Continuous Experimentation: Selected Case Studies from Literature

Veli-Matti Valén
Department of Computer Science
University of Helsinki
Helsinki, Finland
veli-matti.valen@helsinki.fi

***Abstract*—Continuous experimentation is a common way today to test new products, services or planned changes in those. Continuous experimentation also guide companies to make data-driven decisions and help to drive innovation. Web is an optimal platform to quickly evaluate new ideas and innovations. Many internet companies like Microsoft, Google, Amazon, Facebook and Netflix are continuously developing their services through continuous experimentation. In this paper literature study is done to present examples of continuous experimentation at Microsoft and Etsy and how those experiments have affected to the development of new products and services on those companies. Also some examples of the experimentation platforms are presented.**

Keywords—continuous experimentation; controlled experiments; online experimentation; a/b testing

I. INTRODUCTION

Internet provides an ultimate platform for evaluating new ideas quickly and help to make data driven decisions. Internet companies are competing everyday about customers and their attention. This leads to a situation where companies and their products need to evolve at very fast pace in order to stay competitive in the ever changing internet market. To accomplish this, companies need to continuously test new products to drive the development forward. One common way to drive the continuous development in technology companies is through continuous online experimentation. In case of technology companies' continuous experimentation usually means constant execution of controlled experiments on products and services.

The motivation behind controlled experiments for companies is the ability to do data-driven decisions and to move away from the traditional way of making business decision by the loudest person or the highest paid person in the company. The idea is to run experiments, gather the data involved and then analyze the data. The analysis of the gathered data lays the foundation for future decisions. The data can be used to improve products, establish best practices and to resolve debates within the company. When running experiments with thousands or millions of live users, the data amounts are huge. Generally companies are dealing with Big Data and are using business analytics to analyze the data.

This paper is a literature analysis on how today's internet companies are doing continuous experimentation on their customers in the web. Paper will present examples of online experimentations from Microsoft and Etsy. The majority of the source materials for this paper are found from Microsoft's EXP platform archives. EXP platform was Microsoft's experimentation project before the project team was merged into Bing. The idea of the Experimentation Platform was to accelerate software innovation through trustworthy experimentation [3].

The structure of the paper is following. In chapter two is described the idea behind online experimentation and various ways for doing experiments in the web. In chapter three are described the platforms that are used to execute online experimentation in the web. Common experimentation architecture is also described in chapter three. In chapter four is presented case studies from Microsoft about online experimentation. In chapter five is presented same kind of examples of online experimentation from Etsy. Chapter six is for the conclusions and chapter seven is the summary of the paper.

This paper was written as part of the seminar Real-time Value Delivery in Software Engineering (RTVDSE14). The seminar was part of the Software Systems specialization studies and was held in Helsinki on spring 2014. The seminar was organized by the Department of Computer Science of Helsinki University.

II. CONTINUOUS EXPERIMENTATION

Continuous experimentation is a term to describe constant testing of products or services. One common way to do continuous experimentation in web is through online controlled testing [5]. In controlled testing the tests are executed on real users or customers without their knowing. Users are randomly assigned to two or more groups, to the control group, on which the normal existing version of the product is shown and to the treatment, which the new version of the product that is being tested is shown [4]. There can be multiple treatment groups to test multiple versions of the product at the same time [4]. The data from the groups is then collected and analyzed. The basic idea behind controlled testing is described in figure 1. Controlled experiments are called with many names of which

the most common are A/B tests, randomized tests, split tests and parallel flights [5].

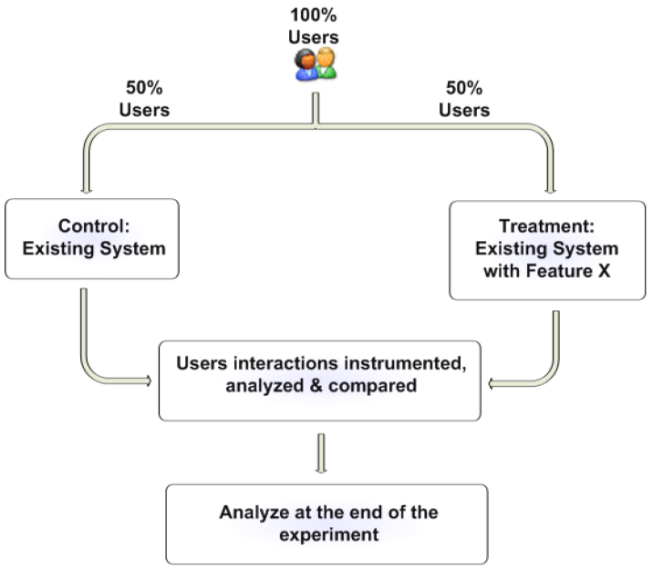

Figure 1. The idea behind controlled testing [5].

The most important part of controlled testing is to define the metrics that are collected from the testing and evaluated afterwards. These are also referred as Overall Evaluation Criteria (OEC) [4]. A good OEC should not be short-term focused, for example number of clicks gained during testing. But a more long-termed focused that predicts long-term goals, like site revisits [5]. These metrics can vary from runtime performance to implicit and explicit user behaviors and survey data [5]. After the tests are run, the data is collected and analyzed to identify if there was a major difference in results between the control and the treatment groups. The difference between the OEC statistics for the control and treatment groups is called the treatment effect [4]. If the experiment was properly planned and executed, the only real difference is the change of OEC values between the groups [5].

When running controlled experiments, there are some known pitfalls that organizations should try to avoid. One of the biggest pain points can be the decision about the used OEC's for the experiments. Organization must be able to find a common understanding what they want to meter [7]. Sometimes people may claim that their goals for the experimentation are "soft" or "intangible" and those cannot be measured [7]. One identified pitfall with controlled experiments is also the "newness" effect. Some users can be sensitive for dramatic changes. For example big changes in the site navigation can degrade the customer experience and turn away users, even the new feature is better and more efficient than before [7]. Because many of the controlled experiments nowadays rely on the internet browser's cookie technology, running tests on users who are browsing web using privacy mode or deleting temporary internet files after surfing the web can contaminate the results or create inconsistency [7].

III. PLATFORMS FOR ONLINE EXPERIMENTATION

The online testing requires an infrastructure on which the tests can be carried out and the tools that are used to control and monitor the test execution and data collection. Many big companies have invested lot of money and resources to build testing platforms so they can more efficiently test their products. These platforms are divided to two categories, private platforms can be used only by the companies that own them and to public platforms that are available for big audiences. Public platforms are provided as a service for companies who want to do online testing themselves, but who do not want to invest to the infrastructure or tooling.

Examples of private platforms are such as Microsoft Experimentation Platform (EXP) and Amazon WebLab. The EXP is Microsoft's internal platform that is used to test various Microsoft's web sites and services and to quickly discover which features and product development activities will help Microsoft best to achieve their business objectives [8]. WebLab is a platform to evaluate the effect of web site development on their customers through scientifically controlled online experiments. Public platforms are such as Google Analytics Content Experiments and Visual Website Optimizer. These platforms are available for the public and charged based on the number web site visits by a monthly fee.

Common architecture of an experimentation platform consists from three main components, which are the randomization algorithm, assignment method and the data path. The randomization algorithm is responsible of mapping the users, which are visiting the website to a control and treatment groups. It is vital that the randomization algorithm works correctly and consistently so that users are divided equally to control and treatment groups and that the assignments are fixed, so that users always end up to the same site. Also if multiple experimentations are run it is vital that there is no correlation between those. The assignment method component enables the experimentation platform to execute different code paths for different users depending the groups that they are divided to by the randomization. Good assignment method can manipulate anything from web site front-ends to back-end algorithms. The data path component is responsible of collecting the raw data gathered during the experiment and converting it to metrics that can be quantified. [1]

Microsoft EXP's experimentation architecture is described in the following pictures. In the figure 2, the user is requesting a web site, the EXP will respond to user's request and return a control or treatment version of the requested web site. In figure 3, user submits Page Request (PR) to the server. After that the server will record Unique User ID (UUID) of the user, the URL of the web site the user is requesting and the Page Request (PR) and stores those to the EXP database.

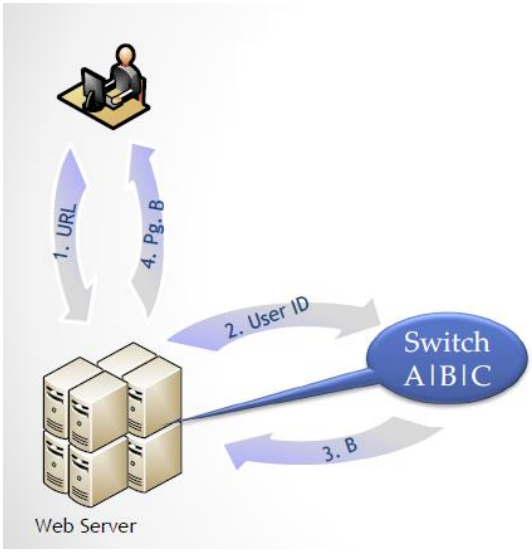

Figure 2. Microsoft EXP assign treatment [9].

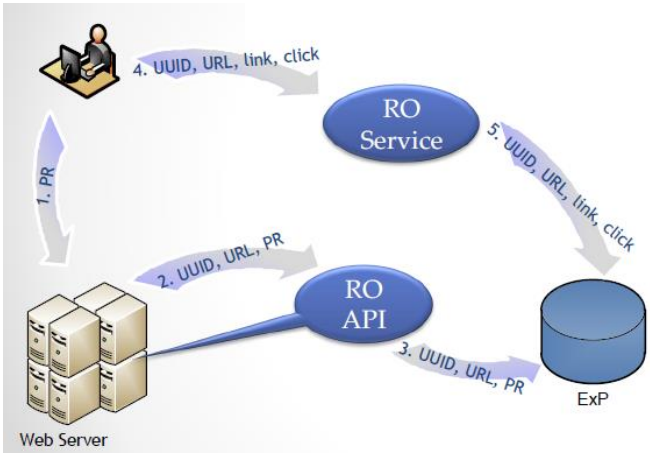

Figure 3. Microsoft EXP record observation (RO) [9].

IV. EXPERIMENTATION AT MICROSOFT

Microsoft is one the world's biggest software companies and is responsible of products like Windows and Office. In addition to this, Microsoft is also running hundreds of web sites that serve different aspects of their business. The most important ones of these are MSN web portal and Bing search engine. Microsoft has its own EXP platform to execute continuous testing on their websites and services. It took three years for Microsoft to build the EXP platform that is capable of running and analyzing multiple controlled experiments at a very large scale. The controlled experiments that Microsoft has done with the platform vary from 10K to 100M users per experiment [2]. Some of the case studies of Microsoft's experimentations are described in this chapter.

The first case study is about MSN Real Estate's search a home widget. The search widget lets user to search apartments from Microsoft's partner sites. Microsoft gets referral fee from every search that users execute with the widget. Microsoft wanted to redesign the widget from better user experience. The OEC for this experiment was the number searches users executed with the widget. For each search, Microsoft got referral revenues from partners. The control and the treatment versions of the widgets are presented in the figure 4. After the experimentation was run, the results showed that the treatment 5 increased referral revenues for MSN Real Estate almost 10%. It was estimated by the Microsoft's EXP team that the simplicity of the treatment 5 was the reason why the users did more searches by using it. [3]

In this experimentation the OEC was properly defined and it was focusing on the long-term goal, which was the referral revenue gathered from the searches.

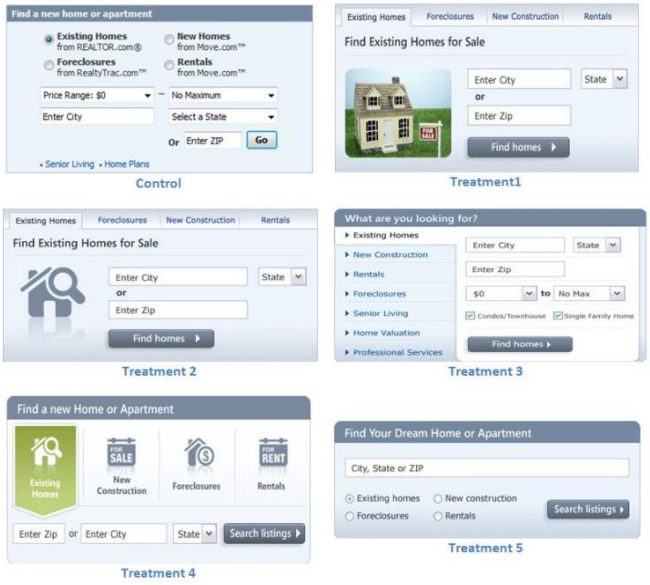

Figure 4. Tested widgets for the MSN Real Estate [3].

The second case study from Microsoft is about the redesign of the Office Online store. Office Online is a central place to get more knowledge about Microsoft Office products. Users can also buy Office through the Office Online web store. Microsoft wanted to redesign the Office Online home page to make it more appealing to users and to get them to buy more copies of Office by clicking the buy button. Any clicks within the red rectangles in control l and treatment would be counted as click on the buy button (figure 5 and 6). The OEC for this experiment was the amount of clicks for the buy button. After the experiment was run it was seen from the results that control had more clicks than the treatment. The treatment had 64% fever clicks. It was argued that because the control was not showing the price of the product, the users were clicking it more than the treatment, in which the price is showing. Usually users want to know the price of the product before making the decision to make the purchase. [3]

In this experimentation the OEC was badly defined. The number of clicks of the buy button does not yet tell if the user is making the purchase or just viewing the price of the product. The experimentation team should have used

more thought to define a better OEC that would have predicted long-term goals. The OEC could have been the metered revenue from the Office 2007 product purchases. Poor definition of the OEC can make the results of experiments worthless in a long run.

Figure 5. Control version of the Office Online experiment [3].

Figure 6. Treatment version of the Office Online experiment [3].

V. EXPERIMENTATION AT ETSY

Etsy is a web site that focuses on selling handmade or vintage items. It is known as the "world's handmade and vintage marketplace" [6]. Etsy's idea is to provide a web store platform for its customers so they can use it sell their products. The items that are sold in Etsy cover art, photography, clothing, jewelry, food, bath and beauty products [10]. Etsy is well known for its presentations and focus on the idea of continuous online experimentation. Etsy has put a lot of effort to develop an infrastructure and tools on which they use execute their experiments [6]. Etsy runs hundreds of controlled experiments on their site continuously. The idea of running the experiments is to do small, measurable changes without breaking anything [6].

The first case study from Etsy is the infinite scroll experiment. The idea was to implement infinite scroll to a search results page. Infinite scroll is a common feature in today's web sites. The idea behind is to always show more items when users scroll the page down. The OEC's that were defined for this experiment were the number of items users saw in their search results, the number of results the users clicked from the search results, the number of items users saved as favorite items and the number of items the users bought from the search results. The control group was the normal version of the scroll feature and the treatment was the infinite scroll version. The results from this experimentation were shocking. The control group saw double the amount of search results than the treatment. The treatment group clicked 10% less items, they saved 8% less items as favorites and bought 22% less items than the control group. [6]

The OECs for the infinite scroll were properly defined and they metered the user behavior from which the user experience can be derived from. After the experiment the project team was able to see clearly which one of the designs for the search results would be better for Etsy's and their customers' business.

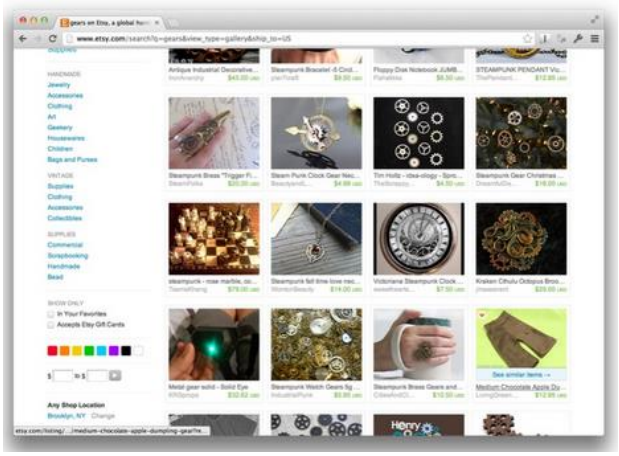

Figure 7. Infinite scroll feature [6].

The second case study from Etsy is the redesign of the search field which purpose was to get rid of the old search function. The old search field had a dropdown list of the different sections of the marketplace. So by selecting one of the items in the dropdown list, the search function would only do search of items from the selected marketplace as seen in the figure 8. The project wanted to remove the feature so that the search would always do a search towards all items. They feared that when users were searching items from the store, they did not know that the search results would be limited only to the selected marketplace. The project team wanted to remove the fixed categories and to add auto-suggestion to the search field as seen in figure 9. The auto-suggestion means that when user starts typing something to the search field the algorithm will make suggestions for the

user based on his/hers typing. The OEC that was defined for this experimentation was the amount items bought by using the search function. The treatment group with the new auto-suggestive search bought 3,7% more vintage items than the control group with the old search function. By running the experiment, the project team also verified their hypothesis about the common users who are not familiar about the function of the old search dropdown list, which limited the searched items according the selected category. [6]

As with the earlier case study from Etsy, the OEC was precisely defined and it was focused on the long-term goal, the gathered revenue from products sold at the web site. The OEC follows the core business idea behind Etsy, which is to provide a marketplace for Etsy's customers to sell their products in the web.

Figure 8. Old dropdown search field [6].

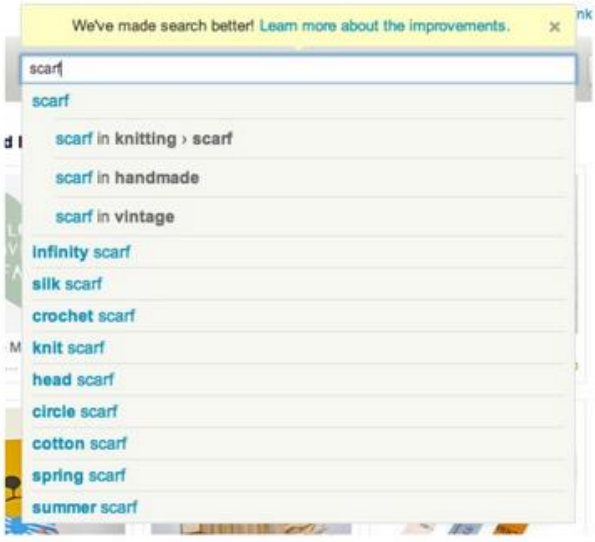

Figure 9. Auto-suggestive search field [6].

VI. CONCLUSIONS

Today, running controlled experiments in the web is a rather easy task for any company that wants to develop web-based products or services. There are multiple public services available that can be used for running continuous experimentation through controlled online experimentations. These tools are easy to use and do not require the person to be professional on web technologies such as HTML. The challenge when running experiments is not the technological aspects. The challenge and the risk lie in the definition of the values that are metered in the experiments. For every experiment, the OEC must be defined and agreed by the organization whose product is being experimented. The most important thing when defining OEC for an experiment is to focus on the long-term goals. The long-term goals should be derived from the business idea of the company running the experiments. As seen in this paper, there is one bad example of improper OEC definition, which literally made the experiment that was run worthless. One important factor about controlled experiments is that the experiments usually tell which of the tested changes of the product or service is better or worse than the other according users. The one thing that these experiments do not tell is why the control is better than the treatment(s) or vice versa. Sometimes running experiments in the web can lead to some very puzzling outcomes.

VII. SUMMARY

Continuous experimentation is a constant way to develop new products or enhance the existing ones for better revenues and better user experience. One common way of executing continuous experimentation is through controlled online experimentations. Controlled experimentation usually means executing A/B testing on web sites or other services. In this paper the idea behind continuous experimentation was described. Also the platforms for executing such experiments were described through examples and also the common experimentation architecture was presented. In the case study chapters examples about controlled experiments were presented from Microsoft and Etsy. In addition to these the applicability of the OEC's on those experimentations were analyzed.

REFERENCES

[1] Kohavi R., Longbotham R., Sommerfield D., Henne R.M., Controlled experiments on the web survey and practical guide. Springer Science+Business Media, published 30 July 2008

[2] Kohavi R., Deng A., Frasca B., "Walker T., Xu Y., Pohlmann N., Online Controlled Experiments at Large Scale" [19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) 2013]

[3] Kohavi R., Crook T., Longbotham R., Online experimentation at Microsoft [Third Workshop on Data Mining Case Studies and Practice Prize]

[4] Kohavi R., Henne R.M., Sommerfield D., "Practical guide to controlled experiments on the web: Listen to your customers no to the HiPPO" [14th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) 2007]

[5] Kohavi R., Longbotham R., Walker T., Online experiments: Practical lessons. IT Professional September 2010 p. 82-85.

[6] McKinley, D., Design for Continuous Experimentation: Talk and Slides. http://mcfunley.com/design-for-continuous-experimentation, accessed February 9, 2014

[7] Online Experimentation at Microsoft (slides)

[8] Experimentation Platform Group Jobs http://www.microsoft-careers.com/go/Experimentation-Platform-Group-Jobs/217323, accessed February 14, 2014

[9] Eliot S., "Testing with real users (slides)" [Better Software Conference, June 9, 2010]

[10] Etsy. www.etsy.com, accessed February 15, 2014