebuild - Machine Learning based private customers payments classification	Payment classification engine
Project 6 - Graph Analytics Tool	1 geography	Private customer	POC - Development, test and productionization of of customized graph algorithms vizualization library	Automated downloadable graph analytics library for vizualization of very broad customer datasets (patterns, behaviours, etc.)

Table 1. Case study final selection of projects and their key characteristics.

Projects Planning For *Documentation Analysis Phase* six projects from the total portfolio of more than 50 projects available in the financial institution were chosen. The selected projects set has been a representative sample and has covered approximately 15% of the total portfolio. The selection principles are outlined above (*2.3* Case Study Research Design). Further, the selection covers four different project types. The first type is Business Delivery that concern development of various models for different banking products, or complex algorithms for analysis of bank's customers, such as private customers micro-segmentation algorithm. The second type is Model Rebuild, which shares the commonality of

rebuilding, retraining, and re-deploying models and algorithms. The third is Proof of Concept (POC) projects that explore the use of new analytics types and techniques, e.g. process mining for improving the lending process. Finally, the fourth type is Capability Development projects which aim to develop advanced competencies and tools for repeatable usage in other data mining projects. The selected task here is the exploration of advanced graph analytics methods and development of visualisation algorithm library. The selection covers four different project types. The first type is Business Delivery that concern development of various models for different banking products, or complex algorithms for analysis of bank's customers, such as private customers micro-segmentation algorithm. The second type is Model Rebuild, which shares the commonality of rebuilding, retraining, and re-deploying models and algorithms. The third is Proof of Concept (POC) projects that explore the use of new analytics types and techniques, e.g. process mining for improving the lending process. Finally, the fourth type is Capability Development projects which aim to develop advanced competencies and tools for repeatable usage in other data mining projects. The selected task here is the exploration of advanced graph analytics methods and development of visualisation algorithm library. It is intended to be used repeatedly and support discovery of customer behaviour patterns in the majority of other data mining projects. Selected projects taxonomy with key characteristics is presented in Table 1 above.

Interview Planning [7] highlighted three types of interviews - unstructured, semi-structured and fully structured; we have opted for semi-structured approach. The latter is a common interview technique in software engineering case studies ([1]) with the mix of open and closed questions. The mixed approach is the most applicable in our case as it allows us to pursue both descriptive and explanatory objectives. In particular, descriptive goals will be attained by receiving additional relevant and contextual information on the projects conducted that will also cover information deviation gaps across projects documentation. Further, explanatory goals will be achieved by obtaining information from participants perspective on the specifics of the conducted data mining process and its comparison to benchmark (CRISP-DM framework).

Based on [1] guidelines, initially formulated interview questions will be piloted where in addition to a traditional approach, participants will be asked feedback regarding understandability of questions, how easy or hard to answer them, etc. Interview questionnaire itself is presented in the chapter *4 Research instruments*.

Based on [1], timeglass model for interview sessions has been chosen. The choice is motivated by the need to address both structured questions for deriving contextual projects information (exploratory part), as well as unstructured part to solicit in-depth perspective from a participant (explanatory part).

Participants of interviews have been identified in cooperation with the organisation and based on projects documentation. The cohort consists of Data Scientists and Projects Managers - all mature or advanced experts with at least 3+ years experience in data mining field and 5+ years of overall relevant working experience. Data Scientists are with Computer Science or Quant backgrounds while Project Managers possess solid technical backgrounds and advanced complex project management experience. All interviewees are provided assurance on complete external and internal anonymity.

Further, interviews are to be recorded and additional notes to be taken during an interview. The interview transcripts are to be confirmed with the interviewee on the correctness and completeness of information recorded.

4 Research instruments

4.1 Initial framework for describing context

As proposed by [18], preliminary framework for describing case study context is introduced (primarily based on [18]). It concerns both institutional aspects where research is conducted as well as individual projects. Framework is to be adjusted on iterative basis.

Aspect	Key Elements
Research Organization	Business Model, Structure, Analytics Maturity, Data-driven decision making maturity, Data Analytics/Data Science Work Practices (incl. data science process, workflows, artifacts)
Projects	Timespan, Team and Roles, Domain, Documentation Status, project status (Proof of Concept, Proof of Value, business-oriented project, etc.), additional facts of interest (eg. novelty)

Table 2. Initial framework for describing institution and projects context.

4.2 Seed (initial) codes for documentation coding

Based on [16, 17], initial set of codes are originated from Research Questions (High Level Codes) and pre-defined variables of interest (Medium Level Codes). Lower level coding includes Comments primarily intended for unstructured expressions of solutions, process changes, remedies, etc. introduced within the case process or retrospectively. Coding instrument and scheme changed iteratively.

Nr	Research Questions	Code	Descriptions
1	RQ1 - What are the CRISP-DM process methodologies associated gaps when applied in the context of financial services industry?	GAPS	Absence of elements across all spectrum - on phase, tasks, activities, deliverable level, also includes absence of "soft aspects", project management associated aspects, etc.
2	RQ2 - What are practical implications (consequences) of these gaps in the settings of financial services industry?	IMPACT	Negative impacts on business value realization, project process, non-compliance with certain internal and external requirements. Problems, difficulties including ones for for project execution, project team, etc.
3	RQ 3 - How are the gaps address? i.e. what are the typical solution patterns/scenarios to address these gaps undertaken by financial institutions?	SOLUTION	Practices adopted in each project if the gap perceived on the way and practices adopted based on retrospect. Modifications, additions, deletions to applied methodology.

Table 3. Initial, High level seed codes derived from Research Questions.

Nr	Code Name	Descriptions
1	Process and Deliverables - Business Understanding (BU) Phase	Description of BU phase, focus on process and deliverable
2	Process and Deliverables - Data Phase	Description of Data phase, focus on process and deliverable
3	Process and Deliverables - Modelling Phase	Description of Modelling phase, focus on process and deliverable
4	Process and Deliverables - Testing Phase	Description of Testing phase, focus on process and deliverable
5	Process and Deliverables - Deployment Phase	Description of Deployment phase, focus on process and deliverable
6	Quality	Anything related to quality of conducted process and project outcomes, deliverables
7	Technology and Enterprise Architecture	Anything related to technology aspects of the project from hardware/software perspective (platforms, tools, environments), infrastructure and competences
8	Internal Stakeholders	Anything related to stakeholders (especially, business users) - stakeholder functions, routines, management in the course of project, adoption, satisfaction, etc.
9	External Stakeholder - Customers	Anythings related to customer/clients perspective and solutions adaptations
10	Compliance	Anythings related to internal and external regulations and requirements
11	Business process change/impact	Anything related to business process change due to results of the project
12	Business value realization	Anything related to business impact and value realizaton of the project, actionability of the results, solution

Table 4. Initial, Medium level seed codes derived from Research Questions.

4.3 Interview Guide

4.3.1 Interview Guide Structure

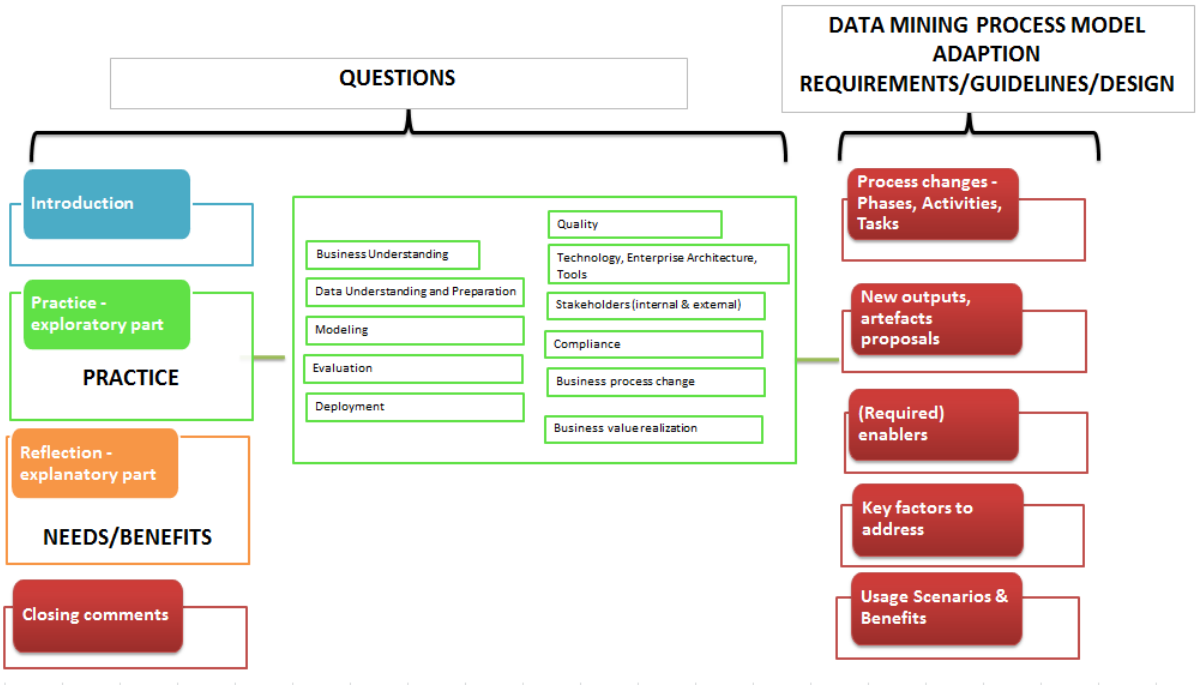

Figure 1. Interview Guide Structure

4.3.2 Interview Instrument

Phase 1 General (10-15 min)

The objective of this interview phase is to introduce the research project and get understanding and short background of interviewee. Further, Phase 1 of interview aims to discuss, provide motivation and anchor with interviewee the particular use case to be focused upon in the further interview part.

1. Questions

1.1. Let me provide You with explanation about the research project – read out from Consent Form (provided in Appendix 8.1)

1.2. Based on the Consent Form presented to You, do You have any questions on the research project and study background? Is everything clear?

1.3. Please briefly comment on Your background, experience in the company and role(s)

1.4. Let me introduce You to the particular use case - we have selected project X. The primary motivation for the selection of the project [depending on the use case] was:

o Completion of the project (End to End)

o Significant degree of adoption and realization of results in business practices

o Complexity (if applicable)

o Geography

1.5. How do You feel with the given selection of the use case? Is that good example of Your usual data mining practices? Please comment

Internal validity checkpoint - are questions understandable so far? Any concerns, ambiguities from Your side?

Phase 2 Exploratory part - overview of data mining practices and adopted approach in the context of selected use case

The objective of this phase is to obtain information from interviewee and discuss:
- his/her overall methodological practice with data mining, data science projects – how projects are conducted usually in respective organization?
- his/her methodological practice in the context of the concrete use case identified and agreed upon (in Phase 1) – i.e. how this particular use case was conducted?
- contextualize given practices based on overall methodological process experience of interviewee.

2. Questions

2.1. Have You applied/have had experience with some form of approach or methodology for data mining, data science projects? Please explain

2.2. Let's deep dive in the project X. Referring to the use case documentation, You have applied key phases of the structured data mining approach (elements of KDD/Crisp-DM) in this particular data science project/use case? Please comment

2.3. In the context of our use case, let us go through each key phases of data mining process and let start with Business Understanding phase. Please consider which were the standard data mining approach gaps, their impact and solutions adopted by You and Your colleagues? Please explain

2.4. In relation to Data Understanding and Preparation phase, which were the standard data mining approach gaps, their impact and solutions adopted by You and Your colleagues? Please explain

2.5. In relation to Modeling phase, which were the standard data mining approach gaps, their impact and solutions adopted by You and Your colleagues? Please explain

2.6. In relation to Evaluation phase, which were the standard data mining approach gaps, their impact and solutions adopted by You and Your colleagues? Please explain

2.7. In relation to Deployment phase, which were the standard data mining approach gaps, their impact and solutions adopted by You and Your colleagues? Please explain

Internal validity checkpoint - are questions understandable so far? Any concerns, ambiguities from Your side?

2.8. In the context of our use case, let us discuss and reflect on set of additional aspects of data mining process and its outcomes. What can You reflect on the quality aspect of conducted data mining process as well as use quality of use case deliverables, outcomes? Please explain

2.9. What can You reflect on the Technology, Enterprise Architecture, Tools aspect in the given use case? Please explain

2.10. What can You reflect on internal/external stakeholders aspect in the given use case? Please explain

2.11. What can You reflect on compliance aspect in the given use case? Please explain

2.12. What can You reflect on business process change, meaning changing how business process/practices by implementing data mining model/solution in the given use case? Please explain

2.13. What can You reflect on business value realization in the given use case? Please explain

Internal validity checkpoint - are questions understandable so far? Any concerns, ambiguities from Your side?

Phase 3 Use Case Analysis – Reflective and Summary Part (20-30 min)

The objective of this phase is to obtain interviewee reflections (in the context of use case experience) and retrospect opinion/view on the use of structured approach/methodologies for data mining.

3. Questions

3.1. Do You consider that use case have achieved its goals? Yes/Partially/ No

3.2. If:

a. Yes - could You elaborate on the reasons for success?

b. If partially - what have been successful and what was not? What have been the associated reasons/causes?

c. If No - what have been the reasons?

3.3. Could You please elaborate:

o to which extent application of the structured approach we discussed contributed to achieving data mining goals and business value realization success? Why? Please explain

o or alternatively to which extent application of standard, structured data mining approach has hindered or not supported achieving data mining goals and business value realization? Why? Please explain
3.4. Do You consider that more systematic and more strict application of methodology could have been more beneficial for the use case? Or on the contrary, do you think that a more lightweight and less strict application of the methodology is preferable? Why? Please explain
3.5. In Your view, what could be adjustments in data mining approach/standard methodologies that could make use cases execution better? Please explain
Internal validity checkpoint - are questions understandable so far? Any concerns, ambiguities from Your side?
Phase 4 Closing comments
Expressing gratitude for interviewee, informing on the transcript sign-off routine, informing how the results of the study will be summarized and shared within research community.

5 Data analysis guidelines

In our research, we focus on qualitative data analysis method as it is referred as most common in case study research (see for example [16]) with the main steps presented in the figure below. The main aspects of these research approach are outlined below.

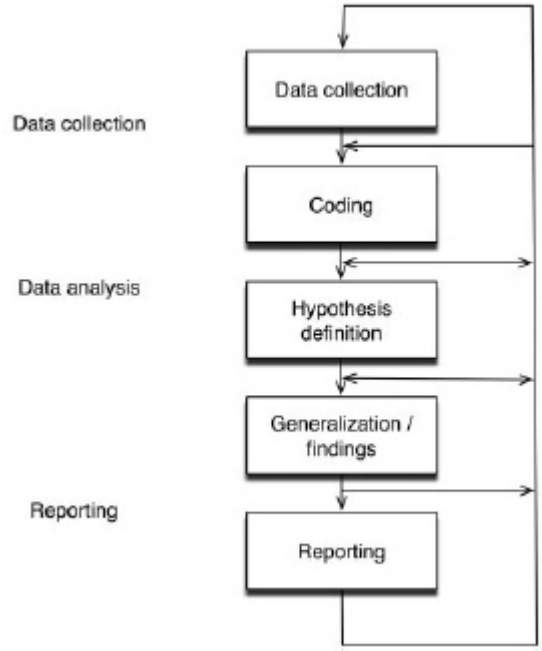

Figure 2. Main steps of data analysis in flexible research, reproduced from [1],p.63

Iterative aspect We take into consideration specifics of qualitative data analysis which include analysis in parallel with data collection as well as iterations. As revealed in [1], new insight are expected to be found during data analysis that could trigger need for additional data to investigate the insights. Therefore, we expect series of update towards research instruments, in particular, interview questionnaire.
Systematic aspect To adopt systematic approach to maximum possible extent, we use data coding to support discovery of recurrent patterns. That in turn will cater to formulation of the number of generalisations (or body of knowledge as referred in [1]) which will be one of the key outcomes of the research. The initial set of codes (as per [16]) are originated from research questions and pre-defined variables of interest and described in 4 Research Instruments. We adopt combination of *template* and

editing approaches mentioned in for example [7] which are also suitable in software engineering case studies and ensure sufficient level of formalism([1]). We start with *template approach* with a priori codes based on research questions while *editing method* is adopted in the research process with codes added or modified based on findings during the analysis. Further, as mentioned earlier we expect iterative research cycle of "coding"-"analysis of coded data"-"results formulation"-"interviews" steps.
Data synthesis aspect Guided by [3], the most suitable generalization technique is *Cross-case synthesis* due to portfolio of multiple projects serving as profound comparative information basis. Further, *Explanation building* technique is expected to be adopted to highlight potential cause-effect relationships and discover explanations which could be systematized with *Pattern matching*.

6 Execution, Results and Reporting

6.1 Evidence collection - Documentation and Interviews

Overall, 6 case study documentation has been retrieved and analyzed during period of Nov - Dec 2019 in 2 iterations. Case documentation was in line with expectations in terms of quality and richness of information. Collection strategy was not adjusted. Initial analysis of documentation supported the next step which were interviews. According to initial plan, 8 interviews were conducted during period (Dec 2019 - Feb 2020). Interview instrument versions were logged; overall, there were 3 iterations of interview instrument. Recordings of interviews saved and transcribed in full resulting in 115 pages of text. Each interview transcript confirmed with interviewee and if necessary corrected. There have been no significant deviations to initial research plan.

6.2 Results and Reporting

As per case study plan, results were analyzed according to 5 Data Analysis guidelines. There have been no significant deviations, there were additional coding methods adopted (based on [20]). The reporting of the case study results has been executed in respective research paper. The reporting paradigm has been guided by approaches indicated as most recognized and adopted in software engineering (eg. [1, 7]). The relevant approaches to consider are:

- portrayal - in-depth description, especially applicable in case of software engineering process ([1]), used frequently (final selection used in the research papaer)
- adversarial statements - adoption of two positions (very rarely used)
- dialogue - presentation in the form of discussion (rarely used)
- stereotype - recently emerged reporting paradigm whereby identified stereotypes are used as basis for reasoning about the choices and behaviours of organization (eg. [1]).

7 Validity

Validity for this research is classified based on [3, 19] principles commonly applied in flexible design studies. In the given classification, four validity aspects are distinguished and we address them accordingly.
Construct validity refers to what extent what is studied correspond to what is intended and defined to be studied by research questions (eg. defined in [1]). In our case, the biggest risks associated with construct validity could stem from interview instrument (questionnaire) and its application, i.e. questions could be interpreted differently by us (researchers) and interviewees. We have addressed construct validity and by the following mitigative actions (integrated in research design):

- including respective checkpoints in questionnaire (please refer to internal validity checkpoints for each interview part reflected in Interview Instrument),

- iterative approach, and
- reconfirming interview information and conclusions with interviewee.

Internal validity risk in the context of case studies research could emerge when casual relationships are examined (as defined eg. [1]). The examination of the latter is not the key purpose of this research. However, the respective risk will be taken into account and mitigated should such analysis facet arise.
External validity concerns generalization of the case study findings, in particular, to what extent findings are of relevance to other cases. We have addressed this aspect by careful selecting target institution for the research, as well as case studies sample portfolio subject to in-depth examination. Further, adoption of triangulation for data sources and theory, quality assurance routines for research execution, as well as maintaining chain of evidence addresses external validity aspect.
Reliability is concerned with data and analysis independence from researcher and replicability. This aspect is addressed by transparent, logged coding procedures and interviews which are improved via iterative research process. Replicability of study is further reinforced by constructing and maintaining Case Study Protocol.
Measures for improving validity for this research Based on [7], the following approaches have been adopted to improve overall validity:

1. triangulation - applied for data sources and data collection procedures (discussed in 2.2)
2. member checking - review of interview results, study findings and conclusions with participants
3. negative case analysis - formulating alternative explanations in the course of analysis
4. audit trail - systematic approach to keeping track of the research materials and process

Bibliography

[1] Runeson P, Host M, Rainer A, Regnell B. *Case study research in software engineering: Guidelines and examples*. John Wiley and Sons; 2012.

[2] P. Brereton, B. A. Kitchenham, and D. Budgen. Using a protocol template for case study planning. In Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering, University of Bari, Italy, 2008.

[3] R. K. Yin. Case Study Research: Design and Methods, 3rd edition. SAGE Publications, 2003.

[4] 105 B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. Hoaglin, K. El Emam, and J. Rosenberg. Preliminary guidelines for empirical research in software engineering.IEEE Transactions on Software Engineering, 28(8):721–734, 2002.

[5] H. K. Klein and M. D. Myers. A set of principles for conducting and evaluating interpretive field studies in information systems. MIS Quarterly, 23(1):67, 1999.

[6] J. W. Anastas and M. L. MacDonald. Research Design for the Social Work and the Human Services. Lexington, New York, 1994.

[7] C. Robson. Real world research 2nd edition. Blackwell, 2002.

[8] R. E. Stake. The Art of Case Study Research. SAGE Publications, 1995.

[9] R. Kyburz-Graber. Does case-study methodology lack rigour? The need for quality criteria for sound case-study research, as illustrated by a recent case in secondary and higher education. Environmental Education Research, 10(1):53–65, 2004.

[10] D. E. Perry, S. E. Sim, and S. Easterbrook. Case studies for software engineers. In 29th Annual IEEE/NASA Software Engineering Workshop—Tutorial Notes, 2005, pp. 96–159.

[11] J. M. Verner, J. Sampson, V. Tosic, N. A. Abu Bakar, and B. A. Kitchenham. Guidelines for industrially-based multiple case studies in software engineering. In Third International Conference on Research Challenges in Information Science, Fez, Morocco, 2009, pp. 313–324.

[12] B. Flyvbjerg. Five misunderstandings about case-study research. In Qualitative Research Practice, concise paperback edition. Sage, 2007, pp. 390–404.

[13] T. C. Lethbridge, S. E. Sim, and J. Singer. Studying software engineers: data collection techniques for software field studies. Empirical Software Engineering, 10:311–341, 2005.

[14] B. Flynn. Empirical research methods in operations management. Journal of Operations Management, 9(2):250–284, 1990.

[15] R. van Solingen and E. Berghout. The Goal/Question/Metric Method. A Practical Guide for Quality Improvement of Software Development. McGraw-Hill, 1999.

[16] C. B. Seaman. Qualitative methods in empirical studies of software engineering. IEEE Transactions on Software Engineering, 25(4):557–572, 1999.

[17] K. M. Eisenhardt. Building theories from case study research. The Academy of Management Review, 14(4):532, 1989.

[18] K. Petersen and C.Wohlin. Context in industrial software engineering research. In Proceeding of the 3rd International Symposium on Empirical Software Engineering and Measurement, 2009, pp. 401–404.

[19] C. Wohlin, M. Host, M. C. Ohlsson, B. Regnell, P. Runeson, and A. Wesslen. Experimentation in Software Engineering: An Introduction. International Series in Software Engineering. Kluwer Academic Publishers, 2000.

[20] Saldaña, J., *The coding manual for qualitative researchers*, 2ns Edition, Sage, 2013.

8 Appendix

8.1 CONSENT FORM

1. Names of researchers and contact information.
PhD researcher
2. Purpose of the research project and case study.

- Investigation of data mining methodologies application practices across various industry/research domain
- Given case study focuses on financial services industry

3. Procedures used in the study, short description of what the participant should do
Participants participate in semi-structured interviews on data mining methodologies application practices based on completed use cases they participated in.
4. The study and what steps the researcher will carry out during these activities
Semi-structured interview and documentation of results. Further, consolidation of findings and publishing.
5. A text clearly stating that the participation is voluntary, and that collected data will be anonymous. Participation is voluntarily and interviewing is fully anonymous.
6. A list of known risks.
No any risk to participants as results and interviewing is anonymous.
7. A description of how confidentiality will be assured. This includes a description of how collected material will be coded and identified in the study.

Interview will be recorded based on consent of participant or alternatively documented via notes taken in electronic format. During interview participant will remain anonymous and will not called upon or associated with personal details which would allow for participants identification.
None of primary information is to be shared. The findings will be consolidated on higher level of abstraction ensuring further anonymization. Any transcribes of interview will be sign-offed by participant. The conductor of the study (is the only person on the project directly handling data collection. Supervisors of the projects do not participate in the data collection, processing and transcribing. Supervisors direct the PhD student in applying scientific methods and interpretation of the documented findings, results analysis and synthesis. Further, supervisors have access only to final publications draft, case study protocol which contains only planning, design of case study. Supervisors do not access or review any raw sourced data (interviews, and their recordings in any format as well as non-consolidated, intermediate analysis of collected evidence).

8. Information about approvals from participating organization.

PhD researcher has received approvals from direct manager in organization.