THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Systematic Analysis of Engineering Change Request Data

Applying Data Mining Tools to Gain New Fact-Based Insights

Ívar Örn Arnarsson

CHALMERS

Department of Industrial and Materials Science
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2020

Systematic Analysis of Engineering Change Request Data
Applying Data Mining Tools to Gain New Fact-Based Insights
Ívar Örn Arnarsson
ISBN 978-91-7905-310-9

Doctoral thesis at Chalmers University of Technology
New serial no 4777
ISSN 0346-718X

Department of Industrial and Materials Science
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone + 46 (0)31-772 1000
Email: varo@chalmers.se

Cover illustration:
High level data mining process illustration
Created by Ívar Örn Arnarsson

Printed by
Chalmers Reproservice
Gothenburg, Sweden 2020

Data is the new science.
Big data holds the answers.

- Pat Gelsinger

Systematic Analysis of Engineering Change Request Data
Applying Data Mining Tools to Gain New Fact-Based Insights

Ívar Örn Arnarsson
Department of Industrial and Materials Science
Chalmers University of Technology, Gothenburg, Sweden

ABSTRACT

Large, complex system development projects take several years to execute. Such projects involve hundreds of engineers who develop thousands of parts and millions of lines of code. During the course of a project, many design decisions often need to be changed due to the emergence of new information. These changes are often well documented in databases, but due to the complexity of the data, few companies analyze engineering change requests (ECRs) in a comprehensive and structured fashion. ECRs are important in the product development process to enhance a product. The opportunity at hand is that vast amount of data on industrial changes are captured and stored, yet the present challenge is to systematically retrieve and use them in a purposeful way.

This PhD thesis explores the growing need of product developers for data expertise and analysis. Product developers increasingly refer to analytics for improvement opportunities for business processes and products. For this reason, we examined the three components necessary to perform data mining and data analytics: exploring and collecting ECR data, collecting domain knowledge for ECR information needs, and applying mathematical tools for solution design and implementation.

Results from extensive interviews generated a list of engineering information needs related to ECRs. When preparing for data mining, it is crucial to understand how the end user or the domain expert will and wants to use the extractable information. Results also show industrial case studies where complex product development processes are modeled using the Markov chain Design Structure Matrix to analyze and compare ECR sequences in four projects. In addition, the study investigates how advanced searches based on natural language processing techniques and clustering within engineering databases can help identify related content in documents. This can help product developers conduct better pre-studies as they can now evaluate a short list of the most relevant historical documents that might contain valuable knowledge.

The main contribution is an application of data mining algorithms to a novel industrial domain. The state of the art is more up for the algorithms themselves.

These proposed procedures and methods were evaluated using industrial data to show patterns for process improvements and cluster similar information. New information derived with data mining and analytics can help product developers make better decisions for new designs or re-designs of processes and products to ensure robust and superior products.

Keywords: Product Development, Engineering Change Request, Design Analytics, Design Structure Matrix, Markov Chain, Machine Learning.

ACKNOWLEDGEMENTS

This research was performed in collaboration with the Department of Industrial and Material Science (IMS) at the Chalmers University of Technology and the Volvo Group Trucks Technology (Volvo) in Gothenburg. All the work from 2015 has been financially supported by Volvo. From 2017 to 2019, the project has received financial support from the Swedish Governmental Agency for Innovation Systems (Vinnova). These supports are gratefully acknowledged.

I have been surrounded by great people during the process of obtaining my PhD degree, and they have inspired and supported in times of need. First, I want to thank my main supervisor, Professor Johan Malmqvist, who has continuously provided me with guidance, ideas, visionary sketches, and feedback. Similarly, I am grateful for the support of my co-supervisor Professor Rikard Söderberg. Next, I want to thank Mats Jirstrand, Emil Gustavsson, and Otto Frost at Fraunhofer-Chalmers Research Centre for Industrial Mathematics for their collaboration and co-authorship. My sincere thanks also goes to everyone at the Product Development Division at IMS for providing me a friendly and supportive environment.

I wish to express my deepest gratitude to my industrial sponsor Anders Ydergård, my industrial supervisor Lena Borg, and my previous industrial supervisor Lars Börjesson for their corporate insight, guidance, and on-site support. I am also grateful to all my colleagues at Volvo whom I have had the great pleasure working with during the past four years.

Finally, I am grateful for having my family and friends' unconditional support at all times and in all situations.

Ívar Örn Arnarsson

Gothenburg, Sweden, 2020

APPENDED PUBLICATIONS

The following research papers form the foundation of this PhD thesis.

Paper A

Arnarsson, Í. Ö., Gustavsson, E., Malmqvist, J., & Jirstrand, M. (2017). Design Analytics is the Answer, But What Questions Would Product Developers Like to Have Answered? In *21st International Conference on Engineering Design* (Vol. 7, pp. 71-80). Retrieved November 29, 2019, from https://pdfs.semanticscholar.org/af88/3bff37952945413aa4efd8edd65314eccc94.pdf

Paper B

Arnarsson, I. Ö., Malmqvist, J., Gustavsson, E., & Jirstrand, M. (2016). Towards Big-Data Analysis of Deviation and Error Reports in Product Development Projects. In *Proceedings of NordDesign* (Vol. 2, pp. 83-92). Retrieved November 29, 2019, from http://publications.lib.chalmers.se/records/fulltext/240205/local_240205.pdf

Paper C

Arnarsson, Í. Ö., Gustavsson, E., Jirstrand, M., & Malmqvist, J. (2020). Modeling industrial engineering change processes using the design structure matrix for sequence analysis: A comparison of multiple projects. *Design Science*, *6*, 1-17. doi:10.1017/dsj.2020.4.

Paper D

Arnarsson, I. Ö., Frost, O., Gustavsson, E., Stenholm, D., Jirstrand, M., & Malmqvist, J. (2019). Supporting Knowledge Re-Use with Effective Searches of Related Engineering Documents-A Comparison of Search Engine and Natural Language Processing-Based Algorithms. In *Proceedings of the Design Society: International Conference on Engineering Design* (Vol. 1, No. 1, pp. 2597-2606). Cambridge, United Kingdom: Cambridge University Press**.**

Paper E

Arnarsson, Í. Ö., Frost, O., Gustavsson, E., Jirstrand, M., & Malmqvist, J. (2019, in press). *Natural language processing methods for knowledge management: Applying document clustering for fast search and grouping of engineering documents*. Article Submitted to Journal Concurrent Engineering in December 2019 for publication.

DISTRIBUTION OF WORK

The work for each paper was distributed among the authors as follows:

Paper A — **Arnarsson** planned and coordinated the study. He also conducted interviews with AB Volvo engineers and performed the literature review and analysis in collaboration with Malmqvist. Gustavsson performed statistical analysis. **Arnarsson**, Gustavsson, Malmqvist, and Jirstrand co-wrote and reviewed the paper.

Paper B — **Arnarsson** coordinated the paper, performed the literature review, analyzed the data, and contributed to writing of the paper. Malmqvist performed the literature review and contributed to writing of the paper. Gustavsson coded the software demonstrator used to perform statistical analysis and wrote parts of the paper with the help of Jirstrand who also reviewed the paper.

Paper C — **Arnarsson** coordinated the paper and performed most of the literature review, together with Malmqvist and Gustavsson. Gustavsson coded the Markov Chain DSM with domain knowledge support from **Arnarsson**. **Arnarsson** and Gustavsson wrote the paper with help of Malmqvist and Jirstrand, who both reviewed the paper.

Paper D — **Arnarsson** coordinated the paper and performed most of the literature review, together with Stenholm. Frost and Gustavsson coded the pipeline and the search service. **Arnarsson**, Stenholm, Frost, and Gustavsson drafted the paper with help of Malmqvist and Jirstrand, who both reviewed the paper.

Paper E — **Arnarsson** coordinated the paper and performed most of the literature review with help from Frost and Gustavsson. Frost and Gustavsson coded the pipeline and the search frontend. **Arnarsson,** Frost, and Gustavsson drafted the paper with help of Malmqvist and Jirstrand, who both reviewed the paper.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

DRM	Design Research Methodology
DSM	Design Structure Matrix
DG	Design Guideline
ECR	Engineering change request
IMS	Industrial and Material Science
MC	Markov Chain
ML	Machine Learning
NLP	Natural Language Processing
PD	Product Development
R&D	Research and Development
RQ	Research Question
Volvo	Volvo Group Truck Technology

1 INTRODUCTION

The fast development of data collection and storage requires data expertise and analysis. Companies confront large volumes of complex data from multiple sources and increasingly rely on analytics to improve opportunities for their business and their products (Wu, Zhu, Wu, & Ding, 2013). The product development area is no exception, with large, complex development projects lasting several years. Product developers make tens of thousands of parts and millions of lines of code that lead to data growth. Such data are difficult to analyze manually, and this offers an opportunity to research the application of data mining and analytic models on product development data to identify patterns and meaningful outputs according to the needs of product developer and support decision-making for new designs/re-design of products.

1.1 Background

Product changes in product development projects are often logged and stored in databases where structured data (i.e., numerical data) are mixed with unstructured data (i.e., text inserted by engineers). Along with progress in both machine learning and data mining, new techniques for retrieving insights from complex datasets have emerged.

This PhD thesis focuses on engineering change requests (ECRs) or engineering change reports that reside in product development databases. The data comes from a large development project with a duration of several years at the Volvo Group Trucks Technology (Volvo). ECRs contain variables, such as title, part name, part number, problem description, root cause, solution, and test results. Organizations continue to improve their products to stay ahead of the competition and retain quality. Product development projects often need to enhance a product or a procedure to initiate such a change process, which is well known under the name of ECR. The ECR data must contain information permitting identification of the types of errors and changes made. Engineering changes are opportunities to improve, enhance, or adapt a product, changes in external circumstances, regulation, etc (Pikosz & Malmqvist, 1998). ECRs within organizations contain information about desired product changes. The effects of such a change can then be evaluated to select the best solution. Changes can occur throughout the entire product life cycle, from the concept phase to the after-market phase.

Volvo wanted to identify the information product developers need related to ECRs to select and test established data mining algorithms. Product developers' information needs were studied to gain domain knowledge on what kind of data analysis is beneficial (Arnarsson, Gustavsson, Malmqvist, & Jirstrand, 2017). Case studies were then conducted, levering ECR data with machine learning and data mining algorithms to test and validate their usefulness, more specifically natural language processing (NLP) (Arnarsson, Frost, Gustavsson, Jirstrand, & Malmqvist, 2020) and Markov chain Design structure Matrix (DSM) (Arnarsson et al., 2019). The benefits for Volvo would be to understand organization needs for data mining, possibilities to streamline the ECR

process with its states and shorten time it takes product developers to retrieve historical information on designs.

Machine learning is a form of analytical model in which algorithms are utilized to explore data and to make predictions from data (for a survey, see Kotsiantis et al. (2007). Data mining is another closely related research field where the aim is to detect patterns and knowledge in datasets (for a survey, see Berkhin, 2006).

Machine learning and data mining offer quantitative methods for performing data analysis with any system that generates data. Once an overview has been created, data can be identified and collected from databases, and data analytics can be employed to extract insights from historical data within companies (Zheng & Dagnino, 2014). The traditional way of analysis involves finding answers in data through manual exploration, for example, users export data to a spreadsheet software tool and examine it. Now, there are opportunities to make these explorations more effective and faster through automation. Moreover, data analysis can help clarify a variety of complex issues that are, otherwise, not obtainable through "manual" inspection and analysis.

In the seminal book *Competing on Analytics: The New Science of Winning*, Davenport and Jeanne (2007) define data analytics as using statistical and quantitative analysis of data, combined with explanatory and predictive modeling. The models and analyses provide the basis for fact-based management and decision-making. For a survey on data mining and knowledge discovery on a general level, see Han, Kamber, and Pei (2011), and for a more technically oriented survey describing techniques and machine learning tools for data mining, see Witten, Frank, Hall, and Pal (2016).

The term "design analytics" has recently been proposed by Van Horn, Olewnik, and Lewis (2012) to refer to the area of research that focuses on processes and tools to enhance the transformation of design-related data into formats suitable for design decision-making. Examples include Tucker and Kim (2011) who applied analytics to consumer trend data to inform product design. Meanwhile, Bae and Kim (2011) conducted a study on how to improve the development process of a digital camera by using data mining techniques on customer information. Similarly, Lewis and Van Horn (2013) explored customer behavior profiles and reflected on customer needs in the late stages of the development process.

In conclusion, design analytics (i.e., data mining, machine learning, and modeling) can provide insights into information needs of many companies.

1.2 Focus on engineering change request data

During a project, many design decisions need to be changed due to the emergence of new information. Notably, changes late in the development process are costly and may cause delay in the project (Clark & Fujimoto, 1991). Unfortunately, the bulk of engineering needs for changes is often discovered late in the process, as shown in Figure 1 that depicts the amount of changes recorded per month during a complex development project (Giffin et al., 2009). These changes are well documented in databases, but due

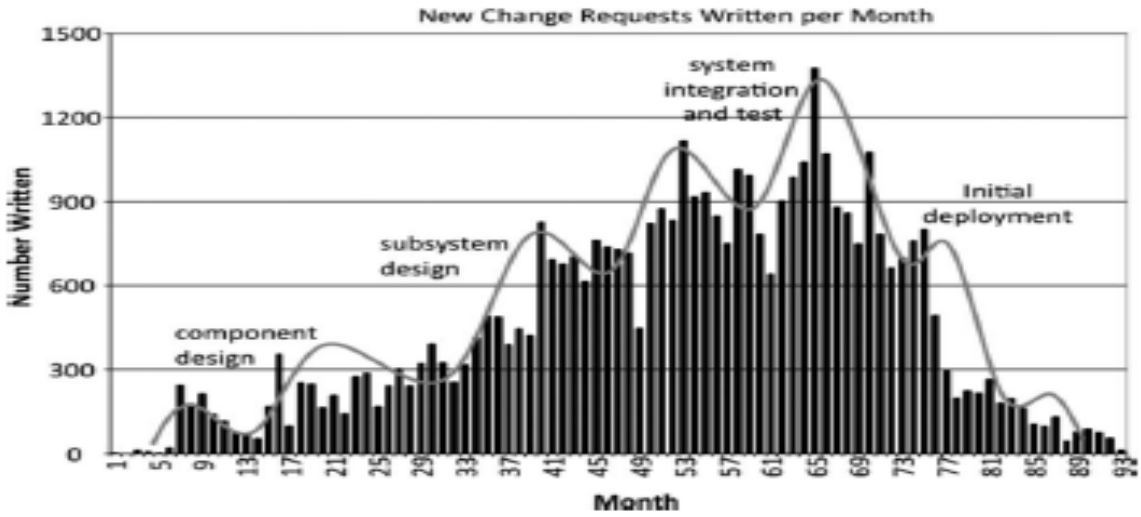

Figure 1. Frequency of change requests during a complex system development project (Giffin et al., 2009).

to the complexity of the data, few companies analyze engineering changes in a comprehensive and structured fashion.

Late changes in product development projects may result in failure to meet objectives for budgeting, scheduling, or technical performance. Weak leadership, lack of planning, and rigid processes play important roles (Thomke & Reinertsen, 2012). The problem with late product changes has been known for several decades, but recent studies confirm that it remains a challenge associated with high costs, quality problems, and development lead time delays (Giffin et al., 2009; The Standish Group, 2014; Fernandes, Henriques, Silva, & Moss, 2015). The root causes of these changes are still poorly understood, and mitigations proposed in the 1980s, such as concurrent engineering and quality function deployment, are not sufficient to solve the problem.

1.3 Research context

The research was performed at AB Volvo, one of the world's leading manufacturers of trucks, buses, construction equipment, and marine and industrial engines. The company employs about 95.000 people, has production facilities in 18 countries, and sells products in more than 190 markets. Volvo performs research and product development of complete vehicles, powertrains, components, and services. The data used in this research are mostly based on trucks.

1.4 Purpose and goals

Although research within design analytics for product development based on customer, manufacturing, and project data have been conducted, they did not specifically consider ECR data.

1.4.1 Purpose

The purpose of this research project and PhD thesis is to examine how historical ECR data can be analyzed to gain new insights and identify patterns to improve the ECR process and support product developers in their daily work. This PhD thesis can help companies explore design analytical models on product development data by building prototype tools and gain new insights to support work performed in product development.

Wish respect of the research golds listed below Blessing and Chakrabarti's (2009) Design Research Methodology was chosen to help collect data to answers and later discuss the research questions.

1.4.2 Scientific goals

The scientific goals of this thesis are as follows:

- develop and implement methods to analyze product development data,
- evaluate the needs of product developers for data mining and identify beneficial outcomes,
- propose, develop, and implement a data analytics tool that matches the identified organizational needs, and
- evaluate the effectiveness of the tools with product developers working in industrial development projects.

1.4.3 Industrial goals

The intended industrial goals are as follows:

- identify improvement areas based on the analysis of data,
- identify product developers' needs for data mining related to ECR data,
- perform data analysis using insights from product developers to improve IT systems, processes, instructions, and data related to ECRs,
- conduct case studies and validate them in workshops with company experts.

1.5 Delimitations of the research

This study was conducted in a large multinational firm that develops and manufactures commercial vehicles. The firm's management of ECRs is typical for large firms that develop complex systems and products, for example, in the aerospace and defense industries. The ECR data structures and processes are similar as other firms'.

The research project is a partnership with the case company; hence, we were able to conduct a detailed inquiry on the topic in a realistic setting. However, the single-case

study research design, including the analysis of the databases, may limit the transferability of the results.

ECR data produced in a product development project can be obtained from different sources of information. This research focuses on a subset of data from a number of projects with duration of several years, excluding any attached documents. The most common documents attached to an ECR are pictures, drawings, and tests results.

Within the overall scope proposed, we were able to investigate the effects only in a limited way due to the setup and time constraints of the study. Data were analyzed, visualizations were created, and workshops were done with experts, but the loop was not closed despite the identification of the root causes. Nevertheless, we still argue that the findings related to needs and possible solutions can be transferred to other contexts that deal with data with a similar structure.

1.6 Thesis structure

The content of each chapter of this PhD thesis is outlined below:

Chapter 1 introduces the topic, analyses the problem, purpose, and goals.

Chapter 2 provides a framework for this research, reviews the state of the art literature on the research topic, and identifies research gaps in the area.

Chapter 3 describes the research approach and methodology used in this research and states the research questions.

Chapter 4 summarizes the results and findings of each paper that is appended.

Chapter 5 discusses the results in relation to the research goals and research questions.

Chapter 6 outlines the conclusions from earlier chapters and the future direction of this research.

The **Appendix** contains the full versions of the five published papers that form the basis of this PhD thesis.

> **Paper A** – Design analytics is the answer, but what questions would product developers like to have answered?
>
> **Paper B** – Towards big data analysis of deviation and error reports in product development projects
>
> **Paper C** – Modeling industrial engineering change processes using the Design Structure Matrix for sequence analysis: A comparison of multiple projects
>
> **Paper D** – Supporting knowledge re-use with effective searches of related engineering documents – A comparison of search engine and natural language processing-based algorithms
>
> **Paper E** – Natural language processing methods for knowledge management – Applying document clustering for fast search and grouping of engineering documents

2 FRAME OF REFERENCE

This chapter presents a theoretical framework for this research, which critically assesses the state of the art to identify research needs and gaps. It also provides the general definitions and describes the work performed in the fields of product development, engineering changes, and design analytics.

2.1 Research area overview

The relevant research topics areas are presented in an Areas of Relevance and Contribution (ARC) diagram based on Blessing and Chakrabarti's (2009) work (Figure 2).

The main research subject is presented in the center of the model: "Systematic Analysis of Engineering Change Request Data." Relevant research areas include product development process, design process models, and design analytics, were identified as the main foundation for the research as it progressed. The top left clusters are the cornerstones of this research. Meanwhile, the other two clusters provide useful insights about the data at hand.

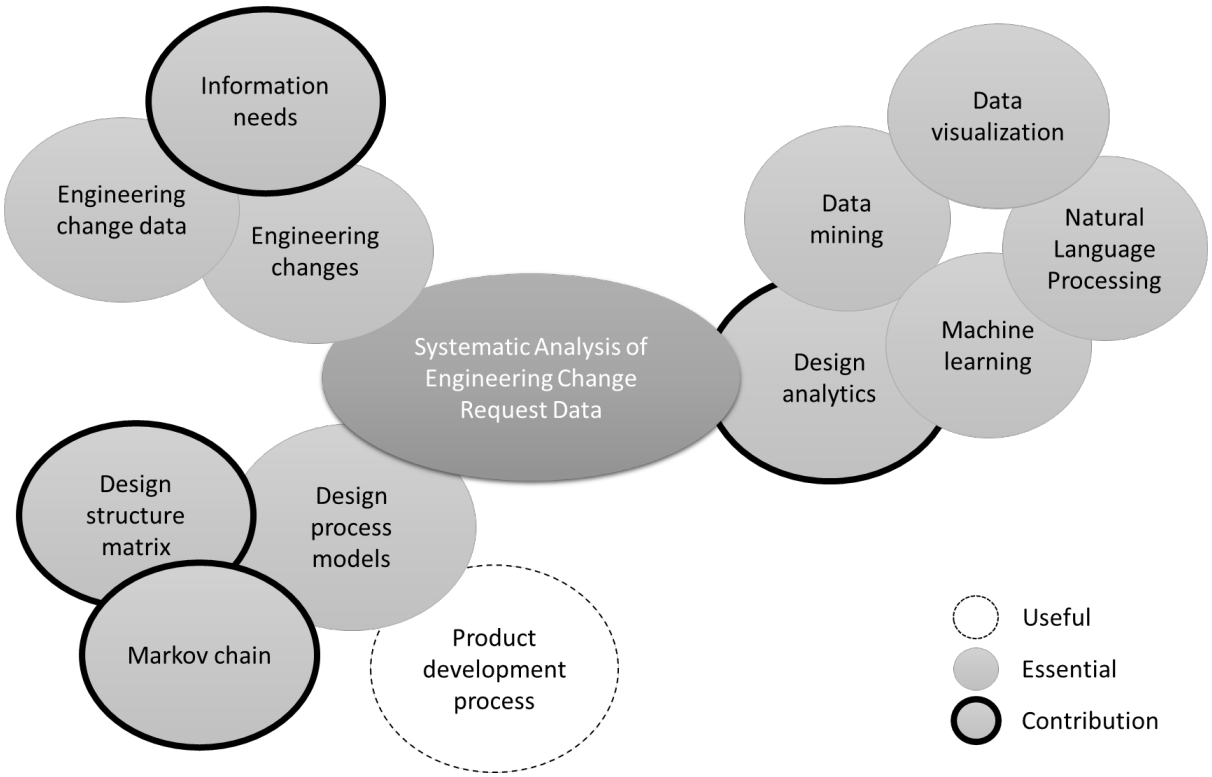

Figure 2: ARC diagram of the research overview.

2.2 Product development process

In engineering, product development covers all processes for creating a new product to or modifying existing products in the market. The incentive for product development is often the customers, ensuring their satisfaction through new or additional benefits. Ulrich and Eppinger (2012) define product development as a sequence of activities, beginning with the identification of a market opportunity and ending with production, sales, and delivery of the product. Ulrich and Eppinger's (2012) proposed generic product development process is presented in Figure 3, where major activities in the process include planning, concept development, system design, detailed design, testing and refinement, and production ramp-up. Similar methodologies for product development have been proposed by Hubka and Eder (1996), Pahl and Beitz (1996), Roozenburg and Eekels (1995), Ullman (1992), and Andreasen and Hein (1987).

Pahl and Beitz (1996) outlined four main phases (Figure 4) in engineering design: product planning and clarification of the task, conceptual design, embodiment design and detail design. Hubka and Eder (1996) proposed a design process with a series of stages that creates information about the design.

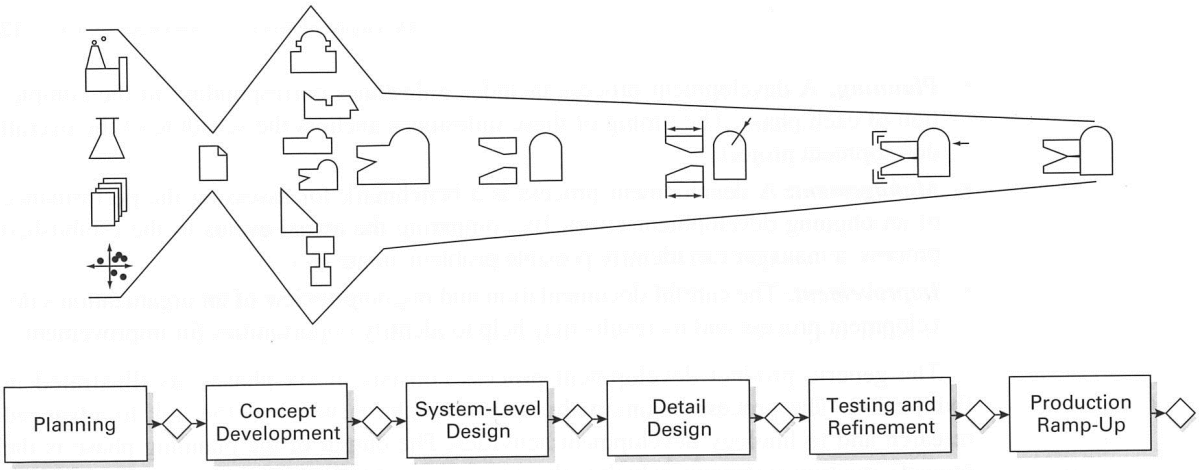

Figure 3. Generic product development process with six phases (Ulrich & Eppinger, 2012).

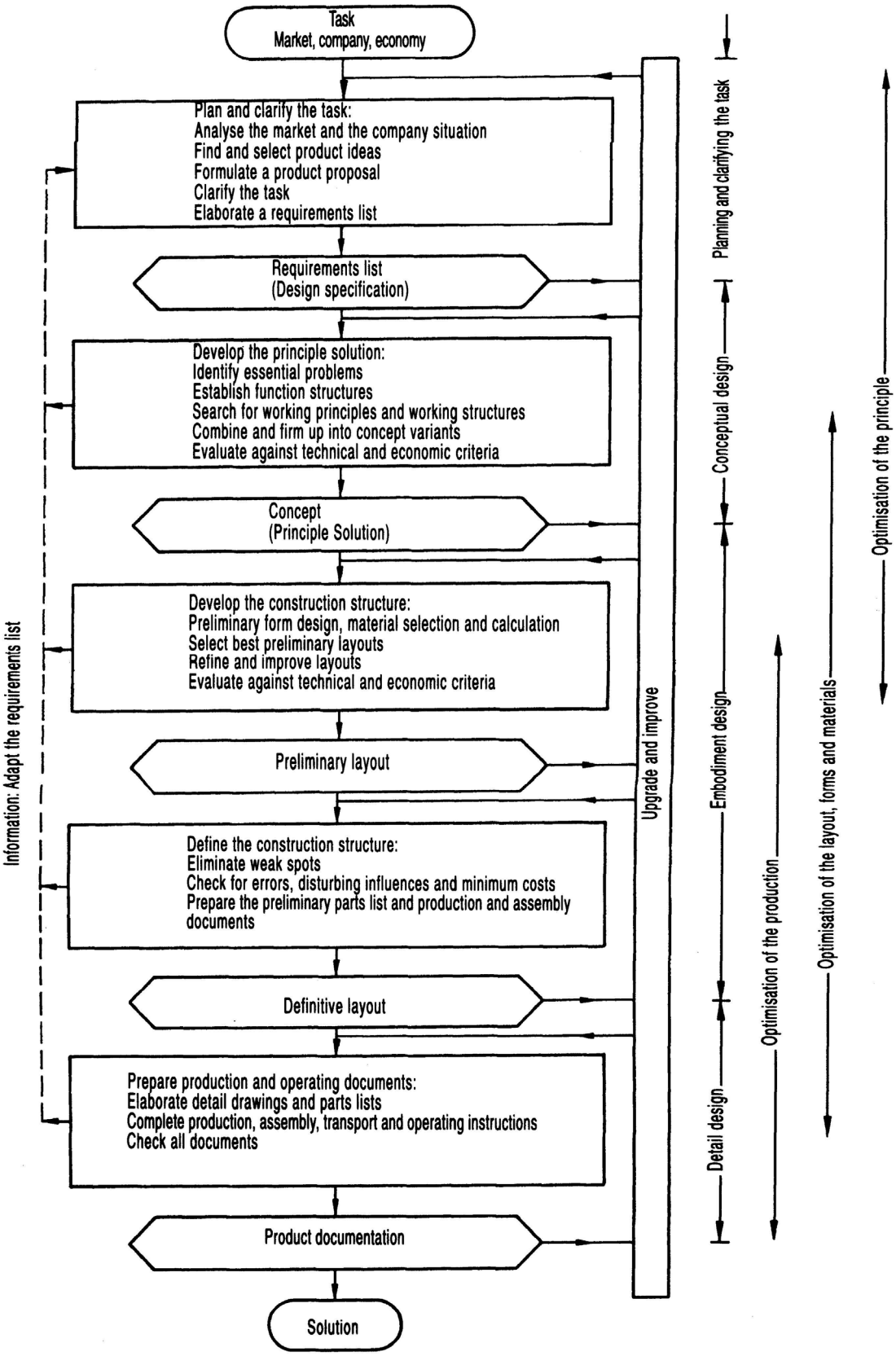

Figure 4. Four-phase design process model (Pahl & Beitz, 1996).

Other design process methodologies have also been proposed by Roozenburg and Eekels (1995) who argued that the design process should be described as a chain of tasks to performed to develop, test, refine, and market a new product. Andreasen and Hein (1987) have a widely known approach for integrated product development, which involves interviews of marketing, design, and production activities during the development stages. Ullman (1992) examined the manufacturing problems that may arise if manufacturing is not included during the design process. New data-driven process models for new product development has recently been proposed by Li, Roy, and Saltz (2019), which considered part of Ulrich and Eppinger's model. The model recognizes the importance of incorporating new information and communication technologies into hardware development. Key information flows that must transpire during concept design are identified to create a data-driven product and propose a process model that can help structure the development of products and features using data. The model is called New Product Development 3 (NPD3), which highlights the interactions between three main categories: physical product development, project management and data product development (Figure 5).

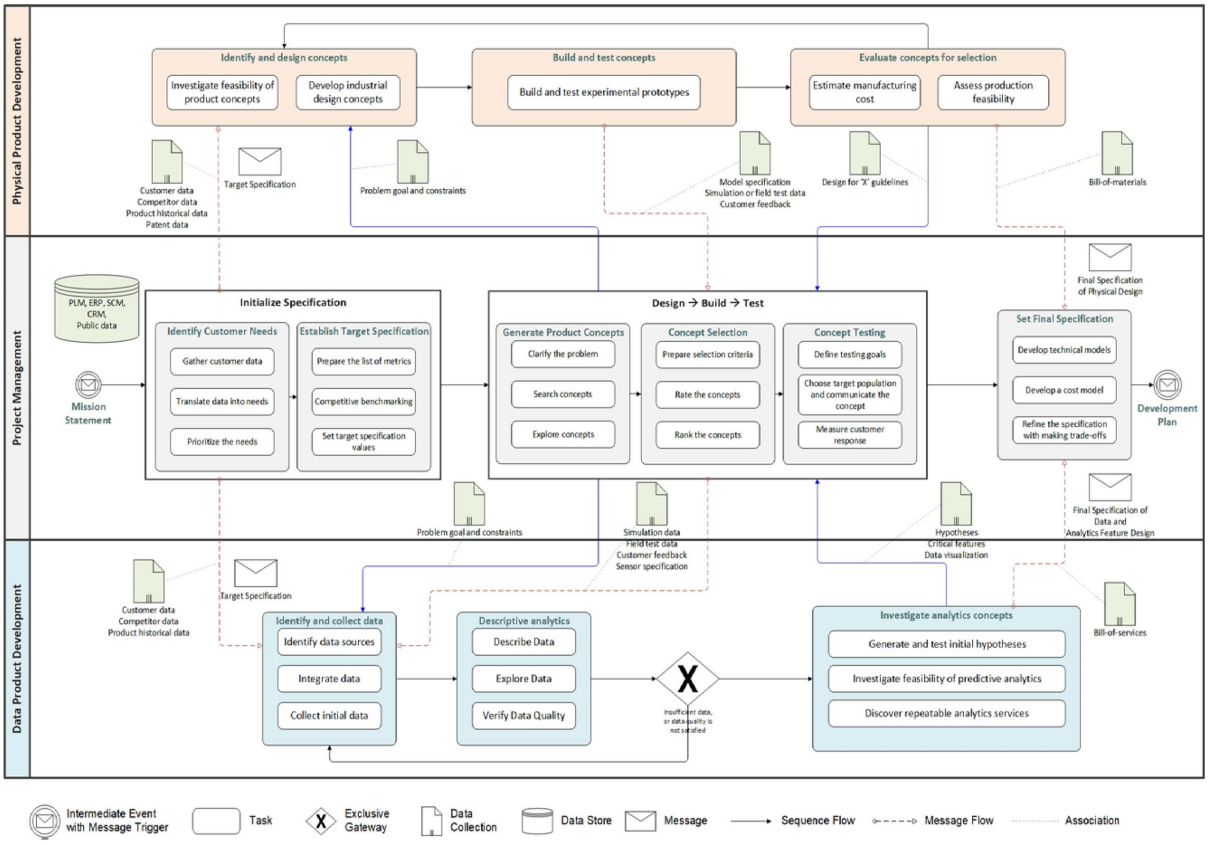

Figure 5: An integrated process model for new product development with data-driven features (NPD3) – concept development (Li et al., 2019).

Iterations are common in product development processes as information flow is constant throughout the problem-solving phase. Pahl and Beitz (1996) highlight information that are processed using analysis and synthesis while developing a solution: concept, calculation, experiment, elaboration of drawing layout, and evaluation of a solution. Iteration is a step-by-step process for approaching a solution from a prototype. Steps are repeated using a higher level of information based on the results of previous loops (Figure 6) so that the solution can be refined and improved continuously.

Prototypes are product iterations where the time and cost of building and evaluating the prototype must be weighed against anticipated benefits. Products high in risk and uncertainty due to the high cost of failure, new technology, or revolutionary aspects should be considered for such prototyping (Ulrich & Eppinger, 2012).

Academic design literature used to emphasize on the design of new products, starting from a blank sheet of paper (Ulrich & Eppinger, 2012; Wright, Duckworth, Jebb, & Dickerson, 2005; Pahl & Beitz, 1996; Cross & Roy, 1989). In the last millennium, design reuse has emerged and has been cited more frequently in the literature ever since. Otto and Wood (2001) presented a methodology that highlights the importance of changes in the product development process by applying reverse engineering and redesign. Products are redesigned with the vision for market or evolution adaptation, followed by modeling, analysis, and experimentation on product performance. Accordingly, an alternative sequence for studying reverse engineering and redesign in product development was developed, which included three main phases: reverse engineering, developing a redesign, and implementing a redesign (Figure 7).

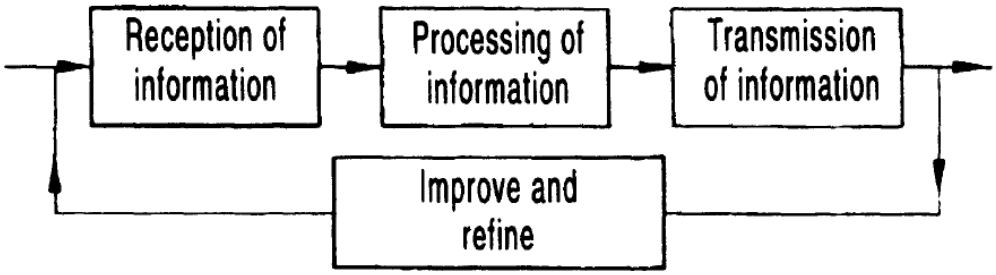

Figure 6. Conversion of information with iteration (Pahl & Beitz, 1996).

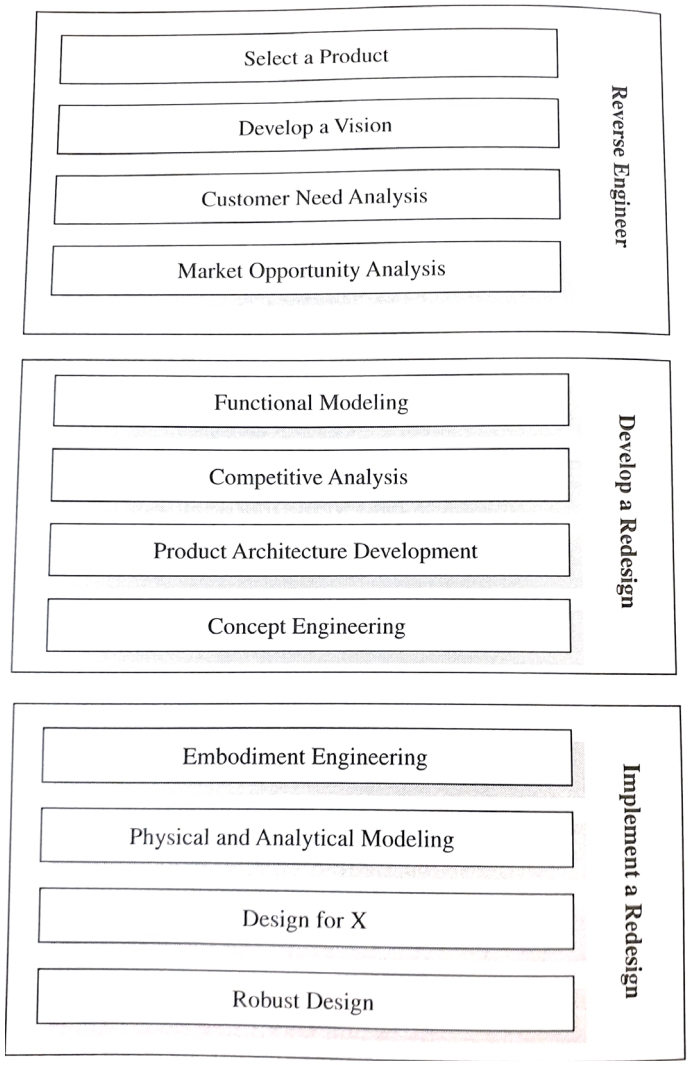

Figure 7: Reverse engineering and redesign product development process (Otto & Wood, 2001).

2.2.1 Engineering changes

The topic of engineering change started to gain popularity soon after the millennium due to the emergence of concepts such as concurrent engineering, product platform design, and simultaneous design.

Product development projects are often meant to enhance an existing product. The documents used to initiate such a change process are known as ECRs. Companies regard engineering changes as sources of problems in the product development process, both during the design and manufacturing (Acar, Benedetto-Neto, & Wright, 1998). Engineering changes aim to improve, enhance, or adapt the product to opportunities or issues identified (Pikosz & Malmqvist, 1998). ECRs are used to specify desired product changes and keep track of the evolution of a requested change from initiation, search for a solution, verification, and decision acceptance. ECRs thus contain both product- and process-related information. Cross and Roy (1989) elaborated on the cost and risk in the engineering design of products when existing products are adapted to new designs.

A high-level overview of an engineering change process was provided by Leech and Turner (1985) who compared this change process to a project that should only be undertaken if the value is greater than the cost. Engineering change processes are similar at a high level, but slight variations can be seen with regard to product characteristics. The change process for safety critical products focuses more on quality than on low cost (Pikosz & Malmqvist. 1988).

The management of ECRs is part of the engineering change process, which corresponds to the first four stages of the generic engineering change management (ECM) process of Jarratt, Eckert, Caldwell, and Clarkson (2011) (Figure 8). The final two stages are known as the engineering change order process. The ECM is a six-stage engineering change process that begins with the ECR, identification of solutions, risk assessment, selection, approval, and implementation of solution, followed by a review of the change. Hamras, Caldwell, Wynn, and Clarkson (2013) reviewed methods for the ECM and identified 25 key requirements, including various components of process model building and use. Maull, Hughes, and Bennett (1992) previously proposed a five-step process, while Dale (1982) suggested two main process phases. Ullah, Tang, Wang, Yin, and Hussain (2018) performed a case study investigating risks in product redesign and found that managing engineering changes in batches rather than in a single cluster can be beneficial in terms of duration.

The detailed steps of Jarratt et al.'s (2011) engineering change process (Figure 8) are as follows:

1. An engineering change is requested on paper or electronic form. The requester of the change outlines the reason, priority level, type of change, and component or system involved.
2. Potential solutions to the change are listed to reduce investigation time or state a known solution. Only one solution is chosen with which to move forward.
3. The impact of implementing the new solution is assessed, considering such factors as design, production, suppliers, and budget. Later in the change process, the selected solution is implemented.
4. The change committee approves the solution before final implementation in which a cost-benefit analysis is performed, and key stakeholders are involved.
5. Implementation of the change takes place immediately or is phased in later, depending on the criticality of the change, for instance, if it is a safety issue or if it can be implemented somewhere in the product life cycle.
6. The change is evaluated to determine if the intended effects have been achieved. Lessons learned are documented for future action.

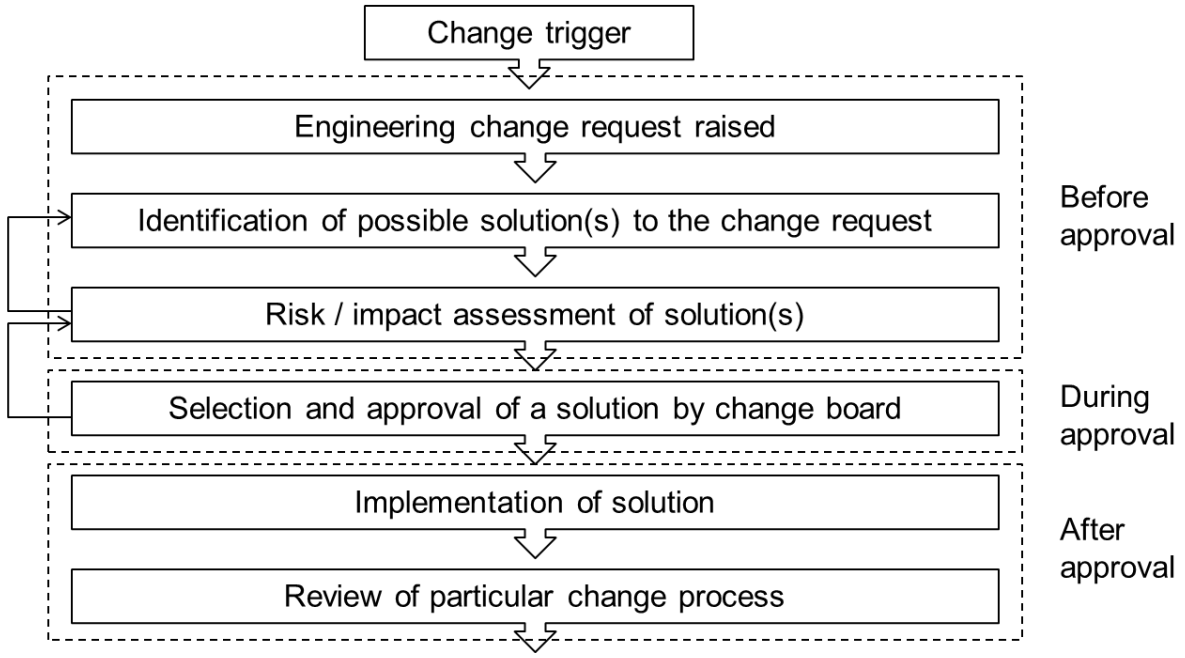

Figure 8. A generic engineering change process (Jarratt et al., 2011).

2.2.2 Engineering change data

As described in the earlier section, the ECR process inputs data during its process. The data itself can be in a form of design input, design output, tests, etc. ECR data in Product Development (PD) projects can be written on paper or in an electronic system. Electronic record logged are stored in databases where structured data (i.e., numerical data and timestamps) are mixed with unstructured data (i.e., free text descriptions). Large development projects can contain tens of thousands of ECRs (Arnarsson et al., 2017).

ECR data have to be collected carefully as key factors must be specified for the data gathering process so that those supplying, validating, and analyzing the data can obtain a consistent view (Basili & Weiss, 1984). ECRs describe the problem and why a change is needed and contain the product/part description, the name and department of the originator, and the date. Basili and Weiss (1984) proposed the following six criteria for data collection to identify troublesome issues and efforts when making changes:

1. Data must contain information that allow the identification of the types of errors and changes made.
2. Data must include the cost of making changes.
3. Data to be collected must be defined according to the clearly specified goals of the study.
4. Data should include studies of projects from production environments.

5. Data analysis should be historical; data must be collected and validated concurrently with development.
6. Data classification schemes to be used must be carefully specified to ensure repeatability in the same or in different study environments.

Change severity level in relation to customer impact is stored in the data. Jarratt et al. (2011) listed four groups of change properties that contribute to understanding the urgency of a change:

- *Error correction:* mistakes discovered during the development life cycle, ranging from minor drawing errors to issues that affect product operation.
- *Change of function:* required when the design does meet its functional requirements. Causes can include incorrect initial assessment or expansion of the operating environment during the design process.
- *Product quality problems:* issues regarding rework and scrap can sometimes be due to poor design, incorrect assembly, or incorrect manufacturing instructions.
- *Safety:* issues with regards to non-commercial boundaries (Inness, 1994). Changes must occur if a product does not meet expected safety or regulatory requirements, which may lead to death, injuries, and property or commercial damage. Hazardous and unintended product usage must also be limited.

Common data stored in ECRs include the change motive, root causes, solutions, parts affected, responsible individual and department, part name and number, report status, severity points, part version, product class, date issued, date of incident, planned closure date, project number, and test information. ECRs also contain transition states, including timestamps (data and time), and have the capacity to handle more than 30 unique states. Each ECR assumes a different state in the resolution process, starting from "ECR created" and ending with "ECR solved." ECR states can be categorized into eight groups (Arnarsson et al., 2018). ECR data include all historical state transitions that ECRs have assumed under the resolution process and state whether an ECR has changed owner (Arnarsson et al., 2018).

Many recent studies have analyzed historical engineering change management data to derive new information. Recent analyses of historic engineering documents affirm potential benefits in the process and workflow management of projects (Snider, Škec, Gopsill, & Hicks, 2017). Interrelations of change information between organizations using structural complexity management and graph-based analysis (Kattner, Mehlstaeubl, Becerril, & Lindemann, 2018).

2.2.3 Opportunities and challenges in ECR processes

Recent studies have identified poor management of requirements (Fernandes et al., 2015) and difficulties in predicting the impact of design changes resulting in the late discovery of problems (Eger, Eckert, & Clarkson, 2007, Giffin et al., 2009) as causes of late changes. According to Thomke and Reinertsen (2012), companies need more time to adjust to the constantly evolving market needs, which can lead to the late detection of product weaknesses. Thomke and Reinertsen further claim that many

companies try to over-utilize their product development resources. When product development employees are nearly fully utilized, speed, efficiency, and output quality decreases (Thomke & Reinertsen, 2012). When resources are highly utilized, queues in projects tend to appear. Queuing may result in the unavailability of resources, longer duration of projects, delayed feedback, and unproductive developers. Conversely, there are many other potential causes, including the lack or poor use of simulation tools (Silow, Rosenqvist, & Falck, 2013), the use of too few physical prototypes or too few milestones, the lack of continuous follow-up, and reporting systems that are too cumbersome, which result in reporting errors.

2.3 Design process models

The ECR process is a type of design process. Wynn and Clarkson (2017) surveyed available design and simulation models to illustrate the rich variety of models. Wynn and Clarkson affirm that detailed, task-based models of design processes can support the design, management, and improvement of "meso-level" processes, including the ECR process.

2.3.1 Design Structure Matrix

Due to the complexities of processes, no single model can fit all. However, DSMs (Steward, 1981; Eppinger & Browning, 2012; Browning, 2016) have been used to successfully construct task-based models of design processes, including stochastic factors. The ECR process has been modeled before with a new product development process, using a stochastic computer model to understand its impact (Huiyan, Gregory, & Thomson, 2006). Design structure matrices support both the qualitative and quantitative analyses of processes (e.g., visualization of processes, computation of process lead times). The main strength of the DSM is its efficient visualization of complex processes characterized by significant amounts of iterations.

2.3.2 Markov chain

The Markov chain is a stochastic process during which the transition probabilities between available states fulfill the Markov property (i.e., the probability of evolving from one state to another depends on the current state). The implication is that the process is "memoryless" and disregards the history of the process. A Markov chain model can be estimated and visualized in a DSM to visually inspect transition pathways for processes.

Markov chain models (Norris, 1998; Gilks, Richardson, & Spiegelhalter, 1995) have many applications in real-life situations, especially when one wants to investigate and understand processes evolving between different discrete states. Markov chain models have previously been utilized for analyzing product development processes (Figure 3) in, for example, Ahmadi, Roemer, and Wang (2001), where the authors employ Markov chains to develop procedures to minimize iterations during the development process, which adversely affect development time and costs. Cho and Eppinger (2001) also used Markov chains to simulate a product development process to ensure better project

planning and control. Meanwhile, Dong (2002) employed ideas from Markov chain models to understand organizational interactions during product development processes. Markov chains have also been used to understand the execution order of subtle signals in a project where workflow is modeled with regards to lead time (Matthews & Philip, 2011).

However, earlier work on DSMs and Markov chains have typically been applied to situation- or system-specific design processes, for example, a brake design process using extended DSM called Work Transformation Matrix (Smith & Eppinger, 1997). According to Smith and Eppinger (1997), generating reliable data for a DSM is challenging and requires additional effort for each new system-specific design process modeled. More recent studies have clustered team attributes in complex product development projects that are modeled through DSM (Yang, Yang, & Yao, 2018).

2.3.3 Opportunities and challenges in modeling process data

Evaluating the ECR process is not trivial as different types of ECRs are routed in different pathways through the system. There is, hence, no single ECR process but rather many. In this regard, ECRs can be characterized as a stochastic process that evolves between discrete ECR states during its lifetime (e.g., under investigation, testing). It is challenging to model sequences in data and almost impossible to do so manually for large projects with almost endless datapoints. The challenge often lies in finding the right method for such modeling to identify best practices and propose improvements to a process. Nonetheless, Markov chain (Gilks et al., 1995) probabilistic models can be used to model how discrete state processes evolve over time and this can help to provide a more holistic insight into the sequences taking place in ECR processes.

2.4 Data mining, machine learning, and design analytics

Focusing on the computational support for ECR analysis, some researchers develop big data mining methods to identify structures or patterns in engineering information.

2.4.1 Data mining application to design information

Fayyad, Piatetsky-Shapiro, and Smyth (1996) provided an overview of data mining and knowledge discovery in databases, elaborating on how the two concepts are related to each other and to other fields, such as statistics, machine learning, and databases. They presented an overall process (Figure 9) for finding data patterns through process iterations to determine which patterns can be considered new knowledge.

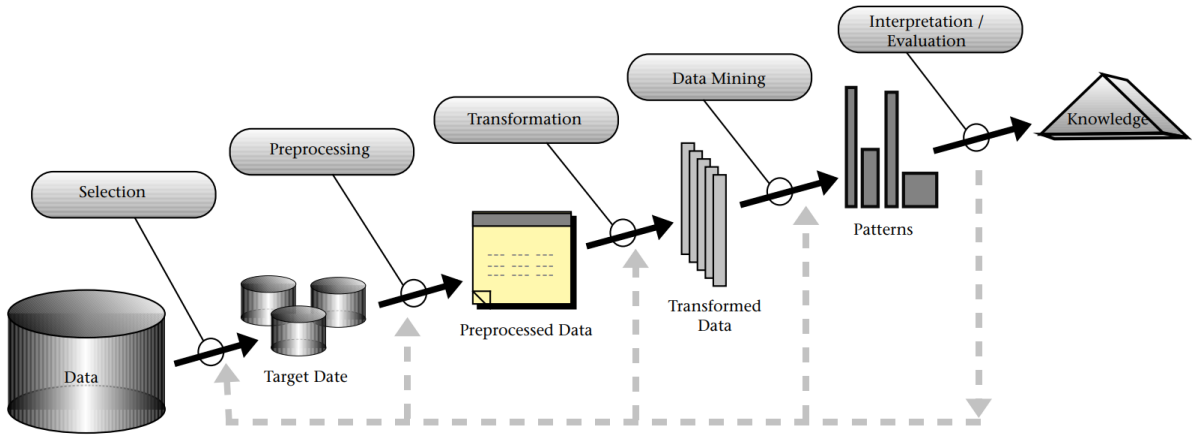

Figure 9. Overview of the process steps that comprise knowledge discovery in databases (Fayyad et al., 1996).

Figure 10 identifies the three components necessary for data mining and data analytics according to Fayyad et al. (1996): (1) data, (2) domain knowledge, and (3) mathematical tools, such as algorithms, optimizations, and statistical models.

Arnarsson et al. (2016) demonstrated how data mining and visualization tools can be applied to explore a database consisting of ECRs from a complex truck development project. The study investigated a process for compiling and cleaning the data, along with methods for numerical and text data analysis, for data visualization and exploration and for pattern identification and analysis.

The application of analytics on product data management systems can enhance the capabilities of these systems as automatic analysis provides information faster for managerial decisions (Snider, Gopsill, Jones, & Hicks, 2018).

Researchers have developed methods to analyze e-mail databases and social media tools to make inferences about project status and connect relevant specialists to queries and issues (Hicks, 2013). Earlier studies have also shown that change requests can be analyzed using network graphs (Giffin et al., 2009). Network graphs can help visualize how change requests are related to one another and show whether they emerge from a single parent or whether they are disconnected. Change analysis during ongoing product development (Eger et al., 2007) use a node-link diagram to allow the designer to monitor the progress of the project. The tool creates a change propagation tree to provide an exploded view of the design links, which help identify change paths. Tree

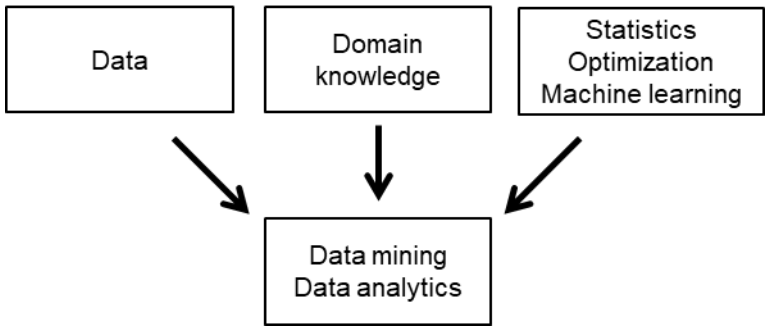

Figure 10. Schematic illustration of data mining and data analytics components (Fayyad et al. 1996).

diagrams and scatter graphs have also been used to analyze data (Giffin et al., 2009). On a more general level, emerging technologies, in particular those for searching and browsing, focus on visualization and other structured presentations of information as key to efficient data analysis. Frameworks, such as d3js (Bostock, 2013), provide building blocks that support tailor-made solutions for the visual representation of text data, word clouds, patent information and the data-driven dynamic manipulation of documents. Kobayashi, Mol, Berkers, Kismihók, and Den Hartog (2018) explained how data mining can be used and how it can help organizational research by allowing the testing of research questions with data to recover useful patterns that were previously not visible due to large amounts to text.

2.4.2 Machine learning

Machine learning is the scientific study of statistical models and algorithms that can be used to perform a specific task to find patterns. Models are trained on specific data to identify patterns much faster than a human could. Extracting design and manufacturing text content have been done successfully using natural language processing (NLP) and node models (e.g., Dong & Agogino, 1997; Catron & Ray, 1991; Kim & Wallace, 2009). Dong (2005) explored design team communication documents using a latent semantic approach.

In general, the development of NLP methods for summarizing and interpreting entire documents have increased in the last few years. Methods for transforming single words into high-dimensional representations have previously been studied extensively (e.g., word2vec, Mikolov, Chen, Corrado, & Dean, 2013). Le and Mikolov (2014) developed this method further by introducing doc2vec, a word embedding methodology where one trains a model to translate entire documents into a high-dimensional numerical representation. This numerical representation of documents can be utilized to compare different documents, find similar documents, and cluster or group documents into different themes. Through methods like doc2vec, two powerful properties that can be achieved: (1) Contextual information can, in some cases, be interpreted, and (2) synonyms to words utilized in a document can automatically be encoded in the numerical representation.

Classical document clustering techniques include k-means (Ahmad & Hashmi, 2016), which uses structured and unstructured datasets to find distance measures between data points, and Newman's (2004) algorithm, which detects and extracts community structure from networks based on the idea of modularity. Latent Dirichlet Allocation (LDA) is more tailored for text-based data and can be considered as an approach to analyze an underlying set of topics in text documents. LDA is a useful method for processing large collections of text and finding short descriptions that can be used to explore statistical relationships for tasks, such as classification, summarization, and judgment of relevance and similarity (Blei, Ng, & Jordan, 2003). LDA is a generative statistical model – a form of unsupervised learning that views documents as bags of words (i.e., order does not matter) and then tries to find clusters that describe the different topics the documents seem to be about (Misra, Cappé, & Yvon, 2008). LDA has been presented as a graphical model for topic discovery, allowing observations to

be explained by unobserved groups. It is useful when dealing with large corpuses and has been shown to outperform other dimension reduction techniques (Blei et al., 2003). Yoon, Seo, Coh, Song, and Lee (2017) used LDA to examine new product opportunities by measuring the semantic similarities between patents and products, creating visual map portfolios that recommend untapped products. Prior studies have applied text clustering to optimize design structure matrices based on PD organizations (Yang, Lu, Yao, & Zhang, 2014) and PD project scheduling (Tripathy & Eppinger, 2013). Sarkar, Dong, Henderson, and Robinson (2014) applied spectral characterization to present a graph theoretic spectral approach that reveals hidden modular layers. Meanwhile, Yang et al. (2018) provided an innovative spectral clustering approach using similarities of team attributes and relationships based on PD organizational structure.

2.4.3 Design analytics

In the seminal book *Competing on Analytics: The New Science of Winning*, Davenport and Jeanne (2007) define data analytics as using statistical and quantitative analysis of data, combined with explanatory and predictive modeling. The models and analyses provide the basis for fact-based management and decision-making. For a survey on data mining and knowledge discovery on a general level, see Han et al. (2011), and for a more technically oriented survey describing techniques and machine learning tools for data mining, see Witten et al. (2016).

Previous information need-focused studies on analytics include Bichsel (2012), who interviewed four focus groups to determine how they relate to analytics. Bichsel's interviews covered data analyses, strategic decisions, decision-making, and culture and politics surrounding analytics. Bichsel highlighted the balance between benefits and challenges that people encounter when working with analytics. Bichsel argued that analytics should start with a strategic question and a plan to address that question using data. Analytics should be considered as an investment and not as an expense. Analytics does not require perfect data, but it should be initiated when there is corporate commitment and readiness. LaValle, Lesser, Shockley, Hopkins, and Kruschwitz (2011) conducted interviews with over 3,000 business managers and analysts from different industries to understand the challenges they deal with and demonstrate how analytics can be used to help their decision-making. The researchers concluded that although the use of analytical techniques has increased, people relate to them in different ways. Similar to Bichsel, LaValle et al. underlined the importance of having a clear question and having organizational readiness and commitment when starting an analytical implementation project as these, rather than perfect data, can guide the entire process.

In the engineering design domain, early studies related to data analytics include Kuffner and Ullman (1991) who identified the need of design engineers for more design information beyond the standard design documents when developing complex products. Reich (1997) proposed a seven-step process for developing machine learning tools that support civil engineering tasks. Similarly, Menon, Tong, and Sathiyakeerthi (2005) developed data mining tools for analyzing textual databases to enable faster product development processes.

The term "design analytics" has been proposed recently by Van Horn et al. (2012) to refer to the area of research that focuses on processes and tools to enhance the transformation of design-related data into formats suitable for design decision-making. Examples include Tucker and Kim (2011) who applied analytics to consumer trend data (e.g., product reviews) to inform product designers. Meanwhile, Bae and Kim (2011) conducted a study on how to optimize the development process of a digital camera by using data mining techniques on customer information. Similarly, Lewis and Horn (2013) explored customer behavior profiles and reflected on customer needs in the late stages of the development process.

Ma et al. (2014) proposed a new demand modeling technique to help design engineers extract knowledge from large-scale data. The model, Demand Trend Mining (DTM), is an analysis tool to "capture the trend of demand as a function of design attributes". DTM can realize Predictive Life Cycle Design for the manufacturing, re-manufacturing, and recycling stages (Figure 11).

Ma and Kim (2014) proposed the application of a Continuous Preference Trend Mining (CPTM) to address challenges in product and design analytics. Similarly, CPTM in Arnarsson et al. (2018) used time stamped data and a predictive model. The CPTM methodology is illustrated in Figure 12.

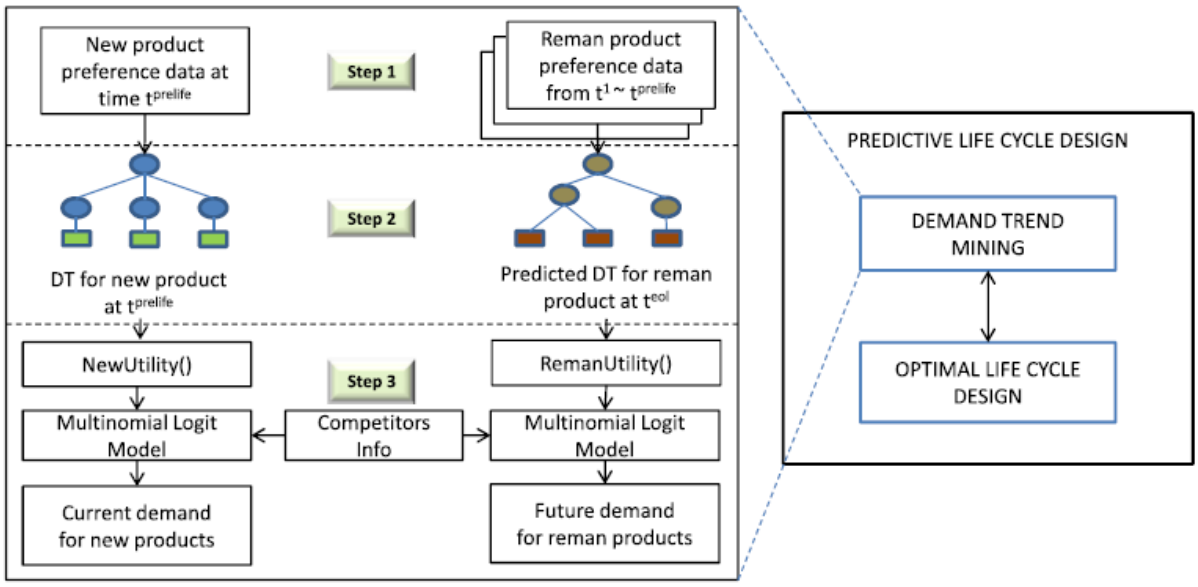

Figure 11. Framework of Predictive Life Cycle Design (Ma, Kwak, & Kim, 2014).

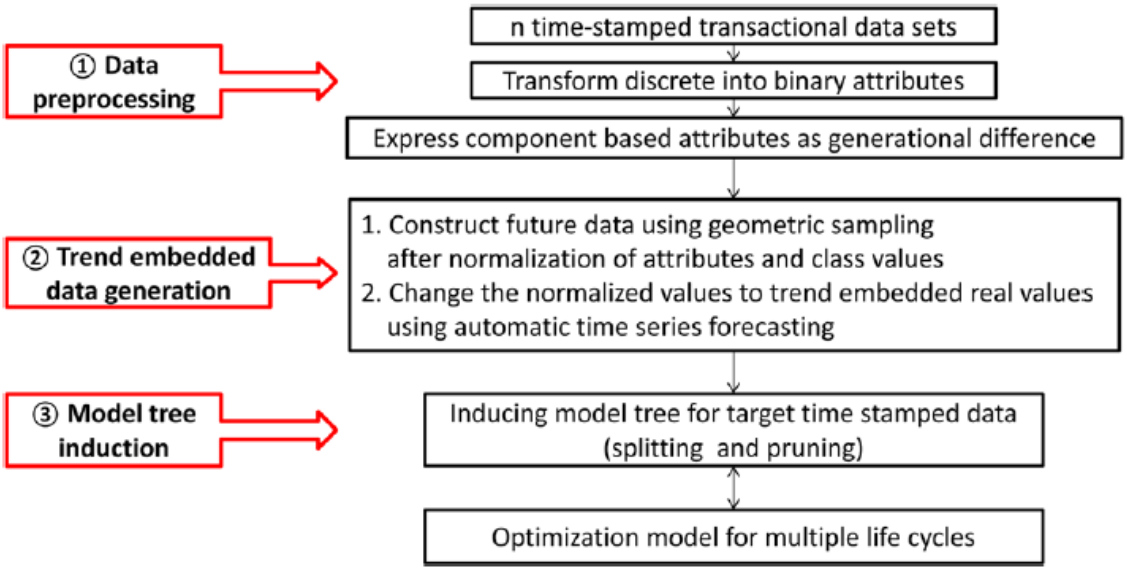

Figure 12. Overall flow of methodology for CPTM (Ma & Kim, 2014).

Zhang, Hao, and Thomson (2015) performed a literature review of big data analytics for product life cycle, particularly product life cycle management and cleaner production. They focused on large amounts of real-time and multi-source life cycle data collected now regarding the manufacturing and maintenance processes of the product life cycle.

2.4.4 Opportunities and challenges in using computational support

Data mining is advancing rapidly, and with developments in the area of machine learning, there are opportunities to develop a method to test search queries that can be used for clustering ECR documents. The challenge is to explore the research gap in performing advanced searches based on NLP and clustering techniques within an ECR database. Such models can ensure better pre-studies for product developers as they can now evaluate a short list of the most relevant historical documents and their topics, which may contain valuable knowledge. Document cluster analysis can be utilized to summarize and group documents with similar content, allowing more effective knowledge management of ECR documents.

2.5 Research gaps

Late changes. Previous work only considered "single" products/systems. Variant rich, platform-based products, such as trucks, introduce an additional level of complexity to the task of understanding the causes and effects of changes and errors. There is thus a need to study this situation in a more detailed and comprehensive way and consider a larger set of causes and mitigations and more complex processes of platform-based development. Accordingly, data from complete projects that include many systems have been used for studies on late changes.

Design analytics. Several studies have been conducted regarding design analytics for product development based on project and customer data, but earlier research did not specifically consider ECR data. As suggested in the literature, analytics should start with a process to determine the relevant strategic questions before performing any analysis. This process is not trivial as there is abundant data, which are not necessarily produced with the intent of ultimately answering a specific question. The research gap is the identification of strategic questions (e.g., hypotheses, information needs) for product developers so that they can apply ECR data in design analytics.

DSM. Statistical ECR DSM analyses have not been performed so far despite the availability of data. Engineers want to have a broader view of the resolution process of an ECR to help them improve the process. No previous research has specifically analyzed data from ECR states in product development, but the literature has identified questions that support the analysis of ECRs. This paper addresses this research gap and aims to apply Markov chain modeling and analysis to ECR databases in product development.

NLP and clustering. Engineering documents (e.g., ECRs and Design Guidelines (DGs)) have not been analyzed before with machine learning tools, such as NLP and clustering techniques. Engineers want to automate searches of documents with related unstructured data (Arnarsson et al., 2017). This paper addresses this research gap and aims to demonstrate how machine learning tools can be utilized on documents generated in product development to identify groups of documents with similar content.

3 RESEARCH QUESTIONS AND APPROACH

This chapter presents the research questions and describes the research approach and methodology applied in this study.

3.1 Research questions

Based on the research focus of this PhD thesis, the following research questions have been formulated with the goal of answering them throughout the PhD thesis. The questions concern the usage of data in product development, the perspectives of product developers, and testing with analytical capabilities.

RQ1. What information needs do product developers have regarding ECR data and what methods can support these needs?

The aim here is to gain domain-specific knowledge about product developer needs from ECR data and the kind of analysis product developers would like to see to help them make better decisions in new product development projects. The research question also aims to elaborate on the methods or tools that can support the analysis of ECR data as described by product developers.

RQ2. What new insights can product developers gain by applying data mining to historical ECR data?

The research question focuses on the process of working with data from extraction to visualization and illustrates some data visualization and exploration ideas that can lead to new insights about ECR data.

RQ3. What are the benefits and limitations of using the Markov chain DSM for ECR process analysis?

The research question focuses on a statistical probability method known as Markov chain that has potential for supporting information needs listed as outcomes in research paper A. These outcomes are then visualized in a Markov chain DSM to draw conclusions on patterns and improvements with product developers.

RQ4. How can NLP and document clustering algorithms be utilized for grouping ECRs, and what benefits can product developers gain from such?

The research question focuses on the use of data mining, NLP models, and clustering of information to support the information needs of product developers listed as outcomes in research paper A. These outcomes are then visualized in a search service frontend for the exploration of related documents and document clusters that product developers can search and filter.

3.2 Design research methodology

Research methodologies serve to address research gaps and questions at hand. Blessing and Chakrabarti (2009) claim that one of the main issues when conducting design research is the diversity of design activities. The chosen research methodology should enable data collection and discuss and answer the research questions. A risk when conducting design research is that topics can lead to multiple pathways and unconnected streams of research (Eckert, Stacey, Clarkson, 2003). The methodology used in this PhD research is related to Blessing and Chakrabarti's (2009) Design Research Methodology (DRM). DRM is a design research methodology used to ensure scientific validity and overcome lack of scientific rigor. DRM is described as "an approach and a set of supporting methods and guidelines to be used as a framework for doing design research" (Blessing and Chakrabarti, 2009). A related research methodology is the qualitative study theory (Maxwell, 2012a). We chose DRM because this research aims to support product developers in their designs, provide insights into their data-driven needs, and test data mining tools in case studies.

DRM strives to fulfill two purposes: to understand the study objective and to propose useful tools and methods to be applied. DRM consists of four stages (Figure 13):

1. *Research Clarification*: Identifies and clarifies research problems and goals that will determine successful research. The main source of information at this stage is a literature study, together with scenarios of desired outcomes.
2. *Descriptive Study I:* Empirical studies are used to create increased understanding of the research problems and goals. The main outcome is the identification of influencing factors and the formulation of models and theories.
3. *Prescriptive Study*: This stage identifies research gaps in the current and the desired situations. The focus of the prescriptive study is to enhance *Descriptive Study I* by using supportive guidelines designed to evaluate previous assumptions and concepts.
4. *Descriptive Study II:* The impact of the proposed study is evaluated using measurable criteria to determine whether the supportive guidelines improve the current situation.

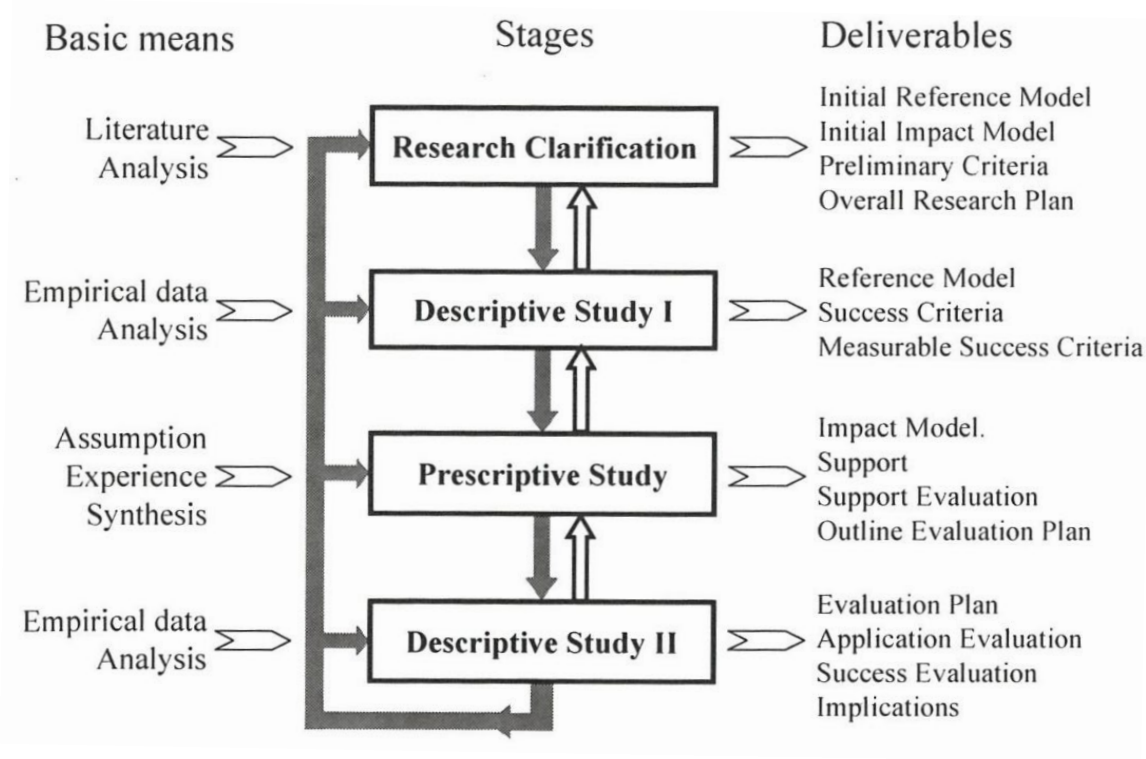

Figure 13. Design Research Methodology framework (Blessing & Chakrabarti, 2009).

All five appended papers are connected by components of data mining and data analytics. Paper A (Descriptive Study I) established an understanding of product developers' information needs to further guide future research. Paper B focused on the first two stages (Research Clarification and Descriptive Study I) of the DRM with literature analysis and empirical analysis. Paper C and D are mainly Prescriptive Studies, with elements of Descriptive Study I, where prototypes are studied with limited test subject based on the outcomes of the previous steps. Table 1 summarizes the methods used for each research question.

A case study is an up-close and in-depth study of a situation over a period of time. It is used to study complex phenomena in their natural setting to gain more understanding. Broad and complex topics are narrowed down to manageable research questions, and in-depth insights are obtained by collecting qualitative or quantitative datasets about the phenomenon. Building on hypothesis for interesting research areas a "case study demonstrator" was developed as a prototype IT tool to evaluate and test feasibility of hypothesis. As we had access to the company and were looking at a technological push this required testing to help with providing early feasibility measures using the company's data. Testing a solution that might work, adapt is and see how well it works. Case study demonstrator can therefore provide an overview of interesting results with a limited timeframe, explore underlying ideas and tools, and demonstrate which of them can be used in real-life scenarios.

Table 1. Research methods and related research questions.

	Research Question	Method	Resulting paper
1	What information needs do product developers have regarding ECR data and what methods can support these needs?	Interview study	Paper A
2	What new insights can product developers gain by applying data mining to historical ECR data?	Case study demonstrator	Paper B
3	What are the benefits and limitations of using the Markov chain DSM for ECR process analysis?	Case study demonstrator	Paper C
4	How can NLP and document clustering algorithms be utilized for grouping ECRs, and what benefits can product developers gain from such?	Case study demonstrator	Papers D and E

3.2.1 Paper A

Paper A is based on semi-structured interviews with experienced product developers. It aims to identify the needs of product developers regarding ECR information. Interviews were conducted to gain an in-depth understanding of product developers' needs and evaluate proposals for data analyses. A heterogeneously purposive-sampled group was selected for the interviews (Figure 14). Twenty interviews were conducted with individuals of diverse experiences in the PD process, ranging from engineering to testing and manufacturing. The interview guide consisted of 16 questions divided into four categories: demographics, behavior, values/improvements, and wrap-up. The demographics section included background questions pertaining to the interviewees' roles and previous projects within the company. Behavioral questions asked how and how often interviewees engage with ECRs, asking them to describe the process for handling ECR reports, including their own involvement. Values/improvement questions were directed towards ECR data and processes, technical difficulties. and recurring errors. Some examples of value/improvement questions are "Is there

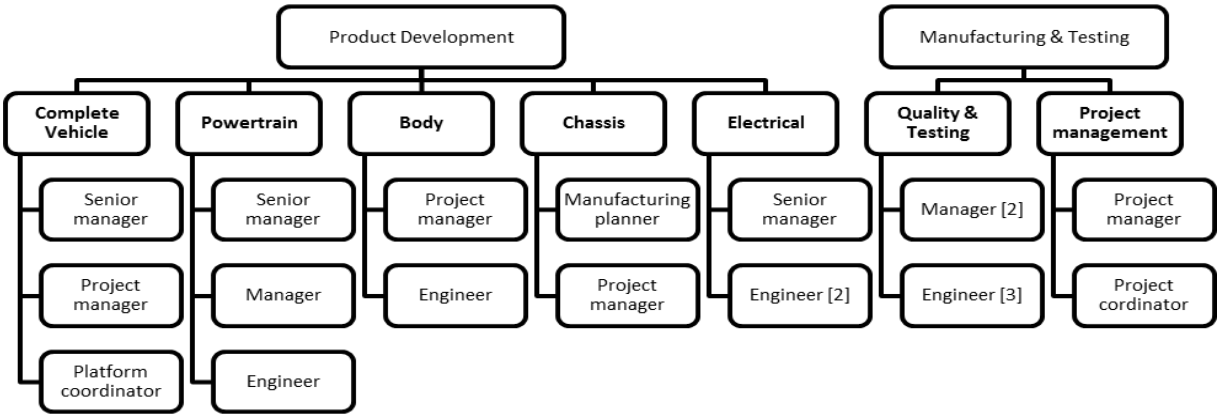

Figure 14. Graph describing the interviewee sample.

something that can be improved in the product development process from which data can be used for learning purposes?" and "Is there a lack of analytics on ECRs from historical or on-going projects to support new decision-making?" The wrap-up section was used for open conversation during which interviewees could speak freely and clarify their own statements.

Interviews were performed in person or over Skype. The interview guide was handed out beforehand to help interviewees prepare. The interviews lasted approximately 60 minutes and were recorded and transcribed. The relative structured interview guide gave the structure to sort the findings into codes, together with addition answers from open questions. Questions formed the bases for the codes that could then be used for counting frequency for answers and draw a conclusion on what topics interviewees agreed upon.

The interview data analyses followed Bryman and Bell's (2018) recommendations and started with codifying and categorizing the topics. Each topic was phrased as a functional information need, such as "identifying repeated ECRs." The topics were then arranged into sections. Finally, two main categories of information needs emerged: (1) needs related to data mining and analytics support and (2) needs related to process and data quality. Results were validated in a workshop with engineers and stakeholders to formulate a strategy for data analysis.

3.2.2 Paper B

In Paper B, a case-based method suitable for investigating complex configurations of events and structures was applied (Brady & Collier, 2010). The case company wanted to know how ECRs can be explored systematically; therefore, an in-depth study was conducted to identify specific behavior in a PD project. The idea was to utilize data mining tools to generate new data-driven insights. We wanted to demonstrate how such approach works and evaluate how well it works.

We selected a single large project to analyze what data are available. Data sources included a prototype build and test report database, a product documentation database, and other individual documents, such as time plans and department descriptions. This test sample was selected due to the relatively large size of data (i.e., around 4,000 ECRs) from a large, recently concluded truck development project.

For the preprocessing, transformation, and mining of data, many data mining tools were considered. Commercial tools included Hadoop, SAS data miner, and IBM SPSS, while open source tools included Python, RStudio, and RapidMiner. We chose Python as it is an open source; it has appropriate data mining libraries, such as Matplotlib (Hunter, 2007), entity extraction (e.g., Maynard et al., 2001), and information retrieval methods (e.g., Nadeau & Sekine, 2008) to analyze and find patterns; and it can perform computational analysis on frequent parts and words (Van Rossum & Drake, 1995). The method used was based on counting frequency of unstructured data (words) in a combination with time stamps, filtered for parts, functional unit or design team. All computational parts were done in Python, using library Matplotlib (Hunter, 2007) was used to create the visualizations and for text mining the entity extraction (Maynard et al., 2001) and information retrieval (Nadeau and Sekine, 2007) methods were used. The research partner is also competent in using the said programming language. Similar or the same results could have been achieved through other IT solutions. Flowchart of the methodology can be seen in Figure 15.

The results were presented and evaluated in a workshop with 10 engineers with a range of roles within a typical PD project to identify which type of analyses have value to them and how to proceed with future work. Feedback were gathered, and results were discussed to answer the following questions:

- Does this approach work?
- Does it enable new insights?
- What patterns can be identified?
    - Peaks in ECRs, causes of failures, or frequent issues

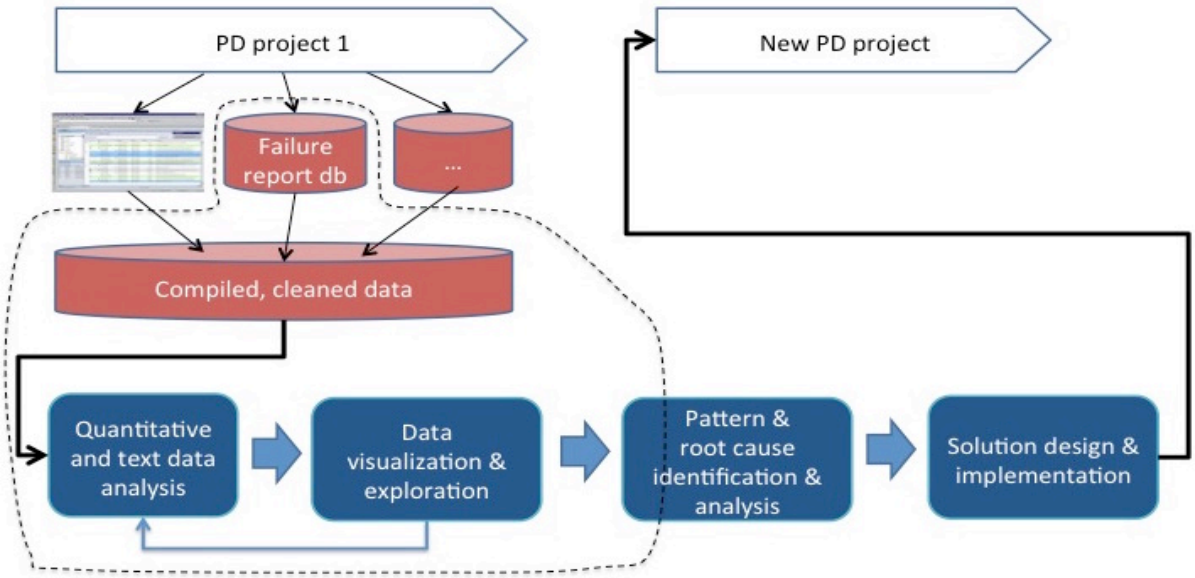

Figure 15. Flowchart of the methodology. Parts within dashed lines were tested in this case study.

3.2.3 Paper C

Paper C contains an empirical data analysis based on a product development database of ECRs from industrial commercial vehicle development projects. It aims to model process pathways for ECRs and compare patterns identified for multiple projects. The Markov chain method was selected to calculate the probability of transition between states and DSM was selected for this study because the approach allows for a scalable way for visualizing ECRs and ease of readability from the row and column structure. As such, the manual analysis of a few ECRs at a time was avoided. Transitions of ECR states are mostly random; therefore, it is suitable to use the Markov chain as it can model randomly changing systems by assuming that the future state only depends on the current state (Gilks et al., 1995).

Data from four product development projects were analyzed, consisting of 620 to 2,949 ECRs with over 100,000 state transitions. A state transition of when an ECR changes from one state to another during the engineering change process. Due to data privacy reasons, fictional status names and numbers were introduced into the matrix although they still reveal similar information. Data within the ECRs contained state transitions that include timestamps (date and time), with the possibility of taking on more than 30 unique states. Each ECR was assigned a state under the resolution process starting with "ECR created" and ending with "ECR resolved." Each ECR contained data about the range of historical state transitions. Table 2 and Table 3 explain in detail the main categories of ECRs, the process structure and the name of each state. The state names are anonymized for commercial confidentiality but fully represent and explain how they are structured in industry.

Table 2. Main phases for the process states used in the Markov chain DSM.

1 ECR created
2 ECR re-issued
3 Identification of solution
4 Verification by factory
5 Verification phase
6 External ECR
7 ECR outside of project scope
8 ECR solved

Table 3. Shows where ECR state names from data fit within the engineering change process.

Engineering change process	ECR state names
Before approval	1 ECR created
	10 ECR distributed
	11 ECR on hold
	13 ECR from external project
	15 ECR without solving responsible
	19 ECR with solving responsible
	2 ECR re-issued
	21 ECR incomplete
	22 ECR not approved
	3 Identification of solution
	31 Assessment of solution
During approval	35 Decision on assessed solution
	35 Solution approved
	37 Testing solution
	39 Solution ready

The method use was based looking at structured data (time stamps) that showed transition paths between states. All computational parts were done in Python, including: cleaning and preparation of the data, Markov chain model was estimated from data and visualized in a DSM.

The data regarding ECRs were stored in corpus databases, extracted into Excel files, and loaded into Python for cleaning and preparation. The Markov chain DSM was computed in Python and then estimated on the data.

For this industrial case study, we decided to limit our scope to the first three categories: ECR created, ECR re-issued, and identification of solution (Table 2) because it is more beneficial to compare sequences until the identification of possible solutions. After which, the verification process started. However, due to the many verification methods used, it was harder to identify process improvement for different methods of verification. Detailed state names for the three first categories are listed in Table 3 where they fit in Jarratt et al.'s (2011) engineering change process. The order of state names was established according to the company's process and were not resequenced. One task in a DSM analysis is to resequencing the initial sequence to an "optimal" sequence (i.e., minimizing the length of iterations), but this was not done here.

The Markov chain DSM was utilized to understand the normal flow of ECRs (i.e., what seems to be the most common transition patterns). The testing of results was done in a workshop with a company process expert who had project experience and process knowledge. Questions were asked, and feedback were gathered regarding which patterns deliver values, thus determining what is normal or a deviation.

3.2.4 Paper D

Paper D uses machine learning parts, such as NLP, to correlate and rank similarity between reports from product development (ECRs) and design guideline (DG) databases. Identifying relationships between documents can help engineers make better decisions in future projects. This paper aims to test two models for search engine and document embedding and show examples of search queries that can be further used for correlating documents and manually evaluating the results. The problem was the difficulty in determining the optimal algorithm for identifying these relationships between documents; hence, comparing the algorithms was necessary.

ECRs and DGs were stored in separate databases. These two documents types have different goals: ECRs focus on achieving a rapid process, from the identification of issues to their resolution, whereas DGs focus on creating a reusable asset to decrease the risk of repeating the same mistake and establish best practice guidelines.

The methodology and the process used in this research involved feeding the data from the two databases into two streams, representing a search engine and document embedding models that were then tested. The DG database had a more complex scheme as the format of these documents is structured more loosely; they therefore needed more complex data extraction, as represented by the "DG data extraction" box. Notably, the two models had a common interface denoted as "search service."

The ECR data contained around 10,000 documents from recent development projects. Each ECR consisted of more than 50 variables that were used to document and follow up on the logged change. The documents contained log files of design- and test-related issues that need to be resolved. They also contained other variables, such as title, part name, problem description, root cause, solution, and test results.

The database of collected design guidelines was structured in a spreadsheet where each row containing a "good-to-know" item was further referred to as Knowledge Element (KE). Each KE contains a rationale, including Declarative Knowledge (Know-What to do), Procedural Knowledge (Know-How to do it), and Causal Knowledge (Know-Why it should be done) (Alavi & Leidner, 2001; Lundvall & Johnson, 1994). Additionally, each KE consisted of illustrative images and references to other individuals or sources that can support the reuse of knowledge. In total, 25 DGs were analyzed in this study, and together, they comprised a total of over 800 KEs.

For the search engine (i.e., standard search methods that rely on matching exact words), the pipeline was somewhat simpler. The search engine used was Elasticsearch (2018) that has all the functionalities required to build a search-based application. Similar to the normalization step in the document embedding module, the input data to the model was lowercased, and all the stop words were removed from the text to eliminate noise otherwise caused by these common terms. The search engine was part of this project, and it provided some benchmark results for validating the document embedding module.

While the search engine serves as the baseline for a typical information retrieval task, such as the one presented in this paper, the document embedding model represents a novel approach to information retrieval tasks of this kind. The document embedding

model is simple to use, and because it outputs a numerical representation of the text, the output can be combined with any other data mining or feature extraction method that also produces numerical data. As such, the document embedding model can be utilized not only for information retrieval tasks but also for other analytical tasks involving ECRs, such as classifying or clustering.

The results were then validated together with domain knowledge experts from the company since the text in documents was case specific. One domain knowledge expert validated a single case. Validation criteria for acceptance were set to three levels: green (good relation, two or more alike), yellow (some relation, one alike), and red (no relation). This research evaluated the possibility to find and analyze correlations between ECRs and DG databases.

3.2.5 Paper E

Paper E uses NLP and document clustering on a PD database containing ECRs from real commercial vehicle PD projects within an organization. The NLP search method enables users to easily search through the data within the database using simple queries. We focused on unstructured data as previous findings suggest that current IT systems are deficient in searching and retrieving relevant documents based on text data. The methodology and the process used in this research started with data from the ECR database undergoing preprocessing in which stop word removal, lemmatization, cleaning, and concatenation were applied.

After preprocessing, data were fed parallel to the Elasticsearch (i.e., search engine) index and then used for the training of a doc2vec model. Doc2vec is an unsupervised algorithm to generate vectors for sentences and can then be used to find similarities between sentences. The resulting doc2vec model was indexed to enable fast retrieval of the document vectors. The doc2vec and Elasticsearch indexes were queried by the search service, which contained the search and the clustering APIs. These APIs served the search frontend application.

The ECR data contained approximately 8,000 documents from recent development projects. Each ECR consisted of more than 50 variables that were used to document and follow up on the logged changes. The documents contained log files of design- and test-related issues that need to be resolved. They also contained variables such as title, part name, problem description, root cause, solution, and test results.

The search application considered for this project consisted of three modules: the Elasticsearch, the doc2vec model, and the clustering application. These modules are linked by a frontend search service that can pass queries to both the Elasticsearch and the doc2vec modules and summarize results using the clustering application.

For the doc2vec module (i.e., the model that translates each document into a high-dimensional numerical representation), a pipeline consisting of three steps was used (Figure 16). First, the data cleaning and normalization step reduced the cardinality of the set tokens in the dataset, thus reducing dataset complexity to improve the robustness and accuracy of the final model. This step included lowercasing, removing non-letter characters, and lemmatizing using the WordNet lemmatizer (Bird & Loper, 2004).

Second, a doc2vec model (Le & Mikolov, 2014) was trained using the gensim library (Rehurek & Sojka, 2010) on the normalized data. Third, this model was used by the search service that was given a query; it then transformed the said query into a document vector that could be matched against the embedded documents of the model.

For the search engine (i.e., standard search methods that rely on matching exact words), the pipeline was somewhat simpler. The search engine used was Elasticsearch (2018) that has all the functionalities required to build a search-based application. Similar to the normalization step in the doc2vec model, the input data to the model was lowercased. Additionally, all the stop words were removed from the text to eliminate noise otherwise caused by these common terms. As part of this project, Elasticsearch provided some benchmark results to validate the doc2vec model. While Elasticsearch serves as the baseline for a typical information retrieval task like the one presented in this paper, the doc2vec model represents a novel approach to information retrieval tasks of this kind. The doc2vec model is simple to use, and because it produces a numerical representation of the text, the output can be combined with any other data mining or feature extraction method that also produces numerical data. The doc2vec model can therefore be utilized not only for information retrieval tasks but also for other analytical tasks involving ECRs, such as classifying or clustering.

After performing the search and generating a list of documents from either the search engine or the doc2vec model, these documents were sent to a clustering algorithm. In this study, we utilized the LDA for clustering the documents. LDA is a generative statistical model – a form of unsupervised learning that views documents as bags of words (i.e., order does not matter) and then tries to find a number of topics that represent the different topics in the document list.

The results from the algorithm generated a set of clusters (i.e., a set of topic words) and each cluster's corresponding documents. The results were displayed in the developed frontend application where the user can easily click around, analyze the document clusters, and determine how they are connected to each other. Results were then validated together with a domain knowledge expert from the case company, as the text in the documents was a unique case. The domain expert examined the documents for each cluster and assessed whether they are related to the cluster. We focused on evaluating whether the results are relevant up to four documents in each cluster.

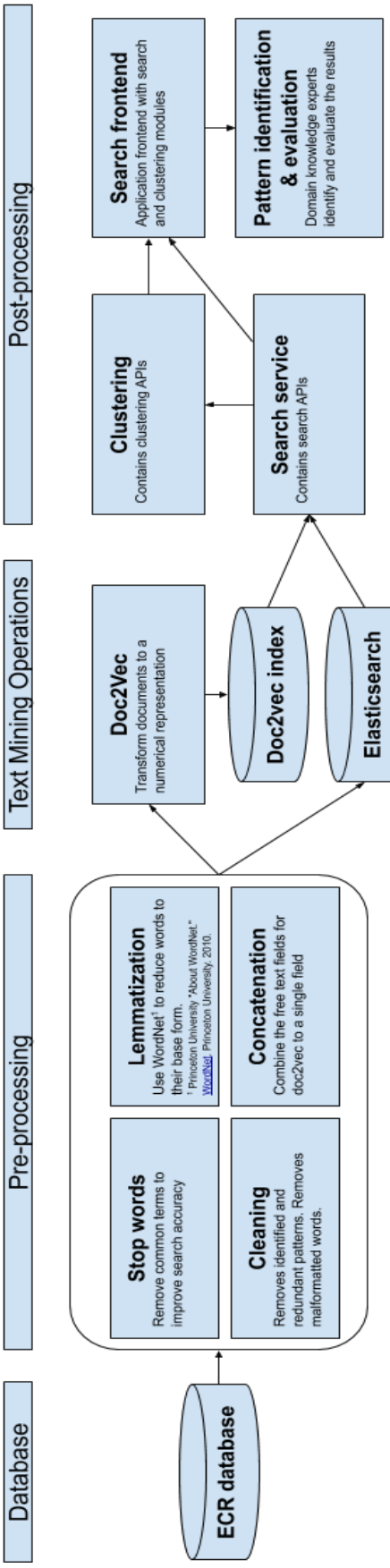

Figure 16. Flowchart of the methodology used.

4 SUMMARY OF APPENDED PAPERS

The five appended papers are parts of a three-part project aimed at utilizing design analytical tools for analyzing ECR data to guide product development projects. Each paper builds on the findings of the previous paper and covers aspects necessary to perform data mining or data analytics.

Paper A identifies the "domain knowledge" or the need of product developers for ECR information during development projects and what kind of analysis can help them make data-driven decisions in future PD projects.

Paper B focuses on the first "data" part and examines a database containing ECRs issued during a product development project to investigate how the data can be explored and visualized.

Papers C, D, and *E* apply the final "mathematical tools" component to perform data mining or data analytics based on the previous domain knowledge.

4.1 Paper A

The **aim** of Paper A was to identify the information needs of product developers through design analytics. Companies look for improvement opportunities within their businesses or products through data analysis to support engineers in their daily work.

The **challenging** part was finding out what specific information needs product developers have since the company lacked studies on the said topic. Employees use currently available tools, so there is a knowledge gap between what they know and what they do not know regarding information needs. Participants selected for the interviews possess broad and diverse knowledge, and this enabled the identification of needs from line organization to managers. We then structured and formulated the interview guide with open-ended questions to ascertain their information needs and guide data analysis. Domain knowledge is important before starting an analytics journey to identify beneficial and meaningful outputs that support developers in making better decisions for new product designs/re-designs.

The **key contribution** of Paper A consisted of findings from the interviews that were divided into two main categories of information needs: (1) needs related to data mining and analytics support and (2) needs related to process and data quality (Figure 17). The main conclusion was that the engineers want to identify related ECRs within projects, perform ECR process analysis, and perform searches across multiple databases.

The **piece of insight** generated from Paper A was a list of the top four information needs based on the frequency of the same topic in each interview (Figure 18). These needs directly guided Papers C and D, which were based on the needs ranked in the "product developers need for data mining and design analytics." Paper C, in particular, focused on second highest ranking need to perform ECR process analysis. Meanwhile, Paper D focused on three other needs, which can be combined: (1) identifying related ECRs, (2)

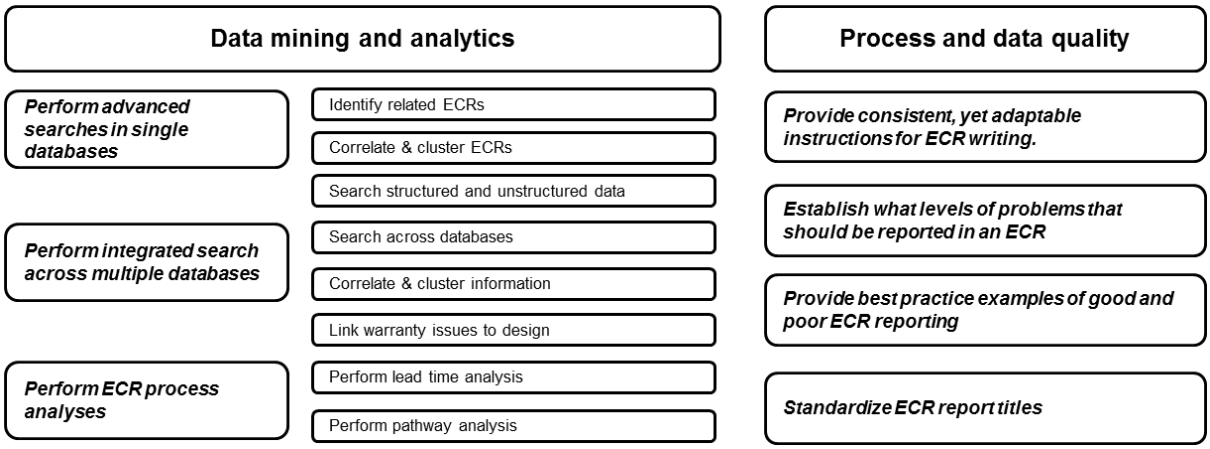

Figure 17. ECR-related information needs and issues identified by interviewees.

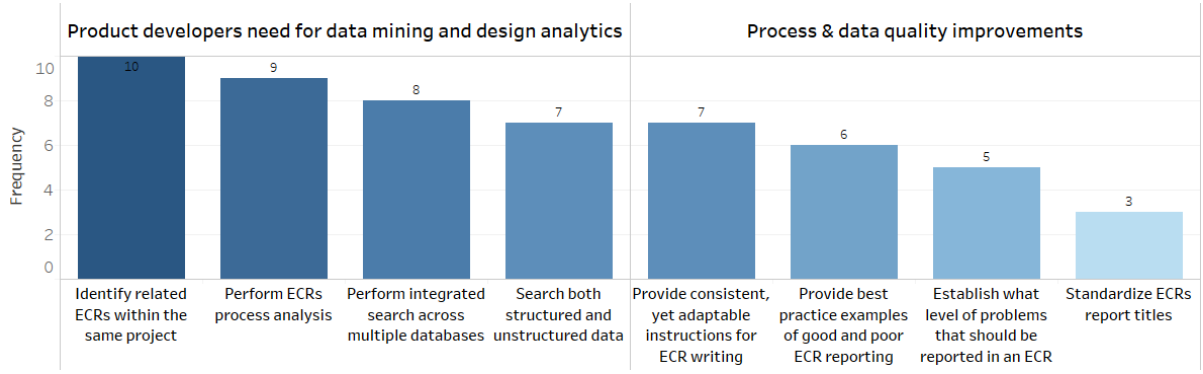

Figure 18. Frequency of information needs topics stated in the interviews.

searching across multiple databases, and (3) searching on both structured and unstructured data. Accordingly, the case company has taken direct actions based on the four findings related to "process and data quality improvements" to improve each one of them.

4.2 Paper B

The **aim** of the paper was to explore the process behind ECR data, from data identification, data quality evaluation, cleaning, and analysis to data visualization and exploration. This process was done to determine how well ECR data lends itself to data mining and machine learning analysis to identify patterns in the data.

The **challenge** was that large, complex system development projects, such as complete truck development projects, take several years to execute. During a project, many design decisions often need to be modified due to the emergence of new information. These changes are often well documented in databases, but due to the complexity of the data, few companies analyze engineering change in a comprehensive and structured fashion. This paper argues that data mining tools can be applied to such analyses and proposes a process for conducting the analysis and using the results for product and process improvements.

The **key contributions** of Paper B were visualizations of quantitative and text analyses, examples of computational text analyses based on report titles, and descriptions of fault/problem/cause/background and additional comments (Figure 19). The figure was produced in Python showing 3 design teams; it shows the frequency of faults (ECRs) over time in a project. To the right of the graph is a list of the most frequent parts mentioned in these ECRs and a list of the most frequent word. The part names and words cannot be displayed due to confidentiality reasons. The fault graph is therefore normalized. The name of parts and words (e.g. A1) are not related but are marked as such to refer to the design team (e.g., design team A).

The **piece of insight** generated by Paper B enabled us to learn which parts and texts were most problematic for each design team. Accordingly, we discovered families of similar parts that reoccur most often in the ECR. The graphs also showed where the peaks of ECRs occurred and how they were distributed during the project. The main conclusion was that text data mining can provide new insights into ECRs by visualizing frequent problematic words and parts and the magnitude of changes on the project timeline.

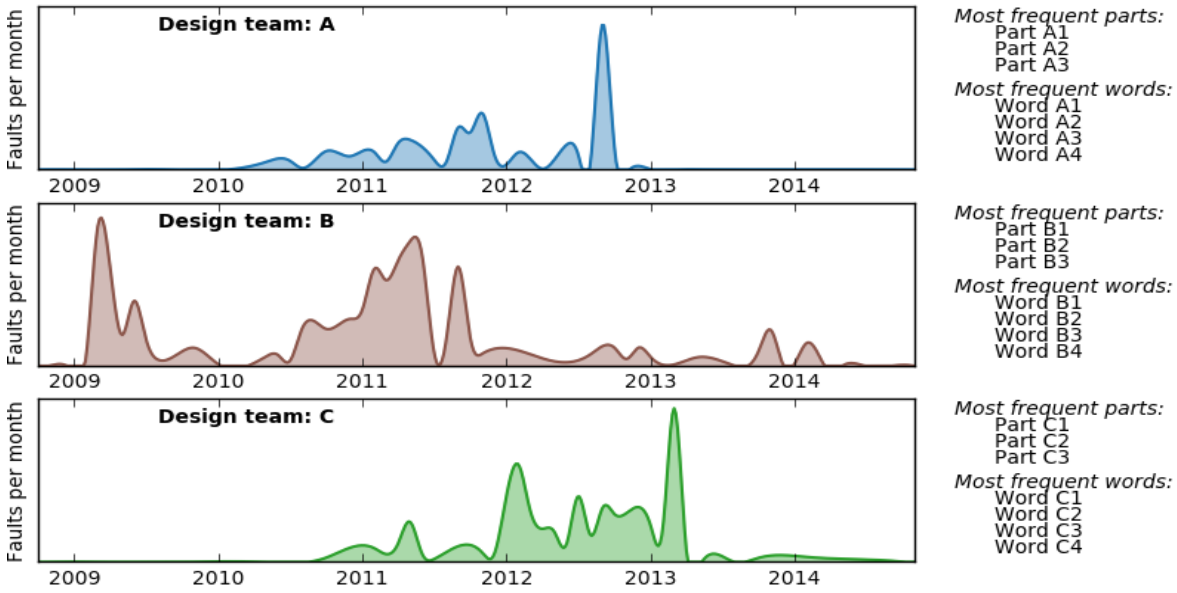

Figure 19. Example of the three design teams displaying a number of issues over time. To the right are the most frequent parts and words listed in descending order.

4.3 Paper C

The **aim** for Paper C was to perform ECR process analysis based on the information needs of product developers and to visualize process sequences.

The **challenge** was that ECRs can be a stochastic process that evolves between discrete ECR states during its lifetime. Markov chain probabilistic models can be used to model how discrete state processes evolve over time. We therefore utilized a Markov chain model on ECR data from four PD projects for process analysis and displayed the results in a Markov chain DSM.

The **key contribution** of Paper C was a Markov chain DSM for each of the four projects, together with an average matrix for the project, which showed the statistical probability of transition pathways for an industrial design process (see example of project B in Figure 20).

The **piece of insight** generated from Paper C was the evaluation of these matrices together with the company's process expert based on case company interest, which enabled us to identify four patterns: most common sequence, prerequisite deviation, process iteration, and deviation from most common sequence (Table 4). Patterns were numbered (e.g. P1) and given a descriptive name and color. Patterns for each project were then compared in Table 5 where one can identify common patterns for all projects, pattern deviation between projects, and unique patterns for specific projects. Table 5 shows a matrix of patterns that can be seen in each project.

Table 4: Descriptive name and color for each pattern.

Pattern	**Name**	**Color**
P1	Most common sequence	
P2	Prerequisite deviation	
P3	Process iteration	
P4	Deviation from most common sequence	

Table 5:Interaction matrix of projects and patterns. X marks an interatction between a pattern and a project.

		Pattern			
		P1	**P2**	**P3**	**P4**
Project	**A**	X	X		
	B	X	X	X	
	C		X		X
	D	X	X		

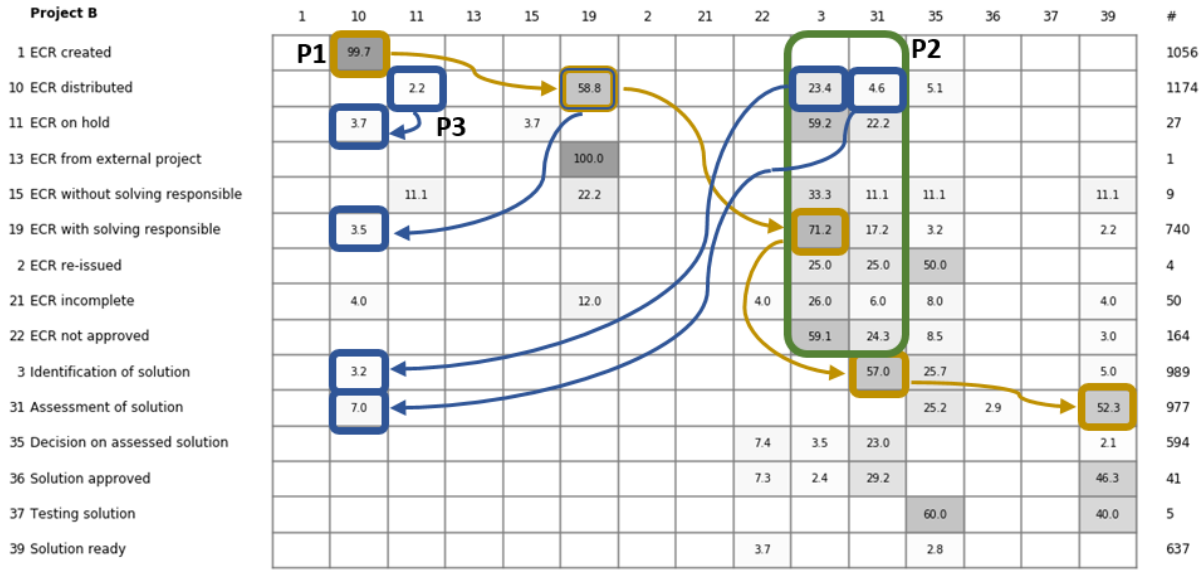

Figure 20: Sequence visualization for Project B with 1,056 ECRs. The most common sequence is indicaded by brown arrows, prerequisite deviation is indicated by green arrows, process iteration is indicaded by blue arrows, while deviation from most common sequence is indicated by red arrows.

4.4 Paper D

This paper was based on the information needs of product developers from Paper A. Product developers wanted a way to obtain a short list of similar documents as they usually use more than one database. The **aim** was therefore to utilize data from two databases and test different applications to determine which one performs better in ranking similar documents.

The **challenge** was selecting the right model that can analyze unstructured data in these engineering documents to effectively create a similarity score and set a threshold to group these documents. Two models were selected and tested: Elasticsearch and document embedding module.

The **key contribution** of Paper D was a flowchart of a methodology building the search pipeline, example how a query can be structured and results showing most similar documents and distance score (Table 6). It is hard to pinpoint a threshold in the score for a relation between documents since the scores for the DEM are cosine values representing a distance, whereas Elasticsearch scores differ greatly between short and long queries. The results were then verified by comparing similar document results to that of the query. Results showed differences when using long and short queries in combination with DEM and Elasticsearch (Table 7).

The **piece of insight** generated from this paper was that the clusters of related documents were considered as an effective way for product developers to gain quick automated insights into previous designs and current DGs. Automated searches and

grouping of documents benefit product developers looking for similar documents by affording them time savings and knowledge management.

Table 6: Example of the results format when the DEM is queried by an entire document.

Case	Title	Description	Score
90304	Difficult to push the amplifier branch through main hole	Fault/problem/cause/background During harness routing, we have to push amplifier branch through a small hole. This operation is damaging to wires and time consuming, the hole is too small. Action: Enlarge hole on amplifier side. Secure the harness by changing the clipping point.	1
92388	Not enough room to push sleeper harness in sleeper main hole	Fault/problem/cause/background When routing the sleeper harness in living area, we have to push branches though small holes in the bunk frame. Action: We need bigger holes to not damage harnesses and win assembly time. Comments: Prime issue #258. Test: Log files attached (if SW problem):	0.531
123457	Adjustment of hole in panel needed for panel fastener	Fault/problem/cause/background Adjustment of hole needed for kick panel plastic fastener Action: Add a ~5mm to diameter of hole for the plastic fastener for part 800001 to allow the plastic pin to come through.	0.454
51345	Difficult to fit harness through pillar	Fault/problem/cause/background During harness routing, we have to "Fish" a branch through a small hole on pillar. This operation would need a special tool due to bad ergonomic position and take too much time. Action: Change harness rout to pillar.	0.454

Table 7: Matrix showing the results of 20 random queries with a few words using DEM and Elasticsearch, respectively and displaying case number, number of words queried, and top three scores for each query.

		Document Embedding Model			Elasticsearch		
Case	**Number of Words**	**Score 1**	**Score 2**	**Score 3**	**Score 1**	**Score 2**	**Score 3**
1126	4	0.47	0.45	0.34	24.30	17.05	15.46
229	3	0.48	0.38	0.38	11.02	8.69	8.51
1116	6	0.51	0.51	0.48	30.12	25.16	15.49
1727	3	0.58	0.57	0.57	18.83	14.09	14.04
1115	3	0.65	0.63	0.59	15.91	15.46	14.97
60	4	0.55	0.53	0.52	20.27	19.70	17.96
1822	2	0.62	0.61	0.57	10.07	9.76	9.11
1117	5	0.55	0.54	0.53	11.73	10.37	10.37
301	2	0.61	0.60	0.54	15.38	12.12	10.90
1905	3	0.61	0.59	0.56	17.16	14.41	14.36
1293	2	0.66	0.63	0.63	14.97	14.97	13.17
334	3	0.44	0.42	0.41	11.39	9.99	9.82
1864	5	0.59	0.57	0.57	17.13	16.38	15.53
1853	4	0.63	0.62	0.60	16.84	13.09	12.75
673	3	0.60	0.59	0.58	15.88	15.51	15.51
1397	3	0.68	0.54	0.52	15.64	11.87	11.64
1120	4	0.61	0.58	0.57	25.28	21.14	11.38
3481	4	0.52	0.51	0.50	24.30	20.28	18.77
2119	3	0.64	0.61	0.61	16.59	14.88	14.62
3440	3	0.51	0.50	0.49	13.59	10.40	8.86

Good relation
Some relation
No relation

4.5 Paper E

The **aim** of this paper was to build partially on the results of Paper D by extending the post-processing pipeline to further work with document clustering but scaling back the databases to only use ECR data. As previously mentioned in Paper A, the primary information need of product developers is that they want a way to obtain a short list of documents similar to those they are working on. This paper tested a method for NLP and document clustering to further identify related documents resulting from searches. Method consists of Elasticsearch, doc2vec and clustering APIs.

The **challenging part** was selecting the right model that can analyze unstructured data and cluster engineering documents to effectively find similar documents. A method for document clustering was tested, together with a frontend interface.

The **key contribution** of Paper E was a process of building the document clustering pipeline, from pre-processing and text mining operations to post-processing. Further explaining the NLP and document clustering algorithms used for the prototype IT tool. The tool is then set up with a frontend interface (Figure 21) for queries to be tested. The results were filtered and later verified by a company expert who examined the similarity of documents in each of the clusters. Results showed that the methodology was able to create clusters of related documents with descriptive labels for each cluster.

The **piece of insight** generated from this paper was that the clusters of related documents were considered as an effective way for product developers to gain quick automated insights into previous designs and current DGs. Automated searches and grouping of documents benefit product developers looking for similar documents by affording them time savings and knowledge management.

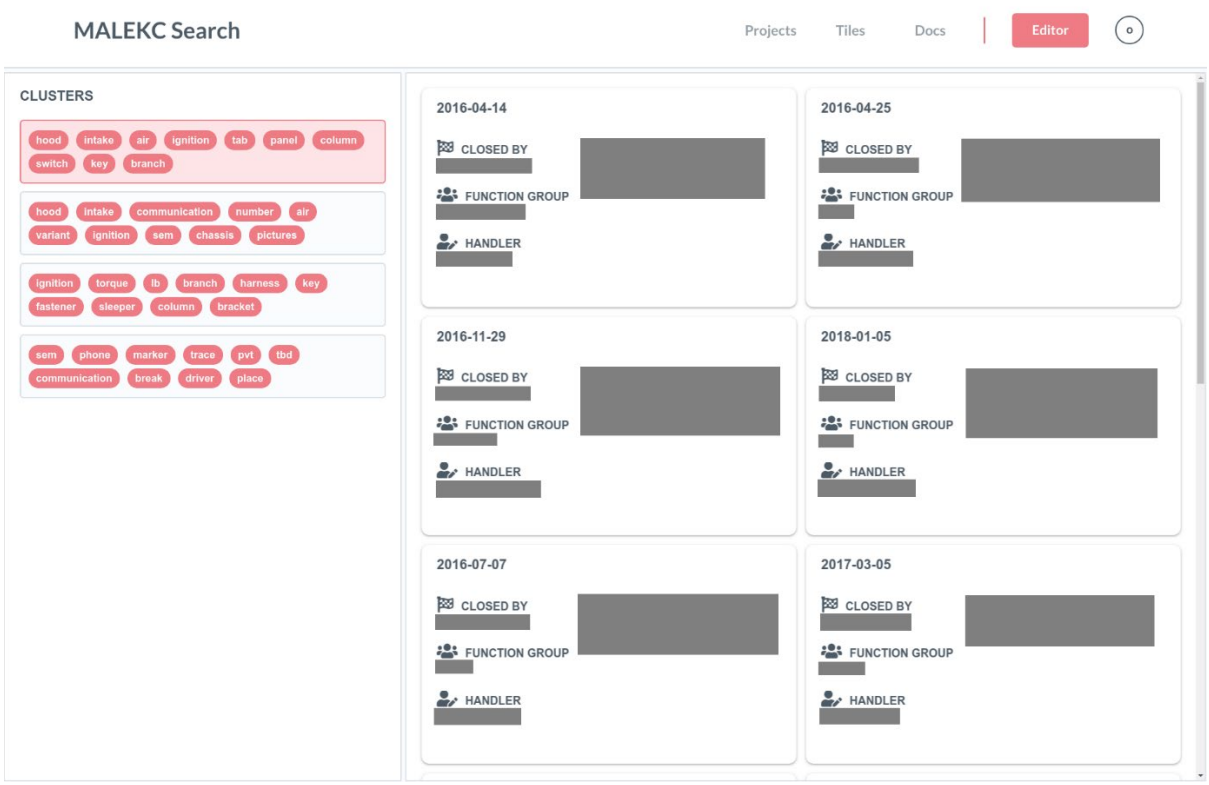

Figure 21: Front end of the clustering application.

5 DISCUSSION

This chapter answers the research questions one at a time. Research quality, validation, and contribution to both science and industry are also discussed.

5.1 ECR information needs of product developers

RQ1. What information needs do product developers have regarding ECR data, and what methods can support these needs?

The **purpose** of Paper A was to gain domain knowledge about the need of product developers for ECR information during development projects. When preparing for data mining, it is crucial to understand how end users or domain experts will use the extractable information. The main **benefits** for product developers is to look for beneficial patterns and meaningful outputs in product development data to support developers in making better decisions for new designs/re-designs and ultimately produce superior and robust products.

We discovered in Paper A that the most frequent request from product developers working with complex PD projects generating ECRs is their desire to connect ECRs with other documents and models related to product development (Figure 18). The interviews support the initial hypothesis to identify product developers' information needs. The findings of this study are slightly different from Bichsel's (2012) interviews, which focused on the benefits and challenges people experience when working with data analysis. Nonetheless, we followed his reasoning that analytics should start with a strategic question and should answer that question with data. In LaValle et al. (2011), business managers interviewed showed a lack of understanding on how they can apply analytics for better decision-making. Our research findings somehow closed this gap and clarified how analytics can support product developers in decision-making.

Results show that product developers want easier ways to find relevant historical information as part of a pre-study before producing new designs. They want to identify related ECRs within the same project and compare ECRs from previous projects. This information can be obtained from various databases from different departments, but comprehensive searchability across databases is missing. This searchability should include structured and unstructured data as there may be hidden information in the systems about previous similar designs and service experiences that current searches do not cover since they are limited to structured data. This searchability can resolve knowledge barriers and enable product developers to conduct their own data. Product developers ultimately want to know what to test for, how much testing is needed, and what potential problems and risks need to be considered before dealing with similar or new designs. They also want to be able to link current warranty issues to the design. They currently perform this work manually, but they want a way that can identify this automatically for them.

Moreover, product developers want to perform a detailed ECR process analysis in which ECR lead times and paths are broken down to and analyzed based on design group levels, groups of parts, or on projects with short and long lead times. Comparisons can be made by examining success factors and outliers and asking why some departments move faster than others in specific project sections. Accordingly, a breakdown of sections can help identify communication issues between departments and departmental cultural differences regarding decision lead times.

Design analytical methods can promote a more systematic and easily accessible use of the data stored in ECR databases, thus improving decision-making in new product development projects. One method for identifying related ECRs is through *information clustering* machine learning algorithms (Ahmad & Hashmi, 2016). Clustering methods can also help identify related information in multiple databases. Meanwhile, *text search and classification tools* can help users access the rich knowledge stored in the text body of reports (Hicks, 2013). Using *pattern identification tools*, which are in a way similar to plagiarism detection tools, can also enable the comparison of completed ECR reports (Giffin et al., 2019; Eger et al., 2007).

Design analytics utilizes available ECR data, such as responsible departments, timestamps and statuses, to understand variation in lead times and pathways when performing ECRs. Analyses can be performed with a DSM using data on departments and ECR process steps. DSMs have been used previously to illustrate complexities in processes where no single model can fit all (Steward, 1981; Eppinger & Browning 2012; Browning, 2016) and have been used to construct task-based models of design processes, including stochastic factors. DSMs offer support for both qualitative and quantitative analyses of processes, such as the visualization of processes and the computation of process lead times.

Path analysis can be performed with data mining to identify slow and fast periods for ECRs and compare these periods between departments to determine best practices. Path analysis can also identify waste in the system, for example, how many ECRs have been written and directly closed thereafter without any action.

As previously noted, there are **limitations** related to the number of processes and data quality issues in the current IT environment. This situation may cast doubt on the possibilities of making accurate assessments of process performance (e.g., if one department wanted to avoid reporting quality issues, thus giving the impression that quality issues do not exist). Researchers argue that effective use of design analytics does not require *perfect* data quality (LaValle et al., 2011; Bichsel, 2012); nevertheless, *sufficiently* good data quality is a prerequisite of efficient design analytics.

5.2 Exploration of ECR data

RQ2. What new insights can product developers gain by applying data mining to historical ECR data?

The study performed in Paper B finds that ECR data lends itself well to data mining since most of the data can be exported as text data. Data mining tools provide various ways of filtering, visualizing, and exploring data. We can identify which parts and departments produce the most ECR reports and which ones have been delayed in the project scope (Figure 19). We can also find patterns, such as repeated ECR reports on the same part, and identify which failure modes (e.g., "wear") are most frequent. The **purpose** of these observations is to provide insights for the identification of the most frequent ECR issues (parts and words) product developers work on and visualize where in the project timeline these ECR issues occur to understand their timing (i.e., whether or not they are late) and explore ideas to help address the root causes of ECRs and late ECR changes (Figures 6–8 in Paper A).

The main **benefit** for product developers is that they are afforded new ways of data visualization and exploration that are not currently done with ECR data at Volvo. As the data patterns are similar to other reports (e.g., Giffin et al., 2009), we argue that the test performed in this study confirms the feasibility and potential of the approach and its applicability to the analysis of complex system development projects in other industries. Such visualizations can assist engineers during projects, giving them an overview of products with the most design changes. Kobayashi et al. (2018) affirmed that data mining provides opportunities to recover potential useful patterns that have previously been inaccessible from larger quantities of text. The findings in Paper B are in line with this assertion – visualizations created from data mining allow us to see new patterns in a project timeline. Another benefit is to review historical projects to analyze if product changes in projects are front or end loaded (Figures 6–8 in Paper B). This is in line with Snider et al.'s (2017) hypothesis that the use of historic engineering documents benefits process and workflow management in projects. For managerial decisions, automatic analysis also provides information faster than before (Snider et al., 2018).

One **limitation** is that further testing is desirable using multiple data sources (e.g., PDM systems, software change request databases, and manufacturing error databases). It would also be interesting to compare ECRs across multiple product development projects to determine whether patterns repeat themselves within design teams. Figure 19 in Paper B shows a visualization of three design teams and the most frequent parts and words in ECRs they have to deal with during a project. The visualization also shows where in the timeline they occurred so that product developers can make estimates (e.g., whether ECRs are early or late). No root cause studies have been conducted to verify if ECRs are early or late. Although data mining tools can provide analytics capabilities and visualization, identifying and analyzing the meanings and benefits of possible patterns still need to be done with the help of domain knowledge experts. These research results fulfill our research purpose and encourage data mining and the exploration of ECR data.

5.3 Application of Markov chain to ECR data

RQ3. What are the benefits and limitations of using the Markov chain DSM for ECR process analysis?

The **purpose** of using the Markov chain DSM was to understand the flow of ECRs based on the different ECR steps. The transition matrix can be used to draw conclusions by noting common patterns or finding anomalies. The Markov chain can also be utilized to compare transition matrices for different projects and mark differences in the workflow.

The **benefit** of applying DSM is that it helps identify the most common transition patterns in the ECR steps. Such insight can be beneficial as it can encourage discussion on, for instance, whether ECRs can deviate from common pathways in the process or take any chosen transition possibility. Early in the process (i.e., before the approval of a solution), ECRs should follow a fairly similar pathway of creation to make sure that they are responsible for identifying a solution. Accordingly, we discovered that the Markov chain DSM has the capacity to identify these deviations. Conclusions are best drawn with the help of engineers who have domain knowledge about the different steps in the process and can identify deviations and areas of interest.

We identified four main patterns in these matrices created for the four projects: most common sequence, prerequisite deviation, process iteration, and deviation from most common sequence. The most common sequence served as a benchmark for patterns across projects as deviations can be detected from it (Pattern 1). Prerequisite deviation was confirmed by the case company process expert as it was a deviation from the typical way of working where a person is assigned as responsible for the solution, but that was not done for part of the transitions (Pattern 2). Pattern 3 indicated process iterations where states return to the previous state. No root causes were identified for this; nevertheless, it is known that product changes can be costly, and the same is assumed for process iterations. Our findings are similar to Huiyan et al. (2006) who also found loopbacks in the ECR control flow. Pattern 4 showed a deviation for the most common process sequence for one project, when compared to the other projects. It is unknown why this deviation exists, but it may be dependent on the project type. Our research findings align with earlier work on Markov chain models, such as Norris (1998) and Gilks et al. (1995) who affirm the model's many applications to real-life situations, especially when one wants to investigate and understand processes evolving between various discrete states. We also agree with Giffin et al. (2009) that managing ECRs is a complex process, and understanding the transitions of ECR is difficult because iterative processes are agile (Beck et al., 2001).

Limiting factors are that it is usually not enough to create a model, domain knowledge should also be applied to the process of analysis. Engineers with domain knowledge need to evaluate possible patterns to understand them and identify pathways that are meaningful for their work. Another limitation of using the Markov chain is that the Markov property is memoryless. However, it is often assumed to be the opposite, which means that when a transition is computed in the model, it only considers its predecessor.

The path to a specific ECR status can affect the probability of transitioning to new states. Notably, an initial analysis indicated that the Markov memoryless property is, in most cases, a valid assumption.

5.4 Application of natural language processing and document clustering to engineering documents

RQ4. How can NLP and document clustering algorithms be utilized for grouping ECRs, and what benefits can product developers gain from such?

The **purpose** was to use document clustering techniques that have been gaining more attention over the last years due to improved algorithms and larger datasets available. In this study, we found that document clustering techniques (i.e., LDA) can be applied to ECRs, and the results are useful. The domain expert validated the clusters found via the LDA method and confirmed that they do summarize the main topics that the ECRs discuss.

The **benefit** of using the NLP and document clustering algorithms tested is that clusters of similar documents can be identified automatically. Traditionally, product developers type in a search query in the form of structured or unstructured text. The search results are then displayed in an ordered list, using a selected variable that can be filtered by structured text. The main benefit of the approach employed in this study is that when product developers type in a search query, they receive a list of documents that is already clustered into groups. This approach saves time for users since similar documents appear in respective clusters. There are several use cases for this approach:

- Knowledge management: When making design guidelines with best practice designs, this approach can help identify related design areas on similar topics so that they can be further documented for future designs.
- Inexperienced product developers: Inexperienced product developers may not always have knowledge of previous work or know where to locate these documents. They can use this approach to gain an insight into similar product features that have been previously designed so that these designs can be considered when they work on a new product or redesign an existing product.
- Design issues: When product developers confront design issues, whether these are related to product quality or development, they can leverage a search like this to quickly identify previous designs in the database that they can use for root cause analysis. The number of documented designs makes this task challenging. The approach we have proposed can help address this challenge by identifying clusters of related designs for users.

Product developers can take advantage of the document clusters to quickly identify documents of value to a specific situation. From an ECR perspective, this approach helps identify historical issues related to a current issue and ensure better pre-studies on historical documents before making a new product or redesigning one. It usually takes

product developers a few hours to make a manual list of related documents, but using a model like this reduces that time to minutes.

Although document clustering techniques have gained more attention recently, **limiting factors** can be the suitability of these techniques for grouping ECRs that contain very domain-specific language is not directly apparent. Moreover, the usual stop words and lemmatization techniques are not optimized. During the study, we also analyzed and tested other document clustering techniques. For example, standard clustering methods (e.g., DBSCAN, k-means) were utilized for the numerical representations derived from the doc2vec model and the latent semantic analysis (LSA). However, the most interesting results were found with the LDA, which is the reason the results presented only the LDA clusters. The results from the other clustering techniques were often very random in behavior, and it was not clear why specific documents belonged to the same clusters. To obtain better results in the LDA clusters, more effort should be spent in cleaning the data to avoid unnecessary information and domain-specific stop words (e.g. "Volvo" or "Truck") in all documents.

5.5 Validity, verification, and transferability of results

5.5.1 Validity and verification

Maxwell's (2012b) eight step checklist was used to strengthen the validity of this research. *Long-term involvement* is a key component in this research as it has been anchored on the close involvement the main research author has had with Volvo over the past four years. By spending most of the research time at Volvo, the researcher has gained insights into the industry and developed a deep understanding of the research subject, which in turn afford credibility to the findings. According to Creswell (2013), the more connected the research is to its environment, the more accurate and valid the research findings will be. *Rich data* have been gathered though interviews with engineers and data experts. Data and results were then reviewed in workshops involving domain knowledge experts at the case company to gain *respondent validation* as findings were discussed and next steps planned. *Interventions* are often performed in field research. The researcher's influence on the study was enriched with discussions with supervisors and company representatives on how to improve, develop, and test the ideas of the topic under study. *Searching for discrepant evidence and negative cases* has been performed through literature reviews of similar studies and analyses of the company's needs. Negative feedback was collected during workshops throughout the research process to guide future work. *Triangulation* was used (e.g. during the study of information needs) by interviewing a broad range of engineers, from line organization to managerial positions, to mitigate the risk of bias regarding stakeholder needs. Information from systems was collected according to the recommendations of engineers involved with product development projects and not selected at random by the researcher. Frequency of topic mentioned in interview have been used as claims from findings were validated by speaking to a broad range of stakeholders in order to validate if the opinions were shared by majority. *Comparison* of multiple cases was performed by comparing results from multiple projects and using more than one result during the

analysis. As such, results can be generalized as more of an average rather than merely presenting possible outliers.

Papers A, B and D have been peer reviewed and accepted by recognized European or international conferences where a public presentation and defense took place and subject experts had the opportunity to comment on the research results. Papers C and E have undergone strict review protocols set by journals. Results have been shared with a broad range of experts in the academia and the industry. Additionally, preliminary findings have been presented before academic and industrial participants at the Wingquist Laboratory.

Verification of design research can be signified by the acceptance of experts (Buur, 1990) when new results are presented and when results mirror reality in the industry. Regarding transferability, this study was conducted in a large multinational firm that develops and manufactures commercial vehicles. The way ECRs are managed is thus typical for other large firms that develop complex systems and products (e.g., aerospace and defense industries). ECR data structures and processes are similar; we therefore argue that the findings related to needs and potential solutions can be transferred to these contexts. However, the product developers interviewed had mechanical or electrical background. It is possible that software developers have somewhat varying information needs. It is likely that the same results will be generated by using similar research methods. The only uncertainty regarding reliability is the industrial focus at that time; interviewees might answer differently due to the shifting focus within the industry.

5.5.2 Transferability

It is difficult to rigorously evaluate any search application since the evaluation criteria depend on the use case. In this research project, manual evaluation was performed on query outputs resulting from the NLP search application and the patterns in the DSMs. The state of the art is more with automated searches into databases. Then the main contribution has been to apply prototype tools to data in an industrial domain. The five papers are connected as components of data mining, starting with user needs, data exploration and ending with prototype IT tools to evaluate the feasibility and effectiveness of them. Nevertheless, throughout the study, there has been no evidence that the method would not be applicable to similar studies involving data analysis where the goal is to find similarities between text-rich documents. Therefore, we believe that the study can be transferred to most cases where unstructured data is accessible. The field of study should not matter since the algorithms query document text.

5.6 Scientific contribution

The general aim of this PhD thesis is to *explore how the rapidly growing field of data mining and design analytics can be applied to ECRs to identify and understand the information needs of product developers for such analysis.* In particular, this thesis has made the following contributions:

- Mapped out product developer needs for data mining and design analytics (Paper A)
    - An important new scientific contribution is the identification of the "needs of product developers for data mining and design analytics" as this has never been identified before with ECRs. The findings suggest that there is a need for text mining in the form of expanding search engines for unstructured data, multiple databases and for clustering of related information. Analysis of lead time in projects and pathways for project states.
- Presented a methodology for the process of ECR data extraction from a database and the steps needed to gain insights from the data (Paper B)
    - Paper B showed that ECR text data mining is possible, and beneficial visualizations can be created to support engineers in their projects. Multiple visualization was created to demonstrate usage of counting frequency of words over time in projects and filter for parts, function units and design teams.
- Demonstrated process modeling for ECRs using the Markov chain DSM (Paper C)
    - Previous studies have utilized Markov chain models to analyze product development processes (e.g., Ahmadi et al., 2001) who employed Markov chains to minimize iterations during the development process that adversely affect development time and costs. Cho and Eppinger (2001) used Markov chains to simulate a product development process to ensure better project planning and control. Meanwhile, Dong (2002) adopted ideas from Markov chain models to understand organizational interactions during product development processes. However, despite the availability of data, statistical ECR DSM analyses have not been performed so far. Engineers want to know how the resolution process of an ECR looks like so that they can improve the process.
    - No previous research has specifically analyzed ECR transition data in PD projects, and we have addressed this specific research gap. The results of this research have guided our efforts to mirror the needs in reality.
- Demonstrated the use of NLP and document clustering of ECR data (Papers D and E)
    - Applying a search application combined with document clustering to identify similar documents has not been done before, specifically on ECR data.
    - The method demonstrated how document clustering of engineering documents can be achieved.

- Results show comparison between using Document Embedding Model and Elasticsearch on short and long queries, ways for ranking cases based on search score and demonstrator for clustering cases.

5.7 Industrial contribution

This PhD thesis identified meaningful data mining and design analytical tools capable of extracting new information from ECRs. Its industrial contributions are as follows:

- Discovered areas of interest to product developers regarding ECRs (Paper A) that outlined
  - Guidance for data mining and analytics innovation activities
  - Activities needed to improve process and data quality
    - Updating ECR instructions for easier use and creating more consistency
    - Creating a best practice guide for writing ECR reports in a form of a one pager to guide the user who felt like the complete user instructions were cumbersome
    - Standardizing ECR report titles was implemented for easier recognition of report content
    - Assessing the data quality of ECR reports and removing several redundant input fields which were not in use anymore according to statistics from database
- Demonstrated how data mining can be used to identify the most frequent parts and words appearing in ECRs over the project lifespan (Paper B)
- Identified tools and algorithms for analyzing ECR processes and clustering ECR reports
  - As-is status for pathways of historical ECR in projects, with ideas for possible improvements (Paper C) that would help product developers to identify earlier design residing in database and save time for future development projects
  - Case study on how ECR can be clustered automatically (Paper D)
  - Demonstrator created based on NLP and document clustering to test the capabilities and effectiveness of such a method (Paper E)

6 CONCLUSION AND FUTURE WORK

This chapter summarizes the results, states the conclusions, and outlines the direction of future work.

6.1 Conclusions

This PhD thesis focuses on how ECR data in the product development process can be used to extract new information to promote product and process improvements for product developers.

The *first research question* identified the information needs and interests of product developers in a large multinational firm regarding ECRs. It demonstrated how questions that are difficult to answer using current IT systems can be addressed by applying data mining and design analytical tools and methods. The interviewees confirmed that the amount of ECR data collected nowadays is not used proactively and that data mining and design analytical tools can help them with this task.

Two main categories of information needs have been identified. The first category is directly related to data mining and design analytical capabilities and includes requests for functionality to conduct comprehensive and flexible database searches. Such databases contain structured and unstructured data for functions that enable integrated searches across multiple databases and support the analyses of lead times and pathways in ECR processes. The second category is related to process improvements and ECR data quality. The developers expressed concern about the quality of and consistency in logging ECR data and wanted to communicate best practices to all groups within the organization. The interviews affirmed that developers want to have more easily accessible tools for analyzing previous projects and gain knowledge of pitfalls and risks before choosing specific product designs. If they are to be used by a majority of product developers, data mining and design analytical tools should be easy to use and learn.

The *second research question* explained the challenges of analyzing large amounts of ECRs containing both quantitative and qualitative data in a systematic manner. Data mining tools provide a means for analyzing large volumes of complex data. Such analyses involve combining data sources and types, cleaning data, charting, analyzing large text datasets, and visualizing and exploring multi-faceted and multi-level data. Potential benefits of such analyses include finding patterns in ECR report data, such as repeated ECR reports on the same part and occurrence of similar failure modes (wear, corrosion, and misfits). This work has so far developed a basic data mining approach to analyze and test ECR data on a limited dataset. The approach has been tested on a dataset consisting of 4,000 ECRs and seems to work well.

The *third research question* focused on modeling ECR data using the Markov chain DSM. Results show that this can be accomplished, and it is useful, particularly in analyzing ECR transitions. We identified four cases as key results of the study. First, the model identified the percentage of ECRs that had been closed directly after creation.

Second, the model confirmed that there is a tendency to skip the identification of the individual responsible for the solution and the most common way of working and instead transition to the identification of a possible solution. Third, the model affirmed another tendency to skip the identification of a possible solution and transition directly to the assessment of a solution. Finally, the model confirmed the possibility of identifying frequently occurring iterations in the transition process (i.e., we see loops in the model where ECRs transition back to an earlier point).

The results show that the Markov chain model is beneficial in analyzing the ECR process. It is possible to learn from these transitions to improve the creation of ECRs so that (1) they are not closed after creation, (2) they can make certain transitions mandatory if they should follow the most common paths, and (3) they can show where iterations in the process can be found.

The *fourth research question* demonstrated how NLP and document clustering approaches, which seem to work well in finding and clustering similar documents in an ECR database. The search queries used consisted of one word that defines a clear direction regarding the information users want to obtain. A demonstrator application was created to demonstrate the method within the industry and evaluate the value of such a model for product developers. The proposed approach of querying words related to engineering documents and clustering them according to similarities through NLP and document clustering algorithms has potential benefits. The approach was performed on an ECR database, and so far, we have found no problems with including additional databases into the pipeline of the search service. The study affirms that NLP and document clustering algorithms work in the cases tested. The company expert performed a manual evaluation of the results and confirmed that the four documents related to each cluster are relevant. More cases need to be explored to determine if there are instances when the algorithm starts to become more inefficient and returns unrelated documents.

6.2 Future work

This PhD thesis offers interesting future research opportunities regarding data in the product development process. Knowledge gained from interviews with product developers can be used to further evaluate existing and develop novel data mining and design analytical tools to support developers. The tools developed are prototypes and more full scale development is needed before they are used by practitioners in the real world. ECR data from previous product development projects are available and can be used to answer other gaps identified.

Observations together with engineers when the current DSM was reviewed underline three opportunities for the use and continuation of the Markov chain DSM and unstructured data searches.

- *Including the quantity of ECRs with the probability of a transition* to enable engineers to identify the volume of ECR traffic transitioning through the matrix

and the size of process deviations. This has been done for a complete row and column of a status but can be broken down further to an individual status on the row or column to allow a more in-depth analysis.

- *Introducing the time element into the model and evaluating lead time between transitions* to break down the analysis of lead time into sections to identify slow and fast project periods. Comparisons can be made on success factors and outliers by asking why some departments move faster than others in specific project sections. Analyzing sections can help identify communication issues between departments, improve the speed with which quality issues are solved, and identify department differences regarding decision periods.
- *Making a risk assessment of ECRs based on the predicted time of solving the ECR issue from creation.* We can use historical information to predict efforts needed for the visual management of ECRs, including resource planning and lead time for solving issues.
- *Further developing the unstructured data search for machine learning in the automotive industry, which was created through a grant from the Swedish innovation agency Vinnova.* The project is called Machine Learning for Engineering Knowledge (MALEK), and this *research proposal* describes how to create knowledge from existing data sources (e.g., ECRs and check sheet data) and utilize this efficiently in the development process. A potential solution is utilizing machine learning algorithms and smart assistants to identify the right knowledge at the right time for the right individual, which, in the context of this project, is an engineer or service technician making an "uninformed" decision. Chalmers, along with the Wingquist Laboratory (WQ), is the main applicant. Fraunhofer Chalmers Research Center is the driving partner of machine learning knowledge, while Rejmes Transportfordon AB and AB Volvo are the main industrial participants and primary users of the research results. The *aim* of this research is to enable a transition from an experienced-based development process to a proactive product management system in which machine learning is used to predict decision points in the development process and provide situation-adapted knowledge customized for these engineering decision points. The tools developed will be introduced to industrial projects for testing and validation. We see further opportunities for such tools by expanding to manufacturing and warranty data sources.

REFERENCES

Acar, B. S., Benedetto-Neto, H., & Wright, I. C. (1998). *Design change: problem or opportunity*. In engineering design conference, Brunel University, UK: Professional Engineering Publishing. Retrieved November 29, 2019, from https://www.researchgate.net/profile/Henrique_Benedetto/publication/322600132_Design_Change_-_Problem_or_Opportunity/links/5a61f2780f7e9b6b8fd41ae8/Design-Change-Problem-or-Opportunity.pdf

Ahmad, A., & Hashmi, S. (2016). K-Harmonic means type clustering algorithm for mixed datasets. *Applied Soft Computing*, *48*, 39-49. Doi:10.1016/j.asoc.2016.06.019.

Ahmadi, R., Roemer, T. A., & Wang, R. H. (2001). Structuring product development processes. *European Journal of Operational Research*, *130*(3), 539-558. doi: 10.1016/S0377-2217(99)00412-9.

Alavi, M., & Leidner, D. E. (2001). Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS quarterly*, 107-136. Doi:10.2307/3250961.

Andreasen, M., & Hein, L. (1987). *Integrated product development.* Bedford, United Kingdom: IFS. Doi:10.1007/springerreference_6775.

Arnarsson, Í. Ö., Frost, O., Gustavsson, E., Jirstrand, M., & Malmqvist, J. (2019, in press). *Natural language processing methods for knowledge management: Applying document clustering for fast search and grouping of engineering documents*. Article Submitted to Journal Concurrent Engineering in December 2019 for publication.

Arnarsson, Í. Ö., Frost, O., Gustavsson, E., Stenholm, D., Jirstrand, M., & Malmqvist, J. (2019). Supporting knowledge re-use with effective searches of related engineering documents-a comparison of search engine and natural language processing-based algorithms. In *Proceedings of the 22nd Design Society: International Conference on Engineering Design 2019*, 1(1), 2597-2606. Cambridge, United Kingdom: Cambridge University Press**.** Doi:10.1017/dsi.2019.266.

Arnarsson, Í. Ö., Gustavsson, E., Jirstrand, M., & Malmqvist, J. (2020). Modeling industrial engineering change processes using the design structure matrix for sequence analysis: A comparison of multiple projects. *Design Science*, *6*, 1-17. doi:10.1017/dsj.2020.4.

Arnarsson, Í. Ö., Gustavsson, E., Malmqvist, J., & Jirstrand, M. (2017). Design analytics is the answer, but what questions would product developers like to have answered? In *21st International Conference on Engineering Design,* 7, 71-80. Retrieved November 29, 2019, from https://pdfs.semanticscholar.org/af88/3bff37952945413aa4efd8edd65314eccc94.pdf

Arnarsson, I. Ö., Malmqvist, J., Gustavsson, E., & Jirstrand, M. (2016). Towards big-data analysis of deviation and error reports in product development projects. In *Proceedings of NordDesign 2016,* 2, 83-92. Retrieved November 29, 2019, from

http://publications.lib.chalmers.se/records/fulltext/240205/local_240205.pdf. Doi: org/10.1109/bigdata.2017.8258521

Bae, J. K., & Kim, J. (2011). Product development with data mining techniques: A case on design of digital camera. *Expert Systems with Applications*, *38*(8), 9274-9280. Doi:10.1016/j.eswa.2011.01.030.

Basili, V. R., & Weiss, D. M. (1984). A methodology for collecting valid software engineering data. *IEEE Transactions on software engineering*, *10*(6), 728-738. Doi:10.1007/3-540-27662-9_7

Bell, E., Bryman, A., & Harley, B. (2018). *Business research methods*. Oxford university press.

Berkhin, P. (2006). A survey of clustering data mining techniques. In J. Kogan, C. Nicholas, & M. Teboulle (Eds.), *Grouping multidimensional data* (pp. 25-71). Berlin, Germany: Springer. Doi:10.1007/3-540-28349-8_2.

Bichsel, J. (2012). *Analytics in higher education: Benefits, barriers, progress, and recommendations*. Retrieved November 29, 2019, from https://library.educause.edu/-/media/files/library/2012/6/ers1207.pdf?la=en&hash=B6E84D1B3A1A0921609BF64F298D741297DA3006. Doi:10.1007/978-3-319-99459-8_2.

Bird, S., & Loper. E. (2004). NLTK: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions* (pp. 214-217). Retrieved November 29, 2019, from https://www.aclweb.org/anthology/P04-3031.pdf. Doi:10.3115/1219044.1219075.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(1), 993-1022.

Blessing, L. T., & Chakrabarti, A. (2009). *DRM: A design reseach methodology*. Berlin, Germany: Springer. Doi:10.1007/978-1-84882-587-1_2.

Bostock, M. (2013). Data-driven documents. *D3. js JavaScript library.* Retrieved November 29, 2019, from http://d3js.org/

Brady, H. E., & Collier, D. (Eds.). (2010). *Rethinking social inquiry: Diverse tools, shared standards*. Rowman & Littlefield Publishers.

Browning, T. R. (2015). Design structure matrix extensions and innovations: a survey and new opportunities. *IEEE Transactions on Engineering Management*, *63*(1), 27-52.

Buur, J. (1990). *A theoretical approach to mechatronics design*. Kgs. Lyngby, Denmark: Technical University of Denmark. IK publication: 90.74 A. Retrieved November 29, 2019, from https://orbit.dtu.dk/files/96900446/Buur_PhD_Mechatronics_Design.pdf

Catron, B. A., & Ray, S. R. (1991). ALPS: A language for process specification. *International Journal of Computer Integrated Manufacturing*, *4*(2), 105-113. Doi:10.1080/09511929108944485.

Cho, S. H., & Eppinger, S. (2001). Product development process modeling using advanced simulation. *Proceedings of DETC'01 ASME 2001 Design Engineering Technical Conferences and Computers and Information in Engineering Conference Pittsburgh, Pennsylvania September 9-12, 2001.* Retrieved November 29, 2019, from https://core.ac.uk/download/pdf/4381444.pdf

Clark, K. B., & Fujimoto, T. (1991). *Product development performance: strategy, organization, and management in the world auto industry*. Cambridge, MA: Harvard Business School Press.

Creswell, J. W. (2013) *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: Sage.

Cross, N., & Roy, R. (1989). *Engineering design methods* (Vol. 4). New York, NY: Wiley.

Dale, B. G. (1982). The management of engineering change procedure. *Engineering management international*, *1*(3), 201-208.

Davenport, T. H., & Jeanne, G. H. (2007). *Competing on analytics: the new science of winning*., Cambridge, MA: Harvard Business Press. Doi:10.5860/choice.44-6322.

Dong, A. (2005). The latent semantic approach to studying design team communication. *Design Studies*, *26*(5), 445-461. Doi:10.1016/j.destud.2004.10.003.

Dong, A., & Agogino, A. M. (1997). Text analysis for constructing design representations. *Artificial Intelligence in Engineering*, *11*(2), 65-75. Doi:10.1007/978-94-009-0279-4_2.

Dong, Q. (2002). *Predicting and managing system interactions at early phase of the product development process* ((Doctoral dissertation, Massachusetts Institute of Technology). Retrieved November 29, 2019, from https://dspace.mit.edu/bitstream/handle/1721.1/16881/51849457-MIT.pdf?sequence=2

Eckert, C. M., Stacey, M. K., & Clarkson, P. J. (2003). *The spiral of applied research: A methodological view on integrated design research*. Paper presented at the International Conference on Engineering Design, Stockholm, August 19–23. Retrieved November 29, 2019, from http://oro.open.ac.uk/13226/1/13226.pdf

Eger, T., Eckert, C. M., & Clarkson, P. J. (2007). Engineering change analysis during ongoing product development. In *DS 42: Proceedings of ICED 2007, the 16th International Conference on Engineering Design, Paris, France, 28.-31.07. 2007* (pp. 629-630). Retrieved November 29, 2019, from https://www.designsociety.org/download-publication/25647/Engineering+Change+Analysis+during+Ongoing+Product+Development

Elasticsearch. (2018). The heart of the Elastic Stack. Retrieved November 29, 2019, from https://www.elastic.co/products/elasticsearch

Eppinger, S. D., & Browning, T. R. (2012). *Design structure matrix methods and applications*. Cambridge, MA: MIT.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37-54

Fernandes, J., Henriques, E., Silva, A., & Moss, M. A. (2015). Requirements change in complex technical systems: An empirical study of root causes. *Research in engineering design*, *26*(1), 37-55. Doi:10.1007/s00163-014-0183-7.

Giffin, M., De Weck, O., Bounova, G., Keller, R., Eckert, C., & Clarkson, P. J. (2009). Change propagation analysis in complex technical systems. *Journal of Mechanical Design*, *131*(8), 1-14.doi:10.1115/1.3149847.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. Cleveland, OH: CRC.

Hamraz, B., Caldwell, N. H., Wynn, D. C., & Clarkson, P. J. (2013). Requirements-based development of an improved engineering change management method. *Journal of Engineering Design*, *24*(11), 765-793. doi.org/10.1080/09544828.2013.834039

Han, J., Kamber, M., & Pei, J. (2011). *Data mining concepts and techniques* (3rd ed.). New York, NY: Morgan Kaufman.

Hicks, B. (2013). The language of collaborative engineering projects. In *DS 75-6: Proceedings of the 19th International Conference on Engineering Design (ICED13), Design for Harmonies, Vol. 6: Design Information and Knowledge, Seoul, Korea, 19-22.08. 2013* (pp. 321-330). Retrieved November 29, 2019, from https://www.designsociety.org/download-publication/35081/The+language+of+collaborative+engineering+projects

Hubka, V., & Eder, W. E. (2012). *Design science: introduction to the needs, scope and organization of engineering design knowledge*. Berlin, Germany: Springer.

Huiyan, N., Gregory, G., & Thomson, V. (2006). Engineering change request management in a new product development process. *European Journal of Innovation Management*, *9*(1), 5-19. Doi:10.1108/14601060610639999.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in science and engineering, 9(3), 90-95. Doi:10.1109/mcse.2007.55.

Inness, J. G. 1994. *Achieving successful product change: A handbook*. Upper Saddle River, NJ: Financial Times/Pitman.

Jarratt, T. A. W., Eckert, C. M., Caldwell, N. H., & Clarkson, P. J. (2011). Engineering change: An overview and perspective on the literature. *Research in engineering design*, *22*(2), 103-124. Doi:10.1007/s00163-010-0097-y.

Kattner, N., Mehlstaeubl, J., Becerril, L., & Lindemann, U. (2018). Data analysis in engineering change management–improving collaboration by assessing organizational dependencies based on past engineering change information. In *2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 617-621). New York, NY: IEEE. Doi:10.1109/ieem.2018.8607469.

Kim, S., & Wallace, K. (2009). An automatic identification of negation in design documents. In *ICORD 09: Proceedings of the 2nd International Conference on Research into Design* (pp. 247-254). Retrieved November 29, 2019, from https://www.designsociety.org/download-publication/32287/an_automatic_identification_of_negation_in_design_documents

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text mining in organizational research. *Organizational research methods*, *21*(3), 733-765. Doi:10.1177/1094428117722619.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, *31*(3), 249-269.

Kuffner, T. A., & Ullman, D. G. (1991). The information requests of mechanical design engineers. *Design studies*, *12*(1), 42-50.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT sloan management review*, *52*(2), 21-32.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188-1196). Retrieved November 29, 2019, from http://www.jmlr.org/proceedings/papers/v32/le14.pdf

Leech, D. J., & Turner, B. T. (1985). *Engineering design for profit*. Sydney, Australia: Halsted Press.

Lewis, K., & Van Horn, D. (2013). Design analytics in consumer product design: A simulated study. In *ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. Retrieved November 29, 2019, from https://www.researchgate.net/profile/Kemper_Lewis/publication/236395420_Design_Analytics_in_Consumer_Product_Design_A_Simulated_Study/links/00b7d517eb1a5a2c8f000000/Design-Analytics-in-Consumer-Product-Design-A-Simulated-Study.pdf. Doi:10.1115/detc2013-12982.

Li, Y., Roy, U., & Saltz, J. S. (2019). Towards an integrated process model for new product development with data-driven features (NPD 3). *Research in Engineering Design*, *30*(2), 271-289. Doi:10.1007/s00163-019-00308-6.

Lundvall, B. Ä., & Johnson, B. (1994). The learning economy. *Journal of industry studies*, *1*(2), 23-42.

Ma, J., & Kim, H. M. (2014). Continuous preference trend mining for optimal product design with multiple profit cycles. *Journal of Mechanical Design*, *136*(6), 1-14. Doi:10.1115/1.4026937.

Ma, J., Kwak, M., & Kim, H. M. (2014). Demand trend mining for predictive life cycle design. *Journal of Cleaner Production*, *68*, 189-199. Doi:10.1016/j.jclepro.2014.01.026.

Matthews, P. C., & Philip, A. D. (2011). Baysian project monitoring. In *DS 68-1: Proceedings of the 18th International Conference on Engineering Design (ICED 11), Impacting Society through Engineering Design, Design Processes* (Vol. 1, pp. 69-78). Retrieved November 29, 2019, from https://www.designsociety.org/download-publication/30408/bayesian_project_monitoring

Maull, R., Hughes, D., & Bennett, J. (1992). Special feature. The role of the bill-of-materials as a CAD/CAPM interface and the key importance of engineering change control. *Computing & Control Engineering Journal*, *3*(2), 63-70.

Maynard, D., Tablan, V., Ursu, C., Cunningham, H., & Wilks, Y. (2001, September). Named entity recognition from diverse text types. In Recent Advances in Natural Language Processing 2001 Conference (pp. 257-274).

Maxwell, J. A. (2012a). *A realist approach for qualitative research*. Thousand Oaks, CA: Sage.

Maxwell, J. A. (2012b). *Qualitative research design: An interactive approach* (Vol. 41). Thousand Oaks, CA: Sage.

Menon, R., Tong, L. H., & Sathiyakeerthi, S. (2005). Analyzing textual databases using data mining to enable fast product development processes. *Reliability Engineering & System Safety*, *88*(2), 171-180. Doi:10.1016/j.ress.2004.07.007.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. Retrieved November 29, 2019, from https://arxiv.org/pdf/1301.3781.pdf%C3%AC%E2%80%94%20%C3%AC%E2%80%9E%C5%93

Misra, H., Cappé, O., & Yvon, F. (2008). Using LDA to detect semantically incoherent documents. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning* (pp. 41-48). Retrieved November 29, 2019, from https://www.aclweb.org/anthology/W08-2106.pdf. Doi:10.3115/1596324.1596332.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26. Doi:10.1075/li.30.1.03nad.

Newman, M. (2004). Fast algorithm for detecting community structure in networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, *69*(6), 1-5. Doi:10.1103/physreve.69.066133.

Norris, J. R. (1998). *Markov Chains*. Cambridge, United Kingdom: Cambridge University Press.

Otto, K., & Wood, K. (2001). *Product design: Techniques in reverse engineering and new product design*. Upper Saddle River, NJ: Prentice-Hall.

Pahl, G.,, & Beitz, W. (1996). *Engineering Design: A systematic approach.* Berlin, Germany: Springer-Verlag.

Pikosz, P., & Malmqvist, J. (1998). A comparative study of engineering change management in three Swedish engineering companies. In *Proceedings of the DETC98 ASME design engineering technical conference* (pp. 78-85). Retrieved November 29, 2019, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.123.2915&rep=rep1&type=pdf

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45-50). Retrieved November 29, 2019, from http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=EEDC158A14A8DE042064F26F8A2048EE?doi=10.1.1.695.4595&rep=rep1&type=pdf

Reich, Y. (1997). Machine learning techniques for civil engineering problems. *Computer‐Aided Civil and Infrastructure Engineering*, *12*(4), 295-310.

Roozenburg, N. F., & Eekels, J. (1995). *Product design: fundamentals and methods*. Hoboken, NJ: John Wiley & Sons.

Sarkar, S., Dong, A., Henderson, J. A., & Robinson, P. A. (2014). Spectral characterization of hierarchical modularity in product architectures. *Journal of Mechanical Design*, *136*(1), 1-12. Doi:10.1115/1.4025490.

Silow, N., Rosenqvist, M., & Falck, A. C. (2016). *Proaktiv monteringsergonomisk och geometrisk kvalitetssäkring för hållbar produktion*. Retrieved November 29, 2019, from https://www.vinnova.se/globalassets/mikrosajter/ffi/dokument/slutrapporter-ffi/hallbar-produktion-rapporter/sr_hp_2013-02416_-pegasus_sv.pdf

Smith, R. P., & Eppinger, S. D. (1997). Identifying controlling features of engineering design iteration. *Management science*, *43*(3), 276-293.

Snider, C., Gopsill, J., Jones, D., & Hicks, B. (2018). Engineering project health monitoring: Application of automatic, real-time analytics to PDM systems. In *IFIP International Conference on Product Lifecycle Management* (pp. 600-610). Berlin, Germany: Springer. Retrieved November 29, 2019, from https://research-information.bris.ac.uk/files/162328322/SUBMISSION.pdf. Doi:10.1007/978-3-030-01614-2_55.

Snider, C., Škec, S., Gopsill, J. A., & Hicks, B. J. (2017). The characterisation of engineering activity through email communication and content dynamics, for support of engineering project management. *Design Science*, *3*, 1-31. Doi:10.1017/dsj.2017.16.

Steward, D. V. (1981). The design structure system: A method for managing the design of complex systems. *IEEE transactions on Engineering Management*, *28*(3), 71-74

The Standish Group. (2014). *Standish Group 2014 CHAOS Report.* Retrieved November 29, 2019, from https://www.projectsmart.co.uk/white-papers/chaos-report.pdf.

Thomke, S., & Reinertsen, D. (2012). Six myths of product development. *Harvard Business Review*, *90*(5), 84-94.

Tripathy, A., & Eppinger, S. D. (2013). Structuring work distribution for global product development organizations. *Production and Operations Management*, *22*(6), 1557-1575. Doi:10.1111/poms.12045.

Tucker, C., & Kim, H. H. M. (2011). Predicting emerging product design trend by mining publicly available customer review data. In *18th International Conference on Engineering Design, ICED 11* (pp. 43-52). Retrieved November 29, 2019, from https://www.designsociety.org/download-publication/30612/predicting_emerging_product_design_trend_by_mining_publicly_available_customer_review_data

Ullah, I., Tang, D., Wang, Q., Yin, L., & Hussain, I. (2018). Managing engineering change requirements during the product development process. *Concurrent Engineering*, *26*(2), 171-186. Doi:10.1177/1063293x17735359.

Ullman, D. G. (1992). *The mechanical design process.* New York, NY: McGraw-Hill.

Ulrich, K., & Eppinger, S. (2015). *Product design and development*. New York, NY: McGraw-Hill .

Van Horn, D., Olewnik, A., & Lewis, K. (2012). "Design analytics: capturing, understanding, and meeting customer needs using big data". In *ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (pp. 863-875). New York, NY: American Society of Mechanical Engineers. Retrieved November 29, 2019, from http://www.academia.edu/download/40269684/DESIGN_ANALYTICS_CAPTURING_UNDERSTANDING20151122-13625-rxqja7.pdf. Doi:10.1115/detc2012-71038.

Van Rossum, G., & Drake Jr, F. L. (1995). *Python tutorial* (Vol. 620). Amsterdam: Centrum voor Wiskunde en Informatica.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Burlington, MA: Morgan Kaufmann.

Wright, I. C., Duckworth, A. P., Jebb, A., & Dickerson, D. B. (2000). Research into the process of engineering change within incremental product design. In *Proceedings of engineering design conference* (p. 449). Hoboken, NJ: John Wiley & Sons.

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2013). Data mining with big data. *IEEE transactions on knowledge and data engineering*, *26*(1), 97-107.

Wynn, D. C., & Clarkson, P. J. (2018). Process models in design and development. *Research in Engineering Design*, *29*(2), 161-202. Doi:10.1007/s00163-017-0262-7.

Yang, N., Yang, Q., & Yao, T. (2018). Clustering organization structure in product development projects using similarity. In *DS 96: The 20th International DSM Conference* (pp. 59-66). Retrieved November 29, 2019, from https://www.designsociety.org/download-publication/40976/CLUSTERING+ORGANIZATION+STRUCTURE+IN+PRODUCT+DEVELOPMENT+PROJECTS+USING+SIMILARITY

Yang, Q., Lu, T., Yao, T., & Zhang, B. (2014). The impact of uncertainty and ambiguity related to iteration and overlapping on schedule of product development projects. *International Journal of Project Management*, *32*(5), 827-837. Doi:10.1016/j.ijproman.2013.10.010.

Yoon, J., Seo, W., Coh, B. Y., Song, I., & Lee, J. M. (2017). Identifying product opportunities using collaborative filtering-based patent analysis. *Computers & Industrial Engineering*, *107*, 376-387. Doi:10.1016/j.cie.2016.04.009.

Zhang, X., Hao, Y., & Thomson, V. (2015). Taking ideas from paper to practice: a case study of improving design processes through detailed modeling and systematic analysis. *IFAC-PapersOnLine*, *48*(3), 1043-1048. Doi:10.1016/j.ifacol.2015.06.221.

Zheng, J., & Dagnino, A. (2014). An initial study of predictive machine learning analytics on large volumes of historical data for power system applications. In *2014 IEEE International Conference on Big Data (Big Data)* (pp. 952-959). New York, NY: IEEE. Retrieved November 29, 2019, from https://www.academia.edu/download/38098820/8.IEEEBigData14.pdf. Doi:10.1109/bigdata.2014.7004327.

APPENDICES

Paper A — Design analytics is the answer, but what questions would product developers like to have answered?

Paper B — Towards big data analysis of deviation and error reports in product development projects

Paper C — Modeling industrial engineering change processes using the Design Structure Matrix for sequence analysis: A comparison of multiple projects

Paper D — Supporting knowledge re-use with effective searches of related engineering documents – A comparison of search engine and natural language processing-based algorithms

Paper E — Natural language processing methods for knowledge management – Applying document clustering for fast search and grouping of engineering documents