CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Continuous experimentation for software organizations with low control of roadmap and a large distance to users

A case study

Master's thesis in Computer science and engineering

ROBIN SVENINGSON

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2019

Master's thesis 2019

Continuous experimentation for software organizations with low control of roadmap and a large distance to users

A case study

ROBIN SVENINGSON

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2019

Continuous experimentation for software organizations with low control of roadmap and a large distance to users
ROBIN SVENINGSON

Supervisors: David Issa Mattos and Jan Bosch, Computer Science and Engineering
Examiner: Robert Feldt, Computer Science and Engineering

Master's Thesis 2019
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2019

Continuous experimentation for software organizations with low control of roadmap and a large distance to users
A case study
ROBIN SVENINGSON
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

Continuous experimentation is a recently popular subject in the field of Software Engineering. There are many resources on how to conduct experimentation with techniques such as A/B tests and canary releases, how to assess organizations in how well they use experimentation and what the benefits and challenges are of using more experimentation. However, there is little differentiation in research regarding the two concepts of control of roadmap and distance to users. The first indicating how much control the company has over the product roadmap and the planning/prioritization of product changes, and the second indicating how easy it is to access the users of the products for data-collection purposes. Not all companies have high control of roadmap and a short distance to users, which is something that needs to be addressed when it comes to continuous experimentation. There exists a clear research gap in this area. This exploratory case study aims to be a starting point in filling this research gap. The thesis work is done together with a single case study company, which is a small-scale software consultancy company. Additionally, four other companies are also involved in the process as part of a static validation process. There are four main contributions resulting from this thesis. First, it presents a deterministic way of deciding how much/little control of roadmap and distance to users a company has. Secondly, it shows that there is a relationship between control of roadmap, distance to users and continuous experimentation. Thirdly, it shows how to assess a software company with low control of roadmap and a large distance to users regarding how well they use continuous experimentation. Finally, the thesis identifies what the perceived advantages, disadvantages and blocking issues are for such a company to use more continuous experimentation. The thesis also makes suggestions for future work, such as if the control of roadmap and distance to users are fixed or if they can be changed, as well as understanding if the control of roadmap and distance to users acts as barriers to evolving the use of experimentation.

Keywords: continuous experimentation, A/B testing, distance to users, control of roadmap, assessment, advantages, disadvantages, blocking issues.

Acknowledgements

I would like to, first of all, give a special thank you to the supervisors of this thesis, David Issa Mattos and Jan Bosch, who have been very helpful throughout this thesis work. I would also like to thank everyone who has participated in the case study; all the employees from the case study company who provided the foundation for this thesis, and the participants from the four other companies who provided insightful information from a different perspective. I hope this thesis can be a starting point for more research in this area, and that companies who normally might not consider continuous experimentation, such as for instance software consultancies, will find the information in this thesis helpful.

Robin Sveningson, Gothenburg, June 2019

Contents

List of Figures

List of Tables

1
Introduction

Using evidence to support decision-making in software organizations is something that is being advocated by many, including large companies like Microsoft [4] and Google [10]. A common problem in a lot of software organizations is that decisions are based on opinions and previous experience of the organization members, rather than collected empirical data and proof [12]. This becomes a problem because humans are bad at making estimations and predicting what will be appreciated and used by software users. The problem is exemplified by for instance Netflix, that according to M. Moran say that as much as 90% of what they try is wrong [5, p. 240]. Another problem software organizations are faced with is the issue of software quality. Quality related attributes in software could, for instance, be performance, security or usability [6]. These are non-functional qualities of the software that are important to consider, and different organizations need to decide which ones to prioritize higher depending on their own context. Low software quality is evidently a problem since it affects how the users experience the software and how successful the software will be, i.e. users will only accept software if its quality is on a good enough level [2].

A topic that has gotten recent attention in research is continuous experimentation [21][43][47][49][38][39], which is a general term used for experimentation in the Software Engineering process [48]. The phrase continuous experimentation was originally introduced by Dan McKinley in a talk about how Etsy uses experimentation in their organization [15]. Continuous experimentation includes experimentation concepts like A/B testing, canary releases, gradual rollouts and dark launches [50]. The use of continuous experimentation is desirable for a software organization because it allows the collection of evidence to support the hypotheses that exist in the organization, in order to make well-informed decisions, rather than decisions based on opinions. This prevents decisions based on opinions and previous experience of the organization members. A/B testing is specifically useful as a continuous experimentation method to prevent the problem of uninformed decision making [50], by enabling the identification of causalities between introduced changes and recognized effects [35]. Furthermore, continuous experimentation is also desirable for dealing with quality issues in software and can be done specifically with the use of canary releases, gradual rollouts and dark launches [50].

In the research of continuous experimentation, it is common to discuss companies like Microsoft [35][45][22][7][20][14] and Google [10], and these two example companies are similar in many ways. A striking similarity is that they are both very

large organizations. Another similarity is that they are in the business-to-consumer domain and presumably own (most of) their own products, and in extension have the responsibility of the software users. It is trivial to hypothesize, by looking at the example companies Microsoft and Google, that a lot of the organizations that have been studied in research related to continuous experimentation have high control of the roadmaps of their products, i.e. the ability to affect what product changes are planned and prioritized. This might be the case because they often own their own products and work directly with consumers, and therefore has the control of the planned product changes.

Control of roadmap is a term refers to how much control of the roadmap the organization has, i.e. how easy it is for them to affect what tasks should be executed and which tasks to prioritize, and it can be used to categorize software organizations. An organization with high control of roadmap has an easier time to carry out their agenda and desired changes, while an organization with a low control of roadmap has a much harder time to do so. Another relevant term that can be used to categorize software organizations is the distance to users, which refers to how easy it is for the company to access the users of the software. An organization with a large distance to users has a difficult time to access the users in order to collect user data and feedback, while an organization with a short distance to users has an easier time to do so. A large distance to users can depend on many things, for instance, that the organization does not own their own product and need permission to access users from the product owner, or that there are legal issues in the way of accessing the users. There are examples in the research of continuous experimentation where the concept of distance to users is discussed [42][31]. However, for companies that are mentioned in the research of continuous experimentation, where the distance to users is not defined, it is less trivial to make informed guesses about if they have a large or short distance to users. This is because it is not always apparent if they, for instance, have legal issues in the way. Since continuous experimentation requires data collected from users as well as the ability to easily implement and deploy experiments to test hypotheses [12], it could become more problematic to use continuous experimentation if the organization has a low control of roadmap or large distance to users.

The two terms control of roadmap and distance to users are not considered in the research of continuous experimentation to a large extent, and there exists a research gap in this area. It is desirable to close this research gap since a lot of software organizations might not have high control of roadmap and a short distance to users, and therefore it is of interest to investigate how this affects their ability to use continuous experimentation. An example of such organizations is software consultancy companies, who develop software products for other client organizations, where the client organizations are the product owners who set the roadmap and have responsibility for the software users.

The overall purpose of this thesis is to explore this research gap and provide initial results, that in the future can be further developed. The focus of this thesis

is specifically on organizations with a *low* control of roadmap and *large* distance to users, and it is executed as an exploratory case study together with a software consultancy company that fits into this description. Data collected from the case study company is collected through, for instance, several interviews, a survey and observations of the company's experimentation system. Additionally, data is also collected from interviews with four other companies for static validation purposes. The data collected is used to identify if there is a relationship between the control of roadmap, distance to users and the use of continuous experimentation. The information is then used to first understand how a software organization with low control of roadmap and large distance to users can be assessed in how well they use continuous experimentation. Secondly, it is used to find what perceived advantages, disadvantages or blocking issues (the *effects*) such an organization identifies in regards to using more continuous experimentation. Finally, the results of the thesis are presented and discussed, which might inspire future work to overcome any issues discovered.

This thesis is outlined as follows: Chapter 2 goes through the background of important concepts, such as continuous experimentation, control of roadmap and distance to users. In the next chapter, Chapter 3, the goals of the research and the research method is explained with detailed information regarding how the empirical data was collected and analyzed. Chapter 4 shows three models resulting from this thesis work. Chapter 5 discusses the case study company and the static validation companies in detail. Chapter 6 provides feedback collected on the three models created. Chapter 7 provides empirical data regarding the relationship between continuous experimentation, control of roadmap and distance to users. Chapter 8 goes through the perceived effects identified of using more continuous experimentation. Furthermore, Chapter 9 contains a discussion on the empirical data and models and aims to answer the research questions of this thesis. Finally, Chapter 10 concludes the findings of this thesis and summarizes what has been discussed in this thesis.

2
Background

This chapter provides background information to some of the concepts that are discussed in this thesis work.

2.1 Continuous experimentation

Ros and Runeson define continuous experimentation as a *"general term which covers a wide variety of experiments and the implications of experiments on the whole software engineering process"* [48]. The use of the term continuous experimentation is first found in a presentation in 2012 made by D. McKinley of the company Etsy, where he discussed how Etsy uses experimentation in their organization [15]. There are multiple experimentation techniques that are included in the term continuous experimentation. According to Schermann et. al. some of the most important techniques are A/B testing, canary releases, gradual rollouts and dark launches [50].

According to Bosch roadmapping and prioritization of new features in many organizations are based on the opinions of the people involved and the previous personal experiences of those people [12]. It is often the more senior people whose opinion weigh the heaviest. He proposes two issues with this approach. The first one is that opinions and personal experiences are not good replacements for actual user data, i.e. empirical data and evidence. The second issue he identifies is that an organization that is welcoming the exploration of new innovative ideas might still choose the safest bets in order to avoid risk-taking. There is also a push from the industry to use evidence in the decision-making process instead of opinions, for instance by Microsoft [4] and Google [10]. Furthermore, according to M. Moran Netflix says that more than 90% of the ideas they evaluate are wrong [5, p. 240]. This is an indication of how bad we as humans are at making estimations and predicting the effect of a change, arguing for the same point as Bosch made, where opinions are not good at replacing empirical data. Bosch discusses in another work how while making decisions based on habit can be useful for humans, we risk that we make decisions are not always the most optimal ones [32]. To conceptualize the lack of empirical data in decision-making Holmström Olsson and Bosch introduce the "Open-Loop" problem, which refers to how companies fail to collect enough user data and where the validation of product management decisions is done after the product has been finished, which is too late in the process [23]. The authors explain how this causes the process to become more opinion-based, rather than data-driven.

It is desirable to use more empirical data and evidence when making decisions, i.e. closing the Open-Loop, in order to prevent making uninformed decisions that might be sub-optimal or negative for the organization and to prevent choosing safe-bet solutions to avoid taking risks when there might be better solutions available. Ultimately by closing the Open-Loop product managers can get their decisions tested and validated and use user data to guide the roadmapping and prioritization of features, instead of blindly guessing based on opinions in the organization and previous experience of the organization members what features will beneficial to launch.

Another issue that is very relevant for software organizations is the issue of maintaining high quality in the software. Some examples of quality attributes, also called non-functional requirements, that can be used to assess the overall quality of software are performance, security or usability [6]. Many software organizations use Agile practices in their development process [51]. However, non-functional requirements in Agile processes are often neglected and assigned lower priority than functional requirements [41][13]. According to Bhatti, users of a software system will not accept a software system with low quality, and therefore the quality is an important aspect of the success of the software [2]. The author also identifies that organizations will not succeed in the market with low-quality software.

It is desirable to prevent the issue of low-quality software to ensure the success of the software, i.e. it is desirable to achieve high software quality. The quality should be based on specific quality attributes that are deemed important for specific software by the stakeholders of that software, and there is no general solution that fits everyone.

Continuous experimentation allows the collection of empirical data and evidence in the form of user data and feedback, that can be used to make informed decisions rather than decisions based on opinions and previous experience, i.e. closing the Open-Loop. It can also be used to assure the quality of software. Schermann et. al. introduce the classification of business-driven experiments [50], which are experiments conducted to evaluate what effects features have from a business perspective, mainly with the use of A/B testing. The authors also introduce the classification of regression-driven experiments, which means making sure a new version does not cause any noticeable regression to the users in regards to non-functional requirements. The regression-driven experiments can according to the authors be done with the aid of canary releases, gradual rollouts and dark launches. The business-driven experiments, in this case A/B testing, solves the issue of decision-making being based on opinions and previous experiences because they allow evaluating hypotheses with actual empirical data. The regression-driven experiments, in this case canary releases, gradual rollouts and dark launches, solves the issue of low software quality because they allow making sure no regression in regards to non-functional requirements are noticeable to the users. Furthermore, there exist several other techniques for experimentation, for instance, the HYPEX model created by Olsson and Bosch that defines how to identify features and release them incrementally through minimal viable features [25]. However, during this thesis the main techniques fo-

cused on are A/B tests, canary releases, gradual rollouts and dark launches.

Fabijan et. al. identify two main prerequisites for experimentation with the help of online controlled experiments: statistical foundations and psychological safety [44]. The first one, statistical foundations, is the need for a good understanding of fundamental statistical concepts, such as power analysis, in order to perform successful experimentation. Power analysis is the method used to ensure as small sample-sizes as possible while still being able to identify changes in metrics [1]. For statistical foundations, Fabijan et. al. also mention the need of a good understanding of hypothesis testing. The second prerequisite, psychological safety, is the need for people participating in the experimentation inside the organization to feel psychologically safe. People participating in the experimentation need to be secure in the fact that a lot of things will be proven to be unsuccessful, so they can learn from it and try again. Although these two prerequisites mentioned by the authors are meant for experimentation in the form of online controlled experiments, it is possible to consider the prerequisites general to continuous experimentation as a whole. Especially since one of the key principles of continuous experimentation is online controlled experiments, i.e. A/B testing.

2.2 Continuous experimentation techniques

In the following sections, the four most important continuous experimentation techniques A/B testing, canary releases, gradual rollouts and dark launches are discussed.

2.2.1 A/B testing

A/B testing, online controlled experiments, split tests and randomized experiments are some of the terms that are used to describe the same concept [40]. The idea is that you first divide the statistical population into two groups: Control group and Treatment group. You then expose the control group to a version A of your software, and the Treatment group to a version B. Then the difference between different metrics are measured, and this will indicate which version A or B is most efficient in relation your own end goal(s). [40] A/B testing is useful because it can identify causal relationships [44], and hence validating hypotheses that an organization wants to test. A/B tests do however require statistical significance, so that randomness can be discarded as a source of the measured effect [44]. A/B tests are used in for instance websites, where different users are shown a version A or B of the website in order to measure which version is more effective based on some predefined goals, i.e. if the proposed hypothesis is indeed valid.

When conducting A/B tests there are several things to consider. Olsson and Bosch show how an organization typically moves through a set of steps in order to reach the final stage where experimentation is done based on user feedback [24]. According to the authors an organization typically start with Agile development, then moving

towards continuous integration, where tests and builds are automated in the development pipeline, and after that using continuous deployment, where new software is released to the production environment more continuously instead of in large releases. After Agile, continuous integration and continuous deployment have been reached, the next and final step is to start experimenting. Kohavi et. al. discuss how there are three main aspects to consider when starting to use A/B testing: cultural/organizational, engineering and trustworthiness aspects [20]. The first one is about how the organization needs to understand why to run experiments and what the trade-offs are. The second aspect is about the need of an experimentation system that facilitates experimentation, especially when it is scaled. The third aspect is about the need to identify experiments that interact with each other, as well as the need to identify false positives in the experimentation. Furthermore, there are several examples of papers that focus on the pitfalls related to running A/B tests and how to avoid them [7] [29].

From the technical aspect of experimentation, the experimentation system is something that many authors discuss [20] [10] [45]. The experimentation system refers to the infrastructure that allows the experimentation to happen, e.g. by dividing users into different control/treatment groups and making sure that the users see the right variants based on if they are in the control or treatment group [45]. One example of an experimentation system is Google Optimize [53], which allows the user to execute and monitor different types of experiments, such as A/B tests. When it comes to the actual collection of user data to feed to the experimentation system, Google Analytics is an example of such a service [52], which works well together with Google Optimize. Google Analytics is used for web sites and provides detailed information such as what pages are visited, how long users stay on pages, how the users move through pages and information about the users themselves. When actually implementing the experiment in the code, there are mainly two implementation techniques that can be used. The first one is feature toggles, which means a conditional statement in the code that chooses what code is executed based on which of the control or treatment group the user is assigned to [49]. A second technique is traffic routing, which means having more than one service ran at the same time, and users will be directed to a specific service based on if they are placed in the control or treatment group [49].

Another important aspect of experimentation, and specifically A/B testing, is the actual metrics, i.e. how to measure success and failure. Defining metrics is not an easy thing. One of the papers that discuss this difficulty is by Dmitriev et. al., who show that defining good metrics is a big challenge, especially for long-term effects [29]. There are also several common mistakes that are made when interpreting the metrics as shown by Dmitriev et. al. in another paper [34]. There is research specifically on how to define the metrics for software experimentation, for instance by Deng and Shi who propose a data-driven approach to developing the metrics [28]. Several papers discuss the Overall Evaluation Criterion [8] [29], which is the measure that quantifies the goal of the experiment. Fabijan et. al. describe four types of metrics that should be included in an Overall Evaluation Criterion: success metrics,

which should be improved, guardrail metrics, which should not move out of a specific range, data quality metrics, which makes sure there are no quality problems with the experiment and that it is set-up in a good way, and debug metrics, which shows details of how the success and guardrail metrics changed [44]. Similar metrics and their lifecycle is discussed by Issa Mattos et. al. in their proposed metric model [46].

2.2.2 Canary releases

Canary releases mean releasing a new version of the software to a smaller amount of people [50]. One of the differences between canary releases and A/B testing is the randomization of users selected for the alternative version. A/B testing randomizes users between treatment and control group, in order to identify causalities, while a canary release can be used on a group of manually selected users to ensure that something works as it should and to reduce the impact of anything that might go wrong. An example of a canary release could be when a specific set of users, perhaps 5-10%, gets the new version, in order to monitor how it behaves. These users could, for instance, be "early-adopters", who are eager to try out new things, who were not selected at random.

2.2.3 Gradual rollouts

Gradual rollouts mean gradually increasing the number of users that are exposed to a new version of the software until the old one can be removed completely [50]. It is best combined with other types of experimentation techniques, such as canary releases [50]. It is useful to ensure that users are safely moved from one version until another until the new version is stable enough to be trusted. If there are fewer users exposed to the new version, the negative effects should be less if something goes wrong.

2.2.4 Dark launches

Dark launches are a way of deploying code without it being seen by the users [19]. It is used to test scalability and performance [19], to see that a new version works under pressure in a real environment with plenty of user activity. The new code is deployed in parallel with the old code, and requests in production are sent to the new code as well. Then the new code will be tested with real traffic, in order to see that it behaves as expected. However, the users do not notice it since the new code is not activated for users to experience.

2.3 Assessment of the use of continuous experimentation

An organization can use experimentation in their projects but how much and how well it is used can vary. Therefore it is relevant to not only determine if experimen-

tation is used but to also assess to which degree it is used. In research in the field of Software Engineering there exist some methods to assess how well experimentation is used in an organization. Three methods are discussed further in this section.

2.3.1 The Stairway To Heaven Model

The Stairway To Heaven model is a model introduced by Holmström Olsson, Alahyari and Bosch in 2012 [16], which describes the evolutionary steps organizations normally take in order to reach the final step "R&D as an Experiment System". This final step is when the organization starts experimenting based on user data and can be considered what we today call continuous experimentation. The steps included in the model are in order from lowest to highest: Traditional Development, Agile R&D Organization, Continuous Integration, Continuous Deployment and finally R&D as an Experiment System. The purpose of the model is to show the steps normally taken when moving towards continuous experimentation and to provide guidance on how to transition between steps. However, in the original paper [16] the transition from continuous deployment towards the final step "R&D as an Experiment System" was not covered, but this transition was later provided in another article by Holmström Olsson and Bosch [24]. The Stairway To Heaven model can be used to assess which step an organization is currently at, which is useful to learn what the next step is and get guidance on how to evolve to the next step.

In 2017 Bosch and Holmström Olsson extended the Stairway To Heaven model [33]. The new extension of the model focuses on three different dimensions: speed, data and ecosystems. Organizations can be evolved to different degrees on the three dimensions, which would not be considered in the original model. This is a benefit of the new model, which allows different degrees of evolution on the three dimensions. In the article, the authors focus on the data-dimension of the new Stairway To Heaven model. The data dimension shows how companies start to improve their use of data to support decisions, in order for their decision-making to become more evidence-based. The steps included in the data dimension are in order of lowest to highest: Ad-Hoc, Collection, Automation, Data innovation and Evidence-based Organization. This shows how companies move from not using data, to starting to systematically collect user data, starting to visualize it and automate the process, to development teams taking over data analysis tasks where the data-scientists now act more like mentors, and to finally an evidence-based organization where data-driven development, evidence-based decision-making and continuous experimentation is the organization's way of working. The new version of the Stairway To Heaven model can like the original version be used to assess an organization current step, what the next step is and how to transition there. However, now this can be done on three dimensions instead of just one, which will better represent the organization's current state if the organization is more or less evolved on some of the dimensions.

2.3.2 The Experiment Growth Model

The Experiment Growth Model is a model created by Fabijan et. al. [44], which answers the question of "how do large-scale online software companies grow their experimentation capabilities?". It is developed to understand how companies currently use experimentation, in order to identify what the next step is and how to improve the experimentation to reach further steps. The authors provide guidance for practitioners on how to use the Experimentation Growth Model to perform this. The Experimentation Growth Model is an extension of a previous Experimentation Evolution Model created by some of the same authors [36]. The Experimentation Growth Model focuses on online controlled experiments only, which means that it does not consider all of the continuous experimentation definition, for instance not including canary releases, gradual rollouts and dark launches.

The Experimentation Growth Model consists of four stages; Crawl, Walk, Run and Fly. Each stage represents how far the organization has progressed in a specific dimension, where Fly is the most progressed stage. There are a total of seven of dimensions in the model, which are Technical focus, Experimentation platform capability, Experimentation pervasiveness, Feature team self-sufficiency, Experimentation team organization, Overall Evaluation Criteria and Experimentation Impact. The different dimensions explain different concepts of the experimentation in the organization, such as the collection of data, the implementation of an experimentation system that is used to run the experiments, the organization conducting the experiments, the metrics that are measured in the experiments and the impact of the experimentation.

The Experimentation Growth Model can be used to assess the current state of an organization in the context of the seven different dimensions. This will provide a view of how the organization is working with online controlled experiments. Then once you know what stages the organization is on the different dimensions, it is possible to understand what the next stage is and what should be done to progress to that stage.

2.3.3 The RIGHT model

Fagerholm et. al. introduce the RIGHT model, which is a model that defines how continuous experimentation should be organized, what the process around experimentation should look like and how to design the software architecture [37]. The authors show how the learning cycle should be a repeated number of Build-Measure-Learn blocks, which is an idea from the Lean Startup by Eric Ries [11]. The idea behind the Build-Measure-Learn loop by Ries is that a business is run with the help of scientific methods and validated learning. It means that a business is run by experimentation where things are built, empirical data is collected and learnings are drawn from it, in order to know if to stay on the same course or change direction. In the model by Fagerholm et. al. the Build-Measure-Learn blocks generate learnings, that are provided to the next Build-Measure-Learn block. This Build-Measure-Learn and learnings cycle is simultaneously supported by a technical

infrastructure according to the model. The authors also provide a process model for continuous experimentation, which shows the process that connects the business vision, strategy, experiments and product, as well as an infrastructure model that shows how to design the technical infrastructure for continuous experimentation. While the RIGHT model is not necessarily an assessment model of how well continuous experimentation is used, it could perhaps still be used for this purpose. Given that the RIGHT model is a validated model for how to organize experimentation, how to design the process and how to design the technical architecture, it could be compared to a given company to see how similar that company works compared to the RIGHT model. If the company's way of working differs a lot from the RIGHT model, perhaps it can be an indication of how well continuous experimentation is used in that company.

2.4 Advantages, disadvantages and blocking issues of continuous experimentation

Fabijan et. al. identify several benefits with using controlled experimentation [35], which partly or fully can be considered the same as continuous experimentation. The authors identify benefits such as "value discovery and validation", i.e. the ability to discover and validate what creates value for the product, "ensuring product quality" and "incremental product improvements". The authors also mention "stabilizing and lowering product complexity", which means not deploying and also removing features that are not successful, which would reduce the product complexity. Yaman et. al. show the benefits of introducing continuous experimentation in an organization [42]. The benefits are for instance "new insights with respect to business goals and customers" and "decisions supported by data". The authors also describe the benefit of "reduced development effort", which similarly to the benefit of Fabijan et. al. refers to how development effort can be reduced by focusing only on features that are actually useful.

Lindgren and Münch identify several disadvantages and blocking issues with using more continuous experimentation [30]. For instance the limited resources and the lack of time for experimentation. Rissanen and Münch identify the lack of competence in experimentation and the need for education as a challenge of continuous experimentation, specifically in the B2B-domain [27]. Schermann et. al. identified several obstacles of continuous experimentation [50]. For instance obstacles such as "lack of expertise", "software architecture", i.e. software architecture is not made to support experimentation, and too low return on financial and time investment. Fabijan et. al. identify several challenges with controlled experiments, which again is considered partly or fully the same as continuous experimentation [35]. Some of the challenges include the need to evolve the technology, such as the experimentation platform, and the need for data-scientists and data-engineers. Yaman et. al. talks about the challenges faced of introducing continuous experimentation in an organization [42], which includes challenges such as the inexperience of experimentation in the organization, i.e. that the organization is a bit hesitant to begin using

experimentation.

2.5 Control of roadmap

In an article by Coram and Bohner, the authors discuss the impact that Agile methods have on product management, and the customers' involvement in the Agile process [3]. According to the authors a representative of the customer that is involved in the process must be authorized to make decisions regarding, for instance, the features to be included in a given release. This reflection by the authors highlights an interesting concept, which is the idea of where the authority of making product roadmap or backlog related decisions resides. In Agile development the Product Owner serves an important role and is in charge of managing requirements, providing a prioritized backlog and communicating with the development team(s) according to Paasivaara, Heikkilä and Lassenius [17]. In this context the Product Owner has some or full authority over the decisions related to the product, even though it might still be shared in a larger product management organization. Which team, group or organization the Product Owner and/or product authority belongs to will vary from company to company. When it comes to B2B-contexts, for instance with a software consultancy, it might differ where the Product Owner resides, and the Product Owner can be either part of the company or the client company in the B2B-relationship. Furthermore, in this context, the authority to decide on what will be planned and prioritized for the product's backlog and/or roadmap might also differ. In some cases, the authority might reside in the company itself, and in some cases in the client company.

In this thesis, the authority to plan and prioritize a product's backlog and/or roadmap is referred to as the *control of roadmap*. To the knowledge of this author, this is a subject that has not been discussed a lot in research. Control of roadmap in this thesis assumes the context of the company that develops the software and the actual product, or a team in that company, and considers how much control the company or the team has over the planning and prioritization of the product's roadmap and backlog. The control of roadmap can be considered for any company, independent of if the company is in one or more client relationships in a B2B-context where the company provides software for their client(s), or if the company itself owns the products and works directly with the end-users. Furthermore, the control of roadmap can be considered for a company as a whole, or for a specific team in a company, depending on what makes sense in a given context. In this thesis, the control of roadmap is considered to be *none* if the company or the team in the company has no ability to control what is planned and prioritized in the product's backlog and/or roadmap. A *low* control of roadmap would indicate the company, or the team in a company, can affect the roadmap and/or backlog somewhat, but that there are difficulties, and finally a *high* control of roadmap means that there are no difficulties to plan and prioritize the backlog and/or roadmap. This definition will be used initially in this thesis due to the lack of discussion of this topic in research today.

2.6 Distance to users

Lindgren and Münch present a challenge of continuous experimentation in the B2B-domain which was accessing the end-users of the product for data-collection purposes [30]. Similar challenges where also reported by Rissanen and Münch [27], as well as by Yaman et. al. [42]. This challenge of accessing the end-users of a product for data-collection purposes is referred to as *distance to users* in this thesis, which refers to an imagined distance between the company developing the product in need of user feedback and the end-users. The feedback collected from the users can be of both qualitative or quantitative type, for instance, feedback from qualitative focus groups or quantitative automatic logging of user behavior in the product. These two types of user feedback should be combined according to Olsson and Bosch [26]. The distance to the users might exist because the development company is in a B2B-context with the client(s), that own the products and therefore are in charge of the end-users, but it could also be for any type of company that experiences some sort of hinder between themselves and the users.

In this thesis, the distance to users is considered to be an organizational challenge rather than a technical challenge. This means that the need for technical investments into creating, for instance, the logging service that collects user feedback is not considered to be a distance to users since this is experienced by any company that wants to begin collecting data. Instead, it is thought of from an organizational point of view, if for instance the development company requires approval from another company to collect data, or if there, for instance, are legal issues in the way of accessing the users. Furthermore, if there currently is no data-collection in a company it should not be considered a distance to users, i.e. if there currently is no access to users it should not be considered a distance to users, but rather if it is not possible to access the users and collect data if that is desired, despite what is currently collected from the users, then there exists a distance to users. Therefore the *ability* to access users should be considered rather than the actual access to the users at this moment for a company.

As an initial way of describing the distance to users, this thesis defines no ability to access the users as an *very large* distance to users. If there is some possibility to access the users, but there are difficulties, the distance to users is considered to be *large*. If there is a full possibility to access the users without any difficulties, the distance to users is considered to be *short*.

3
Research method

This chapter explains the research method that has been used for this thesis work.

3.1 Research goals and methodology description

One part of this thesis is to learn if the concepts control of roadmap and distance to users are related to how continuous experimentation is used in a company. If there turns out to be a relationship between these concepts, another main goal is to learn how a company with specifically large distance to users and low control of roadmap can be assessed in how well they use continuous experimentation, and what perceived effects, i.e. the advantages, disadvantages and blocking issues, of using more continuous experimentation such a company identifies. These goals are turned into the following three research questions.

RQ1 "Are the classifications control of roadmap and distance to users related to the use of continuous experimentation in a software company?"

RQ2 ”How can a software company with low control of roadmap and a large distance to users be assessed in terms of how well they use continuous experimentation?”

RQ3 ”For a software company with low control of roadmap and a large distance to users, what are the perceived advantages, disadvantages or blocking issues of using more continuous experimentation?”

The first research question RQ1, which is whether or not there is a relationship between control of roadmap, distance to users and continuous experimentation, together with the way of classifying companies in control of roadmap and distance to users has been worked on continuously throughout this thesis. The research questions RQ2 and RQ3 have therefore been worked with in parallel with RQ1. RQ2 and RQ3 are not dependent on that RQ1 shows that there is a relationship between the concepts, however, if it appears that there is no relationship between the concepts, then explicitly stating "low control of roadmap and a large distance to users" in RQ2 and RQ3 is redundant. If there is shown to be a relationship however, then the explicit statement is valuable. The case study company is from an initial stage identified as having low control of roadmap and a large distance to users, which is why the research questions RQ2 and RQ3 are specifically targeting this type of company. During this research, this early assessment of the company according to these two classification concepts is trying to be proven with data collected in the

case study.

In the third research question blocking issues refers to things that are actively hindering the evolving of the experimentation usage, where disadvantages are simply the negative consequences of evolving the experimentation usage. While the research question specifies to "use more" continuous experimentation, it is meant both as in increasing the frequency of experimentation, but also in the sense of using experimentation more, i.e. improve the experimentation execution on, for instance, different technical, business and organizational dimensions.

To answer the research questions the thesis work has been designed as an exploratory case study together with a single case study company. Additionally, four other companies were involved in a static validation process. The case study company is a small-scale software consultancy company, with 15-20 employees, that works with multiple clients and have a B2B-relationship to these clients. The case study company provides software solutions, in the form of web services to the different clients, who in turn are in charge of the business side of the software products. The company is the unit of analysis of the case study and by observing how the case study company works and by collecting qualitative and quantitative data from the employees of the company, it is possible to answer the proposed research questions. The case study company is considered a good unit of analysis because it is identified initially to have low control of roadmap and a large distance to users, which for instance allows the relationship between control of roadmap, distance to users and continuous experimentation to be explored. The case study company was selected because it is a good unit of analysis, but also because the thesis author works at the company. The potential threats of validity related to that the thesis author is employed by the company are discussed in the validity threats section below.

For this case study, research about how to conduct a case study made by Runeson and Höst [9] has been used for a sound research methodology. This was done for instance by keeping detailed case study protocols, working with triangulation and chains of evidence between data collected and conclusions drawn. The checklist and instructions provided by the authors have been followed to a large extent in this case study.

The case study was conducted in seven different phases, some of which are dependent on each other. The seven phases are shown in table 3.1, not necessarily in the order of which they were performed. The first phase is a literature review, which is followed by a phase about learning how the case company works and its inner processes. The next phase is the assessment of how well continuous experimentation is used in the case study company. Furthermore, the next phase is about the classifications of control of roadmap and distance to users for any type of company, and the phase after that is about determining if there is a relationship between the three concepts control of roadmap, distance to users and continuous experimentation. This is followed by the phase about identifying what the perceived effects are

Phase	Description
Phase 1	Literature review
Phase 2	Learn how the case study company works and about its inner processes
Phase 3	Assessment of how well the case study company uses continuous experimentation
Phase 4	Classification of control of roadmap and distance to users for a company
Phase 5	Determine if there is a relationship between control of roadmap, distance to users and continuous experimentation.
Phase 6	Identify the perceived effects of using more continuous experimentation
Phase 7	Static validation

Table 3.1: The phases of this thesis work

of using more continuous experimentation. Finally, the last stage is about static validation of this thesis work.

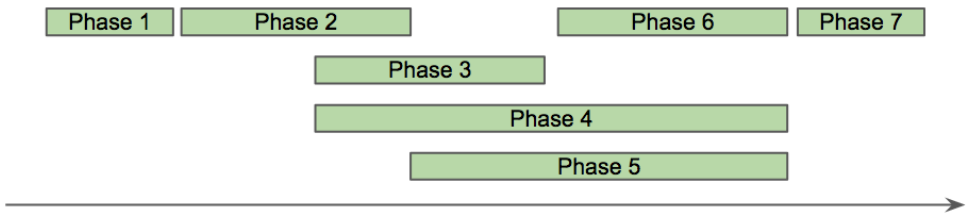


Figure 3.1: Timeline of the phases of this thesis work

Figure 3.1 shows the approximate order and duration of the seven phases of this thesis work. The arrow in the figure indicates time, with the beginning of this thesis work to the left and the end of this thesis work to the right. Phase 1, the literature review, was the first phase to be executed. After that phase 2, learning about the case study company, was executed. While phase 2 was still being finished, phase 3, assessment of the company, and phase 4, classification models for control of roadmap and distance to users, were started in parallel. When phase 2 was finished, phase 5, the relationship between the three concepts, was the next phase started. Once phase 3 was finished, phase 6, the perceived effects of using more experimentation was started. Finally, when phases 4, 5 and 6 were finished the final phase 7 was executed. During the entire thesis work the thesis report has also been worked on continuously. Once the seven phases were finished the final analysis and results were created, and this is when a big part of the report was written.

The order of the phases was dependent on how the thesis work was defined in the beginning, and how it changed during the execution of the thesis work. Initially, the phases 1, 2, 3 and 6 were planned. As the work progressed new information was learned and reflections on the research methodology was made, therefore, during phase 2 the need to add phase 4 and 5 was discovered. Also during the work, the need to add phase 7 was identified. Given that the classification models created were similar to the initial definition of control of roadmap and distance to users,

and given that the relationship between the three concepts was confirmed, the fact that phases 4 and 5 were started quite late in the process did not negatively impact the research.

3.2 Data collection and analysis

This section describes the data collection and data analysis of the seven phases proposed earlier in this chapter.

3.2.1 Phase 1: Literature review

The first phase of this thesis work was to do a literature review in the subject of continuous experimentation and online controlled experiments, as well as how to assess how well continuous experimentation is used and the perceived effects of using more experimentation. The literature review was not done strictly with any systematic reviewing technique, but a large part of the research in this subject was still considered. In the beginning, some papers that were found by the author of this thesis and some recommended by supervisors were studied. Based on those papers a forward and backward snowballing technique was used. Many of the papers identified by Auer and Felderer in their systematic mapping study on continuous experimentation [43] were included in the literature collection used for this thesis work. Several of the resources found in the literature review phase of this thesis work are described in the background chapter of this thesis.

3.2.2 Phase 2: Learn about the company

This phase was meant to uncover important information regarding the case study company, such as what the company does and how long it existed, how the organization looks like in terms of roles, hierarchies, etc., how the organisation works with continuous experimentation and how it implements its experimentation in practice. This information was needed to understand more about the organization and what its main challenges are, as well as to provide a data foundation to the research questions.

Data collection in this phase was done through qualitative interviews, and a total of five interviews were held with a total of five people. The first two interviews were held together with two specific employees in the form of a group interview, where the two employees were selected because of their positions in the company (both were Project Leaders) and because they agreed that they together possess all knowledge about experimentation in the company. The first interview was about the different points of information listed in the previous paragraph. However, in this interview experimentation was discussed in the context of online controlled experiments only. The interview contained a mix of open and closed questions in a structured order, and it allowed the interviewees to give detailed answers if they desired. The second interview with the two employees complemented the first one by adding more open and closed questions in a structured order specifically about canary releases, gradual

rollouts and dark launches. The purpose of the second interview was to complement the information about online controlled experiments, so that the full definition of continuous experimentation was covered. The first interview conducted was an hour long, and the second interview, i.e. the complementary interview, was 10 minutes long since it contained fewer questions. After the initial two interviews were held all of the information needed about the organization had been uncovered.

To triangulate the data collected in the first two interviews, an additional three interviews were held with three different employees, i.e. with one employee per interview. These additional interviews were between 15-20 minutes long. The interviews began with a single open question, which was how continuous experimentation is used in the organization, and a definition of continuous experimentation was given as well. To focus on a single open question allowed the interviewees to elaborate on the use of continuous experimentation in the company, and once they either did not have more to say or when they touched specific interesting subjects some additional questions were asked by the interviewer to guide the conversation. The data triangulation was possible since the open-ended questions allowed the interviewees to give detailed answers on their view of how the company is working with experimentation, and the answers were then used to see if they said the same things and also to see if they added any new reflections regarding experimentation. After performing five interviews, all roles in the company have been represented in the interviews, and also several (most) of the clients the company works with have been represented by an employee of the case study company who works specifically with them. The interview templates that were used for these five interviews can be found in Appendix A.

To further data triangulate, and make sure the information given was correct, company documentation as well as data from the experimentation system were requested. The documentation provided was an internal meeting protocol from a discussion on company values, and an internal communications log from a discussion regarding an A/B test an employee wanted to run. Furthermore, the company also provided access to the experimentation system for one of its multiple clients. This information from both the documentation and the experimentation system was used further to data triangulate things that had been said in interviews.

Analysis of the interview data was done by firstly transcribing all of the interview recordings. Then an analysis of the interview was created and the interviewees were sent the transcript and analysis to give them the opportunity to verify that they had been interpreted correctly or to retract any invalid statements. Analysis of the documentation and experimentation system data was done by going through the content and making notes, and then the observations were sent to the company to verify that none of the observations were sensitive from a confidentiality perspective. A version of thematic analysis was then used, which is defined by Braun and Clarke as "a method for identifying, analysing and reporting patterns (themes) within data." [18], specifically in regards to the analysis of qualitative data. This was done by identifying and highlighting different themes in the interviews, for instance, similar

discussion topics, by going through all of the transcripts. Learnings from different interviews on the same themes as well as data collected elsewhere were joined in a combined document, and for each theme, conclusions were drawn based on all of the information on the same theme.

3.2.3 Phase 3: Assessment of continuous experimentation

The purpose of this phase was to learn how a company can be assessed in how well they use continuous experimentation, or more specifically a company with low control of roadmap and a large distance to users which the case study company is identified as. The goal was to find an existing assessment method that could be used for this purpose or to create a new one. This would enable the answering of the second research question (RQ2).

First of all information about assessing companies in how well they are using continuous experimentation was collected from other research, by looking at for instance the Stairway to Heaven model [16][24][33] and the Experimentation Growth Model [44]. Later in the process the RIGHT model [37] was also considered. None of these assessment methods were appropriate to assess a company with low control of roadmap and large distance to users in how well they use continuous experimentation. The Stairway to Heaven model mainly because its final stage ended at continuous experimentation, rather than assess how well it is used. This was not ideal since the goal was to measure how well continuous experimentation is used, not only if it is used. The Experimentation Growth Model was deemed to not be appropriate for the task because, firstly, it was made for online controlled experiments, i.e. A/B tests, and not the full definition of continuous experimentation, i.e. missing canary releases, gradual rollouts and dark launches. Secondly, because it was specifically made for large-scale companies. This was an issue since, for instance, one dimension in the model expects that there is at least some data-scientist in the organization, which makes sense for a large-scale company, but for smaller companies, this would perhaps not always be the case. Finally, the RIGHT model was deemed to not be appropriate for the task since while it is possible to compare how an organization's way of working with experimentation differs from the RIGHT model, it is not so trivial to say that any differences from the RIGHT model would mean experimentation is used less well. So even if the RIGHT model is a validated way of designing the process and technical architecture for experimentation, it does not automatically mean that all other ways are worse ways of designing the process and technical architecture.

Because the existing assessment models were not appropriate for the task of assessing how a well a company uses continuous experimentation, a custom assessment model was created by this thesis author. However, the Experimentation Growth Model by Fabijan et. al. [44] was considered to be useful for the assessment of continuous experimentation, if the two mentioned issues with only considering A/B tests and only considering large-scale companies could be solved. Therefore, the created custom assessment model was based on the model by Fabijan et. al. A benefit

of this approach was that it allows the custom assessment model to be anchored in existing research from the start. Furthermore, data collected from interviews in phase 2 also helped inspire the construction of the model. Once the model was created the company was assessed by the author of this thesis using the model, and the assessment was made based on what had been said in the interviews of phase 2. To validate that the model had internal validity and to validate that the assessment result was representative of the company, the five employees that were participating in the interviews were asked for feedback in informal discussions in the internal communication channels. They were presented with the model, some background information to the model and the assessment result, and were able to give answers to if they believed the model is a good tool to assess a company in how well they use continuous experimentation, as well as if they felt that the assessment result was a good representation of the company. The employees were also presented with the classification models for control of roadmap and distance to users, as described in phase 4. Out of these five people, a total of four people responded with feedback.

The answers from the employees were relatively short, and therefore no coding techniques or similar were used for the analysis of this data. Instead, conclusions were made directly based on the answers, conclusions which have clear chains of evidence to the actual feedback sentences.

3.2.4 Phase 4: Classification

In order to be able to work with the classification concepts control of roadmap and distance to users it was necessary to define a way to correctly and deterministically classify a company on how much control of roadmap it has and how large the distance to users is. If independent people classify the same company with these classification concepts it is important that they reach the same conclusion. There is little information in research touching on the subject of control of roadmap and distance to users, and how they relate to continuous experimentation, and therefore these deterministic models were constructed by this thesis author.

During the work of this thesis a classification model that establishes how much control of roadmap a company has was defined, and a classification model that established how large distance to users a company has was also defined. These classification models have been defined on an ordinal measurement scale, in order to be able to categorize companies without the need for a highly developed and validated ratio measurement scale. These classification models have made it possible to differentiate between companies in relation to the two concepts control of roadmap and distance to users.

The idea of using classification models for control of roadmap and distance to users came from reflections on initial interviews from phase 2, where the company's relationship to its clients was discussed. So data collected in this phase formed some sort of base for the definition of the classification models, and the data described the problem that seemed to exist in the company (i.e. the problem of working with

clients that are the product owners, rather than themselves being the product owners). Work inspired by these initial interviews and observations enabled the creation of the two classification models. To validate the classification models it was possible to use data from the feedback round that is described in phase 3. In that round the company employees were asked for feedback on the custom assessment model, but also for feedback on the classification models for control of roadmap and distance to users.

3.2.5 Phase 5: Relationship

The purpose of this phase was to answer if there is a relationship between the concepts control of roadmap, distance to users and continuous experimentation. Since continuous experimentation requires data collected from users as well as the ability to easily implement and deploy experiments to test hypotheses [12] there has been a hypothesis from early on in this thesis work by the thesis author that control of roadmap and distance to users affect how continuous experimentation is used in a company. This phase enabled confirmation or refuting of this hypothesis, and gave an answer to RQ1.

The data used for this phase was the interview data collected in phase 2, as well as the documentation data collected at that phase. Furthermore, data collected in phase 6, where perceived effects were uncovered, also formed the base for this phase.

3.2.6 Phase 6: Perceived effects

Phase 6 was meant to uncover what the perceived effects are of using more continuous experimentation in a company with low control of roadmap and a large distance to users. The effects are specifically the advantages, disadvantages and blocking issues. This would give an answer to the third and final research question (RQ3).

Data collection in this phase was done in two parts: 1) a quantitative survey sent to all of the company employees, and 2) qualitative interviews with several selected employees. The idea was to get an overview of what perceived effects the company identifies from the quantitative survey data, and to get a more in-depth understanding of the answers in the survey from qualitative interviews. The focus of this phase was to ask about what effects the participants perceive, but simultaneously the classifications control of roadmap and distance to users were also discussed.

The survey was an online survey, and it started with a short description of the concept of continuous experimentation. After the definition, it asked which advantages, disadvantages and blocking issues the participant identifies if the company was to use more experimentation. These questions had predefined answers with multi-answer checkboxes, but they also had the possibility of filling out a free text "other" option. After these initial three questions, the survey briefly introduces the classification control of roadmap and asked the participant to classify their company with the provided deterministic classification model for control of roadmap created

for this thesis. The participant could choose one of the three stages in the model (Stage 0, Stage 1, Stage 2). Then the participants were asked to rate how much they believe that control of roadmap affects the use of continuous experimentation, on a scale from 0 (Not at all) to 4 (A lot). The next part of the survey was a brief introduction of the classification distance to users. The participants were shown the deterministic classification model for distance to users, also created for this thesis, and asked to rate their company on the three stages (Stage 0, Stage 1, Stage 2). The final question asked to rate how much they believe that distance to users affects the use of continuous experimentation, on a scale from 0 (Not at all) to 4 (A lot). The full survey can be found in Appendix B.

The survey had a total of 13 responses, which is a sufficiently high response rate for a company with 15-20 employees, and it should be considered high enough to be representative of the entire population. The survey responses can be found in Appendix C.

For the qualitative interviews, 4 people were selected on an availability basis, i.e. the people who felt that they had the time. This meant that a total of 4 interviews were held for this phase, with a single person in each interview, and only one of the interviewees had participated in an interview in phase 2 before. The interviews were between 15-20 minutes long and contained open questions in a semi-structured fashion. The topics discussed were similar to the ones in the survey, and the interviewees were asked what perceived effects they identify if the company were to use more continuous experimentation, how they would classify the company on control of roadmap and distance to users as well as how they believe control of roadmap and distance to users are related to continuous experimentation. The participants were also asked if they thought it was desirable to use more continuous experimentation in the company, which was not covered in the survey. The interview template used for these interviews can be found in Appendix D. The interviews were held in Swedish, and any quotes from the interviews that are used in this thesis were translated to English by the thesis author.

External research was used to provide some data-triangulation to the perceived effects identified. There exist several examples of perceived effects of using more continuous experimentation already [27] [35], although they are not specific to low control of roadmap and large distance to users. To gain more confidence in the identified perceived effects, it is reasonable to expect that several of them should probably be existing in the research already.

Data analysis for the survey was made by looking at the survey answers and drawing conclusions directly from the answers and the descriptive statistics. Each conclusion that was drawn maps closely to information found in the survey result, and it is possible to trace what data each conclusion originates from. For the interviews, a similar analysis as done in phase 2 was used, i.e. by transcribing the interviews and using a thematic coding technique to draw conclusions. The participants in these interviews were also sent the transcript and analysis from their interview, to give

them a chance to correct any misinterpretations or retract any invalid statements.

3.2.7 Phase 7: Static validation

Given that this case study only contained a single company it was a challenge to gain generalizable results from the research. To counter this issue and gain more confidence in the results, especially for the first research question about the relationship between the three concepts, static validation was used. The static validation was performed by conducting four additional interviews with other companies than the case study company. These companies were very different from the case study company, both in for instance company size, field, control of roadmap and distance to users. The interviews were about 30 minutes each, and they contained open questions in a semi-structured fashion. These interviews were valuable because they gave the opportunity to learn what other companies think about the topics that are discussed in this thesis, which adds or removes validity from the results, without having to add one or more companies to the case study which would be very time-consuming.

In the four interviews the discussion points were an introduction to their company, how experimentation is used in the company, how they would classify their company or their team in the company when it comes to control of roadmap and distance to users, and if they believe that control of roadmap, distance to users and continuous experimentation are related to each other. For one of the four companies the custom assessment model was also used to assess the interviewee's company, and the interviewee provided feedback on the assessment model itself and if it in a good way assesses a company when it comes to continuous experimentation. This interview content allowed the lack of confidence in the results as described above to be approached. Specifically by discussing the relationship between the three concepts, and by talking to one other company about the assessment model. The interview template that was used for these four interviews can be found in Appendix E. Some of these interviews were held in Swedish and some in English, and any quotes from the interviews that were originally in Swedish and that are used in this thesis were translated to English by the thesis author.

Company	**Size**	**Business**	**Is global**
Company 1	Small-scale	Makes software product(s)	No
Company 2	Large-scale	Makes software and technology	Yes
Company 3	Large-scale	Makes software and technology	Yes
Company 4	Large-scale	Makes software and technology	Yes

Table 3.2: Summary of the static validation companies

In table 3.2 a summary of the four companies used for the static validation can be found. The table shows the company size, what business they do and whether or not they are global. A global company is referred to as a company that operates

and has branches in many countries, where a non-global company would mainly be located in a single country. The four companies are also labeled as company 1-4, which will be referenced in the rest of this thesis. In the table, it is shown that one of the companies is a small-scale company, while the rest are large-scale. All of the companies make some sort of software, but three large-scale companies also make technology. The three large-scale companies are considered global companies, while the small-scale company is not. The companies are described further, based on the data collected in the interviews, in an upcoming chapter.

The static validation interviews were analyzed less formally than the initial interviews, and conclusions were drawn directly from the notes and/or the recordings of the interviews. The analysis was made less formally since the discussion topics were few and since the answers clearly mapped to a specific topic. Therefore it still exists a clear trail of evidence. The interview participants were also given the possibility to give feedback on the analysis made, which increased the trust in that the analysis was done correctly.

3.2.8 Summary of the data collection

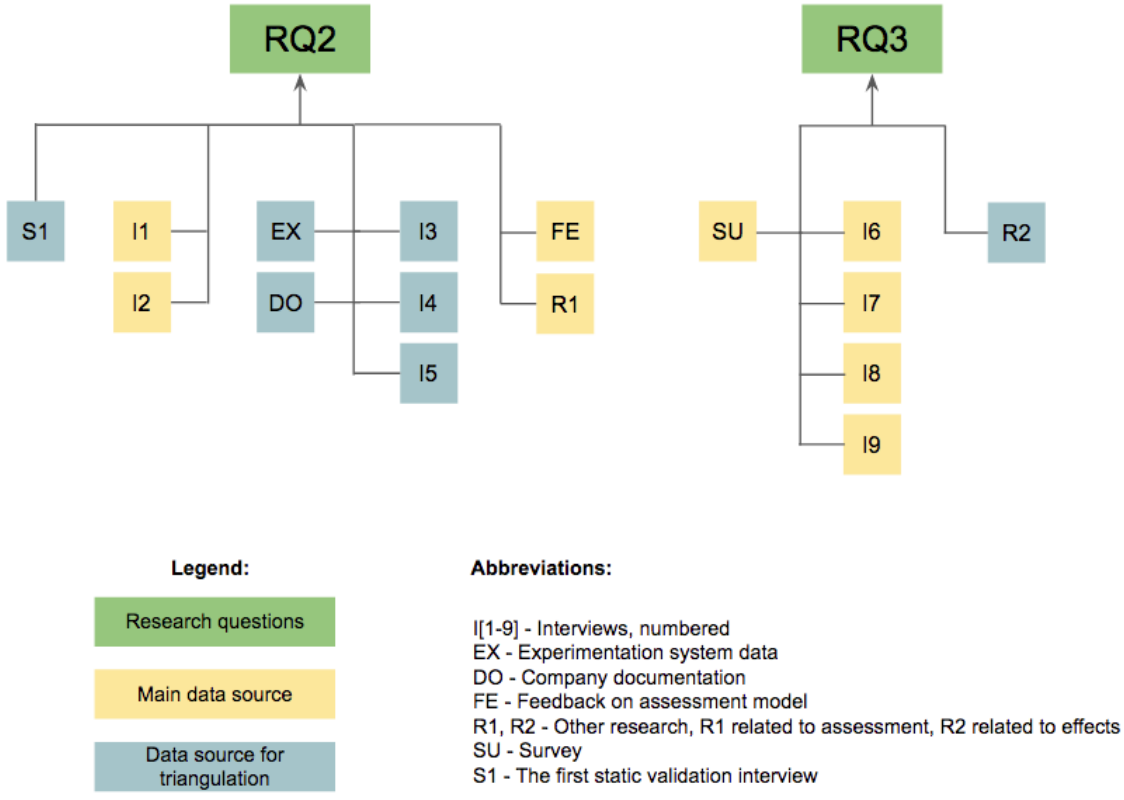


Figure 3.2: Summary of data collected for RQ2 and RQ3

Figure 3.2 shows a summary of the data collected for RQ2 and RQ3. For RQ2, which is the assessment of how well a company uses continuous experimentation,

the main sources of data were the first two interviews that were performed, as well as feedback collected on the model and the other research used. The data for data-triangulation was the data from the experimentation system, the documentation, three additional interviews inside the case study company and a single interview in the static validation phase with other companies. For RQ3, which is the perceived effects of using more continuous experimentation, the survey and the four interviews were the main sources of data. For data-triangulation other research was used.

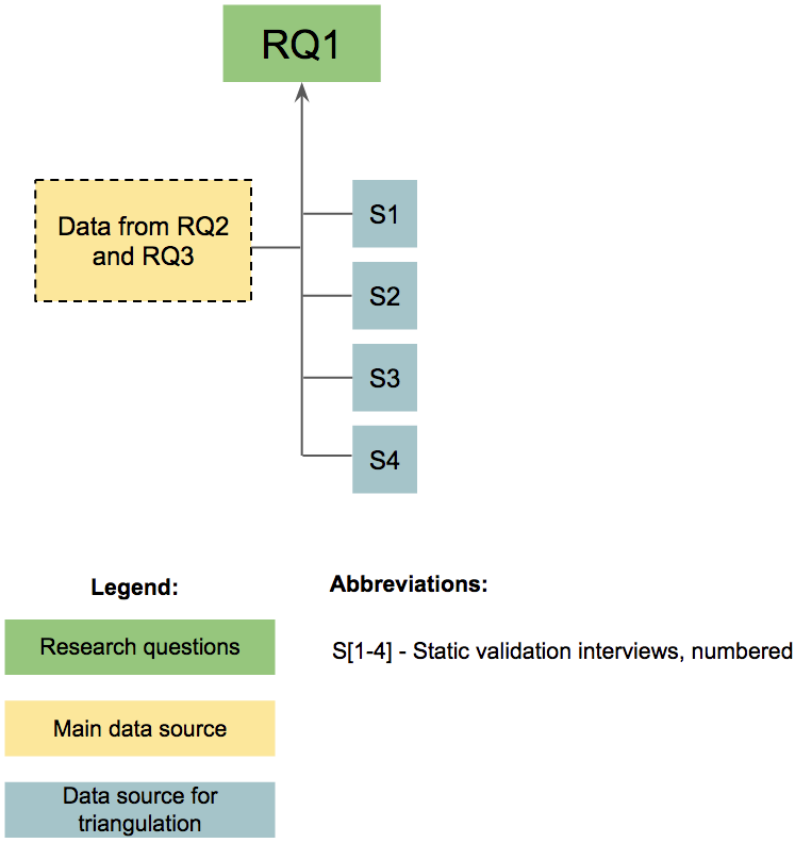


Figure 3.3: Summary of data collected for RQ1

Figure 3.3 shows a summary of the data collected for RQ1, which is the research question about the relationship between control of roadmap, distance to users and continuous experimentation. The main data source was data collected for RQ2 and RQ3, and the four static validation interviews with other companies were used to data-triangulate.

3.3 Validity threats

This section goes through different threats of validity to this thesis work, according to four aspects of validity discussed by Runeson and Höst [9], as well as how those threats have been countered. These aspects are construct validity, internal validity, external validity and reliability. Furthermore, the thesis author's affiliation with

the case study company and how this affects the research is also discussed in this section.

3.3.1 Construct validity

Construct validity refers to how well measures, such as interview questions, actually correspond to what the researcher is looking for, i.e. if it really answers the research questions [9]. One of the measures taken to increase the construct validity has been to provide interview participants and survey participants with brief definitions of concept they should know about before participating. Concepts such as continuous experimentation, control of roadmap and distance to users have been defined to the participants. Another measure that has been taken is that the thesis work has been made in close collaboration with the supervisors, which provided very useful feedback to make sure that the data-collection and data-analysis were done properly. A final measure that has been taken is reflection over the way interview participants use specific terms so that cases when participants use different definitions compared to the interviewer could be identified. The often used semi-structured format of the interview questions allowed additional questions to be added if needed, that helped clarify how interviewees might interpret a specific term.

3.3.2 Internal validity

According to Runeson and Höst internal validity means how for a causal relationship that is being studied there can be external factors causing the observed effect, which the researcher might not be aware of [9]. Since the first research question of this thesis considers if there is a relationship between the three concepts continuous experimentation, control of roadmap and distance to users and because the other research questions are dependent on this relationship, there is some sort of causality studied in this thesis work, even though the direction of the causality is not necessarily considered. To try to achieve high internal validity data triangulation has been used for the data collection, to make sure that what is said is not only a specific individual's belief. The data-triangulation is also used to make sure that all or most of the information appears so that there are no hidden causalities that affect the results. Given that many interviews have been performed, relative to the small size of the case study company, and because experimentation data and documentation have been included as data, there is a reason to believe that most or all of the information that was sought after was discovered in the data collection stages. Furthermore, the feedback received from supervisors has also been useful to assure the internal validity. For instance feedback on what data should be collected. The employees of the case study company have also provided feedback on, for instance, the custom assessment model, and this feedback has been another useful component for assuring internal validity, but also the construct validity. Finally, the thesis author has also tried to anchor the research in existing research and literature, which is useful because it shows if the research methodology or results differ a lot from somewhat similar things others have done before.

3.3.3 External validity

External validity means how much it is possible to generalize from the findings, i.e. if the findings are also useful to others [9]. This research is made as an exploratory case study with a single case study company, and the main purpose of the thesis is not to provide generalized results, but rather to explore the research gap described in this thesis. Further work will be needed to gain generalizable results. However, some generalizability is still possible to achieve, which is described further in this section.

The first research question, which is about the relationship between continuous experimentation, control of roadmap and distance to users, is possible to answer with high external validity. The reasoning behind this is two-fold. First of all the research question is phrased in such a way that it does not necessarily require the direction or the magnitude of the relationship to be defined. To make the other two research questions relevant, specifically with the part about "low control of roadmap and large distance to users", there only needs to be an indication of that there is some relationship between these concepts. Therefore it should be possible to answer this question with high external validity because it does not require the causality to be fully identified. The second reason why the external validity should be high for the first research question is because of the static validation performed with four other companies. The second research question, which is about how to assess how well continuous experimentation is used, should be possible to answer with high external validity because it is based on other research, because it should have high internal validity and because one of the companies in the static validation has provided feedback on the model. Finally, the third research question regarding the perceived effects should have some external validity since it should be possible to identify which of the perceived effects are related to having low control of roadmap and a large distance to users, and to find references of many of the other perceived effects in other literature on the same subject. The perceived effects that are not already identified by other research could have some external validity if it can be reasoned that they are not only specific to the case study company, but to this type of company, for instance if they are believed to exist because the company has low control of roadmap and large distance to users.

3.3.4 Reliability

Reliability means how much the results are dependent on the actual researcher. For the research to be considered reliable, then if another researcher did the same research they should end up with the same results [9]. To enable high reliability, guidelines of cases study work has been followed to a large extent as mentioned earlier in this chapter. Clear case study protocols have been used, interviews have had predefined interview templates and all interviews with the case study company were transcribed and coded. Finally, the supervisors' experience in research has been very useful, since they have been able to provide valuable feedback on the research methodology itself.

3.3.5 Affiliation with the case study company

The author of this thesis is employed at the case study company which comes with benefits but also possible threats to the validity of this research. The main benefit of this affiliation is that it allowed the author to get better access to the case study company for data-collection purposes, e.g. by interview participants being more willing to participate because they know the researcher. However, this also meant certain threats to validity. It was recognized by the author early on that specific measures had to be taken in order to not bias the work. One such measure was to make sure that no data was added to the research based on the author's own experience with the company, and to make sure that conclusions were drawn from an objective point of view. Furthermore, it also helped to follow the case study instructions mentioned previously, as well as working on the research methodology together with the supervisors, in order to get a reliable result. Finally, another measure that has been taken is to make sure to objectively observe the company and not try to present information about the company subjectively, for instance, when doing the assessment of how well the company uses experimentation.

4

Models

This chapter shows the models that resulted from this thesis work. Firstly, the deterministic classification models for the assessment of control of roadmap and distance to users are presented. Secondly, the custom assessment model is presented.

4.1 Control of roadmap and distance to users classification models

Since it was necessary to come up with a way of deterministically determine how much or little control of roadmap and distance to users a company has, two models have been created. One model for the classification control of roadmap, and one model for the classification distance to users. Both of the classification models consist of three stages 0, 1 and 2 to which a company could be assigned. The number of stages was decided to be three in order to keep the models simple, while still enabling enough details to differentiate between companies. Both control of roadmap and distance to users needed a zero-stage, where the control and the distance were non-existing. Furthermore, if it was existing, there needed to be at the minimum two stages to differentiate companies, where three more stages on top of the zero-stage would have created a difficulty in differentiating between the stages. Therefore the number of stages were chosen to be three. The stages allow for transitioning between them, and if the person using the models, for instance, considers high ability to control roadmap a good thing, the desired direction would be transitioning towards stage 2. The same goes for distance to users. In the models the word "team" is used, which is referring to either the company as a whole or a specific product or feature team inside the company, depending on how the organization structure looks like. The person using the model has to set their own context where they decide if the team should refer to the whole company, or a single product or feature team. Furthermore, for both the classification models the transitions between stage 0, 1 and 2 are not necessarily equally large. So it can be possible that transitioning between stages 0-1 is not equally difficult as transitioning between stages 1-2.

	Stage 0	Stage 1	Stage 2
Ability to control roadmap	The roadmap can not be controlled by the team at all. There is no possibility for the team itself to put desired product changes on the roadmap.	Desired product changes can be added to the roadmap by the team, but there are difficulties. The team does not have the full authority to decide on specific changes themselves.	Desired product changes can be added to the roadmap easily by the team itself, and the team is allowed to make the changes they believe are beneficial for the development of the product. Decisions related to changes that might have a very high impact might still be dependent on external parties.

Figure 4.1: Control of roadmap classification

Figure 4.1 shows the model for classifying a company in regards to control of roadmap, or "Ability to control roadmap" as the model defines. The three stages represent no control of roadmap at all (Stage 0), low control of roadmap where the team does not have full authority to decide on specific changes themselves (Stage 1) and high control of roadmap where the team has full authority to decide on specific changes themselves (Stage 2). The model refers to "product changes" or just "changes", which refers to both short-term product changes such as changes put on the backlog, and long-term product changes such as changes put on the product roadmap. For Stage 2 the model defines that although the team has high control of roadmap, the team might still be dependent on external parties to make changes that might have a very high impact. For instance, if the team wants to change something that might seriously impact business metrics, then business management parts of the organization might want to approve the idea first.

	Stage 0	Stage 1	Stage 2
Ability to access users	Users can not be accessed at all. There is no possibility of either qualitative or quantitative data-collection.	Users can be accessed, but there are difficulties with accessing them. Especially when new types of quantitative or qualitative data that is not currently being collected should be collected.	Users can be accessed with ease. To make the decision to collect new qualitative or quantitative user data can be done with close to no delay. Decisions related to any changes regarding data-collection with a very high impact might still be dependent on external parties.

Figure 4.2: Distance to users classification

Figure 4.2 shows the model for classifying a company in regards to distance to users, or "Ability to access users" as the model defines. The three stages represent very large distance to users (Stage 0), i.e. no ability to access the users, large distance to users where the team can access users but has difficulties (Stage 1) and a short distance to users where the team can access users without any difficulties (Stage 2). The model defines in Stage 1 that for new qualitative and quantitative data it might be difficult to access the users. This means that a company that at this moment already has easy access to specific qualitative and quantitative user data might still have a large distance to users if they have a lot of difficulties in collecting new data. For instance, if a company already has worked out the ability to collect data about how users use the application, if they wanted to now start learning who the users are and they have a lot of difficulties in the ability to collect this data, then they would still be considered to have a large distance to users. These difficulties would not be technical ones since everyone needs to invest resources into collecting new user data, but rather organizational difficulties. In Stage 2 of the model, it is defined that the team has the ability to decide to collect new data with close to no delay,

so in the previous example, if there are no organisational challenges when it comes to making the decision to collect the data the team would have a short distance to users. Stage 2 of the model specifies that decisions related to data-collection with a very large impact might still be dependent on external parties, which means that someone else might need to approve the changes if the changes have a very high impact even though the team might still have a short distance to users. An example is if the team, for instance, would start to collect new data that can impact the business metrics a lot, the business management would probably want to have a say in the decision, even if the company in general has a short distance to users. Since the model specifies both quantitative and qualitative user data, the model expects both qualitative and quantitative to be easily accessed for qualifying for a short distance to users.

4.2 Custom assessment model

Experimentation dimensions

		Stage 0	Stage 1	Stage 2
Technical	**Data**	No logging is done.	Logging of basic events is done, such as which views in the product users normally visit or what basic actions they perform (e.g. clicks on buttons).	Logging is done comprehensively with information regarding most of the user activity, such as durations in views, flows through the product or technical details of user hardware.
	Experimentation platform and statistical foundation	No experimentation platform is used, if experiments are run they are run manually.	Custom made or 3rd party experimentation platform exists. Basic features like defining variations, selecting sample size, experiment duration and assigning users to groups exist and are actively being used.	Advanced features, such as A/A testing, power analysis, alerting, automatic shutdown of harmful experiments and interaction detection exist in the platform and are actively being used.
Business	**Metrics**	No metrics are created.	Basic single metrics are created for measuring success.	Success, debug, guardrail and data quality metrics exist. Overall Evaluation Criteria are created that combines several metrics.
	Type and extent of experimentation	No experiments are run.	Experiments are run but not systematically. Experiments are used for changes such as changing existing functionality, deciding if a feature should be removed, adding new functionality or quality assurance.	Experiments are run systematically on all four types of changes in Stage 1.
	Experimentation impact	If experiments are run, the experimentation has no or insignificant impact on the business metrics and the team's way of working.	The experimentation has some impact on important business metrics and it somewhat affects the team's planning and prioritization of product changes.	The experimentation has substantial impact on important business metrics and it completely affects the way the team is working.
Organizational	**Organizational structure**	If experiments are run, no data-scientists or experimentation experts are involved in the creation and execution of experiments.	There are data-scientists and experimentation experts directly involved in the creation and execution of experiments.	Teams and team members are educated in how to conduct statistically sound experiments on their own, and there are data-scientists and/or experimentation experts available to assist when help is needed.

Figure 4.3: The custom assessment model

During this thesis work, a custom assessment model has been created in order to assess a company in how well they use continuous experimentation. This assessment model is heavily inspired by the Experimentation Growth Model, created by Fabijan et. al. [44]. Many of the stages are similar to the ones created by Fabijan et. al., and similar concepts are used, but the custom model is made simpler and made a better fit to answer the research questions in this thesis. The model can be seen in figure 4.3. In the model, there are six different dimensions categorized in three different categories: Technical, Business and Organizational. The model

attempts to describe how well a company uses continuous experimentation on all these three categories. The model uses the word team, which is referring to either the company as a whole or a specific product or feature team in a company. It is up to the reader to decide on an appropriate context for when they use the model. Each dimension in the model has three stages (0, 1, 2) which represents a path on each dimension from not at all to fully evolved. The transitions between the stages are not necessarily equally large, which means that it is not necessarily equally difficult to transition between stages 0-1 and stages 1-2 in the same dimension. Furthermore, the dimensions in the model are not independent of each other, and there are cases where it is necessary to evolve in one dimension in order to evolve in another. One example is when it is desirable to evolve the "Experimentation impact" dimension from stage 0 to 1. Then it is also necessary to evolve on, for instance, the "Type and extent of experimentation" dimension to at least stage 1 because in order to have an experimentation impact there has to be experiments that are run.

When the custom assessment model was created, all of the dimensions defined in the Experimentation Growth Model were added to the custom assessment model. However, since the custom assessment model is for the entirety of continuous experimentation, while the Experimentation Growth Model is only for online controlled experiments (A/B tests), the regression-driven experiments part of continuous experimentation had to be covered by the model as well, i.e. the canary releases, gradual rollouts and dark launches. This meant adding new information to the existing information in the Experimentation Growth Model, and a problem that arose was that the information in the model became too much. Since the Experimentation Growth Model consisted of four stages for each dimension, and new information was added on top of these existing dimensions, the decision was made to only have three dimensions in the custom assessment model. This reduced the amount of information in the model, and it also made the model simpler since it allowed only three stages to differentiate from. There were also zero-stages added to the custom assessment model, which allowed for instance there to be no data-scientists involved or no experiments ran so that the model is appropriate for smaller organizations who might not have data-scientists in the organization or any experiments ongoing. The added zero-stages resulted in that the four stages on each dimension in the Experimentation Growth Model had to be summarized in two stages on the corresponding custom assessment model dimension. Furthermore, some dimensions were combined to create more simplicity, such as the "Experimentation team organization" and "Feature team self-sufficiency" in the Experimentation Growth Model, which became the "Organizational structure" dimension in the custom assessment model. Finally, some additional information was added to the model that was decided to be necessary by the thesis author, such as information to the "Data" dimension in the custom assessment model.

The first dimension in the model is the "Data" dimension, which is about how logging of data is done in the system. This represents which different types of data are stored, and is considered a technical part of the experimentation. Another technical dimension is the "Experimentation platform and statistical foundation" dimension.

This dimension discusses if an experimentation platform exists and what features it includes. It also specifies that the features should not only exist in the experimentation platform to qualify someone for a specific stage but that they also have to be actively used. This is because a lot of third party experimentation platforms might be very advanced and have many features, but if you don't use them you should not qualify for that stage. The first dimension touching the business side of experimentation is the metrics dimension. It specifies single metrics vs. Overall Evaluation Criteria, and the different types of metrics that can be used: success, debug, guardrail and quality metrics. These different types of metrics are defined by Fabijan et. al. [44]. A second business dimension is the "Type and extent of experimentation", that discusses what types of experiments are ran and to what extent they are ran. Finally, the third business dimension "Experimentation impact" specifies what impact the experimentation has both on the business metrics, i.e. if the experimentation actually impacts the important business metrics of the company, and also if it impacts the team's way of working. If experimentation does not actually impact important business metrics in the company it could be considered a to not be a successful experimentation and qualify for stage 0. Finally, when it comes to the organizational part of experimentation, there is the "Organisational structure" dimension. This dimension specifies how data-scientists and experimentation experts are involved in the experimentation. A data-scientist is meant as someone specializing in data collection and analysis, while an experimentation expert would be someone very knowledgeable about the actual experimentation in software products. A single person might qualify for a single one of these roles, none of the roles or both.

5
About the companies

This chapter presents the information learned about the case study company and the four static validation companies. The information resulted from the data-collection practices described in the research method chapter. For the case study company the data collected was from phase 2, which was to learn about the case study company, but also from phase 6, which included information about control of roadmap and distance to users. The information about the static validation companies came from phase 7.

5.1 About the case study company

Information about the case study company was, for instance, learned through several interviews, as described in the research method chapter.

The case study company is a software consultancy and makes web services for several different clients. The company has employees with four different roles; project leaders, frontend developers, backend developers and UI/UX responsible. The company does not have any data scientists in the organization who are specializing in data analysis. Since the company is a software consultancy, the product owners of the services are part of the client organizations. This means that the case study company is in charge of the development of the service and the technical parts, and the clients are in charge of the services themselves and their users. It is the clients who are responsible of the planning and prioritization of tasks, but the case study company has the ability to affect this planning and prioritization. The company has a quick release process and releases new software as soon as it is finished. They work with a process that is similar to Agile, but it does not include all the elements that some formal Agile processes might require. For most of the projects the company uses both continuous integration and continuous deployment, which enables the quick release process.

Both qualitative user data, such as in-app surveys or focus groups, and quantitative user data, such as user behavior in the services, is collected by the company or by its clients. The company has the ambition to be more data-driven and base more decisions on data, but often decisions are based on opinions and they are currently not data-driven as an organization. The company has historically been doing some experimentation, but it is done on a small-scale. The main experimentation technique that has been used is A/B tests. When it comes to canary releases, gradual

rollouts and dark launches the company has generally not used it. There might have been an occasion where something similar to one of those techniques was used, but in general they only do A/B tests. When it comes to the type of experimentation, the company mainly does business-driven experimentation. Some regression-driven experimentation is mentioned, but the main experimentation type appears to be business-driven. The changes that are experimented with are changing and improving existing features, deciding if a feature should be removed and to a small extent testing new functionality in steps. However, removing features is tricky since they are not the product owners. These changes are mainly to the user interface, and rarely functionality on the web services.

The main purpose of the experimentation that is done by the company is to test hypotheses they, or its clients, have. They also do experimentation to educate themselves. When the company has ideas about experiments they want to run, they need an approval from the client first. There is some distrust in whether or not experimentation is useful, and one interviewee believes that they don't gain anything from running the experiments. When performing A/B tests the company does not consider important statistical concepts, such as statistical power. They mainly consider the sample-size of the A/B test. When an experiment is finished is often decided on gut feeling, and the experiment durations vary. Sometimes they test statistical significance of an A/B test with the help of an online tool. Some of the company employees have experience with statistics, but not all, and there are no people specializing in data analysis involved in the experiment process. For the experiments the company works with single metrics, and they do not work with any multi-metrics functions. The effects that are considered with the single metrics are mainly short-term, and the company does not consider long-term effects a lot.

There are different views amongst the interviewees on the impact of the experimentation in the company. Someone says that the experiments have no or minimal impact, while someone else says that they will act on it if they see a clear difference in an A/B test. The reason why the answers differ might be because the interviewees work with, and talk about, different clients. However, there still appears reasons and the desire to use experimentation and to use data-driven decision-making from some employees, but not all.

Quantitative data collection in the company is mainly done through Google Analytics, although the company has some solutions for collecting its own data as well. The experimentation system used is Google Optimize, and for some project(s) the employees of the company had to write some code on their own for making the experimentation work with both Google Optimize and another application they use. The main implementation technique that is used for the experiments are feature toggles in the code, and they do not use runtime traffic routing.

Based on the survey that was conducted and the interviews for phase 6 there is a consensus that the company can be classified with control of roadmap stage 1, i.e. low control of roadmap. There are some in the survey who disagree, but the majority

thinks the control of roadmap is low. The company can influence the roadmap a lot, but it can not decide on changes to the roadmap on its own. This is because even though they discuss things with clients, it is in the end the client who makes the decisions, and in extension controls the roadmap. The client is the product owner, and therefore they are in charge of the roadmap. One interviewee discusses why the control of roadmap is low for the case study company.

> *"... and it is the same with control of roadmap, [...]. A client pays us to do a certain thing. [...] We are pretty limited, it depends a bit on the client, but often we are pretty limited in how much influence we have in defining the roadmap and prioritize the roadmap."* - Case study company employee.

Also based on the survey that was conducted and the interviews for phase 6 there is a consensus that the company can be classified with a distance to users on stage 1, i.e. large distance to users. However, there is more disagreement in distance to users compared to control of roadmap. In the survey a majority classifies the company on stage 1, while one in the survey classifies the company on stage 0 and the rest on stage 2. In the interviews two people classify the company on stage 1 and two people on stage 2. There appears to be a difference in how people classify the company when it comes to qualitative versus quantitative data, and this is probably the reason why the answers vary, i.e. it depends on if the answer is based on qualitative, quantitative or both. Collecting quantitative data, i.e. about user behavior, is easy as long as it is not sensitive, even if it is new quantitative data, but if they wanted to have for instance qualitative data through an in-app survey they would need to ask permissions from the clients. The same interviewee who discussed why the control of roadmap is low also discusses the distance to users is large for the case study company.

> *"We have fairly often a pretty long distance to the user. Because we are a consultancy, and there is... It is not us who directly look at the data, and do more qualitative analyses of the products. It is often our clients who do that, so that is a disadvantage that we have, [...]. We have a longer distance than if we for instance would have been a product company."* - Case study company employee.

5.2 About the static validation companies

For static validation purposes, four more companies in addition to the case study company participated in this research. Information about these companies was learned through interviews, as described in the research method chapter.

Company 1 generally owns its own products and is responsible for its users. It tries

to be data-driven in its decision-making processes, and it collects both qualitative and quantitative data. The employees of the company conduct experiments in their products, which are mainly A/B tests. The interviewee from the company believes that the company has a control of roadmap on stage 2, i.e. a high control of roadmap. The company is not dependent on external parties to control the roadmap. The interviewee believes that the company has a distance to users on stage 2 as well, i.e. a short distance to users. They have access to both user-data of a qualitative and quantitative type.

Company 2 owns its own products. Some experimentation is used in the company, but the interviewee believes that they are not very good at it today. However, they want to do it more and they talk about it. The interviewee mentions that they need to work on the habits regarding experimentation as well as on the experimentation infrastructure. The company has "early-adopters" who are willing to try new versions of the product(s), which is very useful for the company. The interviewee believes that the company (or a team inside the company) has a control of roadmap on stage 1, i.e. a low control of roadmap. The company has many teams, and therefore the roadmap is controlled from a much higher instance in the organization. The interviewee believes that the distance to users for the company (or a team inside the company) is between stages 1 and 2, but probably closer to 2, i.e. a short distance to users. The company needs more data to be collected, but the ability to access the users is not the thing stopping them from collecting more data and it is more about the need to invest the time.

Company 3 generally owns its own products. The employees of the company used to do a lot of experimentation, but today they do less experimentation. The experimentation techniques they used to do included A/B tests. The interviewee believes that the team he/she worked with previously had a control of roadmap on stage 1, i.e. a low control of roadmap. They were able to control the roadmap, but it was not easy. The product owner and his/hers own view influenced the roadmap a lot, and they also had to be accommodating for their different customers. Regarding the distance to users for the company it has varied for different types of data that was going to be collected. The interviewee points out that there are two parts needed for collecting-data: 1) technical part and 2) the approving decision. Getting the approving decision was often hard the first time for a specific area of data, but for the same area of data a repeated request for approval was much more simple.

Company 4 owns a lot of its own product(s). The company works with different clients in a B2B-domain, but the company itself is still responsible for the product(s). It uses some experimentation in some ways, but it is dependent on what is regarded to be experimentation and what is not. It does not however appear to systematically use experimentation techniques such as A/B tests. The company works with a requirement-driven approach, and the development teams gets requirements from product managers who makes the decisions about the product roadmap/backlog. People deciding on the requirements often has an economical/commercial background, and the development team itself is considered to have a low control of the

roadmap according to the interviewee. The interviewee believes that the distance to users is large, and that the company is far away both to the end-users and the clients they work with in the B2B-context. The size of the company is a reason why it is difficult to reach the clients. Since the clients wants to know what data is collected, they experience obstacles such as security when it comes to data-collection.

5.3 Companies on two-dimensional classification space

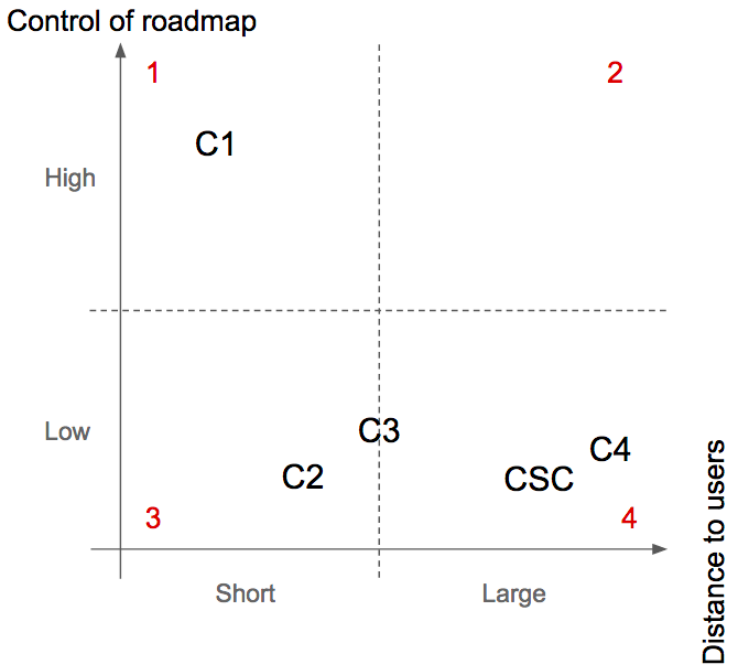


Figure 5.1: The five companies placed on a two-dimensional classification space

Control of roadmap and distance to users can together be classified on a two-dimensional classification space, where the distance to users is placed on the x-axis and control of roadmap is placed on the y-axis, dividing the space into four different quadrants. Figure 5.1 shows the four companies and the case study company placed in such a two-dimensional classification space. Company 1 is believed to have a high control of roadmap and a short distance to users, which is shown in the figure in quadrant 1 and labeled "C1". Company 2 is believed to have a low control of roadmap and a distance to users between short and large, but probably closer to short. This is shown in the figure in quadrant 3, and labeled as "C2". Company 3 is believed to have a low control of roadmap, and the distance to users has varied. Therefore the company is placed between quadrant 3 and 4, and labeled as "C3". Company 4 is believed to have a low control of roadmap and a large distance to users, which places the company in quadrant 4 with the label "C4". Finally, the case study company is believed to have a low control of roadmap and a large distance

to users, which is why it is also placed in quadrant 4 and labeled "CSC" as in Case Study Company.

6
Model feedback

This chapter presents the feedback received from the case study company and the static validation companies on the control of roadmap classification models as well as the custom assessment model. The information presented here comes from the data-collection described in the research method chapter.

6.1 Classification models feedback

This section describes the feedback received on the classification models for control of roadmap and distance to users.

6.1.1 Case study company feedback

When the custom assessment model had been created, it, together with the control of roadmap and distance to users classification models, was sent to the five people participating in the first five interviews. Four out of those people responded, and one of them had specific feedback on the classification models, where that person thought that there could be more stages in both the custom assessment model and the classification models. This was not mentioned by the other three people providing feedback. Furthermore, in the four interviews for the perceived effects phase, where control of roadmap and distance to users were discussed also, there appeared more feedback on the classification models. One interviewee mentions that an "average" control of roadmap might be useful. Another interviewee makes a remark about being in between stages 1 and 2 in control of roadmap, which indicates a potential need for more stages. One of the interviewees reflects on that the measurement scale for the distance to users model is a bit unfair, which is an indication that it should perhaps have more stages. There are several interviewees who do not make remarks about the need for more stages.

6.1.2 Static validation feedback

In the static validation, one interviewee believes that they are between stage 1 and stage 2 when it comes to distance to users, which indicates that there should perhaps be more steps. The same interviewee points out that control of roadmap can differ based on if you refer to the product roadmap or the product backlog, i.e. long-term product changes and short-term product changes respectively.

6.2 Custom assessment model feedback

This section describes the feedback received on the custom assessment model.

6.2.1 Case study company feedback

As part of phase 3 of this thesis work after the custom assessment model had been created some employees in the case study company also had the opportunity to provide feedback on the assessment model. Most or all of the asked employees approve of the model and thinks that it in a fair way assesses a company in how well they use experimentation. One of the employees believes that the data-dimension might not cover everything and that a normal Google Analytics instance qualifies a company for stage 2 automatically. The employee points out that there is more data that can be added on top of what Google Analytics provides, and that this is not included in the model. Another employee believes that the organizational dimension should perhaps not define that a certain type of person is needed and that an employee that has an understanding of how the business works and that has some analytical skills can do this type of work, i.e. there might not be a need for data-scientists and/or experimentation experts. Finally, another employee would like to see more stages, and perhaps also a weighted score that tells you how well you use experimentation based on the result from each dimension. The employee also points out that the "way of working" part of the "Experimentation impact" dimension should perhaps be placed under the organizational category.

6.2.2 Static validation feedback

During the static validation, phase 7 in this thesis work, one of the interviewees from one of the four companies was asked for feedback about the custom assessment model. The interviewee reflected on whether or not it is actually desirable to completely automate the statistical part of experimentation inside the experimentation platform, i.e. is it really desirable to move from stage 0 to stage 2 in the dimension "Experimentation platform and statistical foundation". The reasoning to why it might not be the desirable direction was that it is easy to get it wrong when you automate the statistical part. The interviewee also reflects on whether or not it is desirable that everyone can conduct experiments, perhaps without having any knowledge about statistics at all, without the help of a data-scientist in each team, i.e. whether or not the direction on the "Organisational structure" is the desired direction. However, the interviewee also mentions the idea of a centralized team of data scientists to which a team can send requests when they need help. The interviewee mentioned that he/she believes that the "Data" and "Metrics" dimension are very useful/important dimensions. In general, the interviewee thinks that the model indeed covers all of the parts related to experimentation and that it is very useful to break the experimentation down into the three categories technical, business and organizational. Finally, the interviewee says that he/she prefers a simple model, as it is now, rather than more stages on each dimension.

7

Relationship between continuous experimentation, control of roadmap and distance to users

This chapter presents the information learned about the relationship between the three concepts continuous experimentation, control of roadmap and distance to users. The information was collected in accordance with the data-collection practices described in the research method chapter.

7.1 Case study company result

The interviews and survey with the case study company revealed a lot of information about if there is a relationship between continuous experimentation, control of roadmap and distance to users. This was mainly discussed in the interviews and the survey for the perceived effects, which is phase 6 of this thesis work. Other information was learned in other phases as well, which formed the base of the idea behind the relationship.

The survey result shows that everyone believes that control of roadmap affects the use of continuous experimentation. The respondents believe that the control of roadmap affects the use of experimentation to different degrees, but no-one believes that there is not a relationship. When it comes to distance to users the same result is shown, and everyone believes that distance to users affects the use of continuous experimentation. Some believe it is more and some believe that it is less, but no-one thinks there is no relationship between the two concepts.

In the interviews, the control of roadmap is also considered to be related to how continuous experimentation is used. Someone points out that in the case study company the low control of roadmap is probably not the biggest thing preventing evolving the experimentation in the company, but that it instead is a lack of knowledge and will to do it. However, other interviewees claim that if they had more control over their own roadmap(s) they would have probably used more experimentation. An interviewee says that since the clients are paying them to do a specific thing they are limited in controlling the roadmap. The interviewee describes the low control of roadmap as a blocking issue for using more experimentation. Finally, one interviewee believes that the control of roadmap is not fixed and that it could

perhaps be changed if they wanted to.

When it comes to the distance to users the interviews show that the interviewees from the case study company agree that there is a relationship between distance to users and the use of continuous experimentation. However, not everyone believes that it applies to all types of experimentation or for the company's current circumstances. One interviewee argues that large distance to users is probably not affecting the use of continuous experimentation in the company, because if they wanted to experiment more they could change the distance to users. Another interviewee says that they have good access to a lot of system data, so if they wanted to do experimentation related to for instance infrastructure, it would not be a problem. However, for experimentation like A/B tests the interviewee believes that the large distance to users would be an issue. Another interviewee highlights that a short distance to users makes the experimentation a lot easier to conduct and that user feedback is a precondition for experimentation that should be acquired before the experimentation begins. One interviewee thinks that with a large distance to users experimentation would be possible, but getting results of the experiments would require more time and it would be more difficult to do the actual implementation of the experiments. In one of the interviews, the interviewee says that he/she believes that a large distance to users is a blocking issue to using more experimentation. Finally, more than one interviewee talks about how it perhaps would be possible to reduce the distance to users if they wanted to.

7.2 Static validation result

All of the interviewees in the four different companies believe that there is a relationship, or that there probably is a relationship, between control of roadmap and how continuous experimentation is used in their companies or in a company. One interviewee believes that a low control of roadmap would mean that the development team probably views the product differently and not feel the same extent of ownership. This would have probably meant less experimentation. Another interviewee answers the question "Do you believe that control of roadmap is related to how experimentation is used in your company?" with how being a more autonomous team in a company improves the experimentation, and reasons why this is not trivial to achieve.

> *"Absolutely. It is. [...] The more autonomous you can be, the more control you can have in the team that develops a part of the product, the easier and the better it will become with the experiment. [...] The team defines what the next experiment is, and has a quicker cycle and quicker feedback-loops. That is outermost desirable, but not trivial to achieve. Because of all those reasons that I have mentioned, with organization, culture, process and technology."* - Static validation interviewee.

Another interviewee mentions that he/she believes that having low control of roadmap

will make the ambitions of using more experimentation significantly less. If the experimentation was done but the results not used for the roadmap it would make it less desirable to do experimentation. The interviewee also points out that the control of roadmap will also impact if the product is planned by upfront requirements or based on data-driven decisions through for instance experimentation. The interviewee believes that there is a lot of power in being in charge of the roadmap and trying to foresee what changes will have any effect, and if there is experimentation used for this instead, the person in control of the roadmap can be more in charge of the visions for the product. There is also a risk that needs to be considered where if the wrong experiments are being focused on, such experiments that yield little value, that the negative result of those experiments affects the control of roadmap in a negative way, i.e. that control for the team is reduced again.

The interviewees are all certain or very certain that there is a relationship between distance to users and the use of continuous experimentation in their companies or in a company. If there would not be any data it would be difficult to experiment. One interviewee points out that with a large distance to users he/she believes that experimentation would be more difficult since it limits what experiments can be run, what conclusions can be made and how well the experiments can be validated. One interviewee reflects on the need to ask why the distance to users in a company with a large distance to users actually is large, and that it is probably often possible to access the users better than initially thought. The interviewee mentions that the distance to users is probably not a fixed thing and that you probably can change it if you want to. He/she describes it as a barrier that you can remove. One interviewee reflects on that they have experienced personally how a large distance to users affects experimentation and gives the example that an experiment that takes three days to execute could require two months for approval of the data collection.

8

Perceived effects

The following chapter presents the result from the perceived effects part of the thesis work, i.e. phase 6. The chapter goes through the perceived advantages, disadvantages and blocking issues that the case study company identifies if they were to use more continuous experimentation. This information is based on the results from the survey as well as the interviews, as described in the research method chapter.

8.1 Advantages

There are several perceived advantages identified by the case study company with using more continuous experimentation.

Perceived advantage
Quality assurance
Products are improved
Identify risks that would maybe not have been identified without experimentation
Get new insights
More data-driven arguments provided to the client / less emotional arguments
Building the right things
Clients can make more money

Table 8.1: Perceived advantages of using more continuous experimentation identified in the interviews

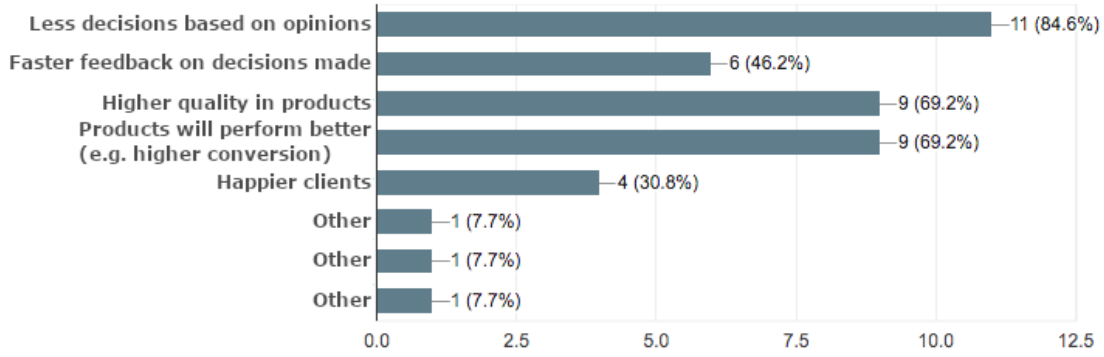


Figure 8.1: Perceived advantages of using more continuous experimentation identified in the survey

Table 8.1 shows the perceived advantages of using more continuous experimentation identified in the interviews, and figure 8.1 shows the survey responses on the same topic. In the interviewees the perceived advantage of having more data to back the arguments the case study company makes towards the clients is mentioned, i.e. more data-driven arguments. The interviewee explains that an advantage of using more experimentation is that they don't have to convince the client(s) to trust them and that they instead can provide evidence for the claims they make. This is also mentioned by another interviewee, that believes an advantage is that less emotional arguments will be used, arguments that instead will be substituted by facts. This would, according to the interviewee, avoid pointless discussions where people have strong personal opinions on specific features. Almost everyone in the survey identifies the advantage of fewer decisions based on opinions, and many in the survey believe an advantage is the feedback on the decisions made is received faster. One interviewee mentions the advantage of building the right things, or at least to have the feeling of building the right things. Furthermore, an interviewee mentions the perceived advantage of the client making more money, and several respondents of the survey believe that a perceived advantage is that the clients become happier.

> *"I think that because there is a lot [specific type of web service] that [case study company name] is doing, it is... Everything that can make the customer earn more money by analyzing how customers [means end users] behave... is a very big advantage."* - Case study company employee.

One advantage that many people in the survey identify is the higher quality in the products, and this is also mentioned by two interviewees as a perceived advantage. One interviewee mentions the ability to identify risks that would perhaps otherwise not have been identified, as well as the advantage of getting new insights from using experimentation. An "other" answer in the survey was that the experimentation could provide a foundation for future decisions that are going to be made. Another "other" answer in the survey revealed that someone perceives the advantage for a better harmony inside the own company since people will not use "vetos". Finally, two interviewees identify the advantage of an improved product, which is also a

perceived advantage by many in the survey. However, one interviewee mentions that because the case study company does not own its own products, it is probably not the one benefiting from A/B testing and learning more about how well the products perform, but the employees of the company are happy to do A/B testing for the clients if they desire it.

8.2 Disadvantages

Several perceived disadvantages with using more continuous experimentation are identified by the case study company.

Perceived disadvantage
More complex code
More complex devops and rollout situation
More difficult situation for new developer
Might require a lot of resources
Clients might not see the benefits of doing experimentation / Have to convince clients

Table 8.2: Perceived disadvantages of using more continuous experimentation identified in the interviews

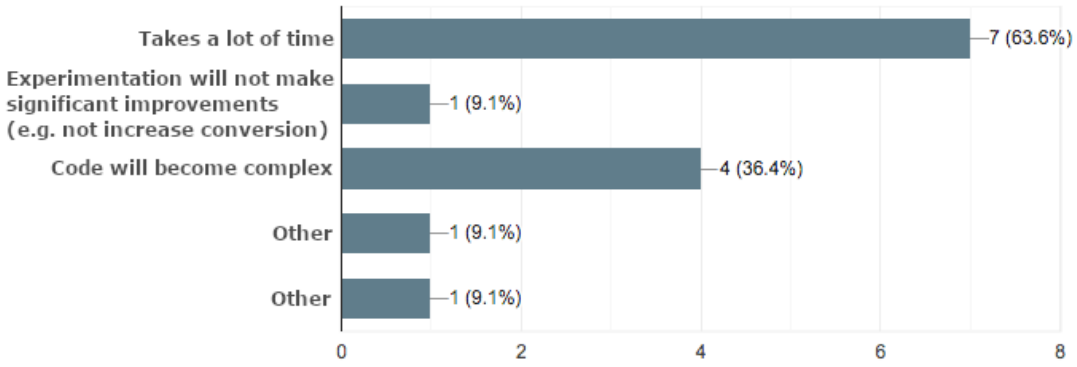


Figure 8.2: Perceived disadvantages of using more continuous experimentation identified in the survey

Table 8.2 shows the perceived disadvantages of using more continuous experimentation that were identified in the interviews, and figure 8.2 shows the survey responses to the same question. One interviewee mentions the perceived disadvantages of more complex code, devops situation, rollout situation and situation for a new developer.

> *"It can be more complex code. More complex situation when it comes to devops and rollouts. Harder for a... Someone who has not been partici-*

> *pating before to enter and be able to deliver, when there are several steps that you have to consider."* - Case study company employee.

More complex code is also something that several people in the survey identify. Furthermore, two interviewees talk about how they have to convince the clients that investments in experimentation are worth it.

> *"Sometimes the client might not understand the importance of doing these experiments, because they might not be as technical as we are, and then we might need to take a conflict, or what you can call it, with the client to argue for that 'Yes but this is actually worth 120 hours, that you spend on this, you might not immediately now see why, but in the long-term it pays off'."* - Case study company employee.

Another perceived disadvantage identified in the interviews is that using more experimentation might require a lot of resources, for instance time, which is also identified in the survey by a large number of respondents. In the survey, an "other" answer is that a perceived disadvantage would be that it would require more of a process in the company. Furthermore, one person in the survey believes that using more experimentation will not make any significant improvements, e.g. not increase the conversion rate. Finally, some general reflections are a survey respondent that says by the "other" answer that using more experimentation takes more time, but that it is almost always worth it. An interviewee also makes a reflection on that the disadvantages that the interviewee perceives are only relevant in the short-term, and that they would not be relevant in a more long-term perspective.

8.3 Blocking issues

There are multiple perceived blocking issues identified by the case study company with using more continuous experimentation.

Perceived blocking issue
Knowledge barrier
Limited resources, especially time
They need to convince client
They have to change their own organization

Table 8.3: Perceived blocking issues of using more continuous experimentation identified in the interviews

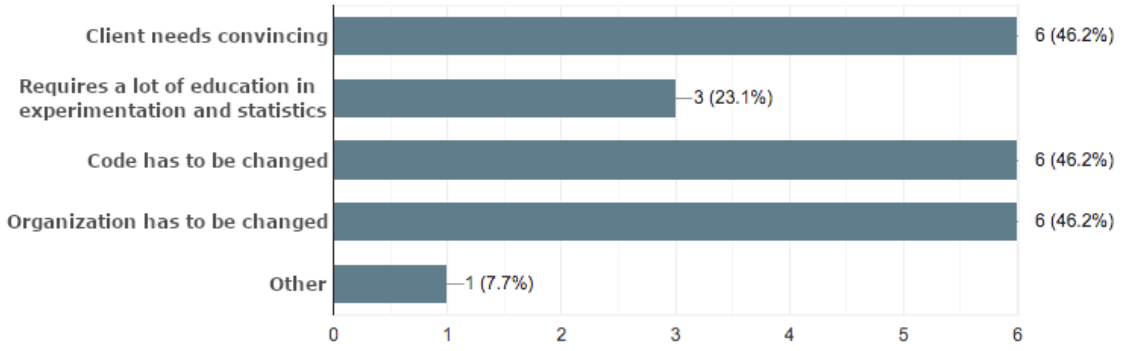


Figure 8.3: Perceived blocking issues of using more continuous experimentation identified in the survey

Table 8.3 shows the perceived blocking issues of using more continuous experimentation that were identified in the interviews, and figure 8.3 shows the perceived blocking issues identified by the survey respondents. One of the identified blocking issues from the interviews is that there is a knowledge barrier of using more experimentation.

> *"Knowledge barrier maybe. But it is always a... barrier until you have the knowledge."* - Case study company employee.

That the need for education is a blocking issue, i.e. that there is a knowledge barrier, is also recognized by some of the survey respondents. One interviewee mentions the blocking issue of having to convince the client, for instance, that it is worth it for the client to fund the experimentation, which is similar to a previously described disadvantage. The need to convince clients is also recognized as a blocking issue by several in the survey. Furthermore, one interviewee reflects on that it is not only the client that has to change, and that the case study company also has to change how they are working.

> *"It is very easy to only say that it is the client, [...]. We must also change a bit how we are working to make it become... we say that we want to work this way, but we have not actually worked that way to some larger extent or in any systematic way."* - Case study company employee.

That the own organization has to change is also something that several in the survey identify. Furthermore, one interviewee recognizes that limited resources, especially time, are blocking issues to using more experimentation, and an "other" answer to the survey is that experimentation needs to be prioritized when it comes to time can be seen as a blocking issue as well. Finally, several survey respondents also identify the blocking issue of that the code has to change in order to use more experimentation.

9
Discussion

This chapter introduces a discussion on the results and methodology of this master thesis. In this chapter, the research questions introduced in a previous chapter are answered.

9.1 Classification models

For this thesis it was early on identified that it was important to find a deterministic way to assess a company on how much control of roadmap and how large or small distance to users they have. As part of this thesis, two classification models were created: one for control of roadmap and one for distance to users. The models contained three stages each: stage 0, stage 1 and stage 2. It was an active decision to only have a small number of stages, so that the models would remain simple to understand and apply to a given company. The intervals of the stages were quite naturally identified since the zero-stage would be "none at all" and the last stage would be as much as possible, and the middle stage would be some sort of in-between value. So for control of roadmap, the natural zero-stage was no control at all, the last stage full control and the middle stage some control but there are difficulties. For distance to users the zero-stage naturally became no ability to access users, i.e. a very large distance to users, the last stage a short distance to users, and the middle stage a large distance to users where it was difficult in accessing the users, but still possible. A problem with the models could be that it is a bit difficult to determine what constitutes as "difficulties", however, it should probably be fairly obvious to most contexts if it is very easy, hard or not at all possible to control the roadmap or access users.

There were multiple people, who provided direct feedback on the models or reflecting in the perceived effects interviews, indicating that there perhaps were too few stages in the models. There was also an indication in a static validation interview about the same thing. However, there were also people who did not mention the need for more stages. The number of stages was, as mentioned earlier in this thesis, a deliberate choice, and it is possible to believe that simplicity should be prioritized over more complex models. It would also be much less trivial to place a company on the model if there were many stages. Consider the example of control of roadmap and if it would have stages no, low, medium and high control of roadmap. To differentiate between low and medium, or medium and high, could perhaps be quite difficult, especially with the goal in mind that the two same people who assess the same company should

arrive at the same conclusion. Therefore the number of stages stayed intact at three. Finally, in a static validation interview, it also appeared that it was important to differentiate between the product roadmap and the product backlog, i.e. short-term versus long-term product changes. This feedback was considered, and in this thesis it has been clarified together with the definitions of the models. Despite how people interpreted the definition of control of roadmap in the interviews, i.e. if despite if they thought of it as short-term or long-term product changes, there is a reason to believe that the result still holds and would not differ if it would have been more clearly defined that both backlog and roadmap should be considered. It is possible to believe so since this was not mentioned by someone until in one of the final phases of the thesis work, and because of the definition of control of roadmap that was introduced to interviewees and survey participants, which defines product changes in general and does not say that they are either short or long-term: "It [the classification of control of roadmap] indicates how much control the company has to change what features and product changes are planned and how they are prioritized.".

9.2 Case study company

The case study company is a software consultancy who provides web services to multiple clients, which means that the clients are the product owners. It is apparent that this company structure and the relationship to the clients causes a lot of special cases, that might not apply to a company that has full ownership of its own products. This is indicated by the many mentions of the client relationship throughout the data-collection part of the thesis work. One of the initial theories early in this thesis was that the case study company could be considered to have a low control of roadmap and a large distance to users. This was important since it affected how the research questions were defined, and the "software company with low control of roadmap and a large distance to users" part of research questions RQ2 and RQ3 would not be relevant in the research questions if the case study company did not fulfill these preconditions. The idea that the company has low control of roadmap and a large distance to users was confirmed by the data collected for this thesis, which is why the phrasing of the research questions still maintain a relevance.

9.3 Static validation companies

The static validation part of this thesis was primarily meant to increase the external validity of the thesis result, especially in regards to any relationship between control of roadmap, distance to users and continuous experimentation. From the interviews with the four static validation companies, it is apparent that the four companies all generally own their own products, but they do differ from each other in how they use experimentation. There is also a difference between the companies when it comes to control of roadmap and distance to users, and only one of the four companies shares the same classification results as the case study company, i.e. low control of roadmap and a large distance to users. The companies together cover all quadrants in the

two-dimensional classification space, with the exception of quadrant 2: high control of roadmap and large distance to users. This diversity over the three quadrants in the two-dimensional classification space is useful because it allows seeing the perspective of other companies that are not exactly like the case study company. The fact that they own their own products is also useful since it allows a perspective from companies that have different ways of working with their products.

9.4 Relationship between concepts

The first research question in this thesis was defined as the following.

> *RQ1: Are the classifications control of roadmap and distance to users related to the use of continuous experimentation in a software company?*

The research question was intentionally defined quite openly, without specifying the direction or the magnitude of the relationship between the concepts, since it was only necessary to learn if there was any relationship identified between the three concepts. If there were no relationship, the control of roadmap and distance to users should not be part of the other two research questions RQ2 and RQ3, which would have meant that the work done in this thesis would not have been unique in the field of continuous experimentation. However, it is clear from the interviews and the survey in the case study company that there is a relationship both between control of roadmap and continuous experimentation, as well as the distance to users and continuous experimentation. This was also confirmed by the four static validation interviews, who believe the same thing. It is, therefore, possible to answer the research question with that there is indeed a relationship between the concepts continuous experimentation, control of roadmap and distance to users. More specifically, it appears that control of roadmap affects how experimentation is used, and that the distance to users also affects how experimentation is used. If the direction and magnitude would be part of the research question it would perhaps require more and deeper data-collection regarding these relationships, but the data collected in this thesis should be enough to answer the research question with that there is a relationship. This effectively makes the control of roadmap and distance to users part of the research questions RQ2 and RQ3 relevant.

9.5 Assessment

This section discusses the assessment of how well continuous experimentation is used, and it also provides an example usage of the assessment model.

9.5.1 Reflections on the assessment model

The second research question in this thesis was defined as the following.

RQ2: How can a software company with low control of roadmap and a large distance to users be assessed in terms of how well they use continuous experimentation?

Based on the Experimentation Growth Model by Fabijan et. al. [44], a custom assessment model was created for this thesis. The model provided by Fabijan et. al. was not sufficient to answer this research question since it was meant for online controlled experiments and large-scale companies. The custom assessment model is however designed to work for continuous experimentation as a whole, and not only online controlled experiments, and it is meant to work for all type of sizes of companies. It fits continuous experimentation in its entirety since it talks about both changing existing functionality, adding new functionality, and deciding when to remove functionality, as well as the quality assurance aspect of continuous experimentation. It is designed to work for all types of sizes of companies since it has added necessary zero-stages to the dimensions, which were lacking in the Experimentation Growth Model. For instance, the need of a zero-stage for "Organisational structure" would be required for a small-scale company, but perhaps not a large-scale company, since a small-scale company might not have dedicated data-scientists, while a large-scale company would more or less always have some data-scientists employed. Furthermore, the custom assessment model was designed to have few stages rather than many stages and more detailed steps for the dimensions. This was a deliberate choice since simplicity was preferred over complexity, and because it was believed that the model would be too big and too complex if it contained six dimensions and more than three stages on each dimension.

The feedback received on the custom assessment model from the case study company was generally good. One employee mentioned that the data dimension might need another step since the normal Google Analytics installation would qualify as stage 2. This was a deliberate design choice, since the Google Analytics data is quite advanced, for instance by including information about who the actual users are, and therefore it is in the opinion of the thesis author that Google Analytics should indeed qualify for the highest stage. Another feedback point received was that the organizational dimension should perhaps not define that a certain type of person is needed, that a data-scientist and/or experimentation expert is not needed. This feedback is, of course, depending on where the company itself puts the desired level of statistical foundation for its experimentation. If assuring a deep statistical foundation is not the highest priority to the company, it is understood why employing data-scientists and/or experimentation experts would not be a priority. However, when the experimentation is scaled, especially the experimentation impact, it should be more important to verify the statistical foundation, for instance as shown in how Microsoft develops its experimentation platform [45] and work with for instance power analysis. Another important reason why this dimension exists, is since Fabijan et. al. points out that a statistical foundation is a prerequisite for experimentation, in this case in the form of A/B testing [44]. The authors mention that power analysis is of a great importance for successful experimentation. Therefore the decision is to let this part of the model to remain, and the individual user can decide if they do

not believe it is desirable to include data scientists or experimentation experts in the process.

More feedback received on the assessment model includes the one employee that points out that there should perhaps be more stages. However, given the previous argument of simplicity over complexity a good amount of stages was decided to be three. This simplicity is also preferred by the interviewee from one of the static validation companies. Regarding adding a score for the assessment result of the entire model it would indeed be a very useful thing to have, but the difficulty lies in the fact that the difficulty in transitioning between stages inside the same dimension might not be equal for each transition, which would make it hard to create a fair score that indicates how much/less or better/worse someone uses experimentation compared to someone else. Finally, in regards to moving the "way of working" part to the organizational category, it is a very good point. However, dividing the "Experimentation impact" into both impact on the business metrics and the impact of the way of working, and placing the first one under the business category and the second one under the organizational category would require another dimension. This is not desirable since adding a new dimension would increase the complexity of the model, which is why they are combined together under the business category.

The feedback from the one static validation interview includes the idea of whether or not it is desirable to completely automate the statistical part of the experimentation in the experimentation platform, and the reasoning why it would not be is because it is easy to get it wrong when you automate it. It is of course up to each individual company to decide on what is desired. However, an argument for automating the statistical part is that it is also possible to make mistakes when it comes to the statistics as a human, and once you have developed an experimentation platform (or integrated a third-party) with high quality and rigorous tests, it should be possible to be more confident in the system than in the manual work by mistake-prone humans. Furthermore, there are many beneficial features that can assist the experimentation that can be automated in an experimentation platform, for instance, the automatic detection of experiments that interact with each other as described by Fabijan et. al. [44]. Therefore the implied desired direction of that dimensions remains, but the reader can ignore the dimension or change it if they do not agree with the desired direction. If it is desirable to educate the team members in how to do experimentation themselves and move the experimentation experts to more supportive roles, is also an individual decision that the user of the model has to decide on. However, based on the "Feature team self-sufficiency" dimension and corresponding discussion in the Experimentation Growth Model by Fabijan et. al. [44], the direction is kept similar in the custom assessment model created for this thesis.

9.5.2 Example usage

Experimentation dimensions

		Stage 0	Stage 1	Stage 2
Technical	**Data**	No logging is done.	Logging of basic events is done, such as which views in the product users normally visit or what basic actions they perform (e.g. clicks on buttons).	Logging is done comprehensively with information regarding most of the user activity, such as durations in views, flows through the product or technical details of user hardware.
	Experimentation platform and statistical foundation	No experimentation platform is used, if experiments are run they are run manually.	Custom made or 3rd party experimentation platform exists. Basic features like defining variations, selecting sample size, experiment duration and assigning users to groups exist and are actively being used.	Advanced features, such as A/A testing, power analysis, alerting, automatic shutdown of harmful experiments and interaction detection exist in the platform and are actively being used.
Business	**Metrics**	No metrics are created.	Basic single metrics are created for measuring success.	Success, debug, guardrail and data quality metrics exist. Overall Evaluation Criteria are created that combines several metrics.
	Type and extent of experimentation	No experiments are run.	Experiments are run but not systematically. Experiments are used for changes such as changing existing functionality, deciding if a feature should be removed, adding new functionality or quality assurance.	Experiments are run systematically on all four types of changes in Stage 1.
	Experimentation impact	If experiments are run, the experimentation has no or insignificant impact on the business metrics and the team's way of working.	The experimentation has some impact on important business metrics and it somewhat affects the team's planning and prioritization of product changes.	The experimentation has substantial impact on important business metrics and it completely affects the way the team is working.
Organizational	**Organizational structure**	If experiments are run, no data-scientists or experimentation experts are involved in the creation and execution of experiments.	There are data-scientists and experimentation experts directly involved in the creation and execution of experiments.	Teams and team members are educated in how to conduct statistically sound experiments on their own, and there are data-scientists and/or experimentation experts available to assist when help is needed.

Figure 9.1: The case study company assessment result

To provide an example usage of the custom assessment model, the assessment result of how well the case study company uses continuous experimentation is shown in figure 9.1. The company uses Google Analytics for data-collection, which provides quite extensive data about the usage of the web service, which puts the company at stage 2 in the "Data" dimension. The company uses Google Optimize as an experimentation platform, which comes with some basic features like selecting sample-sizes and time limits, but since the company does not use any deeper statistical concepts, such as power analysis, they are placed in Stage 1 on the dimension "Experimentation platform and statistical foundation". Furthermore, the company works with single metrics and not any Overall Evaluation Criteria, which puts them on Stage 1 in the dimension "Metrics". For "Type and extent of experimentation" the company mainly does business-driven experimentation, with different types of changes, but they do not run experiments systematically, which places them at stage 1. There are different views on the impact of the experimentation between different employees of the company, but there are indications that the experimentation is not affecting the company's way of working a lot, which places them in stage 0. Finally, the company is placed in stage 0 for "Organizational structure", since there are no experimentation experts or data-scientists involved in the process.

9.5.3 Conclusion of the assessment model

After assessing the case study company with the custom assessment model based on the information gathered about the company, and based on the feedback from some employees of the company about the assessment result, it was found that different people might receive slightly different results from the model. For instance the difference in how the company should be placed in the "Experimentation impact" dimension. However, in general, the employees agreed with the assessment result, which indicates that the model, in general, provides a valid result. It is very difficult to create a model of this size that generates the same results despite who is using it every time. The feedback received on the assessment result was not disagreeing to a large extent, and therefore the assessment result can be considered overall acceptable. The employees from the case study company and the interviewee from the one static validation company did indeed think that the model in general in a good way assesses how well continuous experimentation is used in a company. Since the model was also based on previous research made by Fabijan et. al., it is possible to answer research question number two with that the custom assessment model can be used to assess a software company with a large distance to users and low control of roadmap in terms of how well they use continuous experimentation.

9.6 Perceived effects

The third research question in this thesis was defined as the following.

> *RQ3: For a software company with low control of roadmap and a large distance to users, what are the perceived advantages, disadvantages or blocking issues of using more continuous experimentation?*

To answer this research question with external validity it was previously reasoned that the identified advantages, disadvantages and blocking issues should either already be identified in research made by others, or be specific to this type of company, rather be specific to this case study company.

Perceived advantage	**Other research identifying the same advantage**	**G**
Quality assurance / Higher quality in products	"Ensuring product quality" by Fabijan et. al. [35]	✓
Products are improved / Products will perform better	"Incremental product improvements" by Fabijan et. al. [35]	✓
Identify risks that would perhaps not have been identified without experimentation	-	✓
Get new insights	"Value discovery and validation" by Fabijan et. al. [35]	✓
More data-driven arguments provided to the client / less emotional arguments / less decisions based on opinions	"Decisions supported by data" by Yaman et. al. [42]	✓
Build the right things	"Value discovery and validation" by Fabijan et. al. [35]	✓
Client can make more money	-	✓
Faster feedback on decisions made	-	✓
Happier clients	-	✓
Foundation for future decisions	-	✓
Better harmony inside the own company	-	✓

Table 9.1: Mapping between identified perceived advantages for this thesis and similar advantages identified in other research

There are many perceived advantages of using more continuous experimentation identified in the perceived effects phase of this thesis work. Table 9.1 shows the perceived advantages in this thesis (to the left), similar perceived advantages that are identified by other researchers (in the middle) and whether or not the perceived advantage is considered generalizable (G) to other companies by this thesis author (to the right). The higher quality in products advantage identified in this thesis work is also identified by Fabijan et. al. [35], where the authors show the advantage of ensuring the product quality. The same authors also identify advantages such as the ability to make incremental product improvements and being able to discover and validate what creates value, which maps to the perceived advantages in this thesis work: products are improved and perform better, the ability to get new insights and building the right things. The perceived advantage of having more data-driven arguments to the clients, less emotional arguments and less decisions based on opinions in general, could be considered to be identified by Yaman et. al. [42], who discuss the advantage of decisions supported by data. These five mentioned perceived advantages could, therefore, be considered to have high external validity. The six perceived advantages in the table that are not mapped to advantages identified in other researched could all be argued that they are not specific to the case study

company, and instead that they are specific to the type of company. The perceived advantages that the clients can make more money and become happier could be seen as specific to a company that has these types of client relationships. The ability to identify risks that would perhaps not be identified if experimentation was not used should be an advantage that can apply to any type of company. Getting quicker feedback can also be seen as applicable to any type of company since anyone can benefit from a quicker feedback-loop. Finally, getting a better harmony inside the company and building a foundation for future decisions should also not be specific to the case study company, since better the organization and how it makes decisions should be applicable to any type of company. Therefore all of the perceived advantages could be considered to have high or at least some external validity.

Perceived disadvantage	**Other research identifying the same disadvantage**	**G**
More complex code / Code will become complex	-	X
More complex devops and rollouts situation	-	✓
More difficult situation for new developer	-	✓
Might require a lot of resources / Takes a lot of time	Limited resources and time by Lindgren and Münch [30]	✓
Clients might not see the benefits of doing experimentation / Have to convince clients	-	✓
Experimentation will not make significant improvements (e.g. not increase conversion)	Too low return on financial and time investment by Schermann et. al. [50]	✓
More of a process	-	X

Table 9.2: Mapping between identified perceived disadvantages for this thesis and similar disadvantages identified in other research

Table 9.2 shows the perceived disadvantages identified in this thesis (to the left), the matching disadvantages identified by other authors (in the middle) and whether or not this thesis author believes that the disadvantage could be generalizable (G) to others (to the right). The perceived disadvantage that using more continuous experimentation will take a lot of time and resources, is also shown by Lindgren and Münch who identify that a challenge is that the resources and time are limited [30]. The disadvantage that experimentation will not make a significant impact is also indicated by Schermann et. al. [50], who talk about how experimentation might return too little on financial and time investment is a problem. Both these two disadvantages could, therefore, be considered general to any company. The disadvantage that the devops and rollouts situation might become more complex could also be seen as general, since adding or evolving an experimentation layer

in the development process should probably increase the complexity of this situation despite how it looks like now. It will probably not decrease the complexity in other parts of the devops and rollout situation, and therefore it can be seen as generalizable to any type of company. The same reason applies for the ability for a new developer to understand the code, which should decrease for any company if there is another development process layer added on top of the existing way of working. However, the perceived disadvantages that experimentation might make the code more complex, as well as the perceived disadvantage that it might give more of a process to the company, appear to be specific to the case study company. The disadvantage of more complex code is probably specific to this company since it depends on how much more complex the code will be because of adding or evolving an experimentation layer, as well as how much it will decrease because of how many features that are not used by the users can be removed as a result of experimentation, which is believed to be the case by for instance Fabijan et. al. [35]. For the process it depends on how the process in the company looks like currently since it might reduce some process parts that are no longer needed, for instance, a need to estimate value added by different features. These two disadvantages could be very dependent on how the company looks like, and therefore not necessarily considered generalizable. Finally, the disadvantage related to the client not seeing the benefits and the need to convince the client should be generalizable since they can be considered specific to this type of company, i.e. a company in this type of client relationship, rather than the actual case study company. Therefore all except two disadvantages could be considered to have high or at least some external validity.

Perceived blocking issue	**Other research identifying the same blocking issue**	**G**
Knowledge barrier / Requires a lot of education in experimentation and statistics	Lack of competence and need of education by Rissanen and Münch [27] and "lack of expertise" by Schermann et. al. [50]	✓
Limited resources, especially time / needs to be prioritized when it comes to time	Limited resources and time by Lindgren and Münch [30]	✓
They need to convince client / Client needs convincing	-	✓
They have to change their own organisation / Organization has to be changed	-	✓
Code has to be changed	"Software architecture" by Schermann et. al. [50] and need to evolve technology by Fabijan et. al. [35]	✓

Table 9.3: Mapping between identified perceived blocking issues for this thesis and similar blocking issues identified in other research

Table 9.3 shows the perceived blocking issues of using more continuous experimentation identified in this thesis (to the left), similar blocking issues identified by other researchers (in the middle) and whether or not this thesis author considers the perceived blocking issue to be generalizable (G) to others (to the right). When it comes to the perceived educational/knowledge barrier a similar issue was identified by both Rissanen and Münch [27] as well as Schermann et. al. [50]. For limited resources and time, a similar blocking issue was identified by Lindgren and Münch [30]. Regarding having to change the code, Schermann et. al. [50] as well as Fabijan et. al. [35] identifies how the software architecture is not created to support experimentation and the need to evolve the technology to facilitate experimentation. These three blocking issues can, therefore, be considered general, and not specific to this case study company. For the need to change their own organization it is likely that this is the case for any type of company, and not specific to the case study company. Despite what the organization looks like, if it wanted to move towards experimentation or use it more it probably requires some organizational changes. Finally, the need to convince the client can be considered related to this type of a company, i.e. a company with this type of client relationships, and also be considered generalizable. Therefore all of the blocking issues could be considered to have high or at least some external validity.

The third research question asked what the perceived advantages, disadvantages and blocking issues are for a company with a large distance to users and low control of roadmap of using more continuous experimentation. This has been answered, with some external generalisability, by the result and discussion regarding the perceived effects in this thesis work. The perceived effects can be seen general to a company with low control of roadmap and large distance to users, but it is likely that a lot of them also are relevant for any company despite what its control of roadmap or distance to users looks like.

9.7 Control of roadmap and distance to users as barriers

In this thesis, it was shown that there is some relationship between control of roadmap and continuous experimentation, as well as some relationship between distance to users and continuous experimentation. Although this thesis did not consider the magnitude or direction of these relationships, it is possible to believe from the data collected that control of roadmap and distance to users acts as barriers to using more continuous experimentation and/or improving the existing experimentation on different technical, business or organizational dimensions. There are also several people who believe that the control of roadmap and distance to users are not fixed and that it might be possible to change them.

Barriers to evolving on experimentation dimensions

	Stage 0	Stage 1	Stage 2
Ability to access users	Users can not be accessed at all. There is no possibility of either qualitative or quantitative data-collection.	Users can be accessed, but there are difficulties with accessing them. Especially when new types of quantitative or qualitative data that is not currently being collected should be collected.	Users can be accessed with ease. To make the decision to collect new qualitative or quantitative user data can be done with close to no delay. Decisions related to any changes regarding data-collection with a very high impact might still be dependent on external parties.
Ability to control roadmap	The roadmap can not be controlled by the team at all. There is no possibility for the team itself to put desired product changes on the roadmap.	Desired product changes can be added to the roadmap by the team, but there are difficulties. The team does not have the full authority to decide on specific changes themselves.	Desired product changes can be added to the roadmap easily by the team itself, and the team is allowed to make the changes they believe are beneficial for the development of the product. Decisions related to changes that might have a very high impact might still be dependent on external parties.

Experimentation dimensions

		Stage 0	Stage 1	Stage 2
Technical	**Data**	No logging is done.	Logging of basic events is done, such as which views in the product users normally visit or what basic actions they perform (e.g. clicks on buttons).	Logging is done comprehensively with information regarding most of the user activity, such as durations in views, flows through the product or technical details of user hardware.
	Experimentation platform and statistical foundation	No experimentation platform is used, if experiments are run they are run manually.	Custom made or 3rd party experimentation platform exists. Basic features like defining variations, selecting sample size, experiment duration and assigning users to groups exist and are actively being used.	Advanced features, such as A/A testing, power analysis, alerting, automatic shutdown of harmful experiments and interaction detection exist in the platform and are actively being used.
Business	**Metrics**	No metrics are created.	Basic single metrics are created for measuring success.	Success, debug, guardrail and data quality metrics exist. Overall Evaluation Criteria are created that combines several metrics.
	Type and extent of experimentation	No experiments are run.	Experiments are run but not systematically. Experiments are used for changes such as changing existing functionality, deciding if a feature should be removed, adding new functionality or quality assurance.	Experiments are run systematically on all four types of changes in Stage 1.
	Experimentation impact	If experiments are run, the experimentation has no or insignificant impact on the business metrics and the team's way of working.	The experimentation has some impact on important business metrics and it somewhat affects the team's planning and prioritization of product changes.	The experimentation has substantial impact on important business metrics and it completely affects the way the team is working.
Organizational	**Organizational structure**	If experiments are run, no data-scientists or experimentation experts are involved in the creation and execution of experiments.	There are data-scientists and experimentation experts directly involved in the creation and execution of experiments.	Teams and team members are educated in how to conduct statistically sound experiments on their own, and there are data-scientists and/or experimentation experts available to assist when help is needed.

Figure 9.2: Custom assessment model with barriers

To visualize how control of roadmap and distance to users acts as barriers to evolving the use of experimentation, a new combined model has been constructed. Figure 9.2 shows this combined model, which has the barriers control of roadmap and distance to users placed over the custom assessment model. It is possible to consider control of roadmap and distance to users as barriers to evolving on the rest of the dimensions in the custom assessment model below. The control of roadmap and distance to users, can together with statistical foundations and psychological safety identified by Fabijan et. al. [44], be considered prerequisites that need to be sorted out before using experimentation. However, even with a little control of roadmap and some access to users, it might be possible to evolve on the different experimentation dimensions in the model. But if a company wants to reach stage 2 in each dimension, it might be necessary that control of roadmap and distance to users are evolved as well. Since several people believe that control of roadmap and distance to users can be changed, it should be possible to evolve these two dimensions when

they become problems. While the statistical foundations and psychological safety are prerequisites in the sense that they have to be sorted before beginning to experiment, the control of roadmap and distance to users would by this author be considered barriers that will cause problems but that can be overcome either before or during the evolvement process of experimentation.

A final note on this topic is that it is not definitely answered in this thesis work and that there still is a lot of work that has to be done to prove these ideas with certainty. The idea of control of roadmap and distance to users as barriers is merely a reflection based on the data collected for this thesis. However, it could be useful to consider this combined model with control of roadmap and distance to users as barriers. For a company that wants to improve its experimentation practices, it is possible to use the model to assess where its organization is now in the model, both in terms of control of roadmap, distance to users and experimentation. With that information in mind, it is then possible to see what the next steps are to improve the use of experimentation. Based on the perceived effects of using more experimentation for the case study company uncovered in this thesis work, it is also possible to work out what the advantages, disadvantages or blocking issues would be for the organization under study. Finally, if the organization runs into problems with control of roadmap and/or distance to users during this process of improving the use of experimentation, it could perhaps be possible to move forward on the control of roadmap and/or distance to users dimensions and to remove the barriers hindering evolution on other dimensions related to experimentation.

9.8 Future work

This thesis has been an exploratory case study in a rather unknown part of the continuous experimentation field, which is differentiating companies based on control of roadmap and distance to users, and it has specifically focused on a company with low control of roadmap and a large distance to users. The thesis has explored the relationship between control of roadmap, distance to users and continuous experimentation. It has also explored how a company with low control of roadmap and large distance to users can be assessed in how well they use continuous experimentation, and what perceived effects they identify of using more continuous experimentation. The result of the thesis has been deterministic classification models for the concepts control of roadmap and distance to users, the learning that there is indeed some relationship between the concepts, an assessment model and perceived effects of using more continuous experimentation. However, there is still a lot of work that can be done on the same subject.

One interesting aspect that could require attention is whether or not the control of roadmap and distance to users for a company are fixed. In this thesis work, there have appeared reflections from interview participants on the fact that both control of roadmap and distance to users might not be fixed. It would be very interesting to learn more about if they are fixed or changeable, and how to change them, in order to facilitate the use of continuous experimentation in companies that at the

moment have difficulties related to control of roadmap or distance to users.

Another interesting aspect of future research would be to deeper understand the idea of control of roadmap and distance to users acting as barriers in evolving the use of experimentation, as reflected on in a previous section. This idea is not entirely verified by this thesis and could use some more focus. However, if this barrier idea is valid, there could perhaps be many learnings in understanding it better and knowing how to overcome it.

Finally, other useful future work would be to make a multi-case study where similar research questions to the ones used in this thesis are used but on a larger number of companies. It would be interesting to learn more about how the control of roadmap and distance to users vary between many companies in the business, and to on a more detailed level define the relationship to continuous experimentation and the implications on developing the continuous experimentation when control of roadmap is low or the distance to users is large. With more research in this topic, the field of continuous experimentation could perhaps be applied to, and more relevant to, a larger number of companies, that perhaps today struggle with implementing experimentation in their company because of for instance lack of user data or ability to change the roadmap or backlog.

10
Conclusion

Continuous experimentation has recently been a popular topic in the research of Software Engineering, but for many companies, there might be challenges in implementing continuous experimentation because of how the organization looks like. For instance, a company such as a software consultancy company, who does not own its own products and instead works for external companies that are the product owners. In this thesis, two concepts have been especially studied, which are the ability to control the roadmap and the ability to access the users, called the control of roadmap and distance to users. This thesis has been conducted as an exploratory case study together with a software consultancy company that fits into the description low control of roadmap and large distance to users. The research questions that have been considered are if there is a relationship between control of roadmap, distance to users and continuous experimentation, as well as how a company with specifically low control of roadmap and large distance to users can be assessed in how well they use continuous experimentation, and what perceived effects they identify of using more continuous experimentation. In this thesis the case study company of the thesis has been studied and explained in detail, for instance, in how the organization looks like and how they work with experimentation. There have been four companies described with differences in control of roadmap and distance to users, that has provided an outsider perspective on mainly the relationship between the three concepts.

The contributions of this exploratory case study are, firstly, two deterministic models that have been created to be able to assess how much control of roadmap a company has as well as how large or small its distance to users is. Secondly, that there is indeed a relationship between control of roadmap and continuous experimentation, as well as the distance to users and continuous experimentation, and that it is relevant to differentiate a company with low control of roadmap and large distance to users for the other research questions. Thirdly, a custom assessment model has been presented that allows the assessment of how well a company uses continuous experimentation, which has been applied on the case study company. Finally, the perceived effects of using more continuous experimentation have been elicited and presented.

The thesis provides an answer that there is actually a relationship between the three concepts control of roadmap, distance to users and continuous experimentation. It answers how it is possible to assess a company with low control of roadmap and a large distance to users in how well they use continuous experimentation. Finally, it has also provided an answer to what the perceived effects are of such a company to

use more continuous experimentation. The result of this exploratory case study has also resulted in recommendations for future work in this topic, which for instance includes the recommendation to research if control of roadmap and distance to users can be changed, which perhaps would enable more companies to use continuous experimentation in their development processes.

The overall goals of this thesis were to explore the research gap in this subject, and to provide initial results that can be further developed in the future. Based on the little information that exists in research about especially control of roadmap, but also the distance to users, and how they relate to experimentation, it is possible to state that there is a clear research gap in this area. Since it is shown in this thesis that these three concepts relate to each other, and because there are companies such as software consultancies that might be affected by this relationship, it indicates that there is value in filling the research gap. The custom assessment model created for this thesis, as well as the identified perceived effects of using more experimentation, are two useful starting points in order to provide initial results for the research gap. However, there are a lot of areas of this research that are yet to be discovered, and several of these areas of future work have been suggested in this thesis.

It is this thesis author's hope that the exploratory work that has been done in this thesis is proven useful for people who might struggle with for instance control of roadmap and distance to users but wanting to evolve their use of continuous experimentation.

Bibliography

[1] R. B. Bausell and Y.-F. Li, *Power Analysis for Experimental Research: A Practical Guide for the Biological, Medical and Social Sciences*, 1st ed. Cambridge, UK: Cambridge University Press, 2002.

[2] S. N. Bhatti, "Why quality?", *ACM SIGSOFT Software Engineering Notes*, vol. 30, no. 2, p. 1, Mar. 2005.

[3] M. Coram and S. Bohner, "The Impact of Agile Methods on Software Project Management", *12th IEEE International Conference and Workshops on the Engineering of Computer-Based Systems (ECBS'05)*, pp. 363–370, 2005.

[4] R. Kohavi, R. M. Henne, and D. Sommerfield, "Practical guide to controlled experiments on the web", in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07*, New York, New York, USA: ACM Press, 2007, p. 959.

[5] M. Moran, *Do It Wrong Quickly: How the Web Changes the Old Marketing Rules*. Indianapolis, Indiana, the United States: IBM Press, 2008.

[6] L. Chung and J. C. S. do Prado Leite, "On Non-Functional Requirements in Software Engineering", in *Conceptual Modeling: Foundations and Applications*, 2009, pp. 363–379.

[7] T. Crook, B. Frasca, R. Kohavi, and R. Longbotham, "Seven pitfalls to avoid when running controlled experiments on the web", *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, p. 1105, 2009.

[8] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne, "Controlled experiments on the web: Survey and practical guide", *Data Mining and Knowledge Discovery*, vol. 18, no. 1, pp. 140–181, 2009.

[9] P. Runeson and M. Höst, "Guidelines for conducting and reporting case study research in software engineering", *Empirical Software Engineering*, vol. 14, no. 2, pp. 131–164, 2009.

[10] D. Tang, A. Agarwal, D. O'Brien, and M. Meyer, "Overlapping Experiment Infrastructure: More, Better, Faster Experimentation", *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '10*, p. 17, 2010.

[11] Eric Ries, *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses*. 2011.

[12] J. Bosch, "Building Products as Innovation Experiment Systems", in *Software Business, Third International Conference , ICSOB*, 2012, pp. 27–39.

[13] W. M. Farid, “The NORMAP Methodology: Lightweight Engineering of Non-functional Requirements for Agile Processes”, in *2012 19th Asia-Pacific Software Engineering Conference*, IEEE, Dec. 2012, pp. 322–325.

[14] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu, “Trustworthy online controlled experiments”, *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’12*, p. 786, 2012.

[15] D. McKinley, *Design for Continuous Experimentation*, 2012. [Online]. Available: `https://www.youtube.com/watch?v=qCKj%7B%5C_%7DK5RNfY`.

[16] H. H. Olsson, H. Alahyari, and J. Bosch, “Climbing the "Stairway to heaven" - A mulitiple-case study exploring barriers in the transition from agile development towards continuous deployment of software”, *Proceedings - 38th EUROMICRO Conference on Software Engineering and Advanced Applications, SEAA 2012*, pp. 392–399, 2012.

[17] M. Paasivaara, V. T. Heikkilä, and C. Lassenius, “Experiences in scaling the Product Owner role in large-scale globally distributed Scrum”, *Proceedings - 2012 IEEE 7th International Conference on Global Software Engineering, ICGSE 2012*, pp. 174–178, 2012.

[18] V. Braun and V. Clarke, “Using thematic analysis in psychology”, *Journal of Chemical Information and Modeling*, vol. 53, no. 9, pp. 1689–1699, 2013.

[19] D. G. Feitelson, E. Frachtenberg, and K. L. Beck, “Development and deployment at facebook”, *IEEE Internet Computing*, vol. 17, no. 4, 2013.

[20] R. Kohavi, A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann, “Online controlled experiments at large scale”, in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’13*, New York, New York, USA: ACM Press, 2013, p. 1168.

[21] F. Fagerholm, H. Mäenpää, and J. Münch, “Building Blocks for Continuous Experimentation Categories and Subject Descriptors”, *Proceedings of the 1st International Workshop on Rapid Continuous Software Engineering (RCoSE 2014)*, pp. 26–35, 2014.

[22] R. Kohavi, A. Deng, R. Longbotham, and Y. Xu, “Seven rules of thumb for web site experimenters”, *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’14*, pp. 1857–1866, 2014.

[23] H. H. Olsson and J. Bosch, “From Opinions to Data-Driven Software R&D: A Multi-case Study on How to Close the ’Open Loop’ Problem”, in *2014 40th EUROMICRO Conference on Software Engineering and Advanced Applications*, IEEE, Aug. 2014, pp. 9–16.

[24] H. H. Olsson and J. Bosch, “Climbing the “Stairway to Heaven”: Evolving From Agile Development to Continuous Deployment of Software”, in *Continuous Software Engineering*, Cham: Springer International Publishing, 2014, pp. 15–27.

[25] ——, “The HYPEX Model: From Opinions to Data-Driven Software Development”, in *Continuous Software Engineering*, Springer International Publishing, 2014, pp. 155–164.

[26] ——, "Towards Continuous Customer Validation: A Conceptual Model for Combining Qualitative Customer Feedback with Quantitative Customer Observation", in *Lecture Notes in Business Information Processing*, vol. 210, 2015, pp. 154–166.

[27] O. Rissanen and J. Munch, "Continuous Experimentation in the B2B Domain: A Case Study", *Proceedings - 2nd International Workshop on Rapid Continuous Software Engineering, RCoSE 2015*, pp. 12–18, 2015.

[28] A. Deng and X. Shi, "Data-Driven Metric Development for Online Controlled Experiments", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pp. 77–86, 2016.

[29] P. Dmitriev, B. Frasca, S. Gupta, R. Kohavi, and G. Vaz, "Pitfalls of long-term online controlled experiments", *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*, pp. 1367–1376, 2016.

[30] E. Lindgren and J. Münch, "Raising the odds of success: the current state of experimentation in product development", *Information and Software Technology*, vol. 77, pp. 80–91, 2016.

[31] S. G. Yaman, F. Fagerholm, M. Munezero, J. Münch, M. Aaltola, C. Palmu, and T. Männistö, "Transitioning Towards Continuous Experimentation in a Large Software Product and Service Development Organisation – A Case Study", in *Product-Focused Software Process Improvement*, 2016, pp. 344–359.

[32] J. Bosch, *Using Data to Build Better Products: A Hands-On Guide to Working with Data in R&D - The Basics*, 1st ed. CreateSpace Independent Publishing Platform, 2017.

[33] J. Bosch and H. H. Olsson, "Toward Evidence-Based Organizations: Lessons from Embedded Systems, Online Games, and the Internet of Things", *IEEE Software*, vol. 34, no. 5, pp. 60–66, 2017.

[34] P. Dmitriev, S. Gupta, D. W. Kim, and G. Vaz, "A Dirty Dozen : Twelve Common Metric Interpretation Pitfalls in Online Controlled Experiments", *KDD '17 Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1427–1436, 2017.

[35] A. Fabijan, P. Dmitriev, H. H. Olsson, and J. Bosch, "The benefits of controlled experimentation at scale", *Proceedings - 43rd Euromicro Conference on Software Engineering and Advanced Applications, SEAA 2017*, pp. 18–26, 2017.

[36] ——, "The Evolution of Continuous Experimentation in Software Product Development: From Data to a Data-Driven Organization at Scale", *Proceedings - 2017 IEEE/ACM 39th International Conference on Software Engineering, ICSE 2017*, pp. 770–780, 2017.

[37] F. Fagerholm, A. Sanchez Guinea, H. Mäenpää, and J. Münch, "The RIGHT model for Continuous Experimentation", *Journal of Systems and Software*, vol. 123, pp. 292–305, 2017.

[38] B. Fitzgerald and K. J. Stol, "Continuous software engineering: A roadmap and agenda", *Journal of Systems and Software*, vol. 123, pp. 176–189, 2017.

[39] D. Issa Mattos, J. Bosch, and H. H. Olsson, "Your system gets better every day you use it: Towards automated continuous experimentation", *Proceedings*

- 43rd Euromicro Conference on Software Engineering and Advanced Applications, SEAA 2017, pp. 256–265, 2017.

[40] R. Kohavi and R. Longbotham, "Online Controlled Experiments and A/B Testing", in *Encyclopedia of Machine Learning and Data Mining*, 7/8, vol. 32, Boston, MA: Springer US, 2017, pp. 922–929.

[41] R. R. Maiti and F. J. Mitropoulos, "Capturing, eliciting, and prioritizing (CEP) NFRs in agile software engineering", in *SoutheastCon 2017*, IEEE, Mar. 2017, pp. 1–7.

[42] S. G. Yaman, M. Aaltola, M. Munezero, C. Palmu, F. Fagerholm, J. Münch, O. Syd, and T. Männistö, "Introducing continuous experimentation in large software-intensive product and service organisations", *Journal of Systems and Software*, vol. 133, pp. 195–211, 2017.

[43] F. Auer and M. Felderer, "Current state of research on continuous experimentation: A systematic mapping study", *Proceedings - 44th Euromicro Conference on Software Engineering and Advanced Applications, SEAA 2018*, no. August, pp. 335–344, 2018.

[44] A. Fabijan, P. Dmitriev, C. McFarland, L. Vermeer, H. Holmström Olsson, and J. Bosch, "Experimentation growth: Evolving trustworthy A/B testing capabilities in online software companies", *Journal of Software: Evolution and Process*, vol. 30, no. 12, e2113, Dec. 2018.

[45] S. Gupta, L. Ulanova, S. Bhardwaj, P. Dmitriev, P. Raff, and A. Fabijan, "The Anatomy of a Large-Scale Experimentation Platform", *Proceedings - 2018 IEEE 15th International Conference on Software Architecture, ICSA 2018*, pp. 1–10, 2018.

[46] D. Issa Mattos, P. Dmitriev, A. Fabijan, J. Bosch, and H. Holmström Olsson, "An Activity and Metric Model for Online Controlled Experiments", in *Product-Focused Software Process Improvement*, 2018, pp. 182–198.

[47] R. Ros and E. Bjarnason, "Continuous experimentation scenarios: A case study in e-commerce", *Proceedings - 44th Euromicro Conference on Software Engineering and Advanced Applications, SEAA 2018*, no. 2, pp. 353–356, 2018.

[48] R. Ros and P. Runeson, "Continuous experimentation and A/B testing", *Proceedings of the 4th International Workshop on Rapid Continuous Software Engineering - RCoSE '18*, pp. 35–41, 2018.

[49] G. Schermann, J. Cito, and P. Leitner, "Continuous experimentation: Challenges, implementation techniques, and current research", *IEEE Software*, vol. 35, no. 2, pp. 26–31, 2018.

[50] G. Schermann, J. Cito, P. Leitner, U. Zdun, and H. C. Gall, "We're doing it live: A multi-method empirical study on continuous experimentation", *Information and Software Technology*, vol. 99, pp. 41–57, 2018.

[51] Stack-Overflow, *Stack Overflow Developer Survey Results 2018*, 2018. [Online]. Available: `https://insights.stackoverflow.com/survey/2018`.

[52] Google, *Google Analytics*. [Online]. Available: `https://marketingplatform.google.com/about/analytics/`.

[53] ——, *Google Optimize*. [Online]. Available: `https://marketingplatform.google.com/about/optimize/`.

A

Interview templates pt. 1

This appendix shows the three interview templates used for the five initial interviews that were conducted as part of phase 2 of this thesis work, i.e. the phase about learning how the case study company works. The actual interviews were in Swedish and interview templates translated to Swedish was used. Only the main questions are shown here. There were also several formalities covered in the real interview templates.

Main Questions

- [How organization is working]
    - Do you work with agile? Continuous integration or continuous deployment?
    - What types of roles exist in the organization? Product manager? Business role? Devops? Data scientists?
    - How is prioritization and roadmapping made? Is there hierarchy, i.e. are some people's opinions more important than other's when making decisions?
    - Is active/qualitative user feedback collected?
    - Is passive/quantitative user feedback collected?
    - How is user feedback (data) used to make decisions?
    - Are assumptions that exist in the company normally tested and proven?
    - Is data collection an important aspect in the organization, i.e. using data-driven development?
    - Are any other methods or processes used, eg. Build/Measure/Learn (Lean Startup).

- [What experimentation methods are used]
    - Is experimentation used in the organization, e.g. A/B testing, Online Controlled Experiments, split tests, randomized experiments, control/treatment tests, and online field experiments
    - If A/B is used - is multivariate used as well? (A/B/C testing)

  ALL QUESTION BELOW ARE RELEVANT IF EXPERIMENTATION IS USED IN COMPANY

- [What type of experiments are made]
    - Business-driven vs Regression-driven
      (business = tests functional requirements in a product, to evaluate new functionality from a business perspective)
      (regression = tests non-functional requirements, so that the change does not cause a regression noticeable to the user)
    - What type of changes are experimented with?
    - Are new features experimented with before they are fully implemented, e.g. by creating a small version to see if it is viable?
    - Are features ever removed because they are experimented with and deemed not worth keeping?

- [How is experimentation conducted]
    - *How* is data collected? (If not answered before)
    - How are tests initiated? Are there hypotheses presented beforehand and clear goals with experiment?
    - How are the metrics defined?

    - How is the experiment planned? Are things like sample size, “tracking of right users”, randomization unit, statistical power etc discussed?
    - Are both short-terms and long-terms effects of experiments considered?
    - How is an experiment analysed afterwards?
    - What is the experimentation result used for? (e.g. decision-making? Prioritization?)

- [How is experimentation implemented in code]
    - How is experimentation made in the actual code? E.g. feature toggles and runtime traffic routing
      (feature toggles = basically if-statements saying that if it is this type of user show this)
      (traffic routing = multiple instances of servers and a router that redirects users to different servers/instances)
    - How are experiments deployed and executed in production? Is there some sort of experimentation system?
    - Is some sort of experimentation code language used?

Main questions

- [Canary releases]
  = Releasing new version/feature to a subset of users only (e.g. 5% based on geographical region, or if they perhaps are willing to test new features), while the rest use the normal version, to see that nothing goes wrong.

    - We talked about it a little bit before, but is Canary Releases used in the organization and how is it used?

- [Gradual rollouts]
  = Often combined with canary releases/dark launches. The number of users assigned to the new version is gradually increased (e.g. increase with 5% steps) until the previous version is replaced, or a threshold is reached.

    - Do you perform gradual rollouts and how is it done?

- [Dark launches]
  = New version is launched in production without being enabled/visible to users (parallell to the old one). “Silent” requests are sent to the dark version to see how it behaves in a real environment, to test scalability and performance, but it does not affect users.

    - Do you perform dark launches and how is it used?

Main questions

- How is experimentation used, e.g. a/b testing, canary releases, gradual rollouts and dark launches in the organization?
    - What is the purpose of the experimentation?
    - How is it done practically, i.e. how is it implemented?

Definitions

- [Canary releases]
  = Releasing new version/feature to a subset of users only (e.g. 5% based on geographical region, or if they perhaps are willing to test new features), while the rest use the normal version, to see that nothing goes wrong.

- [Gradual rollouts]
  = Often combined with canary releases/dark launches. The number of users assigned to the new version is gradually increased (e.g. increase with 5% steps) until the previous version is replaced, or a threshold is reached.

- [Dark launches]
  = New version is launched in production without being enabled/visible to users (parallell to the old one). “Silent” requests are sent to the dark version to see how it behaves in a real environment, to test scalability and performance, but it does not affect users.

- [A/B testing / Online Controlled Experiments]
  Simplest definition: Control group, Treatment group, measure effect and see if control and treatment differ on some criteria
  Bigger than the experiment execution: experiment methodology, the implementation in code, etc.

B

Survey

This appendix shows the survey that was used as part of the perceived effects phase of this thesis work. The actual survey was a Swedish translation and hosted with Google Forms.

Continuous Experimentation Survey

Continuous Experimentation is a general term used for experimentation in the Software Engineering process. It includes experimentation techniques such as AB testing, Canary Releases, Gradual Rollouts and Dark Launches. The main purpose of such experimentation is to use more empirical data and evidence to make informed decisions, but also to assure the quality of the software.

AB testing is the main technique to use for making informed decisions, and the other three are used mainly to assure the quality of the software.

For the purpose of simplicity, Continuous Experimentation will be referred to as “experimentation”.

1. Which advantages do you perceive with [COMPANY NAME] using more experimentation? Select all that apply.

☐ Less decisions based on opinions
☐ Products will perform better (e.g. higher conversion)
☐ Faster feedback on decisions made
☐ Happier clients
☐ Higher quality in products
☐ Other _____

2. Which disadvantages do you perceive with [COMPANY NAME] using more experimentation? Select all that apply.

☐ Takes a lot of time
☐ Code will become complex
☐ Experimentation will not make significant improvements (e.g. not increase conversion)
☐ Other _____

3. Which blocking issues do you perceive with [COMPANY NAME] using more experimentation, i.e. things preventing the evolution of experimentation? Select all that apply.

☐ Client needs convincing
☐ Code has to be changed
☐ Requires a lot of education in experimentation and statistics
☐ Organization has to be changed
☐ Other _____

Control of roadmap is a classification that can be applied on a company. It indicates how much control the company has to change what features and product changes are planned and how they are prioritized. The classification can be described with the following 3 stages.

	Stage 0	Stage 1	Stage 2
Ability to control roadmap	The roadmap can not be controlled by the team at all. There is no possibility for the team itself to put desired product changes on the roadmap.	Desired product changes can be added to the roadmap by the team, but there are difficulties. The team does not have the full authority to decide on specific changes themselves.	Desired product changes can be added to the roadmap easily by the team itself, and the team is allowed to make the changes they believe are beneficial for the development of the product. Decisions related to changes that might have a very high impact might still be dependent on external parties.

4. By the above definition of control of roadmap, in which stage would you classify [COMPANY NAME]?

○ Stage 0 ○ Stage 1 ○ Stage 2

5. How much do you believe that control of roadmap affects the use of Continuous Experimentation?

Not at all ○ 0 ○ 1 ○ 2 ○ 3 ○ 4 A lot

Distance to users is another classification that can be applied on a company. It indicates how difficult it is to access the users of a product and collect feedback from them. This feedback can be both qualitative, through for instance focus groups or surveys, or quantitative through for instance automatic collection of user behaviour in the product. The classification can be described with the following 3 stages.

	Stage 0	Stage 1	Stage 2
Ability to access users	Users can not be accessed at all. There is no possibility of either qualitative or quantitative data-collection.	Users can be accessed, but there are difficulties with accessing them. Especially when new types of quantitative or qualitative data that is not currently being collected should be collected.	Users can be accessed with ease. To make the decision to collect new qualitative or quantitative user data can be done with close to no delay. Decisions related to any changes regarding data-collection with a very high impact might still be dependent on external parties.

6. By the above definition of distance to users, in which stage would you classify [COMPANY NAME]?

○ Stage 0 ○ Stage 1 ○ Stage 2

7. How much do you believe that distance to users affects the use of Continuous Experimentation?

Not at all ○ 0 ○ 1 ○ 2 ○ 3 ○ 4 A lot

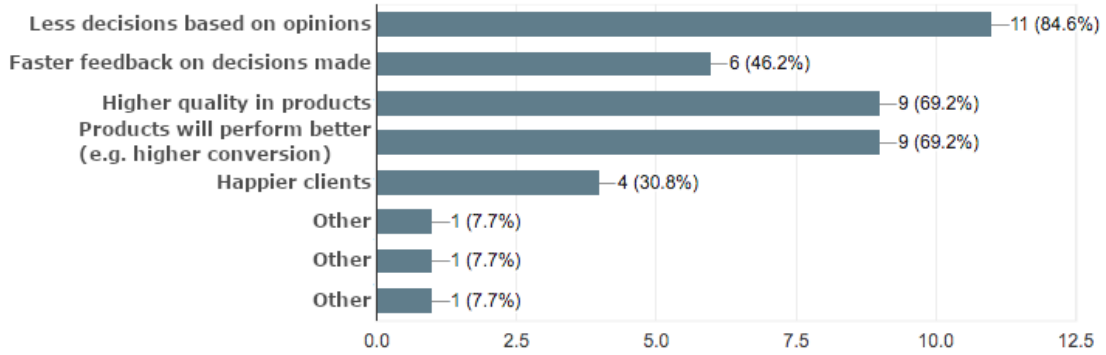
C

Survey responses

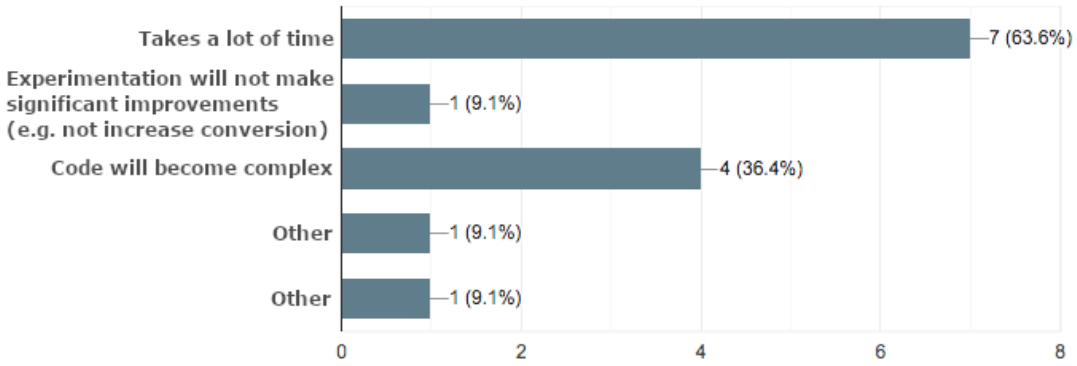
This appendix shows the survey responses from the survey conducted as part of the perceived effects phase of this thesis work. The "Other" answers have not been included in this response summary, and the ones that were useful for the thesis are described in the thesis text.

Total number of participants: **13**

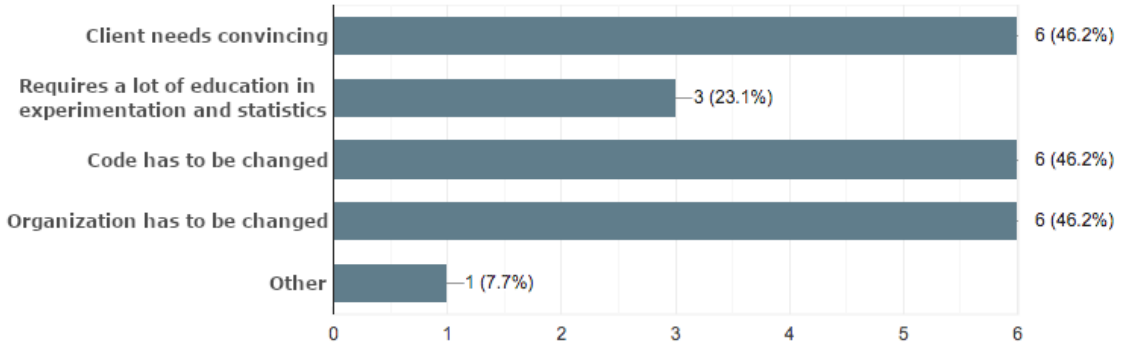
Q1: Which advantages do you perceive with [COMPANY NAME] using more experimentation? Select all that apply. (Responses: 13)



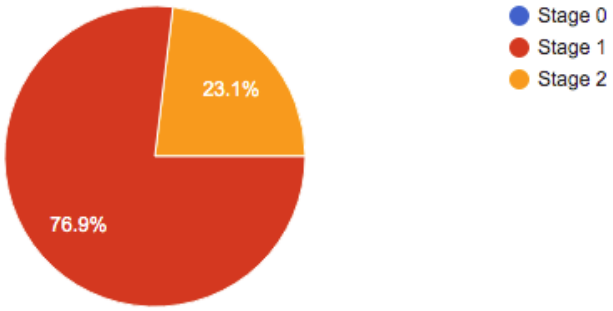
> 2. Which disadvantages do you perceive with [COMPANY NAME] using more experimentation? Select all that apply. (Responses: 11)



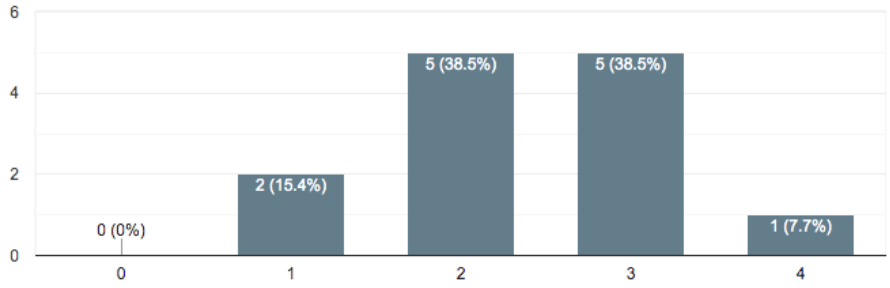
> 3. Which blocking issues do you perceive with [COMPANY NAME] using more experimentation, i.e. things preventing the evolution of experimentation? Select all that apply. (Responses: 13)



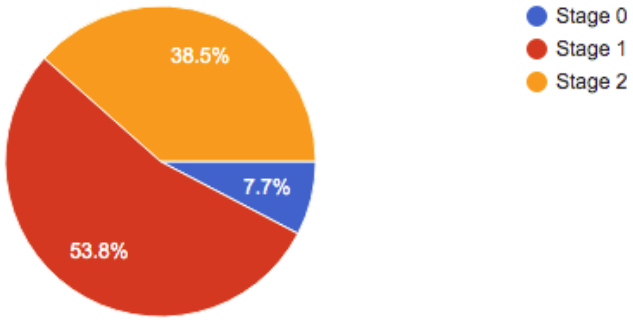
> 4. By the above definition of control of roadmap, in which stage would you classify [COMPANY NAME]? (Responses: 13)



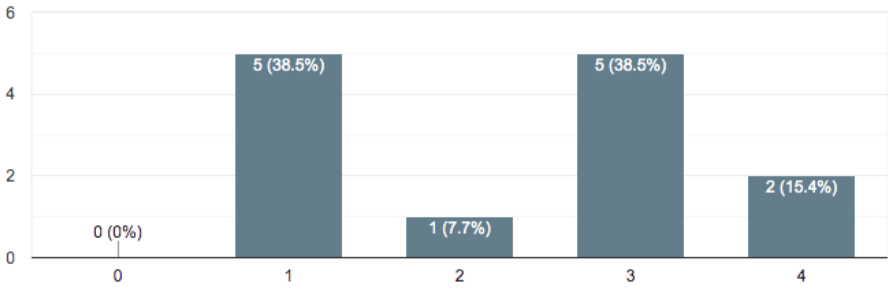
> 5. How much do you believe that control of roadmap affects the use of Continuous Experimentation? (Responses: 13)



> 6. By the above definition of distance to users, in which stage would you classify [COMPANY NAME]? (Responses: 13)



> 7. How much do you believe that distance to users affects the use of Continuous Experimentation (Responses: 13)



D

Interview templates pt. 2

This appendix shows the interview template used for the four interviews that were conducted as part of phase 6 of this thesis work, i.e. the phase about perceived effects. The actual interviews were in Swedish and an interview template translated to Swedish was used. Only the main questions are shown here. There were also several formalities covered in the real interview template.

Main Questions

- What perceived advantages for [COMPANY NAME] do you recognize if it was to evolve its use of experimentation and use more experimentation?
- What perceived disadvantages for [COMPANY NAME] do you recognize if it was to evolve its use of experimentation and use more experimentation?
- What perceived blocking issues for [COMPANY NAME] do you recognize if it was to evolve its use of experimentation and use more experimentation, i.e. things that are hindering the evolution of the experimentation in the company?

- Do you think it is desirable for [COMPANY NAME] to evolve its use of experimentation and use more experimentation? (if you account for the benefits, disadvantages and blocking issues you recognized in previous questions)

[Introduce COR]

- How would you classify [COMPANY NAME] in terms of control of roadmap? (Stages 0-2)
- How much do you believe that [no/low/high, WHAT THEY ANSWERED] control of roadmap affects the use of experimentation in [COMPANY NAME]?

[If answer is **low**]:

- Let's pretend that the control of roadmap for [COMPANY NAME] was instead high. The company now has no difficulties putting something on the roadmap. Do you believe that the use of experimentation in the company would be different? And how?

[If answer is **high**]:

- Let's pretend that the control of roadmap for [COMPANY NAME] was instead low. The company now has difficulties putting something on the roadmap. Do you believe that the use of experimentation in the company would be different? And how?

[Introduce DTU]

- How would you classify [COMPANY NAME] in terms of distance to users? (Stages 0-2)
- How much do you believe that [infinitely/large/short, WHAT THEY ANSWERED] distance to users affects the use of experimentation in [COMPANY NAME]?

[If answer is **large**]:

- Let's pretend that the distance to users for [COMPANY NAME] was instead short. The company now has no difficulties in accessing the users and collecting qualitative or quantitative user feedback from them. Do you believe that the use of experimentation in the company would be different? And how?

[If answer is **short**]:

- Let's pretend that the distance to users for [COMPANY NAME] was instead large. The company now has difficulties in accessing the users and collecting qualitative or

quantitative user feedback from them. Do you believe that the use of experimentation in the company would be different? And how?

Definitions (if needed)
Concepts:

Control of Roadmap is a classification that can be used on a company. It indicates how much control the company has to change what features and product changes are planned and how they are prioritized. The classification can be described with the following 3 stages.

NO LOW HIGH

	Stage 0	Stage 1	Stage 2
Ability to control roadmap	The roadmap can not be controlled by the team at all. There is no possibility for the team itself to put desired product changes on the roadmap.	Desired product changes can be added to the roadmap by the team, but there are difficulties. The team does not have the full authority to decide on specific changes themselves.	Desired product changes can be added to the roadmap easily by the team itself, and the team is allowed to make the changes they believe are beneficial for the development of the product. Decisions related to changes that might have a very high impact might still be dependent on external parties.

Distance to Users is another classification that can be used on a company. It indicates how difficult it is to access the users of a product and collect feedback from them. This feedback can be both qualitative, through for instance focus groups or surveys, or quantitative through for instance automatic collection of user behaviour in the product. The classification can be described with the following 3 stages.

INFINITELY LARGE LARGE SHORT

	Stage 0	Stage 1	Stage 2
Ability to access users	Users can not be accessed at all. There is no possibility of either qualitative or quantitative data-collection.	Users can be accessed, but there are difficulties with accessing them. Especially when new types of quantitative or qualitative data that is not currently being collected should be collected.	Users can be accessed with ease. To make the decision to collect new qualitative or quantitative user data can be done with close to no delay. Decisions related to any changes regarding data-collection with a very high impact might still be dependent on external parties.

Techniques:

- [Canary releases]
  = Releasing new version/feature to a subset of users only (e.g. 5% based on geographical region, or if they perhaps are willing to test new features), while the rest use the normal version, to see that nothing goes wrong.

- [Gradual rollouts]
  = Often combined with canary releases/dark launches. The number of users assigned to the new version is gradually increased (e.g. increase with 5% steps) until the previous version is replaced, or a threshold is reached.

- [Dark launches]
  = New version is launched in production without being enabled/visible to users (parallell to the old one). “Silent” requests are sent to the dark version to see how it behaves in a real environment, to test scalability and performance, but it does not affect users.

- [A/B testing / Online Controlled Experiments]
  Simplest definition: Control group, Treatment group, measure effect and see if control and treatment differ on some criteria
  Bigger than the experiment execution: experiment methodology, the implementation in code, etc.

E

Static validation interview template

This appendix shows the interview template used for the four static validation interviews that were conducted as part of phase 7 of this thesis work, i.e. the phase about static validation. Some of the actual interviews were in Swedish and an interview template translated to Swedish was used in those cases. Only the main questions are shown here. There were also several formalities covered in the real interview template.

Main Questions

- Q1: Is experimentation used in your company and how?

Control of Roadmap is a classification that can be used on a company. It indicates how much control the company has to change what features and product changes are planned and how they are prioritized. The classification can be described with the following 3 stages.

NO LOW HIGH

	Stage 0	Stage 1	Stage 2
Ability to control roadmap	The roadmap can not be controlled by the team at all. There is no possibility for the team itself to put desired product changes on the roadmap.	Desired product changes can be added to the roadmap by the team, but there are difficulties. The team does not have the full authority to decide on specific changes themselves.	Desired product changes can be added to the roadmap easily by the team itself, and the team is allowed to make the changes they believe are beneficial for the development of the product. Decisions related to changes that might have a very high impact might still be dependent on external parties.

- Q2: How would you classify your company in terms of control of roadmap? (Stages 0-2)
- Q3: Do you think control of roadmap is related to how experimentation is used in your company? And how?

[If their company does not have low control of roadmap]

- Q4: Do you think low control of roadmap could be related to how experimentation is used in a software company?

Distance to Users is another classification that can be used on a company. It indicates how difficult it is to access the users of a product and collect feedback from them. This feedback can be both qualitative, through for instance focus groups or surveys, or quantitative through for instance automatic collection of user behaviour in the product. The classification can be described with the following 3 stages.

INFINITELY LARGE LARGE SHORT

	Stage 0	Stage 1	Stage 2
Ability to access users	Users can not be accessed at all. There is no possibility of either qualitative or quantitative data-collection.	Users can be accessed, but there are difficulties with accessing them. Especially when new types of quantitative or qualitative data that is not currently being collected should be collected.	Users can be accessed with ease. To make the decision to collect new qualitative or quantitative user data can be done with close to no delay. Decisions related to any changes regarding data-collection with a very high impact might still be dependent on external parties.

- Q5: How would you classify your company in terms of distance to users? (Stages 0-2)
- Q6: Do you think distance to users is related to how experimentation is used in your company? And how?

[If their company does not have large distance to users]

- Q7: Do you think large distance to users could be related to how experimentation is used in a software company?

(Optional):

Please fill out the following assessment model to determine how well experimentation is used in your company.

Experimentation dimensions

		Stage 0	Stage 1	Stage 2
Technical	**Data**	No logging is done.	Logging of basic events is done, such as which views in the product users normally visit or what basic actions they perform (e.g. clicks on buttons).	Logging is done comprehensively with information regarding most of the user activity, such as durations in views, flows through the product or technical details of user hardware.
	Experimentation platform and statistical foundation	No experimentation platform is used, experiments are done manually.	Custom made or 3rd party experimentation platform exists. Basic features like defining variations, selecting sample size, experiment duration and assigning users to groups exist and are actively being used.	Advanced features, such as A/A testing, power analysis, alerting, automatic shutdown of harmful experiments and interaction detection exist in the platform and are actively being used.
Business	**Metrics**	No metrics are created.	Basic single metrics are created for measuring success.	Success, debug, guardrail and data quality metrics exist. Overall Evaluation Criteria are created that combines several metrics.
	Type and extent of experimentation	No experiments are ran.	Experiments are ran but not systematically. Experiments are used for changes such as changing existing functionality, deciding if a feature should be removed, adding new functionality or quality assurance.	Experiments are ran systematically on all four types of changes in Stage 1.
	Experimentation impact	The experimentation has no or insignificant impact on the business metrics and the team's way of working.	The experimentation has some impact on important business metrics and it somewhat affects the team's planning and prioritization of product changes.	The experimentation has substantial impact on important business metrics and it completely affects the way the team is working.
Organizational	**Organizational structure**	No data-scientists or experimentation experts are involved in the creation and execution of experiments.	There are data-scientists and experimentation experts directly involved in the creation and execution of experiments.	Teams and team members are educated in how to conduct statistically sound experiments on their own, and there are data-scientists and/or experimentation experts available to assist when help is needed.

Q8: Do you think this model in a good way assesses how well experimentation is used in a company?

<u>Techniques</u>:

- [Canary releases]
  = Releasing new version/feature to a subset of users only (e.g. 5% based on geographical region, or if they perhaps are willing to test new features), while the rest use the normal version, to see that nothing goes wrong.

- [Gradual rollouts]
  = Often combined with canary releases/dark launches. The number of users assigned to the new version is gradually increased (e.g. increase with 5% steps) until the previous version is replaced, or a threshold is reached.

- [Dark launches]
  = New version is launched in production without being enabled/visible to users (parallell to the old one). “Silent” requests are sent to the dark version to see how it behaves in a

real environment, to test scalability and performance, but it does not affect users.

- [A/B testing / Online Controlled Experiments]
  Simplest definition: Control group, Treatment group, measure effect and see if control and treatment differ on some criteria
  Bigger than the experiment execution: experiment methodology, the implementation in code, etc.