8-2015

Enabling real time in-situ context based experimentation to observe user behaviour

Kartik MURALIDARAN
Singapore Management University, kartikm.2010@phdis.smu.edu.sg

Enabling Real-Time In Situ Context-Based Experimentation to Observe User Behaviour

by

Kartik Muralidharan

Submitted to School of Information Systems in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Information Systems

<u>**Dissertation Committee:**</u>

Rajesh Krishna BALAN (Supervisor / Chair)
Associate Professor of Information Systems
Singapore Management University

Archan MISRA (Co-supervisor)
Associate Professor of Information Systems
Singapore Management University

Youngki LEE
Assistant Professor of Information Systems
Singapore Management University

Edward CUTRELL
Senior Researcher
Technology for Emerging Markets Group
Microsoft Research India

Singapore Management University
2015

Enabling Real-Time In Situ Context-Based Experimentation to Observe User Behaviour

Kartik Muralidharan

Abstract

Today's mobile phones represent a rich and powerful computing platform, given their sensing, processing and communication capabilities. These devices are also part of the everyday life of millions of people, and coupled with the unprecedented access to personal context, make them the ideal tool for conducting behavioural experiments in an unobtrusive way.

Transforming the mobile device from a mere observer of human context to an enabler of behavioural experiments however, requires not only providing experimenters access to the deep, near-real time human context (e.g., location, activity, group dynamics) but also exposing a disciplined scientific experimentation service that frees them from the many experimental chores such as subject selection and mitigating biases.

This dissertation shows that it is possible to enable insitu real-time experimentation that require context-specific triggers targeting real participants on their actual mobile phones. I first developed a platform called *Jarvis* that allows experimenters to easily, and quickly create a diverse range of observational and treatment studies, specify a variety of opportune moments for targeting participants, and support multiple intervention (treatment) content types. Jarvis automates the process of participant selection and the creation of experimental groups, and adheres to the well known randomized controlled trial (RCT) experimental process.

Of the many possibilities, a use case I envision for Jarvis, is providing retailers a platform to run *lifestyle* based experiments that investigate promotional strategies. Such experiments might entail the platform to provide the experimenter with the appropriate target population based on their preferences. To support this, I developed

a matching and scoring algorithm that accurately factors participants' preferences when matching experiment promotions and is capable of combining structured and unstructured promotion information into a single score. Doing so, will allow the experimentation system to target the right set of participants.

Finally, I developed techniques for capturing and handling context uncertainty within Jarvis. As the opportune experiment-intervention moments are identified from sources such as sensors and social media, which have inherent uncertainties associated with them, it is crucial that such information is recorded and/or processed. More specifically, Jarvis defines a confidence metric for the location predicate as well as dynamically computes the sample size for a given experiment under context uncertainty. In doing so it provides adequate information to the experimenter to process the results of an experiment in addition to maximizing the statistical power.

I validated my dissertation in the following way. Through a series of *live* experiments I showcase the diversity of the system in supporting multiple experiment designs, the ease of experiment specification, and the rich behavioural information accessible to the experimenter in the form of a report. The matching and scoring algorithm was evaluated in two different ways; First, an in-depth analytical evaluation of the ranking algorithm was conducted to understand the accuracy of the algorithm. Second, I ran a user study with 43 undergraduate students to understand the effectiveness of the algorithm. Finally, I validate the context-uncertainty handling capabilities of Jarvis through simulations and show that using overlap ratios to represent location confidence is reliable and that the algorithm to estimate the number of false positives has minimal errors. Both these values are important in understanding the outcome of an experiment and in turn defining it's success criteria.

Keywords: context aware computing, mobile computing, event processing, context uncertainty, mobile application, natural language processing, behavioural experimentation.

Table of Contents

List of Figures

List of Tables

Acknowledgments

A Ph.D. is an epic journey of sorts, and I could never have done any of this without the support and encouragement of a lot of people.

First and foremost I want to thank my advisor Rajesh. I owe you so much. You have taught me, both consciously and unconsciously, how good systems research is done. I appreciate all your contributions of time and ideas to make my Ph.D. experience productive and stimulating.

I was lucky to have a very active thesis committee. Archan Misra, Youngki Lee and Ed Cutrell, thank you for providing me with invaluable advice and comments on my research. I am also very grateful to Narayan Ramasubbu, Srini Seshan, William Tov and Reetika Gupta. They were instrumental at different stages of my dissertation and I have learnt so much from them.

My thesis is a part of the larger LiveLabs ecosystem, relying on many other systems to function. I am therefore indebted to the members of the LiveLabs team. I would particularly like to thank Swetha Gottipati, Nguyen Vu Nhat Minh, Jeena Sebastian, Le Gia Hai, Tan Kiat Wee (William) and Ritesh Kumar. Without them this work could not have happened. They are not just colleagues but my friends too.

I would also like to thank Pei Huan and Chew Hong for making sure my Ph.D. was on track as well as Huang Sipei and Jonathan Wang for handling the nitty gritties of travel claims and the sort. They took care of things so that I could concentrate on my research.

My time at SMU hasn't been all about work. I've been fortunate to have a great group of friends to hang out with. Sougata Sen, Joseph Chan, Nguyen Huynh, Zhang Nairan, Payas Gupta and Swapna Gottipati, thank you for being my sounding board for not just research but other stuff as well. This, I believe, is the key to getting through a Ph.D. program - having good friends to have fun with and complain to.

I would also like to express my thanks to Deependra Moitra. He has been a mentor to me both at a professional and personal level and has encouraged me throughout the Ph.D. program.

Finally, I am especially grateful to my family. My 'attai' and father's constant encouragement towards the pursuit of knowledge is what got me wanting to do a Ph.D. I hope I made them proud. Mummy, without your unending support and love from childhood to now, I never would have made it through this process or any of the tough times in my life. I thank my sister for her support and for 'just being there' when needed. And Darshana, my wife, my best friend. I thank you for all the love, support and patience.

To my father.

List of Publications

Conference Papers

Kartik Muralidharan, Swapna Gottipati, Narayan Ramasubbu, Jing Jiang, and Rajesh Krishna Balan. myDeal: A Mobile Shopping Assistant Matching User Preferences to Promotions. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, London, 2014.

Kartik Muralidharan, Srinivasan Seshan, Narayan Ramasubbu, and Rajesh Krishna Balan. Handling Location Uncertainty in Event Driven Experimentation. In *Proceedings of the 8th ACM International Conference on Distributed Event-based Systems*, Mumbai, 2014.

Workshop Papers

Kartik Muralidharan, Swapna Gottipati, and Rajesh Krishna Balan. Deal or No Deal: Catering to User Preferences. In *Proceedings of the 12th International Conference on Pervasive Computing and Communication Workshops, PerCom Workshops*, Budapest, 2014.

Demo and Poster

Kartik Muralidharan, Swetha Gottipati, Rajesh Krishna Balan and Archan Misra. Jarvis: A Behavioural Experimentation Platform. In *Proceedings of the 15th Workshop on Mobile Computing Systems and Applications (HotMobile '14)*, Santa Barbara, 2014. (Poster)

Kartik Muralidharan, Swapna Gottipati, Narayan Ramasubbu, Jing Jiang, and Rajesh Krishna Balan. myDeal: The Context-Aware Urban Shopping Assistant. In *Proceedings of the 3rd ACM SIGOPS Asia-Pacific Workshop on Systems (APSys '12)*, Seoul, 2012. (Demo)

Chapter 1

Introduction

In 1747 James Lind, whilst working as a surgeon on a ship, was appalled by the high mortality of scurvy amongst the sailors. He planned a comparative trial of the most promising cure for scurvy [64]. Lind selected 12 men from the ship, all suffering from scurvy, and divided them into six pairs, giving each group different additions to their basic diet. Some were given cider, others seawater, others a mixture of garlic, mustard and horseradish. Another group of two were given spoonfuls of vinegar, and the last two oranges and lemons. Those fed citrus fruits experienced a remarkable recovery. While there was nothing new about his discovery - the benefits of lime juice had been known for centuries - Lind had definitively established the superiority of citrus fruits above all other 'remedies'. Although Lind was not the first to suggest citrus fruit as a cure for scurvy, he was the first to study their effect by a systematic experiment that ranks as one of the first clinical experiments in the history of medicine.

Since then, there have been several additions to improve the efficacy of the experimental process. *Randomized controlled trials, placebo-controlled experiments,* all sought to strengthen the validity of an experiment. However, despite these improvements, the experimentation process has several limitations particularly in the field of behavioural sciences. Consider the following scenario, of interest to the marketing sciences, that aims to understand customer needs and behaviour: Information about promotions or deals often reaches consumers at inconvenient times or locations which prevents them from taking advantage of such offers. Today, however, due to technological advances, location has become dynamic and traceable to a specific user. From a targeting perspective, the location-based mobile advertising segment is especially interesting. In the classical location-based advertising scenario, a coffee shop, usually Starbucks, issues coupons to consumers that are near their retail establishments. Of course, any marketer using location-based advertising would prefer to target more accurately than just by location. They would prefer to know more about the customer, including her current needs for the retailers products or services, her past history buying from the retailer and their competition, her present schedule and whether it permits a visit to the retailer, and whether her current context would permit her attention to be directed to an ad [115]. The question therefore arises, is there a right time and right place for consumers to receive and respond to such offers? More specifically, are advertisements, which are more personalized in terms of the customer location, more effective than considering additional information such as the customer's current activity? Current behavioural research techniques to answer this question are quite restrictive. User research methods such as shadowing, field interviews, and diary studies, although plausible, have several limitations ranging from generalizability of the outcome to cost and scalability of conducting the research.

An alternative to these techniques is to use the smartphone as a behavioural research tool. These devices have additional sensors for location, motion, sound, and lighting that make it possible for the phone to determine and provide details of the user and the situation in which the phone is being used. This ability of smartphones to observe and interpret the users situations and activities over time means that many of the elements of targeting can be fulfilled more extensively than other technologies have been able to. A key construct that makes this possible is user context.

Context is defined as any information that can be used to characterize the situation of an entity, where an entity can be a person, place, or physical or computational object. Context-awareness or context-aware computing is the use of this context to provide task-relevant information and/or services to a user [38]. Early discussion of context-aware applications focused on mobile personal devices. Such devices are highly contextual, monitoring personal surroundings (such as location, movement, sound) using built-in sensors and employing applications designed to respond in real time to changes in personal situation [48]. However, as mobile devices and applications have become more advanced and more relevant for business, the information generated in these devices has become available to back-end business applications. The personal context that was originally confined to a mobile device becomes available as input to larger enterprise applications and tools. An example of this in play is Google Now [57], an intelligent personal assistant (available for Google's Android operating system) that combines data from users' accounts and sensor data from mobile phones to provide suggestions. For example, it combines the data from Google Calendar and other context such as the location of the user's next appointment along with the time, traffic data and current location to advise the best time to travel. The future of mobile computing will therefore be context-aware computing with mobile applications adjusting to the user's location, identity and past behaviours [80, 132, 137]. Contextual mobile applications will lead to new user experiences that will be simple, visually attractive, compelling and interactive. Additionally, given that mobile phones are part of the everyday life of millions of people, they represent an ideal computing platform to monitor behaviour and movement [3].

Thus, while user context aids personalization of services and content, it also provides relevant information to support in situ behavioural experiments and answer research questions posed earlier. For example, to answer the earlier question of whether only location information is sufficient when targeting consumers, we could send a promotion to a set of consumers whose only contextual relevance to the promotion is their current location (i.e., they are passing the store offering the promotion) and compare them against (in terms of coupon proneness) another set of consumers who have additional relevant context predicates (for example, they've been inside the store for more than 10 minutes).

What we therefore need is a system that facilitates a mapping (with ease) between the required context attributes and the content to be delivered for the different scenarios, provide support for controlled experimentation, as well as capture user reaction to the context-based content. Such a system could then facilitate a better understanding of user behaviour through the process of experimentation - potentially changing how behavioural scientists study human behaviour. The goal of this system would then be to provide experimenters access to deeper, near-real time user context (e.g., location, activity), handle the hassles of experimentation such as subject selection, information uncertainty and so on, as well as support multiple types of controlled experiment design. In addition, by providing adequate information to the experimenter to process the results of an experiment, should help them decide whether to re-run the experiment (with new parameters and constraints), run a new experiment, or declare success.

In the rest of this chapter, I first discuss previous approaches to supporting in situ experimentation. I then briefly describe my solution. I next state the thesis statement and describe the validation plan. Finally, I end with a roadmap describing the rest of this dissertation.

1.1 Previous Approaches to Supporting In Situ Experimentation

Current research methods used in the behavioural sciences make little use of technology. In addition to traditional self-reports, researchers may also rely on one-time behavioural observations of participants in laboratory settings. Such methods can be useful, but the fact that they are based on behaviour in a lab raises concerns about their generalizability to non-lab contexts. Recently, researchers have begun to use new methods in an effort to examine behaviour in everyday life. For example, daily diary methods [18] ask participants to report the social events and psychological states they experienced either at the end of the day or periodically throughout the day. Another method has used devices that take audio recordings (or snapshots) of participants daily lives every few minutes, which are later transcribed and coded by teams of researchers. These methods have advantages over the traditional survey methods, but they nevertheless suffer from issues associated with forgetting events that took place during the day, and carrying an additional obtrusive electronic device. Also, these techniques may lead to biased results since people are aware of being constantly monitored.

Mobile phones offer an unobtrusive means of obtaining information about the behaviour of individuals and their interactions. Mobile sensing technology has the potential to bring a new perspective to the design of social psychology experiments, both in terms of accuracy of the results of the study and from a practical point of view. Mobile phones are already part of the daily life of people, so their presence is likely to be forgotten by users, leading to accurate observation of spontaneous behaviour. Systems such as EmotionSense [123], Cenceme [98] and Betelgeuse [81] have shown the potential of mobile phone sensing in providing information such as user movement and activity for recreational and healthcare applications.

Unfortunately, to enable this computing paradigm of in situ experimentation, there is a pertinent need to provide a mapping between the *3 Ws* - *What* to send, *Whom* to send it to and *When* to send it - and getting this right is not easy. Building such a system poses several challenges; given the potential volume of user context data that is produced by mobile devices, there is a need to build a strategy to derive value from the information [63, 79], a need to support the process of experimentation (enrolment, intervention allocation, follow-up and data analysis), and the ability to create a diverse set of experiment scenarios. Attempts in the past has shown the ability to create this mapping using a limited set of context sources and build applications designed to respond to events (changes in personal situation) in real time [40, 53, 84, 157, 158]. However, these studies are not representative of reality as they are controlled, often with a limited set of users restricted to specific campus/office environments and support only handful of triggering events. What is missing is a forging of this mapping in real time.

1.2 Solution: Jarvis

To fill this missing gap, I propose a solution, called Jarvis, that is based on the well-known randomized controlled trial (RCT) experimental process [26]. By representing context predicates of an experiment as events, I create an event-driven system that enables a real-time in situ participant enrolment phase. The underlying behavioural experimentation system also provides the components necessary to automate the additional phases of the RCT - intervention allocation, follow-up, and data analysis - with the goals of minimizing selection bias and allocation bias while maximizing the statistical power.

Experiment specification in Jarvis is easy and quick and supports both observational and treatment type studies.

1.3 The Thesis

The thesis statement can thus be stated as follows:

It is possible to build a system that enables the observation of user behaviour, through the process of experimentation, by specifying triggering events (a combination of static and dynamic user-context attributes) to target real mobile users in real environments with personalized content.

This dissertation establishes the thesis via the following steps [Figure 1.1]:

1a. First, it designs the architecture of a context-based experimentation system. This step involves clearly identifying the different modules of the system and

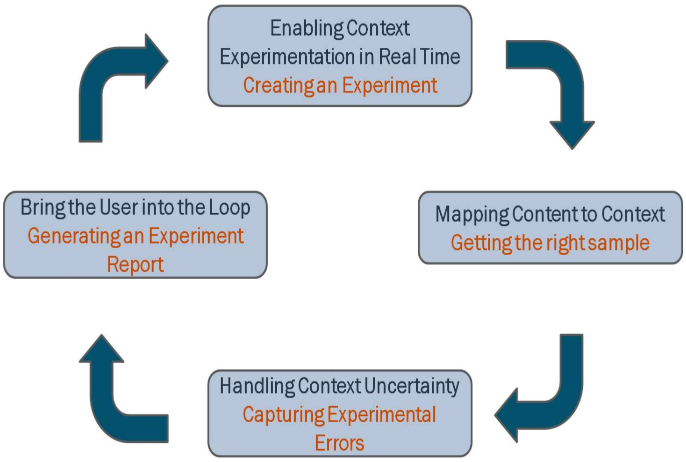


Figure 1.1: Validation Roadmap.

what function they will perform. The architecture will need to support running experiments that can aid inferring what user-contexts are relevant for different scenarios.

1b. Define the ontology of a context-based experiment and database schema to support the process of experimentation. This step involves defining the experiment specification workflow and what user-context predicates and logical operators will be supported. The ontology has to be powerful enough to support a wide range of experiment scenarios.

1c. Develop a fully functional context-based experimentation system. This system should demonstrate that it is possible to specify context predicates that needs to be matched by participants and also send an intervention with the experiment details (e.g., promotions) to the set of matched participants.

2. Design and develop an algorithm that can match and rank content with user preferences and show that users receive relevant content with the help of this algorithm.

3a. Design and develop an algorithm to reliably represent the location confidence for the set of matched participants and estimate the number of false positives within this set with minimal errors.

3b. Design and develop an algorithm to determine the participant sample size appropriate for an experiment subjected to context uncertainty.

4. Collect user feedback based on multiple modes of mobile user interaction, within a context-based experiment, and generate an experiment outcome report that captures the impact of the experiment on user behaviour.

1.4 Dissertation Roadmap

The rest of this dissertation is organized into six chapters and three appendixes as follows:

Chapter 2 describes the characteristics of an in situ behavioural experimentation system. In particular, it describes the challenges in building such a system, the key focus of this dissertation. It then details the requirements any solution geared towards supporting context-based experimental studies as well as motivating experiment scenarios highlighting the need for such a system.

Chapter 3 describes the architecture designed to support the running of context-based experiments. I show how the system adheres to the traditional process of experimentation while abstracting the complexities of participant selection and the handling of experimental bias. Through a series of *live* experiments I showcase the diversity of the system in supporting multiple experiment designs, the ease of experiment specification, conformance to the traditional process of controlled experimentation, and the rich behavioural information accessible to the experimenter in the form of a report.

Ensuring experiment validity requires getting the right sample for that experiment. In Chapter 4 I describe how the system handles the requirement of targeting the right sample by understanding user preferences. Through a user study I show that my preference-based system is more accurate (users can find the most interesting promotions better) and faster to use than two other common system designs.

Many context-aware services make the assumption that the context they use is completely accurate. However, in reality, both sensed and interpreted context is often ambiguous. In Chapter 5 I address how the system represents and handles context uncertainty for an experiment. I show how the module defines a confidence metric for the location predicate as well as how it stochastically estimates additional information such as the number of false positives. I further demonstrate how the module determines the appropriate sample size for a given experiment under the purview of context uncertainty.

Finally, Chapters 6, and 7 present the related work and the dissertation conclusion (that summarizes the main contributions of the dissertation to the mobile community and presents future work) respectively.

Chapter 2

Enabling In Situ Experimentation

Mobile phones represent an ideal computing platform to monitor behaviour and movement, since they are part of the everyday life of millions of people [3]. Recently, systems such as GruMon [143] and SocioPhone [85] have shown the potential of mobile phone sensing in providing information such as group interaction and activity for recreational and healthcare applications. One possible use of these technologies is arguably the support to sociology experiments [95] which involve studying peoples daily life and interactions. In the past, this analysis has been performed with the help of cameras (in home/working environments or in laboratories), by using voice recorders attached to people, and self reports using daily diaries or PDAs [18]. However, these techniques may lead to biased results since people are aware of being constantly monitored. Instead, mobile phones offer an unobtrusive means of obtaining information about the behaviour of individuals and their interactions.

A context-based experimentation platform could then provide the ability to experiment, under varying event (behaviour and interaction) conditions, with real users using their regular phones in real-world environments. Such a system will transform the mobile device from being merely an observer of human context to an enabler of behavioural/sociological experiments and provide greater insight into the user.

In this chapter, I describe what an in situ context-based experiment is. I then describe the kinds of social experiments that this thesis is concerned with. I then present some of the key requirements necessary to support in situ experimentation. Finally, I list the characteristics of an event-driven architecture that can support running such experiments.

2.1 What is an In Situ Context-Based Experiment?

An *event* is an occurrence within a particular system or domain; it is something that has happened, or is contemplated as having happened in that domain [46]. The event concept is simple yet powerful. Suppose you are working on your laptop in a coffee shop, and since you entered this coffee shop several things have happened: a customer walks up to the counter and asks for coffee and pastry; the person behind the counter puts the pastry into the microwave, prepares the coffee, takes the pastry out of the microwave, takes the payment, gives the tray to the customer, and then turns to serve the next customer. These are all events.

Many of the events around us are outside the scope of our interest. Some events are background noise and do not require any reaction, but some do require reaction, and those are referred to as *situations*. A situation is an event occurrence that might require a reaction [4]. These situations often present opportune moments for providing interventions to observe how personality is expressed and perceived in the natural setting of everyday behaviour [95].

In this thesis I support experiments that require sending interventions as a **reaction** to a situation. An in situ context-based experiment is therefore one, where interventions are allocated to participants based on the situation (of the participant) post any enrollment criteria defined by the experiment. Mobile context plays a key role in detecting the situation (with some degree of certainty) as well as assessing the eligibility of the participant for the experiment.

This leads to the central challenge of this dissertation: *Is it possible to build a* *system that allows specifying triggering events (situations) to target users in real environments in real time with interventions? Can the system support the process of experimentation while handling the inherent uncertainty of a situation and yet effectively execute experimental studies and obtain meaningful conclusions?* My design must identify itself with the traditional experimental process allowing the experimenter to specify (with ease) the different facets (participant eligibility criteria, triggering situation, treatment, etc.) of an experiment design while adhering to the principles of a controlled experiment (randomization, selection bias, blinding).

2.2 Motivating Scenarios

This thesis will help to enable the following scenarios. The scenarios are listed along with some of the technology necessary to realize each scenario. In this work, while I will be focussing on targeting consumers with promotional content in urban environments such as shopping malls and university campuses, the platform is diverse enough to permit experimental studies beyond this narrow categorization.

2.2.1 Scenario 1

Paul is a coffee shop owner. He uses a location based advertising service that sends coupons to people passing by his store everyday with the hope that it will improve sales. He wonders whether offering discount coupons to people who have been standing outside the coffee shop for 10 minutes would improve sales as opposed to giving it when they pass by?

This scenario requires the system to be able to do the following things:

- capture the various mobile user-context attributes.
- create an event mapping experiment (with ease) between the required context attributes and the content to be delivered for the different scenarios.

- analyse mobile user-context to identify when the required context-based event occurs. In this case, the location of the participant as well as their activity.
- deliver the content when the required context-based event occurs.
- be able to estimate the number of coupons that were sent in error.
- capture user reaction to context-based content for the different scenarios.
- generate an experiment report.

2.2.2 Scenario 2

William is a social scientist. His area of research is studying personality traits, particularly the dimension of introversion-extraversion. Introverts tend to be more quiet, reserved, and introspective. Unlike extraverts who gain energy from social interaction, introverts have to expend energy in social situations. He would like to setup a study that sends a survey capturing the emotional state of his participants everytime they have been in a social interaction for longer than 30 minutes.

This scenario requires the system to be able to do the following things:

- capture the various mobile user-context attributes.
- create an event mapping (with ease) between the context attributes and the content to be delivered.
- analyse mobile user-context to identify when the required context-based event occurs. In this case, when the participant has initiated a social interaction as well as the duration.
- send the survey to matching users everytime the event occurs.
- capture user reaction to context-based content.
- generate an experiment report.

2.2.3 Scenario 3

Darshana is a professor of marketing. She would like to understand how the consumer's mindset works during the shopping process. Consumers are exposed to a plethora of information (in the form of text, images) during the different decision-making stages. Yet, it is unclear if the information is influencing how they navigate the shopping environment. She would like to conduct an experiment (inside the store of a mall) where the participants are primed to be in one of the mindsets that characterize the two decision making stages (deliberative or implemental) using a promotional offer and track/observe their navigational behaviour in addition to other interaction.

This scenario requires the system to be able to do the following things:

- capture the various mobile user-context attributes.
- create an event mapping experiment (with ease) between the required context attributes and the content to be delivered for the different scenarios.
- analyse mobile user-context to identify when the required context-based event occurs. In this case, the location of the participant.
- create experiment groups and deliver the appropriate content to each group when the required context-based event occurs.
- capture additional navigation behaviour after the intervention is sent.
- capture user reaction to context-based content for the different scenarios.
- generate an experiment report.

2.2.4 Scenario 4

Hai is a mobile games application developer. Like most developers he wants more people downloading his app. In order to increase the click-through rate of the link to his game, Hai is wondering when would be the opportune moment to send it - perhaps while waiting at the bus stand or maybe after checking email at home?

This scenario requires the system to be able to do the following things:

- capture the various mobile user-context attributes.
- create an event mapping experiment (with ease) between the required context attributes and the content to be delivered for the different scenarios.
- analyse mobile user-context to identify when the required context-based event occurs. In this case, the location of the participant, activity as well as mobile interactions.
- deliver the hyper-link when the required context-based event occurs.
- be able to capture the click-through-rate for the different scenarios.
- generate an experiment report.

2.2.5 Additional Scenarios

In addition to the above scenarios, there are other domains where this thesis can facilitate an understanding of user behaviour. One such domain is healthcare. Prior work such as Carrol et al. [21], Hicks et al. [62], Munson et al. [106], Pina et al. [120] and Singh et al. [147] have employed mobile context to identify opportune moments for sending interventions such as appropriate reminders to counteract stress for example.

However, similar to prior work in other domains these applications also work with a limited set of context sources and are designed to send pre-determined responses to events. For example in Carrol et al., the authors rely only on self reporting information to send the appropriate interventions [21]. Such applications can benefit from the additional context sources provided by the platform as well as the flexibility in sending interventions for other events (*automatically*), providing greater insight into the user.

2.3 Building a Real-Time In Situ Context-Based Experimentation System

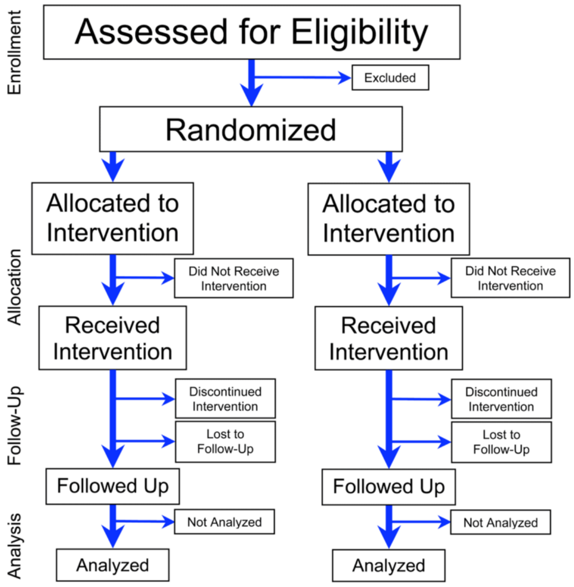
In the section I present some of the key functional capabilities the platform must support in order to run context-based experiments.

2.3.1 Creating an Experiment

An experiment is an orderly procedure carried out with the goal of verifying, refuting, or establishing the validity of a hypothesis. Experiments provide insight into cause-and-effect by demonstrating what outcome occurs when a particular factor is manipulated. Experiments vary greatly in their goal and scale, but always rely on repeatable procedure and logical analysis of the results.

Experiments might be categorized according to a number of dimensions, depending upon professional norms and standards in different fields of study. In some disciplines (e.g., Psychology or Political Science), a 'true experiment' is a method of social research in which there are two kinds of variables. The *independent variable* is manipulated by the experimenter, and the *dependent variable* is measured. The signifying characteristic of a true experiment is that it randomly allocates the subjects in order to neutralize the potential for experimenter bias, and ensures, over a large number of iterations of the experiment, that all confounding factors are controlled for.

A *controlled experiment* often compares the results obtained from experimental samples against control samples, which are practically identical to the experimental sample except for the one aspect whose effect is being tested (the independent variable). A good example would be a drug trial. The sample or group receiving



Flowchart of four phases (enrollment, intervention allocation, follow-up, and data analysis) of a parallel randomized trial of two groups.

Figure 2.1: Randomized Control Trial

the drug would be the experimental group (treatment group); and the one receiving the placebo or regular treatment would be the control one. A controlled experiment where the people being studied are randomly allocated one or other of the different treatments under study is referred to as a *randomized* controlled trial(RCT). RCTs are often used to test the efficacy or effectiveness of various types of interventions and may provide information about its effects. Random assignment of intervention is done after subjects have been assessed for eligibility and recruited, but before the intervention to be studied begins. Figure 2.1 shows a flowchart of four phases (enrollment, intervention allocation, follow-up, and data analysis) of a parallel randomized trial of two groups, modified from the CONSORT (Consolidated Standards of Reporting Trials) 2010 Statement [142]).

The advantages of proper randomization in RCTs include [47]:

- It eliminates bias in treatment assignment, specifically selection bias and confounding.
- It facilitates blinding (masking) of the identity of treatments from investiga-

tors, participants, and assessors.

- It permits the use of probability theory to express the likelihood that any difference in outcome between treatment groups merely indicates chance.

In addition to adhering to the methodological rigor set by RCT protocol, the platform must also handle any complexity in experiment specification. Experiment specification consist of three components - a trigger that specifies a set of contextual predicates that must be matched, an intervention that specifies the resulting experiment treatment, and inputs that specifies the design of the experiment (e.g. sample size). While there are complex ontological models for context specification [154], such specifications would clearly be lost on the non-technical personnel (e.g., behavioural scientists) who might be our primary experimenter base. A key challenge is thus to develop an intuitive, yet sufficiently expressive, GUI that allows experimenters to perform differentiated experimentation.

In this thesis I build *Jarvis*, an experimentation platform that supports running in situ realtime experiments, targeting real participants on their smart phones based on multiple context-specific events. I show that specifying such triggering events using the Jarvis interface is simple for a representative sample set of scenarios and that Jarvis adheres to the traditional process of experimentation. The Jarvis architecture is described in detail in Chapter 3.

2.3.2 Getting the Right Sample

There are various aspects to remember when constructing an experiment. Foremost, the researcher must clearly define the target population. In research, population is a precise group of people or objects that possesses the characteristic that is questioned in a study. To be able to clearly define the target population, the researcher must identify all the specific qualities that are common to all the people or objects in focus. A population can be as simple as all the citizens of Singapore or it can be as specific as all male 17-year old high school students who have been buying comic books since 12 years of age.

Of the many possibilities, a use case I envision for Jarvis, is providing retailers a platform to run *lifestyle* based experiments, examples of which are listed in Section 2.2. Such experiments might entail the platform to provide the experimenter with the right target population. For example, an experimenter might require sending coupons *only* to participants with high coupon usage. Identifying consumer characteristics that indicate such behaviour also then become important. For example, literature on coupon use suggests two consumer characteristics that are strongly associated with coupon use: coupon proneness [87, 151] and price-consciousness [9, 151]. Providing the right sample will ensure the population is free of any sampling bias.

In the case of mobile advertising, an important factor contributing to its success is being able to target audience preferences precisely [86]. Thus an experiment involving sending promotions to participants, factoring in their preferences, will require matching the promotion with participants' needs and preferences to create the best possible sample.

One option to ensure that the right participants are selected, given the experiment promotional content, is to have them indicate their preferences to the system. They could request promotions from a certain company or about a certain kind of product or service. So, only participants that have the deal match with their preferences could be considered as potential subjects, while those that have the deal with lower priority could be ignored. Even if such preferences are not explicitly indicated, it is still possible to guess what they may be interested in. Amazon, for instance makes recommendations of books or music to buy based on your previous buying habits [90].

The need for individual user preferences and their application in decision making in a pervasive environment is now widely accepted. The process of creating, maintaining and applying user preferences in decision making is sometimes referred to as personalization, since it has the effect of tailoring the systems behaviour to the individual needs and wishes of the user so that it appears or acts differently for different users or for the same user under different circumstances. Although most research on personalization has tended to focus on the problem of retrieving Web information (e.g. by adapting user queries to improve the relevance of the answers returned) or on controlling the layout and presentation of output from applications [111], within pervasive systems there is scope for a much wider interpretation of the term.

While identifying a user's deal preferences or intent dynamically is one part of the challenge, matching and prioritizing these preferences with promotional content is also challenging. Promotions typically come in two forms 1) promotions that are offered by the store regardless of payment implement (50% off storewide sale for example), and 2) promotions that are offered if specific payment or discount implements are used (for example, 20% off when using a VISA card, or mileage points if a frequent flyer card is shown, etc.). The key challenge in the matching and ranking of such promotions arises from having to combine both structured (easy to understand numeric discounts (5% off) etc.) and unstructured (free-form text (A free teddy bear) etc.) promotion information with consumer preferences.

How can we combine multiple user contexts (e.g., preferences) with fairly unstructured promotional content? This is a fundamental question that I address in this thesis. My proposed solution adapts natural language semantic techniques and is capable of combining structured and unstructured deal information with user preferences to provide consumers with the most relevant promotions and in turn the right participant sample for the experiment. The algorithm is explained further in Chapter 4.

2.3.3 Capturing Experimental Errors

Many context-aware services make the assumption that the context they use is completely accurate. However, in reality, both sensed and interpreted context is often ambiguous. A challenge facing the development of realistic and deployable context-aware services, therefore, is the ability to handle ambiguous context.

As mentioned previously, of the many possibilities, a use case I envision for Jarvis, is providing retailers a platform to run *lifestyle* based experiments. For example, a coffee shop owner may want to test whether offering discount coupons to people who have been waiting outside the coffee shop for at least 10 minutes, will improve sales. However, a key challenge in running such experiments is that the trigger events are derived from context collected using built-in sensors on the mobile device. These sensors have inherent uncertainties associated with them and as a result can include people who do not satisfy the experiment criteria [6]. Continuing with the previous example, discount coupons could be sent to people who are in fact *not* outside the coffee store but are reported to be by the system as a result of localization error. It is therefore pertinent to arm experimenters with sufficient information of the possible impact of context uncertainty on the outcome of their experiment. For example, informing the experimenter that 2% of the subjects might have falsely satisfied the event conditions will assist them in defining the success criteria of their test. Further, defining a confidence metric for each individual subject, who satisfies the experiment requirements, provides a better understanding of the relationship between the experiment parameters. For example, if subjects considered to have a high context-confidence redeem the discount coupon, we can conclude a strong correlation between the event attributes (standing outside the shop for 10 minutes) and the content delivered. This information is important, not only for understanding user behaviour towards context-based interventions, but also towards building better context-aware systems and applications.

Providing such information unfortunately, is not trivial. The challenges are two fold: 1) Not all context generators provide the necessary information directly. Indoor localization systems such as Radar [11] and EZ [29] for example, do not measure how often the system incorrectly estimates users' location to a given landmark (false positives) and 2) Context uncertainty is highly dynamic and individual. For example, activity classification accuracy is dependent on the activity being classified as well the device being used. It would therefore be incorrect to have a static interpretation of error for a given context source. While techniques of increasing context confidence through redundancy or sensor fusion exist, they do not completely eliminate the need to handle context uncertainty. Given that indoor location is perhaps the single-most important context whilst running such experiments, in my thesis I address the uncertainty-handling capability within Jarvis. In particular, I show how the module defines a confidence metric for the location predicate as well as how it estimates additional information such as the number of false positives. Details of this module is explained further in Chapter 5.

2.3.4 Generating an Experiment Report

Context-aware services provide its users content that is deemed appropriate to the users current environment. People will adopt and use such applications in several settings and for potentially different tasks, so appropriate evaluation techniques must take place in those settings and explore those different tasks. Moreover, many context-aware services make the assumption that the context they use is completely accurate. However, in reality, both sensed and interpreted context is often ambiguous. A challenge facing the development of realistic and deployable context-aware services, therefore, is the ability to handle ambiguous context.

Although some of this uncertainty may be resolved using automatic techniques [23, 58, 108, 126], I argue that the correct handling of uncertain context will often need to involve the user. A de-facto technique from the field of psychology, called the Experience Sampling Method, is found to be effective for learning about person-situation interactions as well as any ambiguity in the situation. The technique compares most closely with recall-based self-reporting techniques such as interviews, traditional surveys, and diaries

The term *mediation* is commonly used to refer to the dialogue that ensues between the user and the system. Mediation techniques are interface elements that help the user to identify and fix system actions that are incorrect, or potentially involve the user in helping the system to avoid making those mistakes in the first place. Further, in addition to resolving any context ambiguity, user mediation can also provide researchers with the possibility to ask usage behaviour specific questions [15, 22, 53, 65]. For example, a user indicating that a location based coupon was sent based on incorrect location can provide feedback to the localization system as well as indicate why the coupon was not used. Thus, explicitly asking users whether the triggering context matched the context in reality will leave no room for ambiguity - which is not the case when using automated techniques.

The need for this explicit interaction however, can often burden the user. As a result, current research effort is towards finding the right balance of user effort in providing feedback and the information needed by the system [14, 88, 121]. One such information source, that can potentially reduce user fatigue, is implicit feedback in the form of user interactions with a device. I claim that users interact with their device differently based on the degree of affinity towards a notification content. For example, users that like a coupon may interact with their device differently (*they might view the coupon for longer durations, scroll the content several times*) as opposed to those that do not like it (*delete the notification within a short span of receiving it*).

In my thesis I concentrate solely on identifying and collecting the multiple dimensions of user mobile interaction as a potential source of user reaction to a context-based experiment. I am not tackling the problem of resolving context ambiguity through user mediation. Interpreting user reaction is however, a step towards this goal. User reaction, coupled with the confidence of the experiment context predicates, can be used to generate the experiment report necessary to draw a conclusion as well as create an optimum dialogue required for explicit user mediation which in turn reduces any response bias. Details of this module is explained further in Section 3.5.

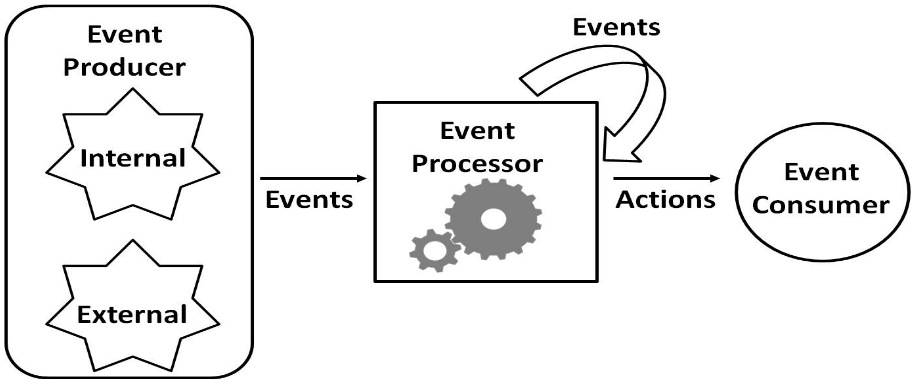


Figure 2.2: Basic concept of an event processing system

2.4 Event-driven Architecture

Event-driven architecture(EDA) [46, 153] is an architectural style in which one or more components in a software system execute in response to receiving one or more event notifications. This architectural pattern may be applied to the design and implementation of applications and systems which transmit events among loosely coupled software components and services [103]. An event-driven system typically consists of event producers (or agents), event consumers (or sinks), and event channels. Producers have the responsibility to detect, gather, and transfer events. Consumers have the responsibility of applying a reaction as soon as an event is presented. Event channels are conduits in which events are transmitted from event producers to event consumers. Figure 2.2 shows the major architectural components of event processing, showing the logical separation of event processing logic from the event producers and event consumers.

As in situ experimentation is naturally centered around events, with the need to identify and react to certain situations as they occur, following an event processing approach is appropriate. An event-driven approach, where changes in state are monitored as they happen, lets an application respond in a much more timely fashion than a batch approach where the detection process runs only intermittently. Further, there are potential scalability and fault tolerance benefits to be gained by using an event-driven approach. An event-driven approach allows processing to be performed asynchronously, and so is well suited to applications where events happen in an irregular manner.

2.5 Summary

Mobile sensing technology has the potential to bring a new perspective to the design of social experiments, both in terms of accuracy of the results of the study and from a practical point of view. However, to facilitate this novel experimentation process there are several requirements and challenges. In this chapter I describe what it means to run a real-time in situ context based experiment. I then list some of the example experiments that this thesis will support. I then present some of the key building blocks necessary for such a system: Foremost the system must ensure the validity of the experiment. This means selecting the right sample, handling the multiple sources of bias, capturing and representing any experimental errors and finally provide sufficient information to the experimenter to draw a conclusion. Additionally, as the system is designed to be used by non-technical experimenters (e.g., sales managers, marketing executives etc.) there must be an intuitive, yet sufficiently powerful way for these nontechnical experimenters to easily specify the kinds of experiment designs that they are interested in.

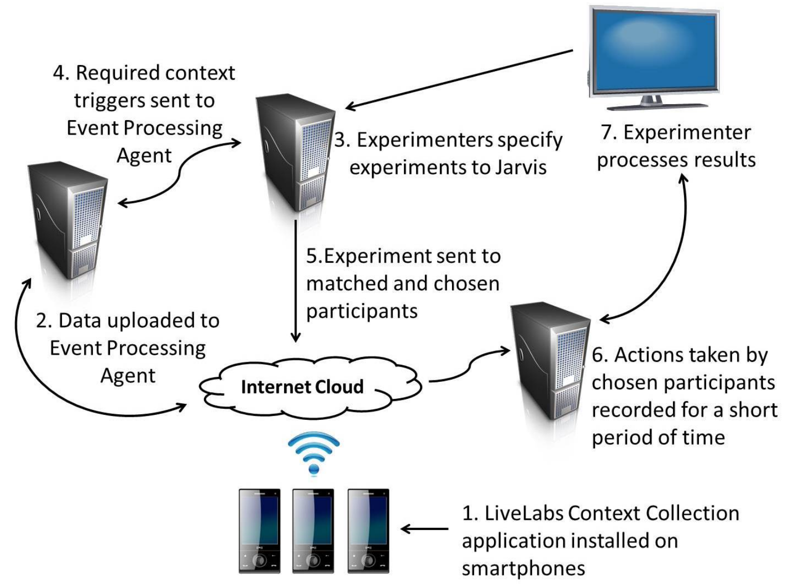
Chapter 3

Real Time Context-Based Experimentation

In this chapter, I describe the goals and the implementation of Jarvis, a platform that supports running in situ context-based experiments. I first describe the life-cycle of a context-based experiment and present the architecture of Jarvis. I then describe the various modules of the system that conforms to the traditional process of controlled experimentation, while abstracting the complexities of handling the multiple types of bias. Through a series of *live* experiments I showcase the diversity of the system in supporting multiple experiment designs, the ease of experiment specification, and the rich behavioural information accessible to the experimenter in the form of a report. Note, Jarvis is part of the larger LiveLabs [13] ecosystem and requires integration with other production grade systems built by my colleagues. Where appropriate, I make clear which is my work and which is work that I am integrating as part of the system.

3.1 The Experiment Life-Cycle

Figure 3.1 shows the sequence of steps necessary to run such above experiments. The sequence is:



The figure is based on the original "How LiveLabs Works" figure from Rajesh et al. [13]

Figure 3.1: An Experiment life-cycle.

1. A context collector application needs to be installed on participant smartphones. We currently support iOS 6+, and Android 3+ smartphones.

2. The collector application collects sensor and context data from the phone and sends it to our real-time Event Processing Agent (EPA) where it is processed to obtain the required context triggers such as location, current activity, group status etc.

3. Experimenters specify their experiments using Jarvis, our Behavioural Experimentation Platform (BEP). Section 3.2 provides more details about the BEP.

4. Jarvis registers the required context triggers with the EPA. For example, "inform me when you find people standing outside the coffee shop for the past 10 minutes". The EPA server will keep track of all these events and call back the BEP when the triggers match the current context.

5. When a callback is received with the list of matched participants, Jarvis will pick a subset and send a notification with the experiment details to each se-

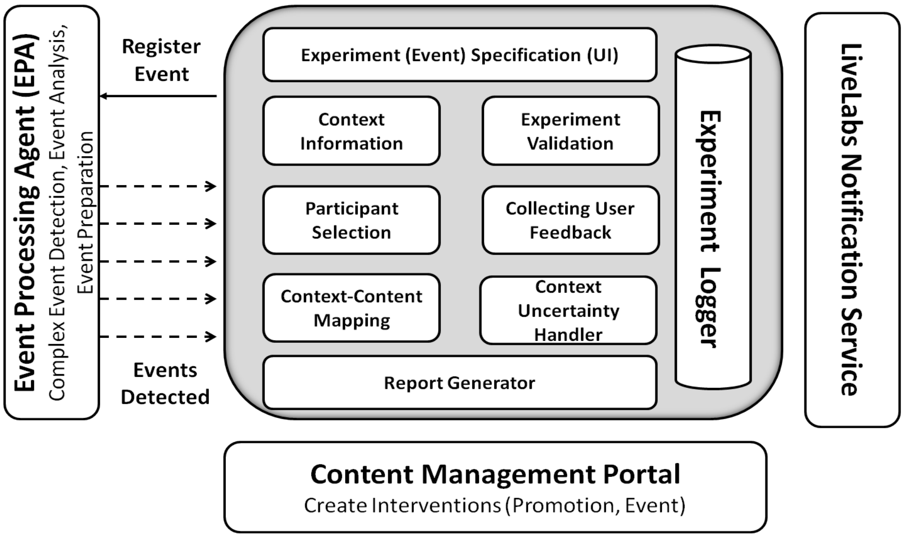


Figure 3.2: System Architecture

lected participant. An experiment could be a discount, a request to run an application, a survey etc. Note that this subset can include participants that in reality have *not* matched all context predicates specified in the experiment but are deemed to have, as a result of error in the context collected. Chapter 5 talks more about how I capture and handle this context uncertainty.

6. Jarvis will monitor the selected participants for a set period of time and record what they did in response to the experiment stimulus. This data is then packaged, in a privacy preserving way, and reported to the experimenter. The ability to observe the entire experimental effect (both positive and negative) is a key unique property and selling point.

7. The experimenter processes the results and determines how to change their experiment (if required).

3.2 Jarvis Architecture

Figure 3.2 shows the various modules of Jarvis needed to support context-based experiments. A web UI allows experimenters to specify a wide variety of predicates consisting of context variables (location, time, gender etc.,) that needs to be matched by the participants. The *Context Information* module provides historical data (if available and applicable) of that context attribute. For example, if the experimenter chooses Starbucks as a target location the module displays the average population density, using a heat map, observed at that location. This information allows the experimenter to make a better judgement of the selected context attribute. Once the experiment is defined the *Experiment Validation* module ensures that the experiment is safe and valid for the participant pool. Events, such as "deliver a specific targeted discount to people at the mall who are moving around in groups of 2", are captured using an SQL-based syntax. The predicates are defined as a set of logical rules, with the context variables chained together using explicit AND operators similar to the *Where* clause of SQL (with each context variable appearing only once). This query is generated and processed by the *Query Generator and Optimizer* module that bears many similarities to that of the Amit event processing tool [4]. When the experiment conditions are satisfied by a participant(s), the *Context Uncertainty Handler* computes the corresponding confidence metric for each of the event attributes. We use an uncertainty model based on a predicate representation of contexts and associated confidence values [125]. In Chapter 5 we will see how this confidence metric is computed for the location attribute. The *Participant Selection* module picks a subset of participants based on a confidence threshold specified by the experimenter in addition to other rules. A further filtering of participants are done through the *Context-Content Mapping* module to ensure that the content to be delivered to participants as part of the experiment (e.g discount coupon) reach the right set of people. Finally, once the participant's response to the experiment has been collected via the *User Feedback* module, the *Report Generator* provides a summary of the experiment that includes an overview of the impact of the different event attributes on the experiment outcome. The complete database schema supporting this architecture is captured in Appendix A.

Jarvis also relies on external systems (built by my colleagues) for its functioning. Primarily these are:

- *LiveLabs Context Collector*: A mobile sensing software, installed on participants personal devices, that collects detailed real-time (or near real-time) context data. The app also provides a channel for sending notifications to participants.

- *SMUddy/EVA*: SMUddy and EVA are in-house built mobile applications catering to our university students. These applications provide students with a host of services such as viewing and subscribing to events happening on campus as well view study room occupancies and heat maps. SMUddy and EVA are also channels for experimenters to send context-based interventions (e.g., promotions, reminder to events) as part of an experiment.

  Note: There are iOS and Android versions of the above three mobile apps. Details of the mobile context and interactions captured by the LiveLabs Context Collector, for each version, is listed in Appendix B.

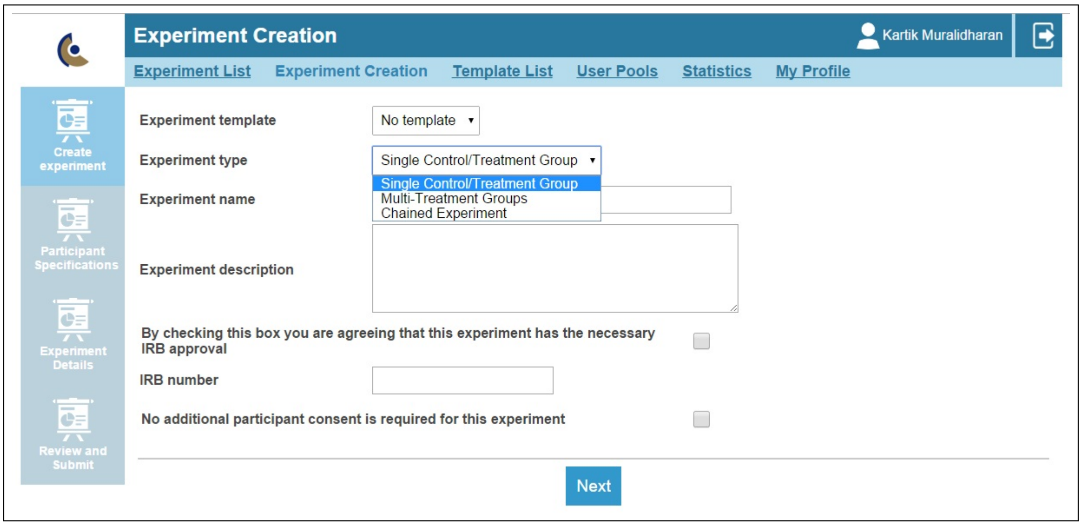
- *Event Processing Agent (EPA)*:Applies advanced analytics on incoming raw data streams to infer a variety of interesting individual and collective context. Jarvis registers the required context triggers with the EPA. The EPA server will keep track of all these events and call back Jarvis when the triggers match the current context.

- *LiveLabs Indoor Localization Service*: Computes the current location of all LiveLabs participants. The EPA uses this service to identify participants satisfying the location context criteria of an experiment.

- *LiveLabs Notification Service*: Provides an API that handles the sending of experiment interventions to the participants' mobile devices.

- *Content Management Portal*: An interface for creating the experiment intervention (Promotion, Event).

Acknowledgement: Given the deployment nature of *LiveLabs* - which includes



Experimenters are required to obtain the necessary IRB approval prior to creating an experiment.

Figure 3.3: Step 1: Experiment Specification

building and operating a large-scale mobile services experimentation testbed, involving members of the general public at multiple public spaces - several personnel were involved in the design of the look and feel of Jarvis. Besides the members of my thesis committee, Swetha Gottipati, the project manager overseeing the development of the various LiveLabs components, assisted me in the design of the database schema that supports running in-situ experiments as well as interfacing Jarvis with the other LiveLabs' systems.

3.3 The Experimental Process

While experiment expressiveness is one aspect of the system design, the ability to input the experiment into the system without any loss of experiment semantics is equally important. Given that the system will eventually be used by non-technical personnel (e.g., psychologists), a key challenge is to develop an intuitive, yet sufficiently expressive, user interface that allow experimenters to perform differentiated experimentation.

Creating an experiment is a four step process done via a web portal[1]:

[1]The web interface was built by my colleague NGUYEN Vu Nhat Minh

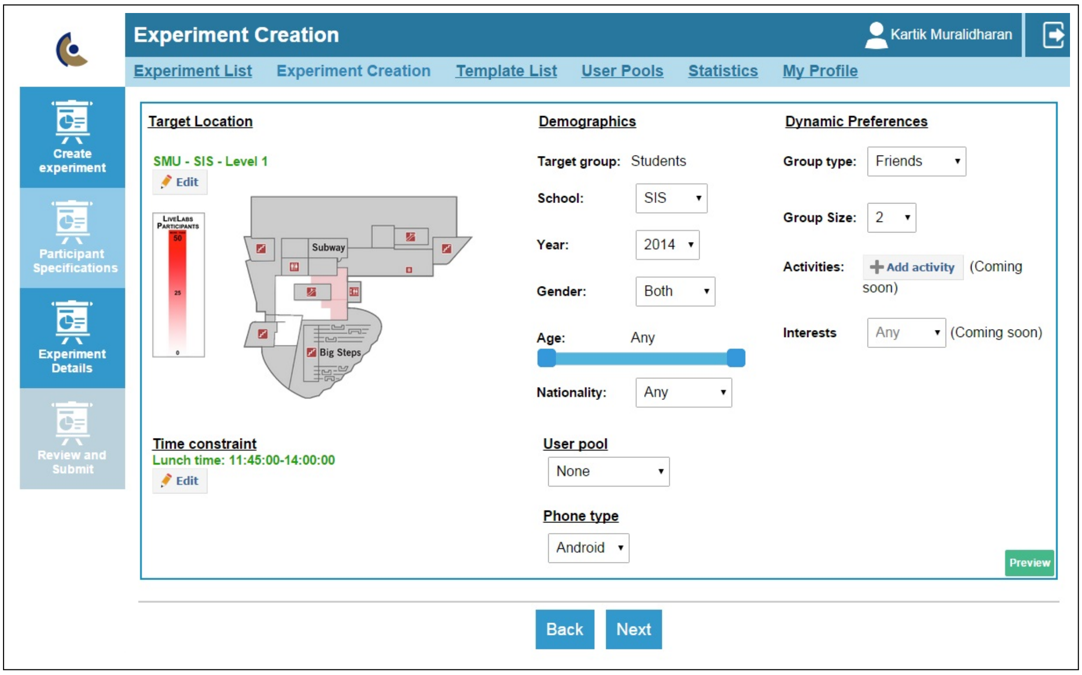


Figure 3.4: Step 2: Participant Specification

Step 1 Experiment Specification The first step in creating an experiment requires the experimenter to select the type of experiment to run as well as key in additional information such as the experiment name and description (Figure 3.3). Currently we support running 3 types of experiments - Single Control/ Treatment experiment, Multi-treatment experiments and Chained experiments. The next section provide more details on these experiment types.

Step 2 Participant Specification

A key feature of the platform is the ability to target participants based on their current context. Figure 3.4 shows the set of static and dynamic context criteria available to the experimenter. The location context is currently restricted to our University campus and be selected up to a room level granularity. The experimenter can either specify a custom time range, or can select from a set of pre-set time ranges. Participants can be sent an intervention *only* within this time range. Besides specifying location and time as conditions, the experimenter can also filter participants on basic demographics such as gender, age and nationality. Future context predicates include targeting groups of a cer-

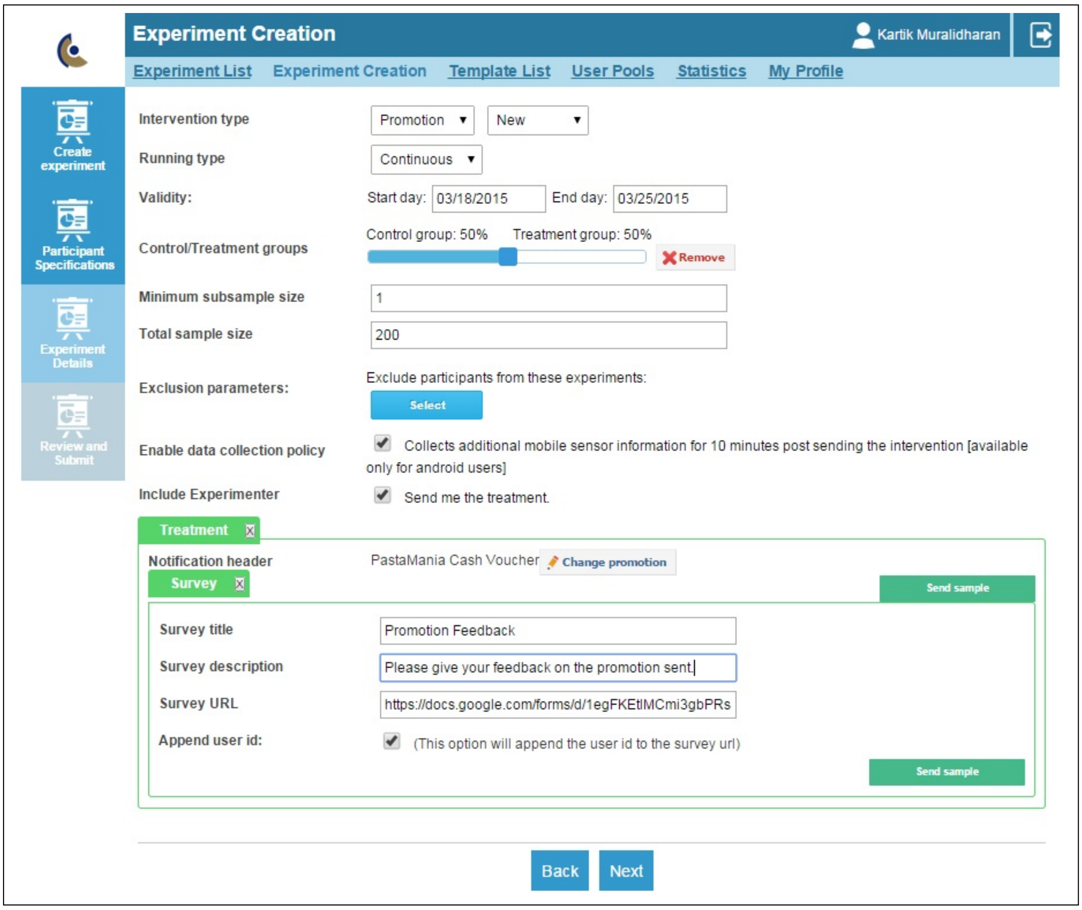


Figure 3.5: Step 3: Experiment Details

tain size as well as activity based triggering. For example, the experimenter could send a promotion to groups of four, sitting in the coffee shop.

Clicking on the preview button provides experimenters a glimpse of how many participants satisfied the context constraints in the past week (Figure 3.7). The goal of this feature is to assist in estimating the reasonable sample size for that experiment.

User Pool: Experimenters often have an enrollment process to recruit participants. For these cases, they would prefer to run their experiments on this hand picked participant pool, rather than have Jarvis select random participants from the LiveLabs subject pool. To facilitate this Jarvis allows experimenters to create their own custom user pool. Figure 3.8 shows the list of participants in a custom user pool titled ‘Test Account’. Experimenters can

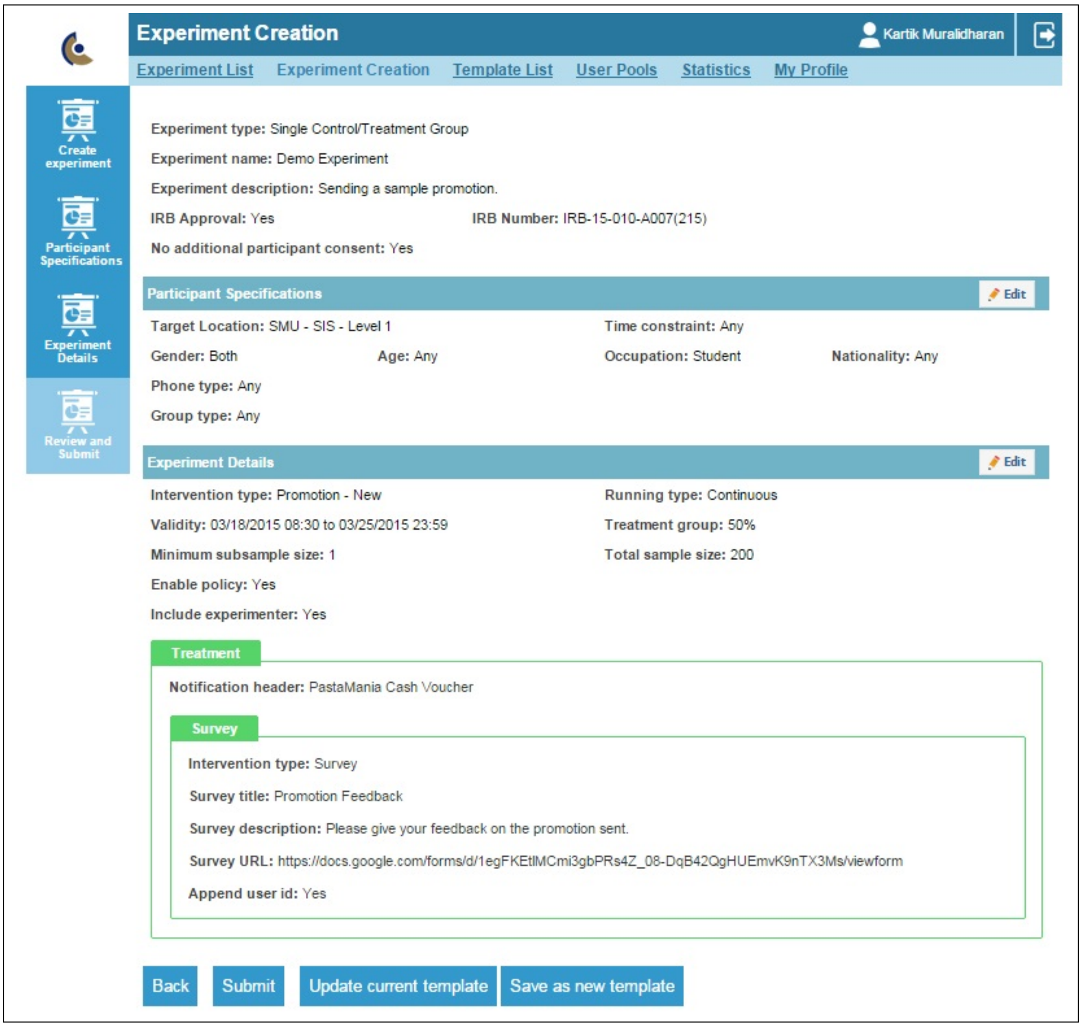
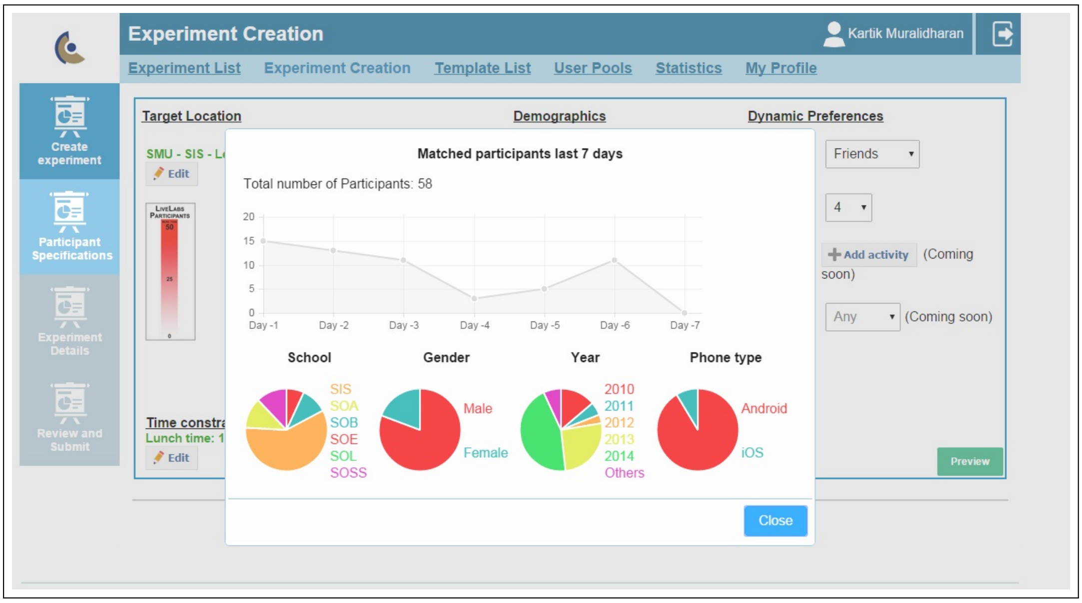


Figure 3.6: Step 4: Review and Submit

add users through their email id and Jarvis will indicate which LiveLabs' applications these users have currently installed on their device. This will allow experimenters to contact the respective participants to install the required app. Once a user pool is created it can be selected in Step 2 (Figure 3.4). If a user pool is selected, experiment subscription will also include the participant ids of the user pool, in which case the EPA will return to Jarvis only those participants within the pool that satisfy the context constraints.

Step 3 Experiment Details This step captures additional details such as the duration of the experiment (how long the experiment should run) as well as the target sample size. Experimenters can also specify if they want to partition the



The total number of participants shown the figure is cumulative over the past one week.

Figure 3.7: Matching Participants

participant sample into control and treatment groups as well indicate the proportion of this breakup (Figure 3.5).

This step also captures the intervention that needs to be sent to participants satisfying context constraints. Details of the supported intervention types is listed in Section 3.3.2.

Step 4 Review and Submit In the final step the experimenter can review the experiment specification and submit it for approval.

Subscribing for an Experiment: Once an experiment is submitted (and approved), it is registered with the EPA using JSON. The JSON request shown in Figure 3.9 captures the context constraints for that experiment. The request also includes additional POST parameters specifying the time constraints of the experiment. All participants matching the context criteria within the time range are sent to Jarvis for further processing. When an experiment has completed it is unregistered with the EPA.

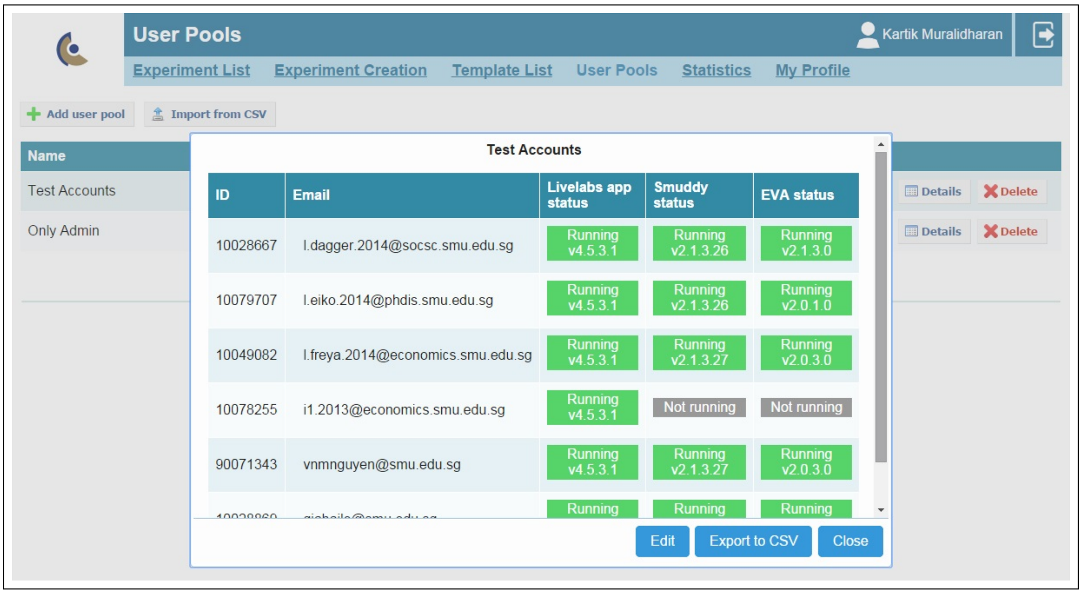


Figure 3.8: Creating a custom participant pool

3.3.1 Experiment Design

A key requirement of the BEP is to support a variety of hypotheses testing. I am currently supporting the running of 3 types controlled experiment designs:

1. Single Control/ Treatment experiment refers to the traditional A/B testing - a randomized experiment with two variants, A and B, which are the control and treatment in the controlled experiment.

2. Multi-Treatment experiment is an extension of the two-sample hypothesis testing to support the inclusion of multiple samples. This type of controlled experiment was designed for testing marketing strategies where all groups are essentially treatment groups and the presence of a control group (where no treatment is given) is optional. Currently the system supports up to 8 treatment groups where each treatment can be any one of the intervention types specified in Section 3.3.2 thereby supporting multivariate testing. Experimenters are given the option of including additional treatment groups and selecting the corresponding intervention type at Step 3 of experiment creation. Figure 3.10 shows a multi-treatment experiment with three treatment groups.

```json
{
    "location": [
        {
            "location_type": "location",
            "location": 1,
            "pid": 177,
            "filter": [
                {
                    "id": "Any",
                    "time_constraint": "Any",
                    "group_type": "Any",
                    "occupation": "Student",
                    "school": "Any",
                    "maxage": "60",
                    "phone_type": "Any",
                    "interests": "{}",
                    "gender": "Both",
                    "year": "0",
                    "group_size": "0",
                    "minage": "18"
                }
            ]
        }
    ]
}
```


Figure 3.9: Experiment Subscription with the EPA.

3. Chained Experiment supports running *multiple experiments* over a selected participant sample. The participant sample is created on running the first experiment in the chain. Subsequent experiments in the chain target participants from the same pool (control and treatment partition is maintained). The participants are triggered with the *same* context criteria for all experiments within the chain. However, the intervention type as well as the duration of each experiment within the chain can be different. Figure 3.11 shows a chained-experiment containing three experiments of different intervention types.

The decision to support these three designs were taken after informal discussions with faculty from the School of Business and School of Social Sciences, our primary users of the experimentation system. Note that in supporting these experiment designs we inherently restrict the complexity of the context predicates that can be expressed by the system. The need to support multiple experiment groups relegates the predicate to simple SQL statements with the context variables chained together using *only* AND operators, with each context variable appearing only once. Using logical rules such as the OR operator between context variables can create confounding experiment groups. To support such operations we would need to create an additional design that targets *only* a single group. Events, such as "deliver

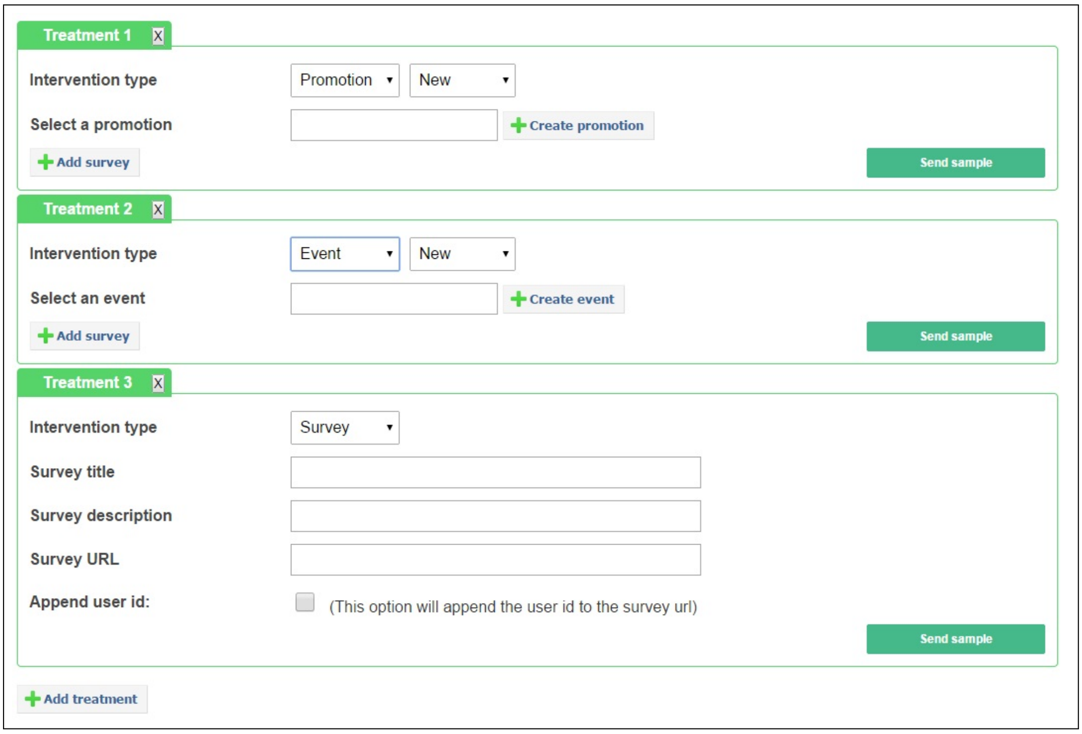


Figure 3.10: A Multi-Treatment Experiment with 3 treatment groups with different intervention types.

a survey to people at the mall who are moving around in groups of 2 *or* standing" could then be supported. However, this will require a more intuitive interface that allows the experimenter to create such a mapping with minimal cognitive load.

3.3.2 Intervention Type

On matching the triggering context criteria, participants are sent an intervention. The system supports 7 intervention types, each linked to one of the three mobile applications, LiveLabs Context Collector, SMUddy and Eva, described at the start of this chapter. Table 3.1 lists the intervention types as well as the target mobile application wherein the intervention will be visible. Interventions of type *Promotion-New* and *Event-New* can also have a survey associated with the intervention to collect additional self-reporting data. The survey is however sent *only* when the participant views the promotion or event. Reminder interventions can only be sent as a followup to an Promotion-New or Event-New type experiment. Further these reminders are sent *only* to those participants that have received the promotion or event

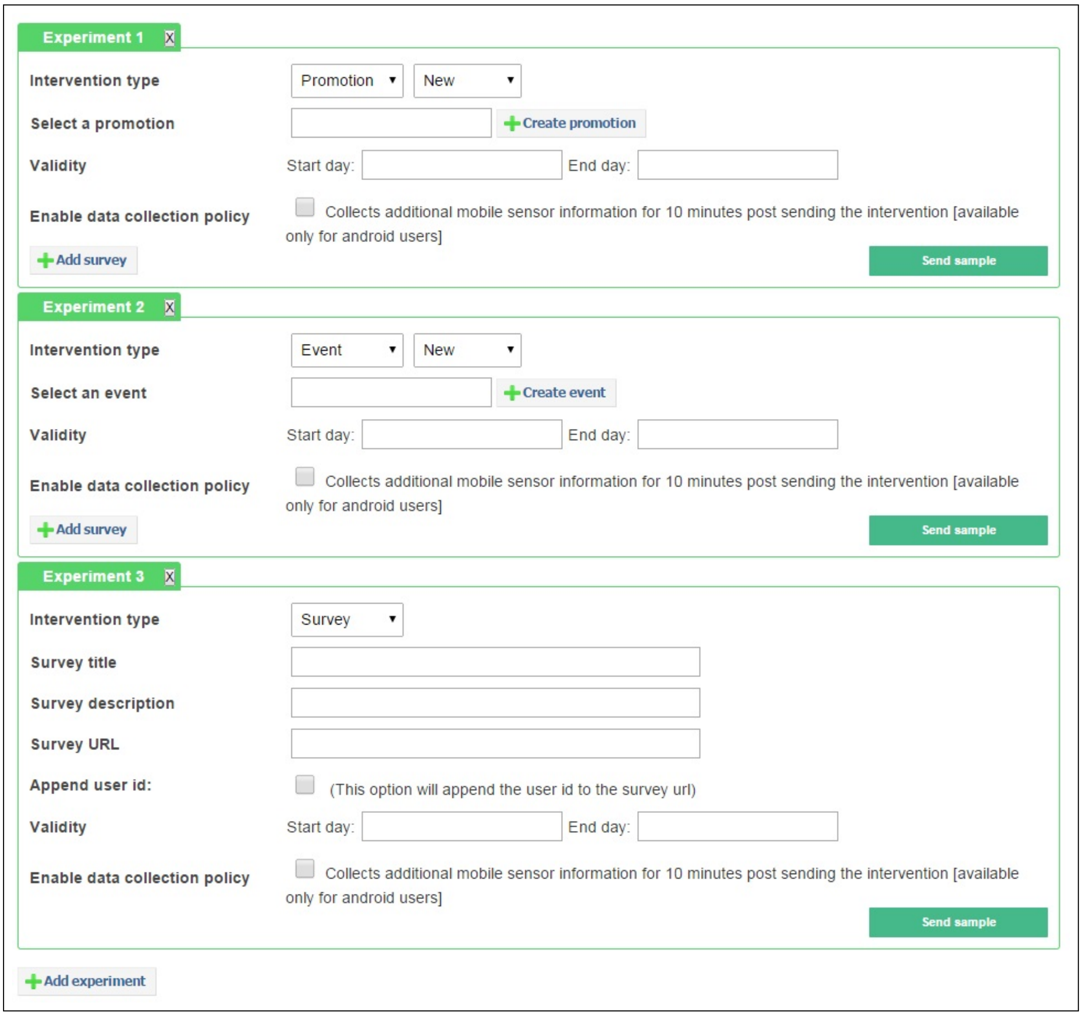


Figure 3.11: A Chained Experiment containing 3 experiments with different intervention types.

respectively. Note, only links to an external survey (hosted by Qualtrics or Google Docs for example) can be associated with the intervention.

For every intervention type the experimenter can also specify the contents of the notification that participants will first see on the notification bar of their mobile device. Clicking the notification opens the corresponding mobile application and displays the intervention. Participants can click on the intervention to further view the details. Figure 3.12 shows the interface for creating a new promotion. Figure 3.13a captures a screen shot of the promotion notification on a mobile device, Figure 3.13b and Figure 3.13c show the promotion tab on SMUddy displaying the intervention and details of the promotion respectively. Figure 3.13d and Fig-

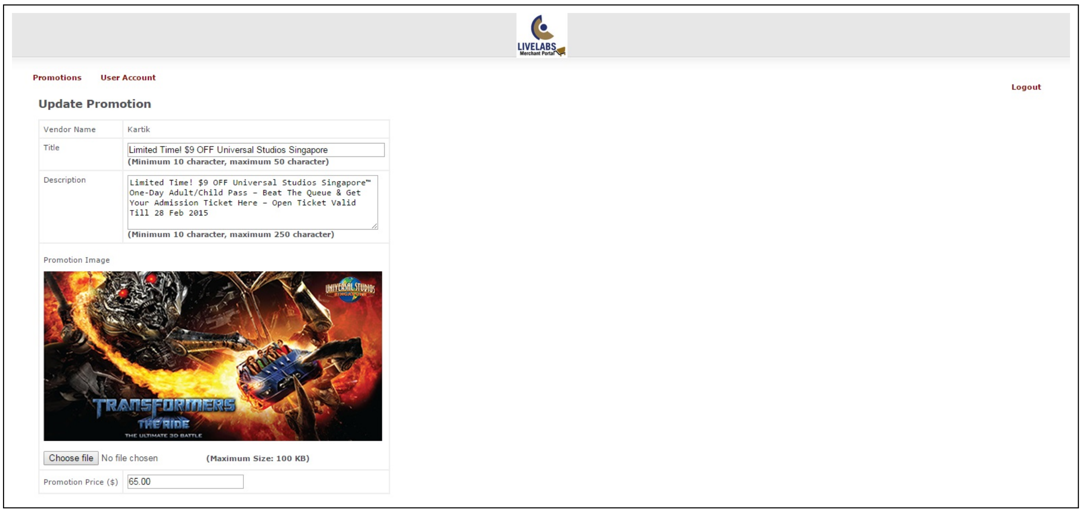


Figure 3.12: Creating a Promotion Intervention

ure 3.13e show the notification requesting feedback for the promotion as well as details of the survey.

Notifications are sent using the LiveLabs notification service that provides programming abstractions for sending messages to both Apple devices via the Apple Push Notification Service (APNS) and Android devices via the Google Cloud Messaging (GCM) service.

3.3.3 Administrative and Housekeeping Options

Once an experiment is submitted it needs to be approved. This due diligence process is to verify that the experiment has been approved by the IRB and does not send any spam content to participants. Administrators are notified when an experiment has been submitted and can approve (or reject) the experiment through the online interface. When an approved experiment runs, administrators also have the option of receiving the intervention sent to participants. This is to ensure there has been no change in content since the approval process.

The web portal also provides several housekeeping options for managing experiments (Figure 3.14). Experimenters can view the list of experiments created by them as well as their details. They can delete an experiment before it has started or stop an experiment once it is running. Frequently used experiment specifications can be

Figure 3.13: The different stages of a Promotion Intervention

saved as a template to reduce turn around time in experiment creation. The portal also indicates the current status of the experiment, whether it is pending approval, if it is running and has completed. In case the experiment is running, experimenters can also view the current number of selected participants for that experiment (Figure 3.15). Log files capture every aspect of the experimentation process for any error handling.

3.3.4 Implementation Details

Jarvis was implemented in about 7000 lines of Java. The web interface is written in PHP and javascript[2]. The database consists of 24 tables created using the postgresql DBMS. Details of the sequence of database access can be found in Appendix A.

[2]The web interface was built by my colleague NGUYEN Vu Nhat Minh

Intervention Type	Description	Target Mobile App	Options
Promotion-New	Send a Promotion	SMUddy	Optional survey to collect self reporting data.
Promotion Reminder	Send a reminder message for a promotion sent as part of an earlier experiment.	SMUddy	
Event-New	Send details of an event	EVA	Optional survey to collect self reporting data.
Event Reminder	Send a reminder message for an event sent as part of an earlier experiment.	EVA	
Survey	Send a link to an external survey.	LiveLabs Context Collector	
Link	Send a link to external content.	LiveLabs Context Collector	
General	Send a message as an intervention.	LiveLabs Context Collector	

Table 3.1: Intervention Types.

3.4 Participant Enrollment

Research on mobile advertising [138] stresses the danger of irrelevance and inappropriateness of messages towards a consumer's willingness to receive mobile advertisements. Consumers are particularly wary on issues such as who sends them mobile coupons, how many they will receive, and when they will receive them [86]. Selecting the wrong participant, for a mobile coupon, can in turn have a negative impact on coupon adoption as well as perceived usefulness and subsequently on the internal validity of an experiment. For example, sending a coupon to a participant, who in the last hour had already received five, may likely consider the next coupon sent as spam. It is therefore imperative that we reduce the impact of selection bias - an error in choosing the individuals or groups to take part in a scientific study.

One such technique of reducing selection bias is using the RFM (recency, fre-

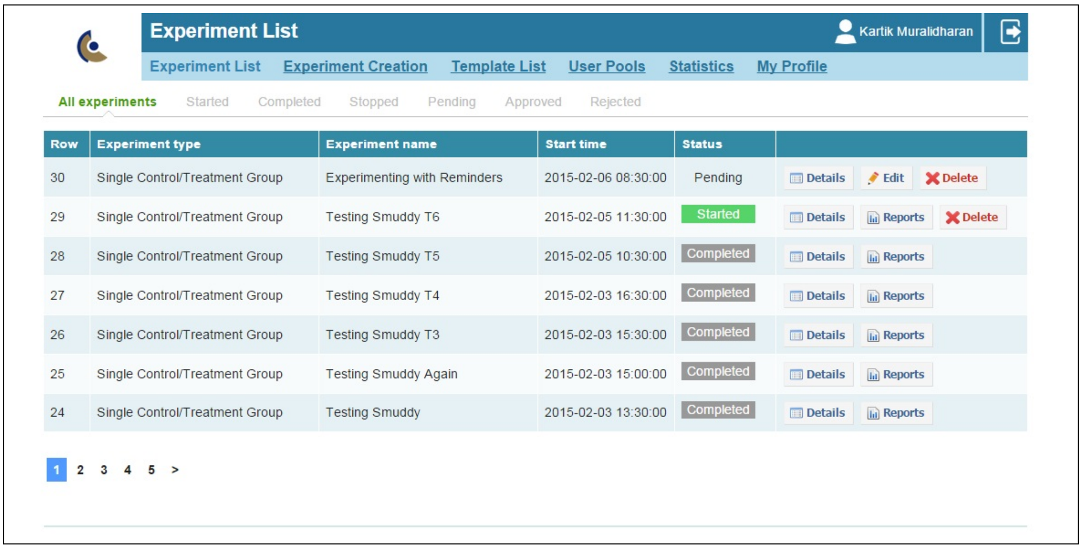


Figure 3.14: List of Experiments created, their details and their current status.

quency, monetary) model [20]. RFM analysis, is a marketing technique typically used to determine quantitatively which customers are the best ones by examining how recently a customer has purchased (recency), how often they purchase (frequency), and how much the customer spends (monetary). Alternatively, the same model can be used to reduce selection bias by ensuring a balance of participants with high and low RFM ranking. While guaranteeing optimum participant selection is beyond the scope of this work, I include additional experiment parameters that can be used for participant selection. I create a variant of the RFM model replacing the *monetary* term with *similarity*. The new RFS model quantitatively determines which participants are appropriate for an experiment by examining how recently was the participant part of an experiment (recency), how often has the participant been part of an experiment (frequency) and does the participant have any bias as the result of being part of a similar prior experiment (similarity).

Participants satisfying the context constraints are sent to Jarvis by the EPA. It is at this point participants who are to be sent the intervention are selected. Participants are then selected by Jarvis in 4 stages:

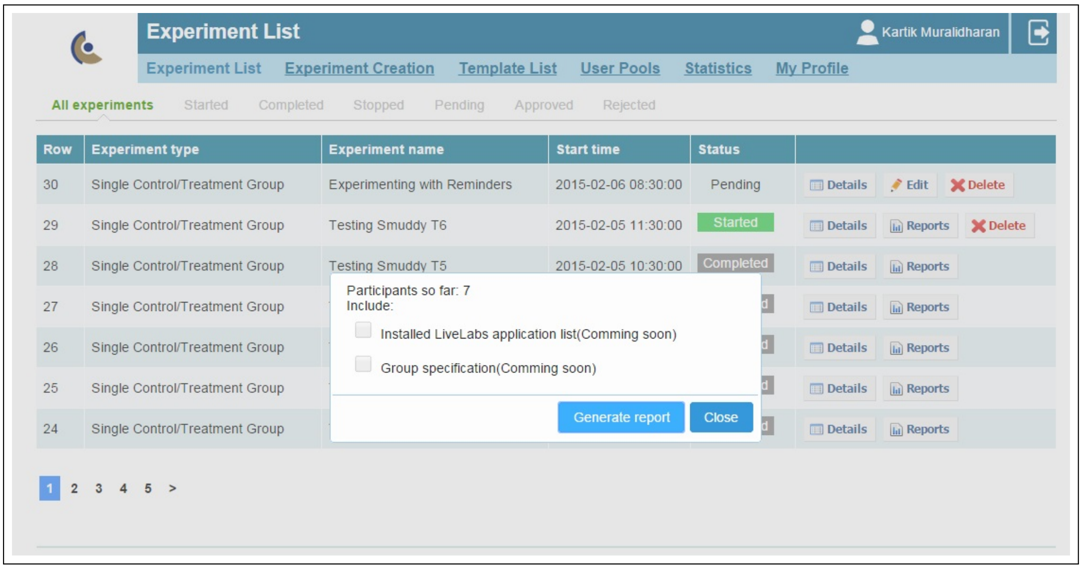


Figure 3.15: Number of participants selected so far.

3.4.1 Selection based on Experiment Details

From the EPA participant pool, the number of participants required for the experiment, or remaining, are randomly selected. Note while subsequent stages will further reduce this number, doing this step at this point eliminates the need to perform additional processing on all participants from the EPA pool.

Next, participants who do not have the target application (SMUddy, Eva, LiveLabs) currently installed on their mobile device or have not uploaded any data to the server in the past week are removed. Note, for multi-treatment experiments, this step can only be performed after participants have been assigned to treatment groups since the target application for each treatment group can be different. The *RFS* model is then applied on these participants in the next stage.

3.4.2 Participant Fatigue

Respondent fatigue is a well-documented phenomenon that occurs with survey participants when they become tired of the survey task and the quality of the data they provide begins to deteriorate. It occurs when survey participants' attention and motivation drop toward later sections of a questionnaire. Tired or bored respondents may more often answer "don't know," engage in "straight-line" responding (i.e. choosing answers down the same column on a page), give more perfunctory answers, or give up answering the questionnaire altogether [82]. Thus, the causes for, and consequences of, respondent fatigue, and possible ways of measuring and controlling for it, should be taken into account when deciding on the length of the questionnaire, question ordering, survey design, and interviewer training.

Therefore, to counter respondent fatigue, the total number of experiments each participant from the previous step was selected for, in the past week, is computed. Participants who were selected for an experiment within the day as well as those who have been selected for more than 6 experiments in the past week are removed from the pool of eligible participants. The time range and the total number of experiments that a participant has been selected for (within the specified time range) are modifiable options in a config file. The optimum values for these numbers is left for future work.

3.4.3 Additional Exclusion Parameters

Another source of selection bias is recruiting participants that have been part of a similar prior experiment. It is possible that these participants may not react favourably to interventions of similar nature or content. This is particularly true of experiments that are run in phases. An experiment in its second phase, should not select participants that were part of phase one of the experiment. Jarvis allow experimenters to select previous experiments whose participants must not be included in the current run (Figure 3.5). Note, since the content type of the intervention is not captured, the system cannot currently automate the exclusion of participants from similar prior experiments. For example, if the intervention type was specified as 'sports' related, participants who have recently received sports related interventions can be removed from the participant pool.

3.4.4 Creating Control/Treatment Groups

Randomised controlled trials are the most rigorous way of determining whether a cause-effect relation exists between treatment and outcome and for assessing the cost effectiveness of a treatment. They have several important features:

- Random allocation to intervention groups.
- Participants and experimenters should remain unaware of which treatment was given until the study is completed.
- All intervention groups are treated identically except for the experimental treatment.
- Participants are normally analysed within the group to which they were allocated, irrespective of whether they experienced the intended intervention.
- The analysis is focused on estimating the size of the difference in predefined outcomes between intervention groups.

Other study designs, including non-randomised controlled trials, can detect associations between an intervention and an outcome. But they cannot rule out the possibility that the association was caused by a third factor linked to both intervention and outcome. Random allocation ensures no systematic differences between intervention groups in factors, known and unknown, that may affect outcome. Double blinding ensures that the preconceived views of participants and experimenters cannot systematically bias the assessment of outcomes. Intention to treat analysis maintains the advantages of random allocation, which may be lost if subjects are excluded from analysis through, for example, withdrawal or failure to comply. Meta-analysis of controlled trials shows that failure to conceal random allocation and the absence of double blinding yield exaggerated estimates of treatment effects [145].

Once the participant pool is devoid of any selection bias Jarvis randomly assigns participants into an experimental group or a control group to achieve a randomized

```
Algorithm Control_Treatment
1. Input set of selected participants.
2. Input the treatment percent.
3. Control percent = 100 – treatment percent.
4. Get the current size of the control and treatment group for the experiment so far.
5. Select a new random seed. Shuffle set of selected participants; //To ensure participants are not
   always assigned to the same group

6. For each participant in the set of selected participants
      a. Compute the current control/treatment participant ratio

      current_control_ratio = control_group.size / (control_group.size + treatment_group.size)

      current_treatment_ratio = treatment_group.size / (control_group.size +treatment_group.size)

      b. If treatment percent > control percent
            i. If (treatment percent > current_treatment_ratio)
                  Add participant to treatment group
               else
                  Add participant to control group
      c. Else If control percent > treatment percent
            i. If (control percent > current_control_ratio)
                  Add participant to control group
               else
                  Add participant to treatment group

7. Return Control-Treatment group
```

Figure 3.16: Pseudo code for partitioning selected participants into Control and Treatment Groups for Experiments of type Single/ Chain

control trial[3]. The assignment of participants into groups is done differently depending on the type of experiment:

Experiment Type: Single, Chain

For experiments of type 'single' and 'chain', experimenters can specify the ratio in which participants are to be divided between control and treatment groups. Therefore, as participants stream into Jarvis, participants are randomly assigned to groups whilst maintaining this ratio. Figure 3.16 shows the pseudo code that performs this partitioning.

For experiments of type chain, initial group partitions are maintained for subsequent experiments in the chain i.e. participants who were assigned to the treatment group for the first experiment in the chain. will continue to remain in the treatment

[3] Although participants are randomly assigned to an experimental group or a control group to reduce sampling bias, our current sample is still a convenience sample (given that our participants are only students at SMU)

```
Algorithm Control_Treatment (For Multi)
1. Input set of selected participants.
2. Input number of treatment groups.
3. Retrieve current treatment groups for that experiment.
4. Determine the current minimun treatment group size
     a. Iterate through the treatment groups
          i. Initialize min= size of first treatment group.
          ii. if (treatment group size <= min)
                  min=treatment group size
              else
                  Next treatment group

5. Select a new random seed. Shuffle set of selected participants; //To ensure participants are not
   always assigned to the same group.
6. Insert selected participants in a round robin fashion starting with the group with minimum size.
7. Return Treatment groups.
```

Figure 3.17: Pseudo code for partitioning selected participants into Treatment Groups for Experiments of type Multi

group for subsequent experiments in the 'chained' experiment.

Experiment Type: Multi

For experiments of type 'multi', experimenters can choose to create up to 8 treatment groups in addition to one control group. Here. Jarvis ensures all groups are of equal size. To do this, participants are assigned to groups in a round robin fashion. Figure 3.17 shows the pseudo code that performs the group assignment for multi-treatment experiments.

3.5 Implicit Participant Feedback

The true power of the experimentation system is in running *in situ* experiments collecting both qualitative and quantitative data. The rich and varied data that can be obtained in situ provides different insights into peoples perceptions and their experiences of using, interacting and reacting to context-based services. Such data can also provide answers to the *why*, such as user motivation, perception, and satisfaction.

However, collecting carefully considered feedback is difficult owing to a number of constraints ranging from user fatigue (a rule of thumb is to ensure participants can complete the questionnaire in less than 2 minutes [112]) to environmental constraints (e.g., a user is unlikely to answer a questionnaire when shopping). Hence, Jarvis was designed to collect user feedback with minimal user participation and yet provide rich and informative user reaction to a context based intervention.

From the point the intervention is sent, every interaction on the mobile device with regards to the intervention is captured - *Notification click time, Notification click location, intervention click time, intervention viewing duration* and plenty more. In addition to physical interactions, fine grained sensor data can also be turned on the moment the intervention is received by the participant. Experimenters can select this option during Step 3 of experiment creation 3.3. Currently all available sensors are toggled on (only on Android devices) for a duration of 10 minutes (configurable option by the admin). However, future work includes allowing the experimenter to decide which sensors to turn on and perhaps also for the duration. This rich sensor data can be used not only to correlate intervention type to user activity - *participants who are seated react to notifications faster*, but also provides additional data to improve context confidence - *sampling wi-fi signal at a higher frequency can improve location confidence*. Details of the mobile context and interactions captured by LiveLabs suite of application are listed in Appendix B.

3.6 Experiment Report

Objective judgement on the contribution of a particular research work can only be achieved on the basis of reproducible experiments. As a result, it becomes important to ensure the reliability and validity of an experiment [152]. Reliability occurs when an experiment measures the same thing more than once and results in the same outcome. Validity is ensuring that the experiment actually measures what needs to be measured.

Given that the experimentation system automates a large part of the experiment life-cycle, it becomes important to capture every decision that the system makes as well as the veracity of the user-context, in order to validate the experiment as

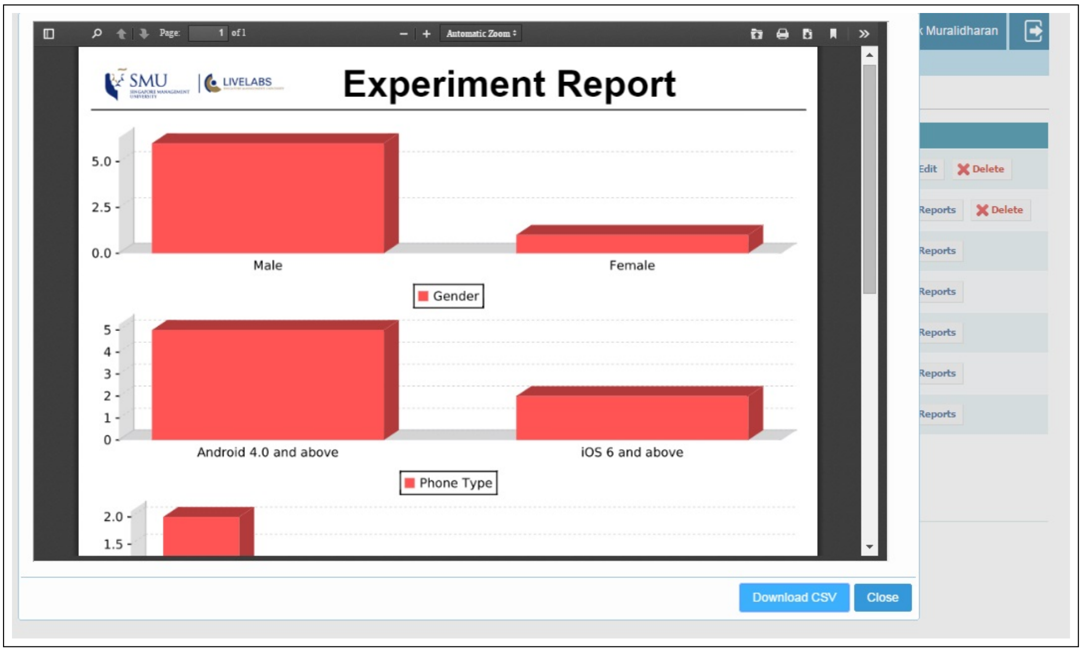


Figure 3.18: An Experiment Report.

well provide sufficient information to verify the reliability. However, instead of providing raw metrics, a report that includes a statistical analysis of these values will perhaps be more readable by an experimenter.

Jarvis therefore provides experimenters detailed experiment output data in addition to some basic statistics in the form of a report. Figure 3.18 shows a screen shot of an experiment report. Experimenters can generate the report on completion of the experiment and can download the report in a pdf format as well download the raw CSV data. Currently the CSV captures a static set of user interactions highlighted in Section 3.5. Information also included are basic demographics details, the time and location of the participant when the intervention was sent as well as which group (control or treatment) they belonged to. Future iterations will allow creating a customizable report where experimenters can select the data fields they require in the output file.

3.7 Validation Plan

A more fine-grained categorization of contextual factors is needed when we want to understand how context affects user behaviour. The goal of the experimentation system is to provide this understanding through the process of experimentation. However, it is necessary to validate the effectiveness of the system in supporting behaviour observation.

I validate the system by executing experiments for known observable social phenomenon and verify whether the system can observe the same. For example, when sending a promotion for 'ladies handbag', the system should observe a statistically significant positive response from female participants as compared to male participants (with all other factors being same). Note, marketing researchers have observed that demographics are generally poor predictors of consumer behaviour and coupon redemption, and that other factors such as *coupon proneness* and *redemption effort* are more strongly associated with coupon use [41, 100]. However, since I only consider perceived usefulness for the current system I hope to observe the intended consumer response for the given promotion-based experiment.

Also, the USP (Unique Selling Proposition) of Jarvis is its ability to run in situ context-based experiments. This feature alone sets it apart from all social lab experiments where the participants are aware of being subjects of an experiment. The experiments[4] for system validation are therefore selected based on two criteria: one, highlight the ability of Jarvis in supporting a diverse range of experiments and in doing so test the different aspects of the system and two, showcase it's ability to capture additional 'useful' social phenomenon within the experimentation process that wasn't possible thus far.

In this section, I describe the four experiments run on Jarvis followed by the results. For each experiment I capture a breakdown of the participant demographics, notification statistics, statistics on how the participants reacted to the intervention and finally the data corresponding to the main hypothesis of the experiment.

[4] SMU-IRB Approval Number: IRB-15-010-A007(215)

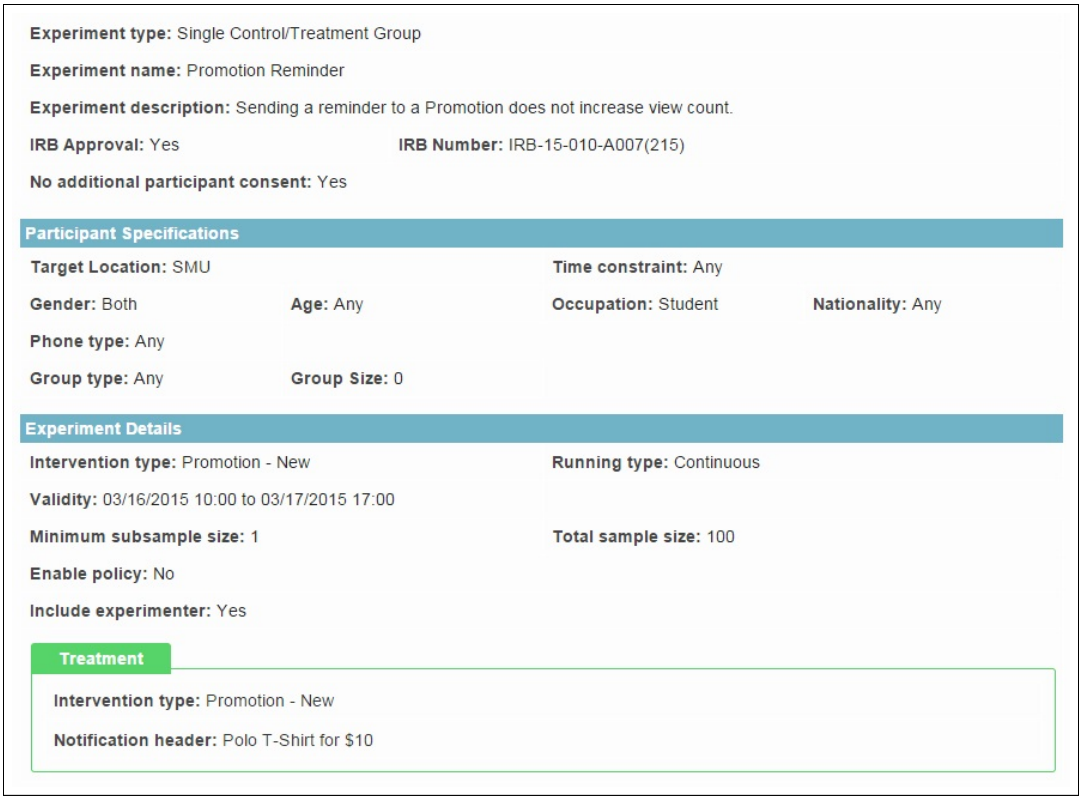
3.7.1 Experiment 1: 'Sending a promotion reminder does not increase view count'

Questionnaires are widely used to collect data in research and are often the only financially viable option when collecting information from large, geographically dispersed populations. Non-response to questionnaires reduces the effective sample size and can introduce bias [155]. As non-response can affect the validity of the study, assessment of response is important in the critical appraisal of any research. For the same reason,the identification of effective strategies to increase response to questionnaires could improve the quality of research. To identify such strategies Edward et al. [43] conducted a systematic review of 75 randomised controlled trials strategies for influencing response to postal questionnaires. of the multiple strategies employed, the authors observed the odds of response were more than doubled when a monetary incentive was used and almost doubled when incentives were not conditional on response.

Analogous to the above study, I recreated the experiment on Jarvis, replacing the questionnaire with a promotion and observing the effect of sending a promotion reminder (*independent variable*) on the view count (*dependent variable*). The experiment was executed in the two phases. In the first phase a promotion was sent out to LiveLabs participants. No contextual constraints was specified for this experiment. Figure 3.19 shows a screenshot of the experiment design of phase 1 as captured on Jarvis.

In the second phase, the set of participants that received the promotion (in phase 1) were divided equally into a control and treatment group, with the treatment group receiving a reminder to view the promotion (the reminder was sent three days after sending the promotion). We then observed how many participants viewed the promotion as a result of the reminder. Figure 3.20 shows a screenshot of the experiment details of phase 2 as captured on Jarvis.

This experiment showcases the ability to run a simple control/treatment experiment using Jarvis. A useful feature of the platform, that this experiment highlights,



Experiment type: Single Control/Treatment Group

Experiment name: Promotion Reminder

Experiment description: Sending a reminder to a Promotion does not increase view count.

IRB Approval: Yes **IRB Number:** IRB-15-010-A007(215)

No additional participant consent: Yes

Participant Specifications

Target Location: SMU **Time constraint:** Any

Gender: Both **Age:** Any **Occupation:** Student **Nationality:** Any

Phone type: Any

Group type: Any **Group Size:** 0

Experiment Details

Intervention type: Promotion - New **Running type:** Continuous

Validity: 03/16/2015 10:00 to 03/17/2015 17:00

Minimum subsample size: 1 **Total sample size:** 100

Enable policy: No

Include experimenter: Yes

Treatment

Intervention type: Promotion - New

Notification header: Polo T-Shirt for $10

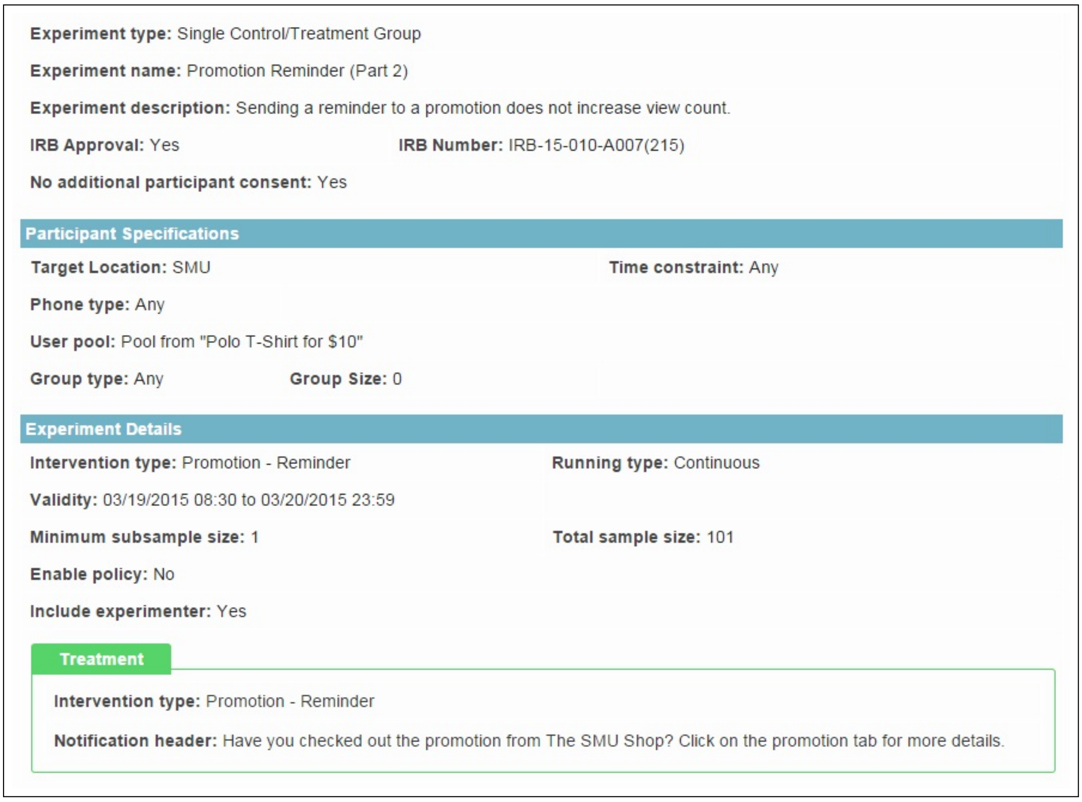
Figure 3.19: Experiment 1: Sending the Promotion.

is the sending of the reminder *only* to those participants that received the promotion thereby eliminating any information bias. Details of the promotion sent is in Appendix C.1.

Results

A total of 100 participants were sent the promotion for phase one of the experiment. Of these participants 8 were inaccessible for the second phase resulting in only 92 participants. Of these 92 participants, half (46 participants) were part of the treatment group receiving a reminder for the promotion while the other half, the control group, did not. Table 3.2 shows a breakdown of the participant demographics for both phases of the experiment.

While the system selects and sends participants' the experiment intervention, not all of them will receive it for multiple reasons - the LiveLabs service has been turned off, notifications has been disabled and so on. While iOS does not provide any additional information on mobile notifications, the system is able to capture



Experiment type: Single Control/Treatment Group

Experiment name: Promotion Reminder (Part 2)

Experiment description: Sending a reminder to a promotion does not increase view count.

IRB Approval: Yes **IRB Number:** IRB-15-010-A007(215)

No additional participant consent: Yes

Participant Specifications

Target Location: SMU **Time constraint:** Any

Phone type: Any

User pool: Pool from "Polo T-Shirt for $10"

Group type: Any **Group Size:** 0

Experiment Details

Intervention type: Promotion - Reminder **Running type:** Continuous

Validity: 03/19/2015 08:30 to 03/20/2015 23:59

Minimum subsample size: 1 **Total sample size:** 101

Enable policy: No

Include experimenter: Yes

Treatment

Intervention type: Promotion - Reminder

Notification header: Have you checked out the promotion from The SMU Shop? Click on the promotion tab for more details.

Figure 3.20: Experiment 1: Sending the Promotion Reminder.

whether the notification was received for Android users. Table 3.3 shows how many of the selected participants eventually received the notification in phase 1.

Table 3.4 captures the set of experiment-intervention interactions such as how long after the notification was sent did the participant click on the notification and for how long did the participant view the promotion. Although 86 participants are considered to have received the notification, only 26 clicked on the promotion.

Finally, the contingency table 3.5 captures the 2x2 frequency distribution matrix of the experiment variables - the number of participants that viewed the promotion as a result of the reminder. Fisher's exact test is used to compute the statistical significance of the contingency table given the small sample size. Since the statistical analysis shows that the significance level is above the cut-off value (α=0.05) we *fail* to reject the null hypothesis.

Promotion			Reminder
Sample Size		100	92
Gender	Male	55	51
	Female	37	34
	Not Specified	8	7
Phone Type	Android	70	67
	iOS	30	25
School	SOA	14	14
	SOB	33	30
	SOE	15	13
	SIS	19	18
	SOL	7	6
	SoSS	12	11

SOA-School of Accountancy, SOB-School of Business, SOE-School of Economics, SIS-School of Information Systems, SOL-School of Law, SoSS-School of Social Sciences

Table 3.2: Experiment 1: Participant Demographics

Promotion Notification			Reminder Notification (Treatment Group)
Total Sent		100	46
Android	Received (Confirmed)	56	24
	Not Received	14	8
iOS	N.A.	30	14

iOS does not provide access to mobile notification information.

Table 3.3: Experiment 1: Notification Statistics

3.7.2 Experiment 2: 'The time spent viewing a promotion is independent of any affinity towards the content'

With the rapid growth of the Internet, online advertising channels, such as sponsored search [4], contextual ads [110], and Behavioural Targeting (BT) [68], are showing great market potentials. However, in contrast to the widely studied general sponsored search, BT, which refers to the delivery of ads to targeted users based on information collected on each individual users web search and browsing behaviours, is still underexplored [159]. To encourage more research on BT and possibly to further develop this market, Yan et al. [159] provide an empirical study on the click-through log of advertisements collected from a commercial search engine to seek the answer to the question: how much can BT help online advertising? From their

Promotion Notification	
Total Notifications Received (Android + iOS)	86
Clicked on Notification	23
Clicked on Notification & Promotion	16
Median Notification Response Time (min)	5.73
Promotion	
Total Promotion Views	46 (26 Participants)
Median Promotion Click Time (min)	62.6
Mean Promotion View Time (s)	22.13
Mean View Count	1.77
Liked the Promotion	4
Transition Type	
Soft Back	26
Hardware Back	12
Other	8
Promotion Reminder	
Total Notifications Received (Android + iOS)	38
Clicked on Notification	7
Median Notification Response Time (min)	1.63
Clicked on Reminder Message	6
Clicked on Notification & Reminder Message	6

Transition type captures the way the participant exits the promotion page, either by pressing the back button within the mobile app (soft) or by pressing mobile hardware back button.

Table 3.4: Experiment 1: Intervention Statistics

experiment results, run over a period of seven days, the authors draw three important conclusions: (1) Users who clicked the same ad will truly have similar behaviours on the Web; (2) Click-Through Rate (CTR) of an ad can be averagely improved as high as 670% by properly segmenting users for behavioural targeted advertising in a sponsored search; (3) Using short term user behaviours to represent users is more effective than using long term user behaviours for BT.

Influenced by the above study, I created an experiment on Jarvis to observe the behaviour of participants towards different kinds of promotions - more specifically is there a relationship between the amount of time spent viewing a promotion and one's affinity towards that promotion. To do this, I created a 'chained' experiment consisting of three experiments. Each experiment sends a promotion (*independent variable*) and observes user behaviour in terms of amount of time spent viewing the

	Viewed the Promotion	Did not View
With Promotion Reminder	7	31
Without Promotion Reminder	4	42

Fisher exact test statistic value is 0.2119 (α=0.05)

Table 3.5: H_0: Sending a reminder to a promotion does not increase view count.

promotion (*dependent variable*). Participants provide self-reporting data through a survey (sent automatically when the promotion is viewed) that captures how much the participant liked the promotion on a 5-point Likert scale. Figure 3.21 shows details of the first experiment in the chain.

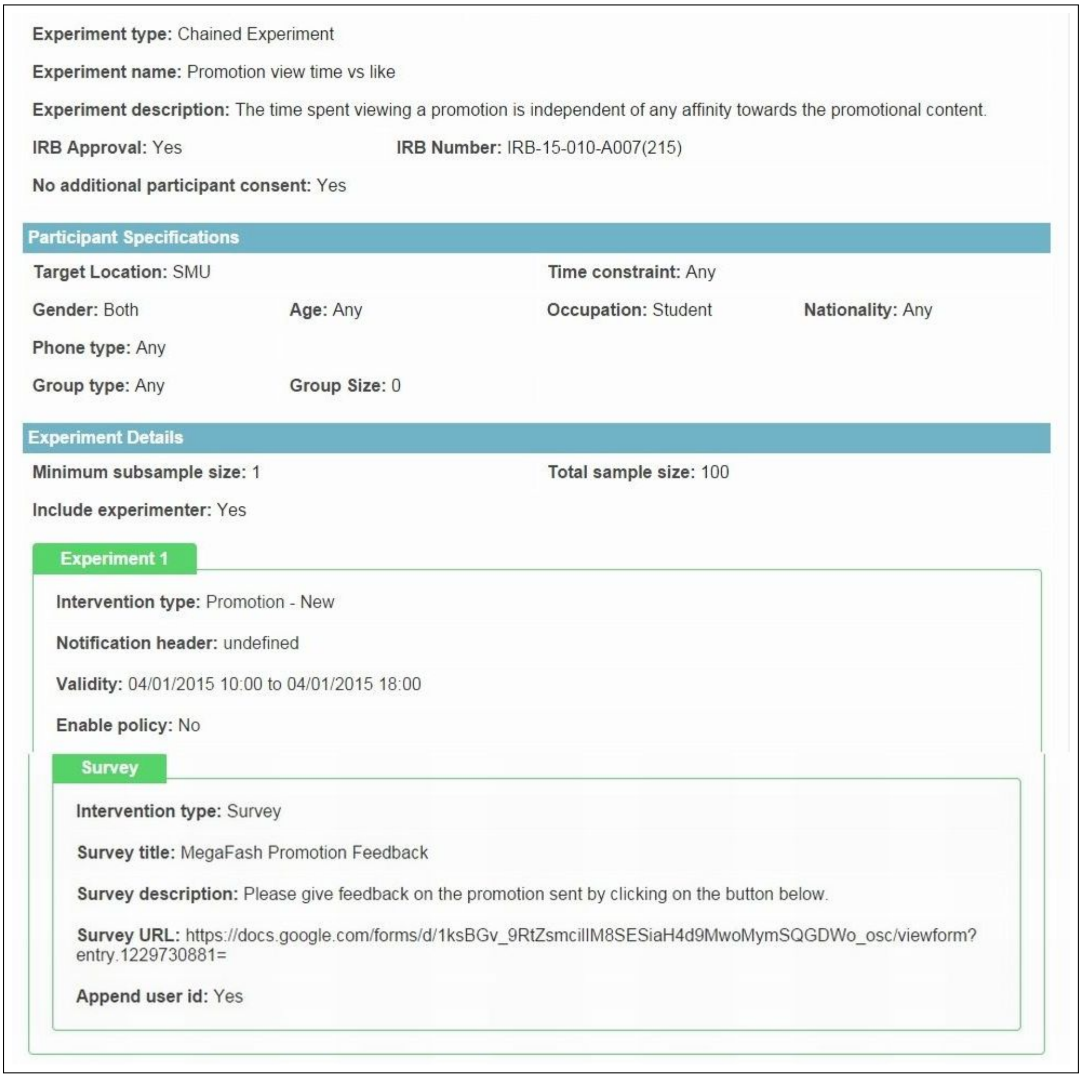
Note, in creating a 'chained' experiment, only those participants that received the first promotion will receive the subsequent promotions in the experiment. This provides insight into the promotion viewing behaviour *within* participants in addition to capturing the viewing behaviour between participants. However, for the purpose of verifying the above hypothesis, three independent experiments of type 'single' would also suffice.

This experiment showcases the ability to run a longitudinal study over a given set of participants, capture user behaviour as well self reporting scores through an external survey. Note while this experiment observes user behaviour in terms of promotion viewing time, other behavioural aspects are also captured such as location, sensor data and so on allowing multiple correlations to be drawn. Details of the promotion and survey sent to participants are in Appendix C.2.

Results

Experiment 2 attempts to capture the correlation between promotion affinity and the time spent viewing the promotion. Table 3.6 shows the number of participants that were selected for each experiment intervention while Table 3.7 captures the number of participants that received the experiment notification.

Similar to the previous experiment, the ratio of the number of participants that received the notification to the number of participants that viewed the promotion



Experiment type: Chained Experiment

Experiment name: Promotion view time vs like

Experiment description: The time spent viewing a promotion is independent of any affinity towards the promotional content.

IRB Approval: Yes **IRB Number:** IRB-15-010-A007(215)

No additional participant consent: Yes

Participant Specifications

Target Location: SMU **Time constraint:** Any

Gender: Both **Age:** Any **Occupation:** Student **Nationality:** Any

Phone type: Any

Group type: Any **Group Size:** 0

Experiment Details

Minimum subsample size: 1 **Total sample size:** 100

Include experimenter: Yes

Experiment 1

Intervention type: Promotion - New

Notification header: undefined

Validity: 04/01/2015 10:00 to 04/01/2015 18:00

Enable policy: No

Survey

Intervention type: Survey

Survey title: MegaFash Promotion Feedback

Survey description: Please give feedback on the promotion sent by clicking on the button below.

Survey URL: https://docs.google.com/forms/d/1ksBGv_9RtZsmcilIM8SESiaH4d9MwoMymSQGDWo_osc/viewform?entry.1229730881=

Append user id: Yes

Figure 3.21: Experiment 2: Observing promotion viewing behaviour.

is small. Surveys are sent only to those participants that viewed the promotion. Table 3.8 shows the intervention interaction statistics.

Finally, Table 3.9 captures the self reported data of the participants that viewed the promotion and the time they spent viewing that promotion. While the data is insufficient to draw any significant conclusion, the one participant that did like the promotion viewed it for the longest time as compared to the other two participants that did not like the promotion.

Promotion 1			Promotion 2	Promotion 3
Sample Size		85	68	43
Gender	Male	47	39	26
	Female	31	22	14
	Not Specified	7	7	3
Phone Type	Android	66	55	35
	iOS	19	13	8
School	SOA	12	9	6
	SOB	26	21	13
	SOE	12	10	8
	SIS	19	15	12
	SOL	8	5	2
	SoSS	8	8	2

SOA-School of Accountancy, SOB-School of Business, SOE-School of Economics, SIS-School of Information Systems, SOL-School of Law, SoSS-School of Social Sciences

Table 3.6: Experiment 2: Participant Demographics

Promotion 1			Promotion 2	Promotion 3
Total Sent		85	68	43
Android	Received (Confirmed)	50	42	27
	Not Received	16	13	8
iOS	N.A.	19	13	8

iOS does not provide access to mobile notification information.

Table 3.7: Experiment 2: Notification Statistics

3.7.3 Experiment 3: 'Response to mobile notifications are independent of activity'

Today's smartphones often use notifications to attract a user's attentionwhen there's an incoming text message or upcoming event, for example. Although some notifications, such as a newly available application update, can be delayed, others require immediate user attention and action. For this reason, most notifications are presented in an obtrusive waywith an on-screen visual, a short vibration, or a flashing LED. If the user doesn't attend to the notification, the phone might continually send reminders.

Although it is valuable to have notifications quickly reach the user, the existing "issue and repeat" strategy can be obtrusive and annoying, given that notifications can occur (repeatedly) in inconvenient situations, such as at night or when the user is

	Promo 1	Promo 2	Promo 3
Promotion Notification			
Total Notifications Received (Android + iOS)	69	55	35
Clicked on Notification	19	11	9
Clicked on Notification & Promotion	12	7	4
Median Notification Response Time (min)	6.4	6.3	75
Promotion			
Total Promotion Views	17	10	6
Median Promotion Click Time (min)	93.58	7.19	100
Mean Promotion View Time (s)	3.68	43.27	6.43
Mean View Count	1	1.3	1
Liked the Promotion	1	1	1
Transition Type			
Soft Back	5	6	1
Hardware Back	5	5	3
Other	7	2	2
Survey			
Total Survey Sent	17	10	6
Clicked on Notification	2	1	2
Clicked on Survey Details	3	2	1
Clicked on Survey Link	2	2	1
Responded to the Survey	2	2	1
Median Notification Response Time (min)	48.32	0.1	23.03

Transition type captures the way the participant exits the promotion page, either by pressing the back button within the mobile app (soft) or by pressing hardware back button. Transition Type & Promotion view time is unavailable for participants who exit the the promotion page any other way.

Table 3.8: Experiment 2: Intervention Statistics

driving. These unwanted notifications often lead to stress and increased frustration, because users feel pressured to address the alerts [93]. One proposed solution is to develop a context-aware phone that could identify whether it is appropriate to trigger a notification [51].

To investigate the contexts in which users typically attend to (or ignore) mobile notifications, Poppinga et al. design MoodDiary, a mobile diary application for mood tracking [121]. After collecting 6,581 notifications from 79 users, the authors developed a model that predicts opportune moments to issue notifications with 77.85 percent accuracy.

Predicting such opportune moments to send notifications is not easy. Several

Participant ID	View Time (s)	Promotion Response	Promotion ID
10061196	N.A.	Somewhat agree	1
10057655	1.72	Strongly disagree	1
10057655	9.02	Neither agree or disagree	2
10028869	N.A.	Somewhat agree	2
10028869	16.29	Somewhat agree	3

Insufficient data to draw a conclusion.

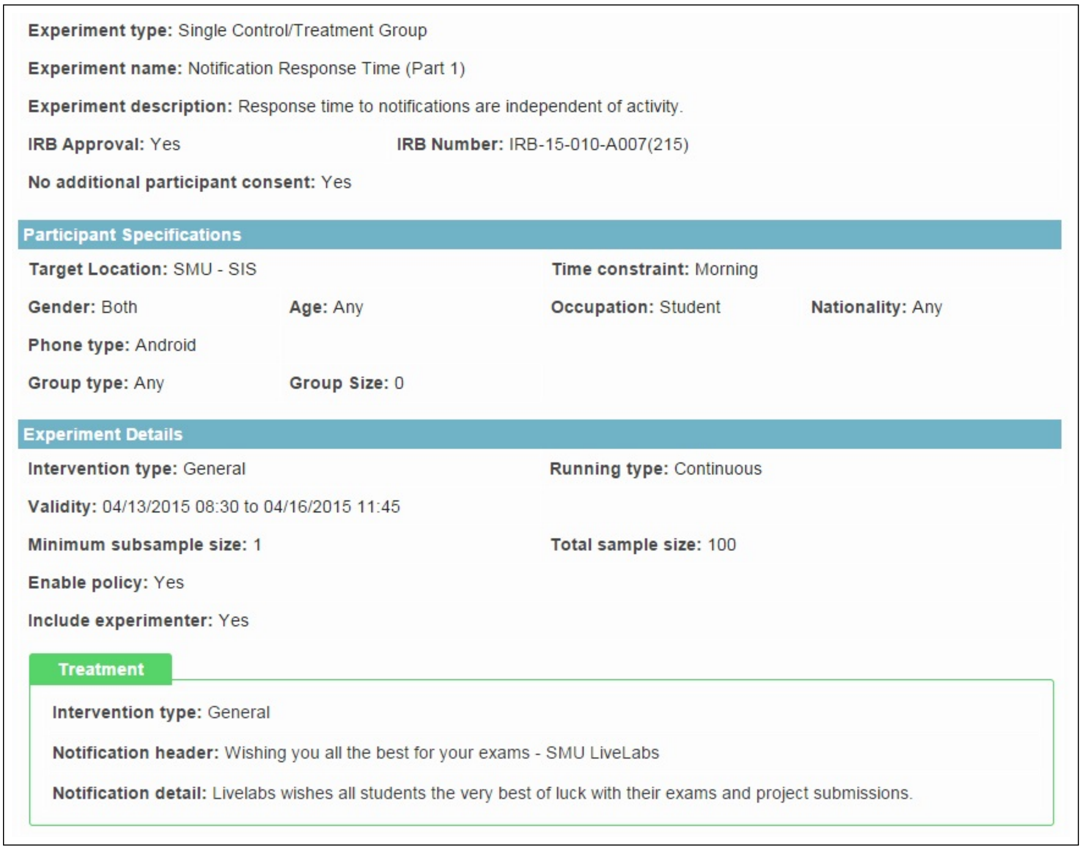
Table 3.9: H_0: The time spent viewing a promotion is independent of any affinity towards that promotion.

studies [49, 50, 44, 83, 118, 119, 133] in the recent past have attempted to do just that. However, a key limitation with these studies is that they do not alter the opportune moments, but instead chose to collect user responsiveness for *all* possible situations and later build a prediction model. This data-driven approach can lead to participant fatigue. In this experiment I recreated a similar study, of observing a participant's response to a notification (or lack of it) and correlate it with his/ her physical activity. With this experiment I showcase the ability of the system not only to recreate a complex study with ease, but also highlight the ability to alter the 'moments' when the notifications are sent to users - something that was not possible with previous studies.

In this 2-part-experiment I observed the notification response time for two time-related opportune moments: *Morning* (8:30am to 11:45am) and *Afternoon* (12:00pm to 3:15pm), while capturing the participant's physical activity at the time of receiving the mobile notification. Details of the notification and experiment specification, of the first part, are shown in Figure 3.22.

Results

In the third experiment, I hypothesize that the response time to a mobile notification is independent of the participant's physical activity. Since this involves capturing additional sensor data, the system selects only participants that have an Android device. Two messages were sent at different times to a total of 78 participants. The demographics breakup and the number of participants that received the notification


Experiment type: Single Control/Treatment Group

Experiment name: Notification Response Time (Part 1)

Experiment description: Response time to notifications are independent of activity.

IRB Approval: Yes **IRB Number:** IRB-15-010-A007(215)

No additional participant consent: Yes

Participant Specifications

Target Location: SMU - SIS **Time constraint:** Morning

Gender: Both **Age:** Any **Occupation:** Student **Nationality:** Any

Phone type: Android

Group type: Any **Group Size:** 0

Experiment Details

Intervention type: General **Running type:** Continuous

Validity: 04/13/2015 08:30 to 04/16/2015 11:45

Minimum subsample size: 1 **Total sample size:** 100

Enable policy: Yes

Include experimenter: Yes

Treatment

Intervention type: General

Notification header: Wishing you all the best for your exams - SMU LiveLabs

Notification detail: Livelabs wishes all students the very best of luck with their exams and project submissions.

Figure 3.22: Experiment 3: Notification Response Time vs. Activity.

are captured in Table 3.10 and 3.11 respectively.

Like the previous experiments, the response rate is again low with only ten participants clicking the notification and five viewing the message out of 57 participants that received the experiment intervention. The interaction statistics are captured in Table 3.12.

Finally, the contingency table 3.13 captures the 2x3 frequency distribution matrix of the experiment variables - the number of participants that responded to the mobile notification and their activity. Here, I consider only those participants that clicked on the notification within 2 minutes of receiving it. I employ an activity recognition model that uses the accelerometer signals from the smartphone to detect the participants locomotive activity such as siting, standing and walking. The signals are collected over a 2 second sliding window and several discriminative features are calculated from the raw signals as mentioned in Yan et al. [160]. These

Message 1				Message 2
Sample Size		29		49
Gender	Male	24		35
	Female	1		9
	Not Specified	4		5
Phone Type	Android	29		49
	iOS	0		0
School	SOA	4		4
	SOB	7		16
	SOE	4		3
	SIS	7		19
	SOL	4		1
	SoSS	3		6

SOA-School of Accountancy, SOB-School of Business, SOE-School of Economics, SIS-School of Information Systems, SOL-School of Law, SoSS-School of Social Sciences

Table 3.10: Experiment 3: Participant Demographics

Message 1				Message 2
Total Sent		29		49
Android	Received (Confirmed)	21		38
	Not Received	8		11

Table 3.11: Experiment 3: Notification Statistics

features vectors are used for classification using a J48 decision tree classifier[5].

I also attempt to observe if there is a relationship between response to the mobile notification and time of day (Table 3.14).

Fisher's exact test is used to compute the statistical significance of the contingency table given the small sample size. Since the statistical analysis shows that the significance level is above the cut-off value (α=0.05) we *fail* to reject the null hypothesis in both cases.

3.7.4 Experiment 4: 'Impulsive people prefer tasks that provide instant rewards as opposed to a delayed one'

Monetary incentives are used to encourage people to perform a variety of different actions. Increasingly, they are used to encourage actions that are directly beneficial

[5] I use the model and source code built by Yan et al. [160] to calculate the features and classify the participants' activity.

	Message 1	Message 2
Message Notification		
Total Notifications Received	21	38
Clicked on Notification	4	6
Clicked on Notification (within 2 minutes of receiving)	1	5
Clicked on Notification & Message	3	1
Mean Notification Response Time (min)	393.52	28.7
Message		
Total Message Views	4	1
Mean Message Click Time (min)	528.18	0.1
Mean Message View Count	1.5	1

Table 3.12: Experiment 3: Intervention Statistics

	Responded to Notification	Did not Respond
Sit	6	33
Stand	0	11
Walk	0	9

Fisher exact test statistic value is 0.2825 (α=0.05)

Table 3.13: H_0: Response to mobile notifications are independent of activity.

to individuals, such as engaging in healthy behaviours or quitting addictive drugs. Interestingly, however, there is little research comparing the effectiveness of different types of incentive schemes. Mullen et al. [104] compares the effectiveness of incentive schemes that offer sure payments versus lotteries of equal expected value. Specifically, two studies examine whether people are more likely to participate in a task if they are offered: (1) a fixed sure payment, (2) a lottery of equal expected value, or (3) a choice between the sure payment and lottery. The author hypothesizes that lotteries will be the most effective incentive and that a choice between a lottery and a fixed payment will be least effective.

However, in order to get a refined understanding of incentive preference it is important to examine some of the characteristics and related behaviours of the individuals. Personality is one of the prominent factors influencing preference in a given situation [56]. For the final experiment we therefore designed a novel study linking personality and incentive preference[6].

[6]This experiment was designed along with Dr.William Tov, a faculty member of the School of Social Sciences in SMU.

	Responded to Notification	Did not Respond
Morning	1	20
Afternoon	5	33

Fisher exact test statistic value is 0.4069 (α=0.05)

Table 3.14: H_0: Response to mobile notifications are independent of time of day.

As part of the experiment design, participants would randomly get one of two incentives for completing a survey aimed at capturing the impulsiveness of the individual. The first incentive is a coupon for a SUBWAY Cookie which can be collected immediately on completing the task. The second incentive is also a coupon for a SUBWAY Cookie, however with the additional clause of having to waiting *at least* one week before collecting the coupon. We hypothesize that the first experiment group will include significantly more individuals that exhibit a personality high on the ABIS (ABbreviated Impulsiveness Scale) as compared to the second group. Figure 3.23 captures the experiment design on Jarvis. Details of the survey can be found in Appendix C.3.

Results

Sample Size		100
Gender	Male	51
	Female	43
	Not Specified	6
Phone Type	Android	40
	iOS	60
School	SOA	14
	SOB	45
	SOE	12
	SIS	18
	SOL	4
	SoSS	7

SOA-School of Accountancy, SOB-School of Business, SOE-School of Economics, SIS-School of Information Systems, SOL-School of Law, SoSS-School of Social Sciences

Table 3.15: Experiment 4: Participant Demographics

The final experiment attempts to understand how personality influences incentive preference. A total of 100 participants were sent one of two interventions,

Total Sent		100
Android	Received (Confirmed)	15
	Not Received	25
iOS	N.A.	60

iOS does not provide access to mobile notification information.

Table 3.16: Experiment 4: Notification Statistics

offering either an immediate incentive or a delayed one for completing a survey. The survey in turn captures the impulsiveness level of the participant. Table 3.15 show details of the participants selected for this experiment. Table 3.16 shows how many of these participants are considered to have received the task notification.

Survey Notification	
Total Notifications Received (Android + iOS)	75
Clicked on Notification	3
Clicked on Notification & Survey	3
Mean Notification Response Time (min)	36.3
Survey	
T1 (Immediate Incentive)	43
T2 (Delayed Incentive)	32
Clicked on Survey Details	12
Clicked on Survey Link	8
Responded to the Survey	7 [T1:6 T2:1]
Mean Survey Response Time (min)	22.4

Table 3.17: Experiment 4: Intervention Statistics

The 75 participants, considered to have received the intervention, were split into two experimental groups. 43 participants received a notification offering an immediate payment for completing the survey while 32 participants received a notification offering a delayed incentive for the same task. Only twelve participants clicked on either of these notifications with seven proceeding to complete the survey. Table 3.17 captures these details.

Table 3.18 captures the ABIS score of the eight participants that completed the survey. Since there is only one participant in the second experimental group, it is not possible to draw a significant conclusion from this experiment.

Participant ID	Treatment Group	ABIS Score
10012979	1	27
10057655	1	29
10076674	1	33
10070561	1	29
10069489	1	22
10060966	1	31
10056829	2	45

Insufficient data to draw a conclusion.

Table 3.18: H_0: The level of impulsiveness has no influence on the choice of incentive (immediate vs. delayed) for completing a task.

3.7.5 Summary of Results

The *live* experiments have shown the ability of the system to target and capture participant behaviour in real time. The set of experiments also display the diversity of the platform in supporting a range of experiments with a multitude of triggering conditions. Further, these tests highlight the ease with which an experiment can be created, leaving most of the heavy lifting of participant selection and bias handling to the platform.

While these *live* experiments with “uncontrolled” participants have the benefit of the results being generated from real participants engaging in real activities in real environments, these experiments have also shown a downside to running in situ studies - low response rate. As the participant is oblivious of the fact that he/she is part of a behavioural experiment there is no compulsion to respond to the experiment stimuli (notification, promotion etc.) - a problem that is not faced by experiments in a lab setting.

While ignoring the experiment intervention is a response in itself, it could also indicate a bias towards notifications received from certain mobile apps, including those provided by LiveLabs.

Notifications are a core feature of mobile phones. They inform users about a variety of events. Users may take immediate action or ignore them depending on the importance of a notification as well as their current context [133]. It therefore becomes necessary to understand a Livelabs participants' subjective perception of mobile notifications and in particular to notifications sent by the LiveLabs mobile apps.

I therefore conduct an online survey to gain greater insight into the perception of and stance towards different mobile notifications in absolute terms and by weighing them against parameters including their related timeframe, application category and social connections. The survey[7] (Appendix C.4) was distributed, via email, to 75 students randomly chosen from the pool of participants selected for at least one of the four *live* experiments. The participants were compensated at a flat rate of $5 SGD for completing the survey.

Survey Discussion

Although the survey did not benefit from a decent sized sample (16 respondents), we none the less observed several differences in the usage and perception of mobile notifications, attributable to parameters we can link to the user, the system or the message.

75% of the respondents receive up to 30 mobile notifications in a day with more than 40% checking the notification immediately. Most respondents (75%) do not follow through the notification if it isn't important enough. A further look into the the influence of application category on notification response reveals that most smartphone users do not deem promotion-based notifications to be important (62.5%) as compared to notifications regarding calls and SMS (0%). With regards to notifications sent by the LiveLabs mobile applications (Section 3.2), more than 50% have turned off notifications or dismiss them upon receiving.

While the results of this survey cannot be generalized to the larger population it none the less provides insight as to why the validation experiments experienced a poor response rate. The results also suggest that for the given sample population alternate intervention types (non-promotion) might be better suited to observe user

[7] SMU-IRB Approval Number: IRB-15-032-A025(415)

behaviour. The detailed summary of responses can be found in C.4.2.

Limitations of System Validation

The four experiments were intended to showcase the ability of Jarvis in targeting participants in real environments under varying context criteria. While this goal was met, there are several known limitations, primarily related to experiment design.

The experiments evaluated the various aspects of the system without considering the finer details of experiment design. For example, I did not take into account aspects such as the design and selection of the promotions for the given target population and environment. A more appropriate evaluation would involve experiments that take these design elements into careful consideration. For example, promotion-based experiments might be more suited in a conducive environment like a shopping mall as opposed to a university campus. Further, working professionals might be a more appropriate target population for such experiments compared to students. Failure to consider these attributes might have also contributed towards the low-response rate.

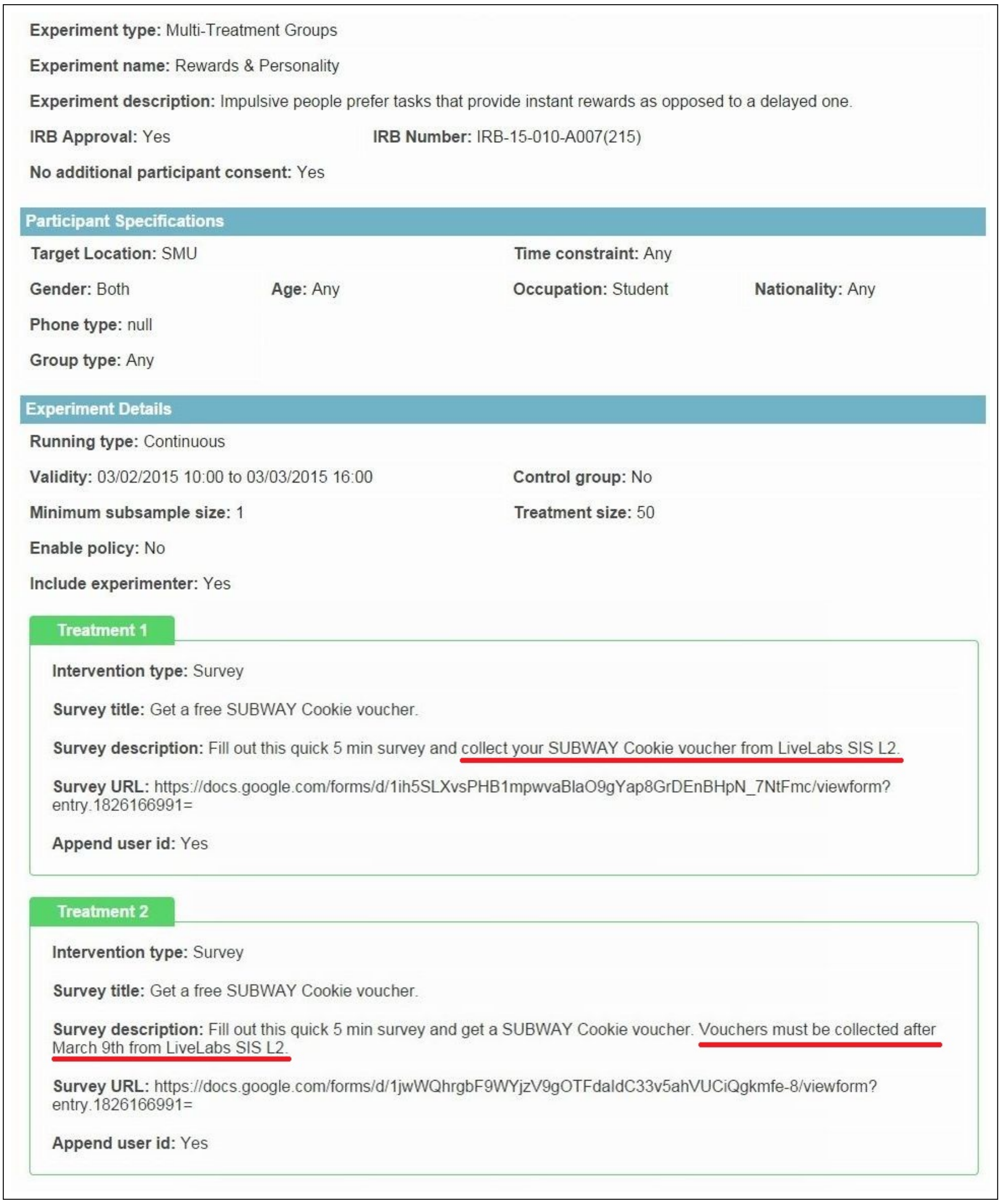
Also, the validation plan did not include experiments created by external experimenters. Given that the system is intended to be used by non-technical personnel (e.g., behavioural scientists, marketing researchers) it is important to verify the usability and usefulness of the system by these members. This evaluation would also provide insight into potential system improvements.

Finally, while the system functionality was verified, it is also important to evaluate the system performance under varying load. This is particularly crucial when running a large number of concurrent experiments.

3.8 Summary

In this chapter I present the architecture of Jarvis and list the components needed to support running in situ context-based experiments. I then describe the process of creating an experiment and explain in detail the role of the different components that are part of this process. Of these, a key module is the participant selection module that is responsible for minimizing the multiple sources of experiment bias. Finally, I present the validation plan, and its results, for evaluating the efficacy and effectiveness of the system through a set of live experiments.

While the experiments showcase the diversity of the platform in specifying a range of experiments, the results were largely inconclusive due to a low response rate. A followup survey indicated that the low response was due to the sample population being non-receptive to LiveLabs mobile app notifications. Although these results do not diminish the potential utility or functioning of Jarvis, it does raise an issue largely faced by the mobile application developer community - a need to understand the target audience. Building trust in users is imperative before establishing loyalty in them. And, to do so, the app must deliver content, information or entertainment in ways the target audience wants them. I believe with the right app, participants will respond to notifications in the way that will be useful for experimenters.



Experiment type: Multi-Treatment Groups

Experiment name: Rewards & Personality

Experiment description: Impulsive people prefer tasks that provide instant rewards as opposed to a delayed one.

IRB Approval: Yes **IRB Number:** IRB-15-010-A007(215)

No additional participant consent: Yes

Participant Specifications

Target Location: SMU **Time constraint:** Any

Gender: Both **Age:** Any **Occupation:** Student **Nationality:** Any

Phone type: null

Group type: Any

Experiment Details

Running type: Continuous

Validity: 03/02/2015 10:00 to 03/03/2015 16:00 **Control group:** No

Minimum subsample size: 1 **Treatment size:** 50

Enable policy: No

Include experimenter: Yes

Treatment 1

Intervention type: Survey

Survey title: Get a free SUBWAY Cookie voucher.

Survey description: Fill out this quick 5 min survey and collect your SUBWAY Cookie voucher from LiveLabs SIS L2.

Survey URL: https://docs.google.com/forms/d/1ih5SLXvsPHB1mpwvaBlaO9gYap8GrDEnBHpN_7NtFmc/viewform?entry.1826166991=

Append user id: Yes

Treatment 2

Intervention type: Survey

Survey title: Get a free SUBWAY Cookie voucher.

Survey description: Fill out this quick 5 min survey and get a SUBWAY Cookie voucher. Vouchers must be collected after March 9th from LiveLabs SIS L2.

Survey URL: https://docs.google.com/forms/d/1jwWQhrgbF9WYjzV9gOTFdaldC33v5ahVUCiQgkmfe-8/viewform?entry.1826166991=

Append user id: Yes

The difference in treatments are underlined in red.

Figure 3.23: Experiment 4: Personality vs. Incentive Preference

Chapter 4

Mapping Content to Context

Several factors make it incredibly hard for consumers to identify stores of interest to them in any particular mall. However, it is not sufficient to just identify interesting stores by name as consumers also want to identify stores that are having deals of specific interest to them [102].

Existing representative mobile advertising i.e., location-based advertising, is limited in effective targeting. This is mainly because it delivers promotions based on just the user's current location without considering any other user context [17]. For example, deals for a nearby restaurant promoting group dining offers do not attract people who are dining alone. Thus in order to target consumers more effectively, mobile advertising will need to consider additional user context, promotion preferences being one [86].

Similar to mobile advertising, promotion-based experiments may also entail targeting the right set of participants for that experiment. For example, an experimenter may want to send a promotion offering discount on coffee to participants that not only like coffee but also prefer discount-based promotions. Thus, in addition to using contextual triggers to actuate specified experiment interventions on matching subjects, Jarvis might also need to match the experiment promotional content with user preferences in order to get the best sample. This requirement of targeting the right sample can be done by understanding user preferences as well as prioritizing deals [135].

The key challenge in this preferences-to-promotion *matching algorithm* however, arises from having to combine both structured (easy to understand numeric discounts ("5% off") etc.) and unstructured (free-form text ("A free teddy bear") etc.) promotion information (1). The algorithm also needs to factor in consumer preferences such as "prefer discounts over free gifts" etc (2). Also, most consumers own multiple payment and discount cards and this multiplies the complexity as each individual card (and combinations of) has its own set of promotions and deals (3). Our final algorithm adapts natural language semantic techniques and combines all three factors to provide experimenters with the most relevant participants for a promotion.

An additional challenge arises in evaluating the matching algorithm. To measure the accuracy of the algorithm we need to match experiment-promotions with participant-preferences and verify whether the participant received a promotion appropriate with their preference. This requires knowledge of the promotion-preferences of our participants - which we are yet to learn. Alternatively, we could perform an analytical evaluation of the matching algorithm to understand its accuracy, devoid of any participants.

To get around this challenge of evaluating the algorithm accuracy whilst still incorporating users, I build a mobile application that takes promotion preferences as *input*. This provides the necessary information to perform the matching. However, instead of providing a single promotion matching the preferences I provide the user with a ranked list of relevant promotions (similar to a search engine). Building such a system serves two purposes: 1) the matching algorithm can be evaluated without prior knowledge of participant promotion-preferences and 2) the application allows us to learn user preferences (over time) and serves as an additional channel for sending in situ interventions similar to the apps listed in Section 3.2.

Building such a mobile application also introduces an additional challenge from having to display hundreds of promotions and store information on a small mobile screen without burdening the user. This challenge is the main reason why the ranking algorithm is so vital; as a good algorithm allows us to immediately identify the most relevant promotions. I therefore focus on an innovative solution for the matching and ranking aspects.

In this chapter I illustrate how Jarvis handles the requirement of mapping preferences to content through myDeal [107], a context-based mobile shopping application. In the myDeal system, in addition to factoring the user's current location, we allow the user to specify other context such as *preferences* as well as *payment cards* owned; all important pieces of information needed to improve promotion relevance.

myDeal was evaluated in two different ways; First, an in-depth analytical evaluation of the ranking algorithm was conducted to understand the accuracy of the algorithm. Second, we ran a user study with 43 undergraduate students to understand the effectiveness of the myDeal system. The user study results show that our preference-based system is more accurate (users can find the most interesting promotions better) and faster to use than two other common system designs (1. all promotions just listed alphabetically, and 2. promotional information hidden under various categories) for these types of shopping applications.

Note that although the process of matching promotions to participants - to get the right participant sample - is different from ranking promotions based on user preferences, the individual components of understanding the different parts of a deal as well as factoring user preferences are still necessary. In this thesis I build and evaluate the latter process, with the former as future work.

4.1 The myDeal Shopping Assistant

The system architecture is shown in Figure 4.1. I first explain our design decisions followed by the solution details of each component of the system.

4.1.1 Design Considerations

Factor in User Preferences

"I prefer discount deals over vouchers" "Is there a lunch promotion with free dessert?" "I'm looking for group discount deals" - the list goes on. Understand-

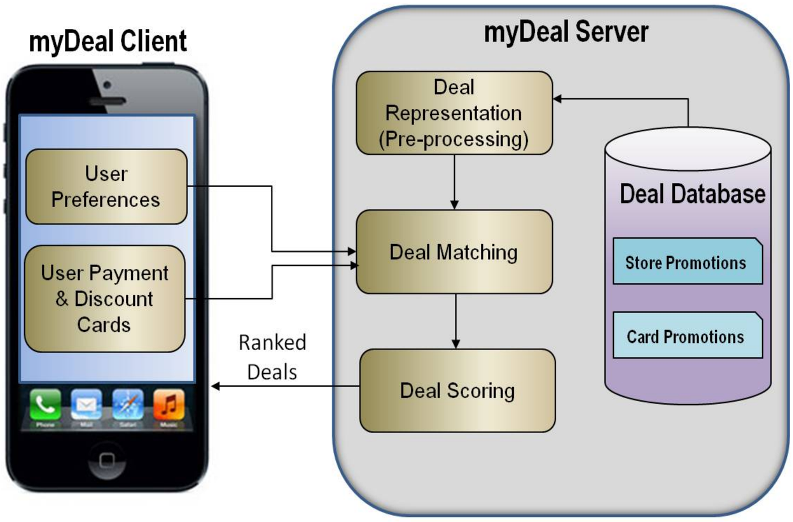


Figure 4.1: myDeal System Architecture.

ing promotion details is not enough — we also have to factor in the consumer's preferences. Does she want immediate discounts over vouchers? or does she prefer frequent flyer points over everything else?

A major limitation with all existing shopping assistance programs (that I know of) is that they do not really take into account user preferences. At best, the user is allowed to specify categories or keywords and the programs sort the results based on that. However, a single mall can have hundreds of stores with many hundreds of deals between them. As such, even keyword searches and categories break down in this type of rich data environment (this claim is validated in the user study presented later). This problem is made worse when we consider that there are multiple malls usually within walking distance and that the consumer has multiple payment implements (which usually offer promotions on top of those already offered by the stores themselves). A pre-study demographics survey showed that several users carried between 4-6 cards in order to avail such offers.

A key consideration was to therefore integrate user preference into the system so that only the most relevant promotions and stores are brought to the consumer's immediate attention. However, this is a non-trivial process as promotions are stated using a combination of structured (easy to understand numeric discounts. e.g."5%

off") and unstructured (free-form text. e.g."Free teddy bear") components. A naive approach would be to just rank promotions based on the easy to sort structured components. This is not good as many promotion details, of high interest to the consumer, tend to be located in the unstructured components.

Has to be a Mobile Application

A fast growing smart phone market suggests that consumers really want to have this capability on their mobile phones allowing them to plan their shopping experience anytime and anywhere they wanted to.

However, the smart phone is not without its limitations; chief of which is their small display screen (at most 4-5 inches). This makes it important to design an application that only shows users the promotions and stores that are of highest value to them (hiding the rest away for the user to browse through manually if so desired). This is particularly important when we factor in the hundreds of possible store/deal/card combinations that could be applicable to any given consumer.

4.1.2 Building myDeal

Keeping in mind the design considerations discussed in Section 4.1.1, building myDeal requires the following three components:

An electronic representation of deals: Currently, most deals are not stored in an electronic form. The first task (*Deal Representation*) was thus to devise a schema that could capture both stores deals and promotions as well as deals and promotions offered by the various payment and discount cards from *Deal Database*.

Finding top deals: Given a set of cards carried by the user, the user's preferences, and the deals offered by retailers and card issuers, *Deal Matching* and *Scoring* components find the "best deals" for that user.

Presenting deals to the user: Ultimately, it is the user who has to decide which deal is the best for them. Hence, it is crucial that relevant information is presented to the user in a way that makes it easy for them to find the deal that maximizes their

```xml
<Deals>      <!-- Repeats -->
   <ID>RDSC008</ID>
      <Range>
         <Desc>20% Discount on bill</Desc>
         <Amount>
            <Type>Percentage D</Type>
            <Value>20</Value>
            <AmountSpend></AmountSpend>
            <Card>American Express</Card>
            <Retailer>Walmart</Retailer>
         </Amount>
      </Range>
      <Day>
         <Date>any</Date>
         <Time>any</Time>
      </Day>
   <isStackable> <!-- Repeats -->
      <CardName>any</CardName>
      <Retailer>any</Retailer>
      <ProductCategory>any</ProductCategory>
   </isStackable>
</Deals>
```

Figure 4.2: XML Schema for Describing Deals

needs. In addition, users must also be able to specify their shopping preferences easily on the *myDeal* mobile application.

4.1.3 Representation of Deals

Deals are offered by two main entities; retailers and payment/discount card operators. In the retailers case, the deal is likely to be valid only at that specific retailer whereas payment/discount card deals are likely to be valid across many retailers. For example, a supermarket could offer a 50% deal on laundry detergent. That deal is likely to only be valid at that supermarket and possibly its branches. On the other hand, a bank could offer a 2% cash rebate on its premium VISA credit card on *all* purchases. This cash rebate would apply no matter where the detergent was bought. Note: it is also possible for deals to be constrained to a particular retailer and a particular card. Our schema, described below, can handle this case as well.

I manually inspected a few hundred deals, from various retailers and payment cards, and identified four components that were used to create all deals - every deal encountered was some combination of these four components with different values and attributes assigned to each component. The four components are:

1. *Cash back*: These are specific cash refund. For example, 3% cash back of

entire bill. These discounts are in the form of percentages or fixed values ($5 cash back).

2. *Discounts*: These are specific cash discounts. For example, 5% cash discount of entire bill. These discounts are in the form of percentages or fixed values ($10 discount).

3. *Vouchers*: These are vouchers that can be accrued and then exchanged later for either cash or products. Frequent flyer miles, store loyalty points, etc. are some examples.

4. *Rewards*: These are deals in the form of real products. For example, get a teddy bear free with every $10 purchase.

An XML-based schema (Figure 4.2) is used to describe deals comprising of these four components. In addition to capturing basic deal information, a *Stackable* tag specifies whether this deal can be combined with other deals. For example, a deal offered by a loyalty card may only be usable with a deal offered by a credit card issued by a specific bank. I describe just the main parts below, omitting the rest for brevity:

$< Range >$: This block is used to specify the exact deals offered by this discount. It consists of the following tags:

- $< Desc >$: This tag contains the full description of any applicable reward. This will be shown to the user to help them understand exactly what products they will receive.
- $< Amount >$: This is the amount of the deal. It consists of $< Type >$ and $< Value >$. $< Type >$ can be set to Percentage (for 5% total discounts etc.), Fixed (for $10 discounts etc.), Cumulative (for loyalty points type deals), or One Time (used to provide the "cash" value of rewards. The $< Value >$ tag contains the actual amount of the deal.

$< Stackable >$: This tag group specifies whether this deal can be combined with other deals. For example, a deal offered by a loyalty card may only be usable with a deal offered by a credit card issued by a specific bank. The *any* keyword can be used to match every possibility. In addition, the *anycc* keyword (that matches any credit/debit card and nothing else) can also be used for the $< CardName >$ field. This simplifies the common case where reward cards can only be stacked with payment cards and not with other reward cards.

4.1.4 Finding the Best Promotion

The matching and scoring subsystem rank orders the best deals available for the user. These ranks determine how good a deal the user would get if they shopped at a particular retailer, possibly for a particular product, using particular payment and discount cards. Rank ordering the best deals involves two major steps: 1) Matching deals preferred by the user to those offered by the retailer and card issuers and 2) assigning scores for each of these valid combinations.

Matching algorithm

The matching algorithm is dependent on the parameters of four entities involved in most shopping scenarios; namely the deals, the cards carried by the user, user preferences and location. The four entities are mapped in two steps. First, users are mapped to retailers based on their location and the kind of deals they are looking for (e.g., Dining) and their deal preference. Second, retailers are mapped to the cards carried by the user. The matching algorithm filters out those retail outlets that do not match the above criteria.

Scoring algorithm

Deals are generally a combination of structured and unstructured content. Those that only consist of simple numerical values can be scored easily and ranked. The real challenge however is in how we score the following type of deals:

- Deals with multiple numeric values (e.g., Get a complimentary S$30 gift voucher and additional 3% rebate).
- Deals with non-numeric values(e.g., 1-for-1 free lunch)
- Deals with multiple non-numeric values.
- Deals with numeric and non-numeric values.
- Deals with multiple numeric and non-numeric values.

In order to score any deal, we first need to extract values for the following four categories from the deal description: *Discount*, *CashBack*, *Voucher* and *Reward*. For each category we use regular expressions to match and extract the corresponding values. For example, if the deal description is "Get a 20% Discount on the Total Bill and enjoy a complimentary Voucher of $10" we extract the following values *Discount* = 20 and *Voucher* = 10.

The following formula is then used to score a deal:

$$Score = \alpha \cdot Discount \ + \ \beta \cdot CashBack \ + \ \gamma \cdot Voucher + \ \theta \cdot Reward,$$

where α, β, γ and θ are weights to adjust the importance of each deal category. The value of these weights are specified by the user as deal preferences.

Deriving a value for rewards

Consider the following deal description: "Enjoy a Free Ice Cream or Cake with every meal purchased". To apply a score to this deal we must effectively assign a value to both reward items *Ice Cream* and *Cake*. We[1] propose a 2-step machine learning algorithm that uses semantics to determine the corresponding value of the reward.

[1]This part of the work was done in collaboration with Swapna Gottipati, a PhD candidate from the Data Management & Analytics group at SMU.

Step 1: Get a list of all reward items and their corresponding value (if specified) from all deals. This is done using standard NLP techniques such as part of speech (POS) tagging. For our system we use the Brill POS tagger from CST [24]. In our example, the rewards *Ice Cream* and *Cake* are extracted and added to the list.

Step 2: Using a semantic similarity method described by Lin et al. [89] and Pedersen et al. [117], we cluster rewards in the list into a semantic space in which rewards that are closely associated are placed in the same cluster. Several existing algorithms compute relatedness only by traversing the hypernymy taxonomy and find that *Ice Cream* and *Cake* are relatively unrelated. However, WordNet provides other types of semantic links in addition to hypernymy, such as meronymy (part/whole relationships), antonymy, and verb entailment, as well as implicit links defined by overlap in the text of definitional glosses. These links can provide valuable relatedness information. If we assume that relatedness is transitive across a wide variety of such links, then it is natural to follow paths such as ice cream-frozen dessert-dessert, sweet-dessert and find a higher degree of relatedness between *Ice Cream* and *Cake*.

Lin's similarity measure uses the information content (IC) of the words/concepts, and the least common subsumer (LCS) of the concepts in the WordNet taxonomy. LCS is the common ancestor of two concepts which has the maximum information content. The similarity measure between concepts w_i, w_j is defined as follows.

$$Sim(w_i, w_j) = \frac{2 \times IC(LCS(w_i, w_j))}{IC(w_i) + IC(w_j)} \tag{4.1}$$

where

$$IC(c) = -log(P(c)) \tag{4.2}$$

$LCS(w_i, w_j)$ is a common subsumer of w_i, w_j , $IC(c)$ is the information content of the concept c and $P(c)$ is the probability of c. In our example, the rewards *Ice Cream* and *Cake* are likely to be clustered together as *'Dessert'*. The median value of the cluster is then assigned to rewards that currently do not have any value associated with them. Going ahead with our example, if the median value of the cluster *'Dessert'* is $5, the rewards *Ice Cream* and *Cake* are also assigned the same value.

Where to perform the match and scoring

A question that arose when building myDeal was deciding where to perform the ranking. In particular, the matching and scoring could be performed on the user's mobile phone or using an external service. Each of these options had its strengths and weaknesses.

Rank ordering on the user's mobile phone provides the highest amount of privacy — no card information is sent anywhere. However, the user's mobile phone is computationally limited and requires access to deal information across multiple retailers and card issuers. While resource utilization and privacy are a concern, as these issues were not the primary focus of this work, it was decided to use a backend service to perform the filtering and scoring, at the risk of users card information(only card name) revealed to the hosting site.

4.1.5 Integrating the User

The final component of the system is the user interface for presenting deals to users. It was designed to satisfy the following key properties:

- **Clear indication of combination of cards**: Many deals require combining multiple cards together to attain them. It is thus crucial to point out to the user which cards need to be used. This was achieved by using colour coding to distinguish the different pieces of information.

- **Display deal breakdowns**: This was achieved by showing the complete description of discount percentage, cash back, vouchers and rewards.

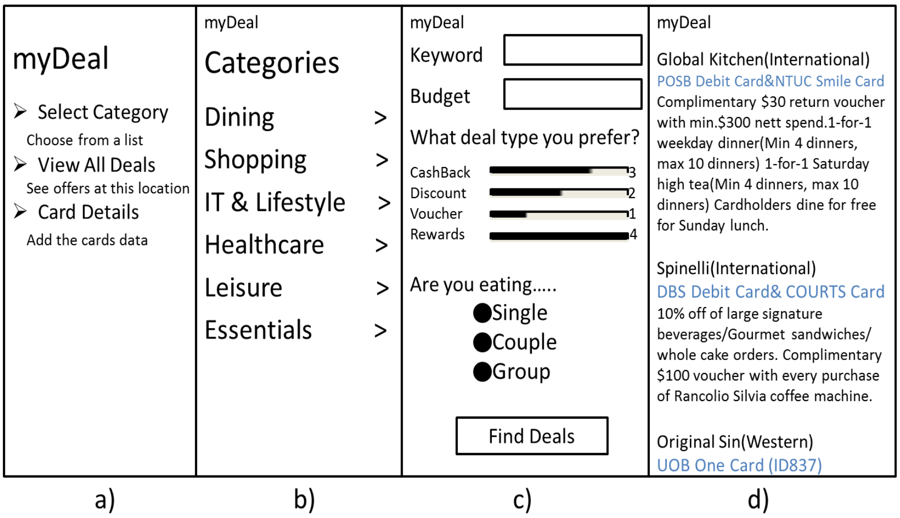


Figure 4.3: myDeal Usage Sequence on the Window Phone

- **Ordering/positioning deals**: This was achieved by applying a score to each deal and displaying deals in order of this score.

4.1.6 End-to-End System

The system works as follows: The user enters a shopping mall and loads the myDeal application (Figure 4.3a). myDeal presents several options that the user can navigate through. The user can either choose to view all deals offered at that location or may choose to specify a particular category of interest (e.g.Dining, Healthcare, etc.) as shown in (Figure 4.3b). The user then proceeds to input additional optional details like type of deal preferred, keywords if any, and desired amount to spend and so on as shown in (Figure 4.3c).

myDeal will extract user's card details from the secure storage area of the phone, append any additional user input and location, and send them to the backend service over an existing wireless communication channel. The ranking service will compute a score which are then ordered and sent to the user's mobile application (Figure 4.3d).

4.2 Validation Plan

In this section, I describe the validation approach used to evaluate myDeal.

4.2.1 Success Criteria

The goal of the validation was to test if myDeal was successful at presenting appropriate deal information to end users in an easy to use manner. To focus the validation, the following criteria were identified as being crucial for myDeal's success:

1. The scoring algorithm is accurate.
2. Users can make deal decisions quickly.
3. Users can find the best deal accurately.

4.2.2 Dataset

The dataset for the algorithm evaluation consisted of real-world dining related deals manually extracted from multiple sources (10 shopping malls and 4 major credit card providers). A total of 842 deals from 610 restaurant covering 7 cuisine types were used in the user study. Also, it was observed from a detailed breakdown of deal component combinations (cash back, discount, vouchers and rewards) 5% of the deals offered cash deals while 25% of the deals offered rewards— validating the decision to specifically handle unstructured free-form rewards in the ranking algorithm.

4.2.3 Participants and setup

The participants were a mix of Accountancy,Business, Economics and Information Systems students. Their demographics are shown in Table 4.1.

The participants were compensated at a flat rate of \$20 SGD ($\approx$ \$17 USD)for completion of the entire set of tasks. The participants were compensated for completion of the entire set of tasks. It was emphasised that they were not under time pressure, and could take as long as they needed to complete the task. This was a deliberate bias against the goal of fast transaction times.

Total Number	43
Gender	Male (18), Female (25)
Major	Accountancy(4),Economics(3), Business(20), Information Systems(12), Social Science(4)
To what extent do you browse for promotions/deals on the Internet from your phone?	Not at All(17), Rarely (9), Sometimes (12), Very Often (5)
To what extent do you use applications that show you promotions/deals near you on the phone?	Not at All(22), Rarely (10), Sometimes (8), Very Often (3),
How important is your phone to you?	Not very Important (2), Somewhat Important(18), Very Important (23),

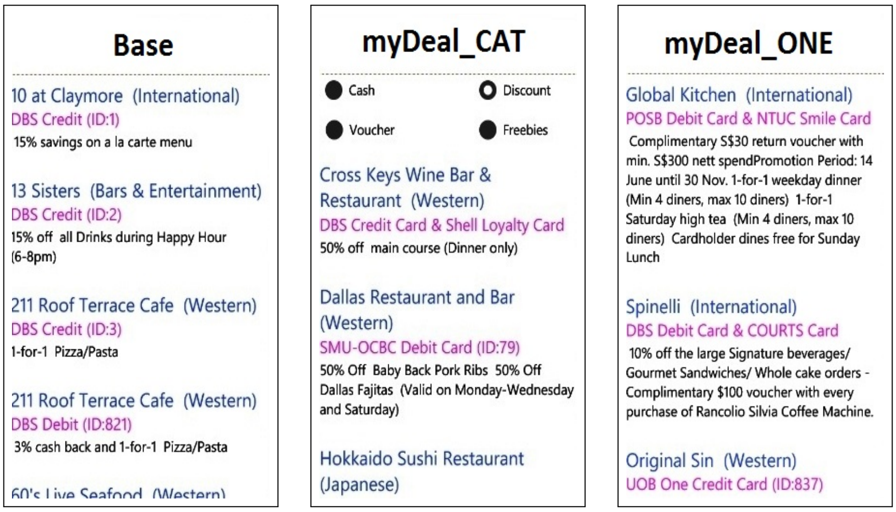
Table 4.1: Demographic Statistics

4.2.4 System Variants

Three variants were used in order to effectively validate the system. The first variant (Base) displays the deals alphabetically (Figure 4.4a) by the retail name. An exact keyword search option is also provided. This baseline is representative of the options available with current state of art applications such as CitiShopper [31] and Mobiqpons [101].

The second variant (myDeal_CAT) builds on top of the first and allows users to view deals categorically by the deal components (e.g view deals that offer discounts, deals that offer vouchers etc.) (Figure 4.4b). Further, within each category deals are sorted according to the numerical value of the deal using only one particular deal component (e.g deals offering 50% discounts will appear higher then those offering 20%). Note that in the myDeal_CAT system we do not calculate aggregate deal score of any kind. Deals that do not include any numerical value would appear below those that do. This variant represents a natural transitional progression between the baseline and the full myDeal system.

The third variant (myDeal_ONE) displays deals (Figure 4.4c) ordered by the



a) b) c)

Figure 4.4: System Variants

aggregate deal score calculated by the algorithm. In addition to a keyword search option users can also input their preferences on the deal components they prefer (e.g., they prefer discounts twice as much as vouchers etc.) as well as provide other information such as their budget. The purpose of the myDeal_CAT variant was to evaluate whether a total rank-ordered scoring of deals was necessary or just a categorical ordering of views was sufficient.

4.2.5 Experimental Procedure

A total of 43 undergraduate students were recruited for the tests. The participants worked alone in a lab for the duration of the study. They were provided a Windows Phone containing all three of our system variants. Each participant was then given the instructions for the study[2] and provided with basic training in how to use the phone. The training period lasted for at most 5 minutes, and consisted of having the participant start the baseline application on the phone and understand the various navigation options.

All 43 students completed the same set of tasks (shown in Table 4.2). Their goal in all tasks was to select the best deal in terms of the overall savings achievable. The

[2]SMU-IRB Approval Number: IRB-10-0088-A0087

Expt Code	Description	Effect Studied
Single	Going for lunch alone. No choice of cuisine.	Ability to identify the most rewarding deal.
Couple	Going for lunch with a friend. No choice of cuisine.	Ability to identify the most rewarding deal for a couple.
S-Focus	Going for lunch alone. Choice of cuisine is ‘Western’. Looking for deals offering rewards.	Ability to identify the most rewarding deal with semi-focused options.
V-Focus	Going for lunch in a group. Choice of cuisine is ‘Chinese’. Looking for deals offering discount and voucher and prefer voucher value over discount.	Ability to identify most rewarding deal when the social setting is a group, with very focused options.
T-Time	Going for lunch alone. No choice of cuisine. Time constraint of 1 minute.	Ability to identify the most rewarding deal in a time constrained situation

Table 4.2: myDeal User Study Experiments

participants were told that they should take into account each component of the deal when calculating the overall savings possible for that deal. The procedure for each task was as follows; first, the participant was given a scenario that they had to follow (e.g., you are eating lunch alone and feel like having chinese food and perhaps an ice-cream cone afterwards). They were then provided with deal information (for several cards and retailers) and were asked to pick the best deal in their opinion. They were free to select any deal (I noted down their final selection) and they were not allowed to ask for any help in the selection process. Also, the overall deal scores computed by the algorithm for each deal were not available to the user. The overall score was intentionally omitted as we did not want to bias the user’s perception. Instead, I wanted to test if the deals that the users thought were the best matched what the algorithm considered to be the best.

During each task, I observed the time taken for selecting a deal for the presented scenario. It was emphasised that they were not under any time pressure, and could take as long as they needed to complete the task. This was a deliberate bias against the goal of fast transaction times. The experiments were first performed on the base system and then followed up with myDeal_CAT and myDeal_ONE. Any learning effect was minimized by randomising between myDeal_CAT and myDeal_ONE.

After each task was completed, participants were presented with a brief questionnaire (that used an easy 5-point Likert scale) that captured their perceived ease of use and accuracy of the completed task. After completing all tasks, users filled an end of experiment questionnaire (Appendix C.5).

Note that in a real scenario, the number of deals presented to the user will be significantly less than that displayed for each of the above test scenarios. Recall that in a real scenario deals are first filtered based on location and would typically correspond to those available in a single mall. For the study however, it was deliberately chosen to ignore location and present the participant with *all* deals that matched the test scenario. This was done to increase task complexity and test the ranking algorithm for a larger subset of deals. Ignoring location also allows us to effectively benchmark our system against applications (such as the Base variant described earlier) that show all available deals irrespective of location.

4.3 Experimental Results

In this section, I present the results of our evaluation. The goal of the evaluation was to determine if myDeal satisfied the success criteria.

4.3.1 Results: The Algorithm is Accurate

The ranking algorithm was evaluated by comparing its scores for the top 10 deals versus the scores of 3 experts using twelve different scenarios. In each scenario, the weights allocated to each of the four deal components was changed . Table 4.3 shows the average difference in rank position and the average error magnitude (for all the errors that were made, what was the average error) for the top 10 deals ordered by the experts with that of the algorithm for all twelve scenarios. The total time taken by the algorithm to score all 842 deals used in the user study was 312ms - well within reasonable limits.

The values in the "Rank Difference" columns indicate the average difference in the score ranks assigned by each entity for each set of deals while the numbers in

Scenario	Weights (%)				Rank	Error
No.	C	D	V	R	Diff.	Mag.
1	100	0	0	0	0.0 (0.0)	0.0 (0.0)
2	0	100	0	0	1.2 (1.03)	2.0 (2.0)
3	0	0	100	0	0.1 (0.0)	1.0 (0.0)
4	33.3	33.3	33.3	0	1.0 (1.05)	2.0 (0.0)
5	50	50	0	0	1.0 (1.05)	2.0 (0.0)
6	50	0	50	0	0.0 (0.0)	0.0 (0.0)
7	0	50	50	0	0.5 (0.7)	1.7 (0.5)
8	0	0	0	100	2.5 (2.0)	3.6 (1.3)
9	25	25	25	25	1.6 (1.26)	2.0 (1.07)
10	25	25	0	50	1.8 (1.8)	2.6 (1.6)
11	25	0	25	50	5.1 (4.2)	5.1 (4.2)
12	0	25	25	50	1.0 (1.4)	2.0 (1.4)

C=Cash Back, D=Discount, V=Vouchers, R=Rewards. Values in parentheses are standard deviations.

Table 4.3: Accuracy of Algorithm Relative to Expert

the second column show that in the event of a rank variation what the average difference would. A lower number in both columns indicate a higher level of agreement between the two entities in terms of scoring.

The first seven scenarios were evaluated without considering the unstructured data part of the deal (Rewards in the form of free text). The results show that the algorithm was effectively able to extract values from the structured data part of the deal. In particular, the largest average difference between the experts and the algorithm was just 1.2 (scenario 2) and even in that case, the average magnitude of the error was just 2 scoring positions (i.e., if the expert ranked a deal 3rd, if the algorithm made a mistake, it would, on average, rank that deal 1st or 5th).

In the last 5 scenarios (scenarios 8 to 12), the deal scores included the free-form text of the deal. The average rank difference and error magnitude in this case is relatively higher. This result was expected as scoring would involve assigning a value to the free text which is subjective. The algorithm uses similarity matching to assign this value while the expert uses heuristics to do the same. Overall however, the average difference is not large enough to invalidate the accuracy of the algorithm.

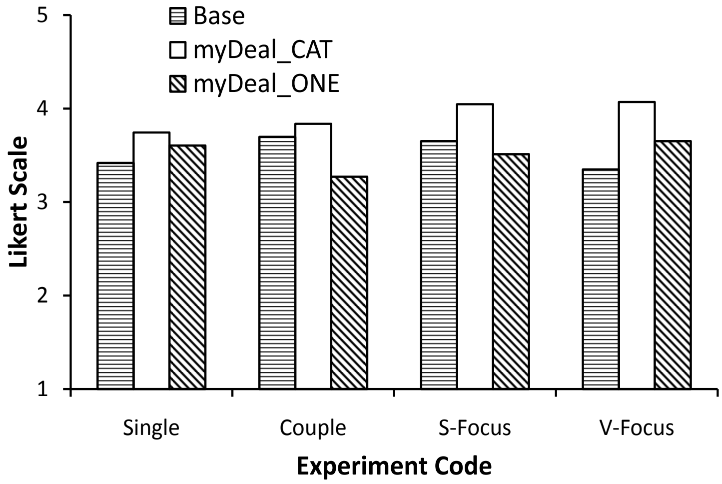


Figure 4.5: Is myDeal Perceived to be Easy to Use?

	Base	myDeal_CAT	myDeal_ONE
Mean (s)	129.35	85.78	110.16
Variance	5034.65	1406.59	1944.76
	Base, myDeal_CAT	Base, myDeal_ONE	myDeal_CAT, myDeal_ONE
P[two-tail]	1.06158E-11	0.00282825	6.73423E-08

t-Test: Two-Sample Assuming Unequal Variances (α=0.05)

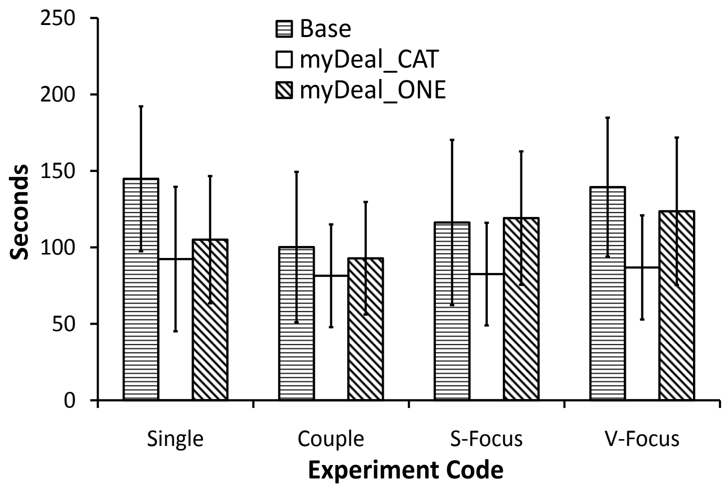
Table 4.4: Mean Time taken to select a deal across all Experiments for each system variant.

4.3.2 Results: myDeal is Easy to Use

Figure 4.5 shows the perceived ease of use of the two myDeal variants and the baseline system. The graph shows the averages of the self-reported Likert score. From the Figure, we see that all three systems are perceived to be easy to use with myDeal_CAT being slightly better than the other two. This is perhaps indicative from the fact that all participants were undergraduate students who are quite mobile savvy and thus quite likely to be comfortable using these types of applications.

4.3.3 Results: myDeal is Fast to Use

Figure 4.6 shows the measured times taken for users to finish each experiment. We observe that the myDeal_CAT times are always lower than the corresponding times



Bars represent standard deviation of the mean.

Figure 4.6: Measured Task Time.

	Base	myDeal_CAT	myDeal_ONE
Mean	35.22965116	36.76744186	44.80813953
Variance	458.6443203	517.1034952	636.9161907
	Base, myDeal_CAT	Base, myDeal_ONE	myDeal_CAT, myDeal_ONE
P[two-tail]	0.518944585	0.000175179	0.002068973

t-Test: Two-Sample Assuming Unequal Variances (α=0.05)

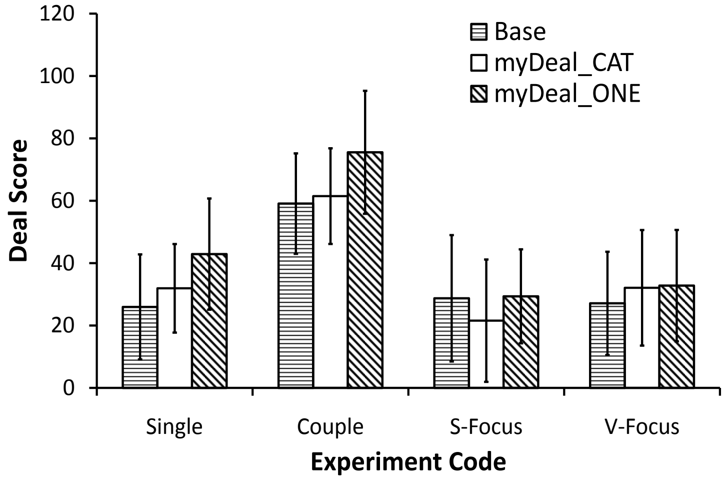
Table 4.5: Mean of the deal scores selected across all Experiments for each system variant.

taken for the other two systems; with myDeal_CAT being significantly lower than the baseline system and myDeal_ONE (Table 4.4). The significant time differences between the myDeal_CAT variant and the myDeal_ONE variant can be explained by the time needed for users to input additional information such as deal preferences in myDeal_ONE.

4.3.4 Results: myDeal is Accurate

The accuracy of each system was measured by comparing the score of the deal chosen by each user for each experiment. Figure 4.7 shows the results of that comparison. These results show that user perception and reality can be very different.

In particular, we see that overall the average deal scores of myDeal_ONE are



Bars represent standard deviation of the mean.

Figure 4.7: Measured Accuracy.

significantly higher than myDeal_CAT and the baseline (Table 4.5)— indicating that the users obtained better deals (higher value overall deals) using myDeal_ONE than the other two systems.

Note that the absolute accuracy of myDeal_ONE is low in certain experiments (S-Focus and V-Focus) as a few users chose alternate, poor scoring deals, to the best deal shown on the mobile device for reasons unknown to us. This is further shown by the histogram in Figure 4.8 that shows that almost 70% of the deals chosen in myDeal_ONE were within the top 20 deals overall. The lower than expected accuracy was due to just a few tragically bad choices where the chosen deals were more than 200 positions away from the best deal.

4.3.5 Results: Best Deal is on Top

A key decision taken was not to display the deal score in myDeal_ONE. This was to prevent any bias in the selection of deals. As the user was free to select any deal that matched the given experiment scenario, the position of the selected deal is compared to the top deals chosen by the algorithm. This provided the second phase evaluation of the algorithm.

Figure 4.9 shows the absolute difference in position of the deal selected by the

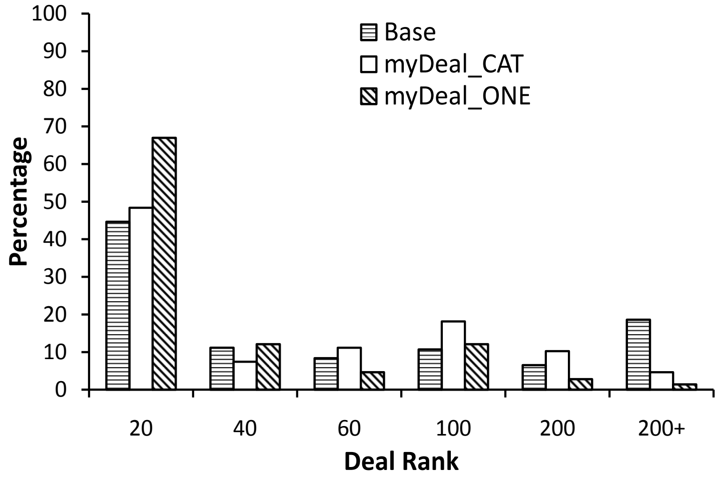


Figure 4.8: Rank Distribution of Deals.

participant from from the top ranked deal for that experiment and Figure 4.10 shows the relative distance between the deal selected by the user and the top ranked deal as displayed on the screen. Clearly most users selected the deals closest to the top most deal (within top 20) in myDeal_ONE.

I analysed the results to understand why most users were only able to pick the 20^{th} or so top deal relative to the algorithm's top choice and identified two main factors; First the users did not really know how to extract the various components of the deal and assign scores to them. Many users were assigning values for Vouchers to Rewards for example. This was a result of our intentional decision to not train users in deal ranking methods and to not show the overall score of myDeal. Second, users frequently picked deals that they thought were good even if the deal itself was not that good. This was to be expected as without an objective guide (like our overall score), humans are quite likely to go with a "gut" feel over what seems to be better. I believe that each of these issues will be corrected when using the full version of myDeal that shows the actual scored ranks with a detailed breakdown of the score components.

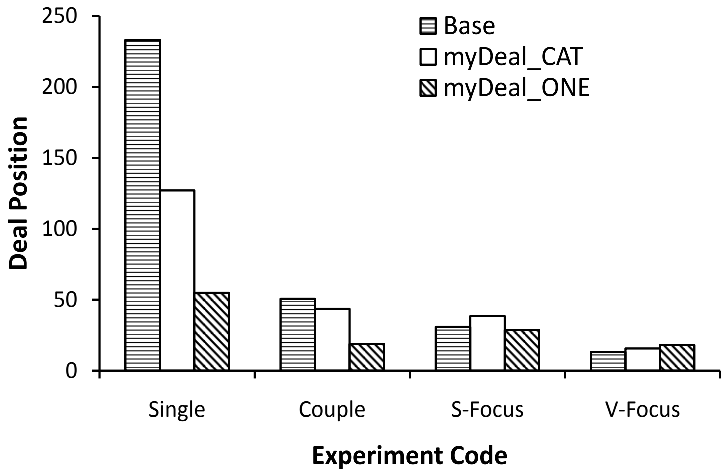


Figure 4.9: Absolute Positioning of Deals.

4.3.6 Summary of Results

Overall, the user study demonstrated that users preferred a combination of the myDeal_ONE and myDeal_CAT UI variants to obtain the best deal information. This is reinforced by the self-reported user scores, shown in Figure 4.11, where every user indicated that both myDeal_ONE and myDeal_CAT were useful systems to have – with most users having an overall positive opinion of myDeal_ONE. This inclination towards myDeal_ONE, as stated by users, was more to do with the additional preference inputs provided, giving a perception of improved performance over the other two variants. Those who preferred myDeal_CAT, preferred the simpler interface over the myDeal_ONE variant.

4.4 Discussion

4.4.1 Time Pressure

In addition to the experiments listed in Table 4.2 each participant was asked to repeat the baseline experiment “Single”. However, this time, they were given a hard time limit of 1 minute. This was to test the system variants in a real-world situation where users do not have much time to look for deals. Figure 4.12 shows the average

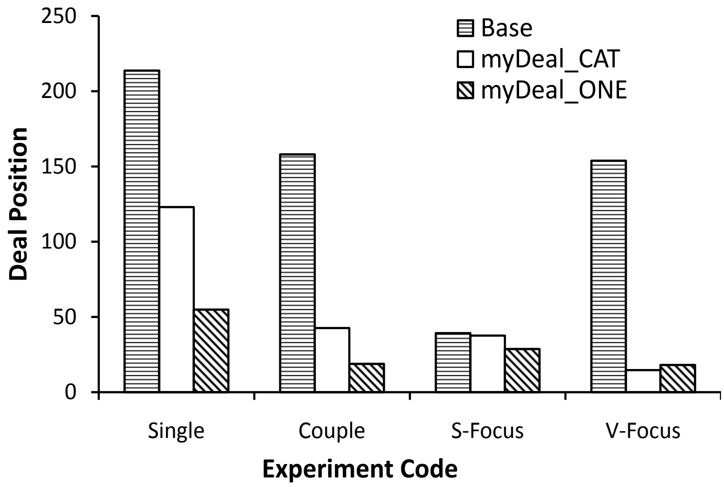


Figure 4.10: Relative Positioning of Deals.

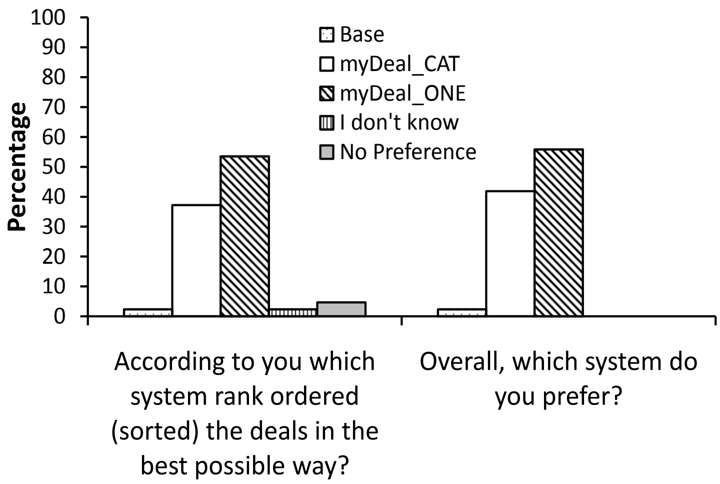
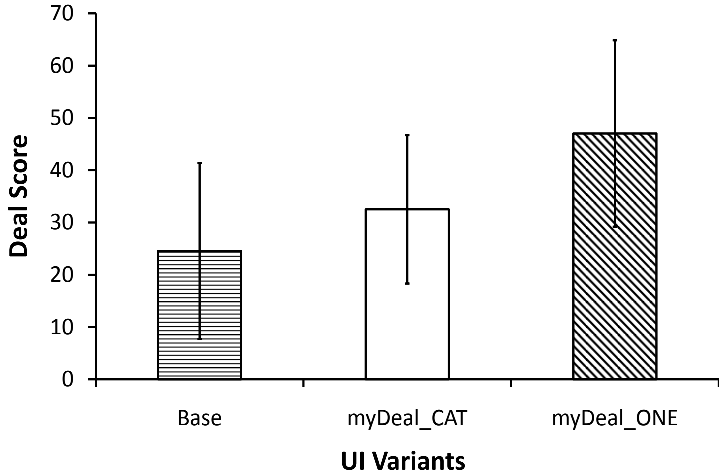


Figure 4.11: Overall Usefulness of myDeal.

score of the deals selected by the participants for each UI variant under this time pressure. It is clear that myDeal_ONE works very well under a time constrained scenario as compared to the other two variants.

Figure 4.13 validates this in terms of the absolute and relative position of the selected deal as compared with the absolute best deal. Figure 4.13 shows that deals selected using myDeal_ONE in a time constrained environment are far more likely to be closer to the best possible deal as compared to the other two systems.



Bars represent standard deviation of the mean.

Figure 4.12: Accuracy in a Time Constrained Scenario

4.4.2 Limitations of the Ranking Algorithm

A key part of the algorithm complexity lies in extracting both the structured and unstructured information parts of the deal. As description of deals are not standard, extraction algorithms are limited by the current state-of-the-art in natural language processing rules and technologies. The current prototype, currently cannot handle deals that include elaborate free-form conditions to satisfy the deal requirement. For example, several deals require purchasing additional items in order to redeem the primary offer. Any additional purchase should effectively reduce the overall score and the algorithm would need to determine the value of this additional item and reduce the overall score by the appropriate factor. My system currently cannot handle these types of deals in a general way.

Secondly, there are several instances where deals are assigned the exact same score. This is obvious as payments card issuers and retailers tend to offer similar deals through a joint promotion (similar to code-sharing in airlines). In this case, deals within the same score cluster are listed alphabetically. This ordering could perhaps bias the user. I propose in subsequent versions to collapse deals with the same score in a manner that would allow the user to make a more informed decision.

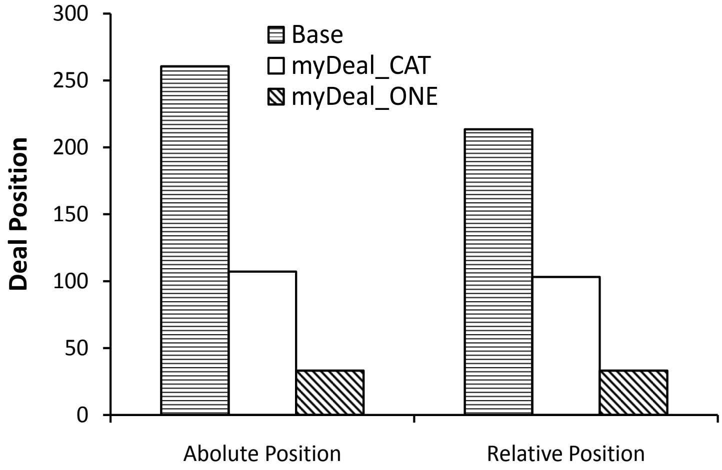


Figure 4.13: Deal Position — Time Constrained Scenario

4.4.3 Real World Deployment Issues

Deployment of myDeal in real world situations is likely to face several challenges. First, a successful deployment of myDeal depends on the extent to which issuers of payment and reward cards along with retailers are willing to share accurate information about their individual promotions. Because myDeal is designed to reduce end user deal searching costs, it is likely that competing agencies might withhold sensitive information rather than willingly share them; creating an operational hindrance for myDeal — an effect observed by previous empirical studies [12]. Next, there is a need for efficient dispute resolution mechanisms between users and retailers when there is a mismatch in the deal information they possess. Such overheads are likely to be a deterrent for the deployment of myDeal.

On the other hand, myDeal can also have potential positive effects on businesses. For example, the positive experience of getting the best deal might boost consumer's satisfaction – possibly resulting in repeat purchases. Similar to the use of other online recommendation systems [5, 129], retailers can utilise myDeal, to improve customer loyalty and differentiate themselves in the market place through customer empowerment.

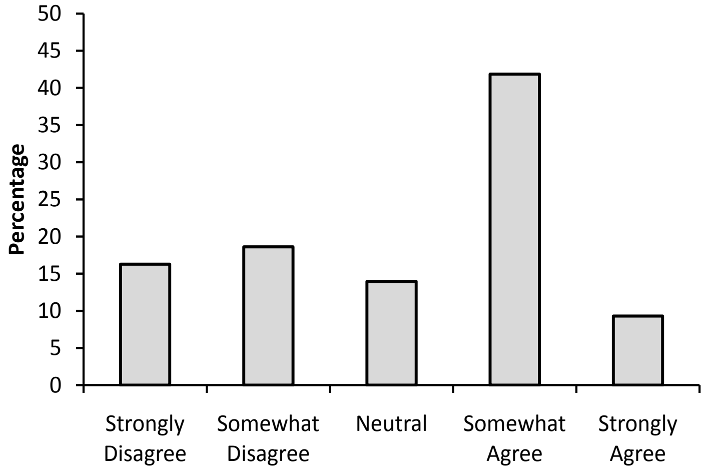


Figure 4.14: Will Users Share Personal Information?

4.4.4 Improving the System

It was intentionally decided not to show the final deal value (as scored by the ranking algorithm) to the participants in the user study. This was an intentional bias against the system as we did not want to influence the user's final deal choice (in ways other than the position of the deal in the system variants). The omission of the final score was particularly noticeable when free-text deal components had to be accounted for as different users assigned very different values to these free-text elements (many users reported not having a basis to assign the values; hence they just made up something). However, the final deployed version of myDeal will present the score values and we believe that will make myDeal significantly better in practice than the results obtained from the user study. The final version will also allow the user to see the detailed breakdown of the final score to determine which parts of the score was provide by which components of the deal in question.

Also, the current prototype only asks users for their weights in relation to the four deal components used in the algorithm (Cash Back, etc.). However, it is possible to improve the accuracy of the algorithm if the user is willing to provide additional information such as their past shopping history, club memberships, frequent flyer preferences, etc. At the end of our user study, we asked each user if they would be willing to provide additional personal information if the deals they received were better for them. The results [Figure 4.14] of the survey indicated that most users were willing as long as they obtained tangible benefits. However, this additional information must be balanced with the need to build a simple to use interface that does not require the user to spend a long time configuring their choices.

4.4.5 Other Limitations

This work has a number of limitations that I am aware of. First, the card descriptions have only been validated qualitatively with a number of real-world cards. It is quite possible that I will need to edit the schema to support deals from a previously unknown real card. However, I feel that the schema is quite complete and is capable of handling most of the deals available today.

Finally, the user study used 43 undergraduates in a controlled lab environment. This leads to a clear bias as a) the sample size is small from a social science perspective, and b) undergraduates tend to be more tech-savvy than the general older population. However, I feel that the results will still be fairly indicative of a large slice of the shopping public. In addition, as the study was performed in a controlled environment it is possible that a real field study will generate different trends and results.

4.5 Summary

A common conundrum in pervasive computing is deciding how to present information to users of these systems. On one hand, users should be provided as much information as possible so that they can make better decisions. However, providing this much information tends to overload the user and have negative usability impacts.

A common technique to reduce this information overload is to use automation (in the form of AI or similar techniques) to process the data and then provide the user with only the pertinent information.

In this Chapter, I attempted to find the sweet spot between automation and user intervention in the specific context of finding the best deal that matched the user's criterion. Within the purview of the experimentation process, this matching algorithm will ensure obtaining the best participant sample given a promotion based intervention. To evaluate this technique, I created myDeal, a system that automatically ranked deals according to user preferences and presented to the user in an efficient way on their mobile device. I first created a XML schema to describe deals and then performed a matching and scoring between the deals preferred by the user and the deals offered by the retailer. The results of this algorithm were then presented to the user on their mobile device. The user was able to scan the deals and pick the one they actually wanted — possibly choosing an option other than the top ranked deal. The results of the user study were very promising and showed that users liked myDeal and were more accurate in picking the best deal when using the system.

Chapter 5

Handling Context Uncertainty

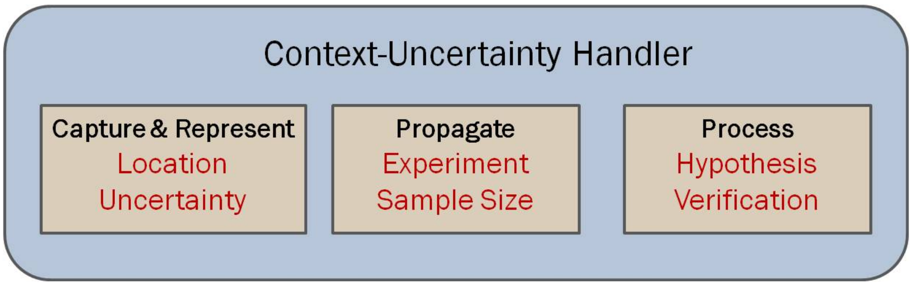
A Gartner report [17] stated that customers and not marketers are driving demand for context-enriched content. Context-enriched content is nothing but the information, data and other content, ranging from articles to advertising to applications, that is based on the user's context. The context here being, relevant facts about current conditions that are true in the moment but may not be in the future.

Early discussion of context-aware applications has shown the ability to use mobile sensing to infer a variety of context and build applications designed to respond in real time to changes in personal situation [48, 53, 84, 158]. However, these studies are not representative of reality as they are controlled, often with a limited set of users restricted to specific campus/office environments. To bridge this gap I build Jarvis, a platform that facilitates a better understanding of user needs through large-scale, in situ, real-time experimentation that require context-specific triggers. The goal of Jarvis is to provide experimenters access to deeper, near-real time user context (e.g., location, activity) without the hassles of experimentation such as subject selection, bias and so on.

Of the many possibilities, a use case I envision for Jarvis, is providing retailers a platform to run *lifestyle* based experiments. For example, a coffee shop owner may want to test whether offering discount coupons to people who have been waiting outside the coffee shop for at least 10 minutes, will improve sales. However, a key challenge in running such experiments is that the trigger events are derived from context collected using built-in sensors on the mobile device. These sensors have inherent uncertainties associated with them and as a result can include people who do not satisfy the experiment criteria [6]. Continuing with the previous example, discount coupons could be sent to people who are in fact *not* outside the coffee store but are reported to be by the system as a result of localization error. It is therefore pertinent to arm experimenters with sufficient information of the possible impact of context uncertainty on the outcome of their experiment. For example, informing the experimenter that 2% of the subjects might have falsely satisfied the event conditions will assist them in defining the success criteria of their test. Further, defining a confidence metric for each individual subject, who satisfies the experiment requirements, provides a better understanding of the relationship between the experiment parameters. For example, if subjects considered to have a high context-confidence redeem the discount coupon, we can conclude a strong correlation between the event attributes (standing outside the shop for 10 minutes) and the content delivered. This information is important, not only for understanding user behaviour towards context-based interventions, but also towards building better context-aware systems and applications.

Providing such information unfortunately, is not trivial. The challenges are two fold: 1) Not all context generators provide the necessary information directly. Indoor localization systems such as Radar [11] and EZ [29] for example, do not measure how often the system incorrectly estimates users' location to a given landmark (false positives) and 2) Context uncertainty is highly dynamic and individual. For example, activity classification accuracy is dependent on the activity being classified as well the device being used. It would therefore be incorrect to have a static interpretation of error for a given context source. While techniques of increasing context confidence through redundancy or sensor fusion exist, they do not completely eliminate the need to handle context uncertainty.

In this chapter, I describe the design of the Uncertainty-Handling module within Jarvis (Figure 5.1) and show how it handles context uncertainty at the different stages of an experiment. First, I show how the module defines a confidence metric



Handling context uncertainty at the different stages of the experimental process

1. Capture & Represent: A location context confidence metric is defined for every participant returned by the Event Processing Agent (Section 5.1).
2. Propagate: During an experiment run, the context confidence metric is used to ensure no statistical bias exists between experiment groups with regards to context confidence. Participants are selected till this goal is reached, impacting the experiment sample size (Section 5.3).
3. Process: Once the experiment is completed, the experiment hypothesis is verified incorporating the context uncertainty information (Section 5.5).

Figure 5.1: Sub-components of the Uncertainty Handling module.

for the location predicate as well as how it stochastically estimates additional information such as the number of false positives. Next I describe how the module uses this confidence metric to determine the appropriate sample size for an experiment, ensuring that there is no statistical bias between experiment groups with regards to context uncertainty. Finally, I show how the module verifies the experiment hypothesis given the additional context uncertainty information. In doing so, I provide adequate information to the experimenter to process the results of an experiment - allowing them to either re-run the experiment (with new parameters and constraints), run a new experiment, or declare success.

5.1 Handling Location Uncertainty

Event processing systems often have to deal with various types of uncertainty, such as the following [7, 8]:

- Incomplete event streams: Consider, for example, the case in which an activity recognition system fails to detect a human activity at some point in time (say, due to occlusion).

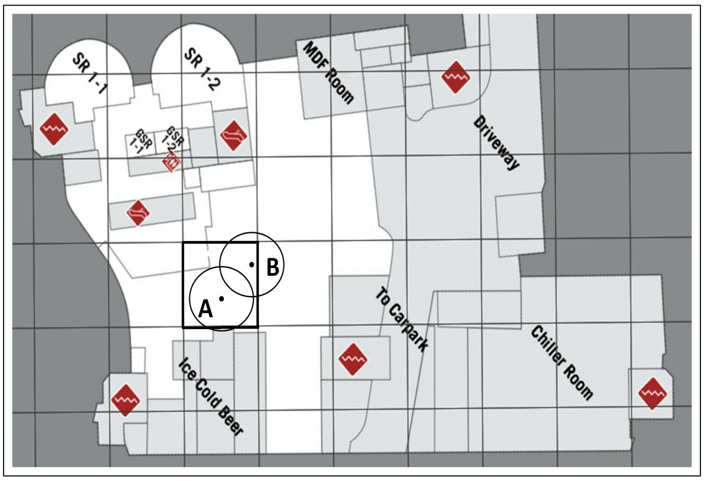
- Insufficient event dictionary: The recognition of a composite event may require the detection of some other events that cannot be detected by the event producers or EPAs available at an application domain. Continuing with the previous example, the activity recognition system may not be able to detect all types of activity necessary for detecting a cooking activity.

- Erroneous event recognition: For example, a persons current location may be approximated, and as a result the interpretation of a situation may be mistaken.

- Inconsistent event annotation: Inconsistencies in the annotation of composite events in the training dataset used for machine learning.

- Imprecise event patterns: In many application domains, we only have an imprecise account of the pattern of a composite event. For instance, it may not be possible to precisely identify all conditions in which an activity of a particular type is said to take place (For example, when do you eat popcorn?).

As a result, traditional event processing systems need to be enhanced to handle these uncertainties. Uncertainty handling methods can be roughly divided into two main approaches - the first approach is uncertainty propagation, according to which the uncertainty of the input events is propagated to the derived events in a coherent way from a mathematical (probabilistic) perspective. In contrast, the second approach is to eliminate the uncertainty, whenever it arises, before the derivation is carried out (This process should not be confused with data cleansing.). Uncertain attributes are replaced by deterministic equivalents, and uncertain events may be screened out, according to some predefined policy. An exemplary policy is to apply a threshold, such that any event with a certainty smaller than this threshold should



A floor map of our building overlayed with the location of two participants that are potential targets for a coupon from Ice Cold Beer. The inner dot represents the system detected coordinates of the participants while the outer ring represents the location error radius.

Figure 5.2: Individual location confidence.

be discarded, or otherwise treated as certain. When the uncertainty is removed, the events can be processed regularly.

The first step in handling uncertainty is to therefore represent context attributes by probabilities and distributions rather than standard native types. Replacing uncertain context attributes by deterministic equivalents provides for a more efficient handling of event uncertainty. In this section, I describe how I associate a probability (in terms of a confidence metric) with the location context attribute.

5.1.1 An overview of the LiveLabs Indoor Localization system

As part of LiveLabs, we currently (or aim to) track participant location indoors at four venues - an international airport, a university campus, a resort island and a convention centre. In order to support multiple mobile OS platforms, our localization system employs a 'reverse fingerprinting' technique by Khan [74]. In their approach, rather than relying on the Wi-Fi AP signal strength readings reported by a mobile device, they use an infrastructure-assisted solution based on querying the commercial Wi-Fi controller infrastructure.

Every floor is divided into multiple zones (e.g., Shops, Classrooms) and each zone contains multiple landmarks. Identifying a participant's location means associating the participant to a landmark and in turn to a zone. Participants are considered to have satisfied the event location condition if their location is *contained in* or *touches* the zone defined in the experiment [46]. The inter-landmark distance is approximately 3 meters for the university campus and 6 meters for the shopping mall.

Using this localization technique, we observe an average location error radius of two landmarks approximately 70% of the time. There is however a caveat when computing the confidence of the location attribute. We observe that the location error distribution depends not only on the venue (and zone), but also varies with time of day and day of week. As a result, using the *static* system defined accuracy is not sufficient. Further, each individual has a certain location confidence based on their current position and error radius as reported by the system. For example in Figure 5.2, participant *A* reported to be at the center of a store, should have a higher probability of actually being within the experiment defined location area (even with location error) as opposed to participant *B* reported to be at the edges. It therefore becomes necessary to compute individual location confidence based on realtime observations and environment conditions. In the next section I will describe how my algorithm defines the location confidence for each participant as well as how it estimates the number of false positives within the set of participants that satisfy the location condition.

5.1.2 Capture & Represent Location Uncertainty

When generating a report, there are two pieces of information needed to process the outcome of an experiment. For every participant satisfying the event we need:

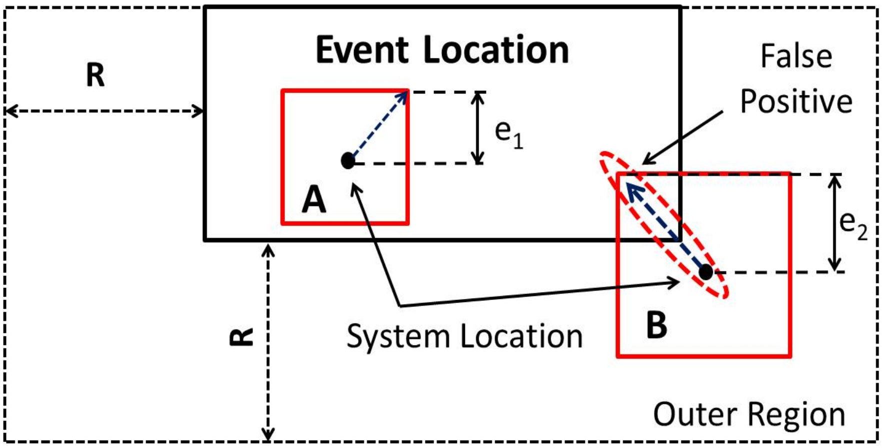
1. The confidence of each event attribute specified as part of an experiment.

2. The number of cases in which the event did not occur in reality (false positives).

In particular when dealing with the location attribute, we want to know the location confidence of each participant as well as the number of participants that *might not* have satisfied the location requirement i.e, their location is reported incorrectly to be within the event location. Note, it is not possible to infer the set of false positives based on the confidence value alone i.e, participants whose location confidence is low does not necessarily imply that the participant is originally from outside the event location. The number of false positives depends on several factors such as the area of the event location, the population density of participants within and outside the event location, the distribution of these participants as well as the location error distribution.

The input to the algorithm is sensor data about the location of people that *satisfy* the location condition. This data is in a (x,y) co-ordinate format, with respect to the given building, along with the error radius detected at that location. Similar to Ranganathan et al. [126] all locations are expressed by the algorithm as minimum bounding rectangles. While approximating sensor regions with minimum bounding rectangles decreases the accuracy of location detection, the advantages in terms of performance and simplicity far outweigh the loss in accuracy. We then compute for each participant, what fraction of their minimum bounding rectangle intersects with the event location. We define this overlap ratio (ranging between 0 and 1) as the location confidence of the participant with the intuition that, larger the overlap, more likely is the participant to have satisfied the location condition specified in the experiment. Thus Participant A in Figure 5.3 has a location confidence of one, as the bounding rectangle is fully contained within the event location. This confidence value also serves as a first step to filter participants that do not meet the confidence requirement set by the experimenter. Note that the algorithm does not compute the location confidence for Participant B in Figure 5.3 as it is not within the set of participants that satisfy the event location condition.

The second part of the algorithm uses Monte-Carlo methods to estimate the number of false positives - the number of participants that in reality were not within



R is the maximum error radius observed at that location. e_1 and e_2 are the respective location error radius of Participant A and Participant B such that $e_1 \leq e_2 \leq$ R .

Figure 5.3: Setting up the environment to compute the location confidence and the false positives of an event.

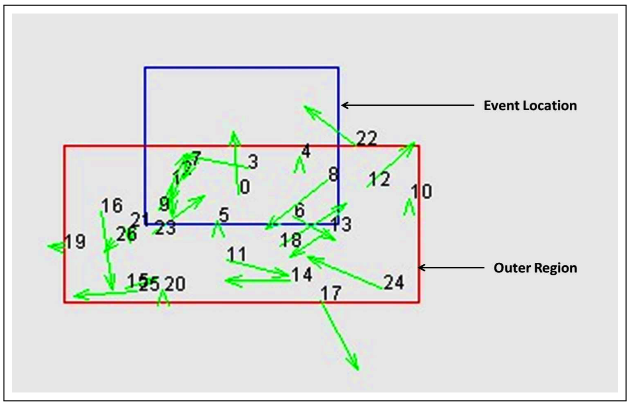
the event location. This is done through the following steps:

1. Define a region outside the event location. This is shown as a dotted line in Figure 5.3 surrounding the event location. The dimensions of this region is proportional to the maximum error radius R observed at that location by our indoor localization system.

2. Retrieve the system location of all participants within these two regions.

3. Apply a location error with a given error distribution, across all participants, thereby shifting the participants from their system location. As a result of this location shift, participants that were outside the event location can now be within. This is shown in Figure 5.3 with Participant B moving into the event location as a result of this shift. Such participants constitute the false positives of the system.

The number of false positives are estimated by emulating the test environment and observing all possible permutations of participants' location, under the given conditions. To do this, The third step is repeated 1000 times, each time shifting a participant from their system location and capturing the number of false positives



The arrow indicates the transition of a participant from their true location to the system location as a result of error. The length of the arrow represents the location error magnitude.

Figure 5.4: Recreating the event environment using a simulator.

as a result of this shift. For a given event location, error distribution and participant spread within and outside this region, the algorithm estimates the number of false positives as the average across these iterations. Thus if the system reports ten participants to be within the event location and the algorithm estimates two as the average number of false positives, I report that 20% of the participants were likely not within the event location. Note, as this step does not filter any participants, it is done post the experiment (e.g., after the coupon has been sent to participants and the behaviour is observed) when generating the final report for the experimenter.

5.2 Experimental Results

5.2.1 Experiment Setup

To evaluate the algorithm I require the ground truth information of each participant i.e, was the participant within the event location in reality. To get around this requirement I built a simulator to recreate the experiment environment. To do this, the simulator takes multiple inputs such as the event location dimensions, the number of participants, population density as well as the location error distribution.

Based on these inputs a set of participants are generated and placed uniformly across the simulated environment. The current position of each participant consti-

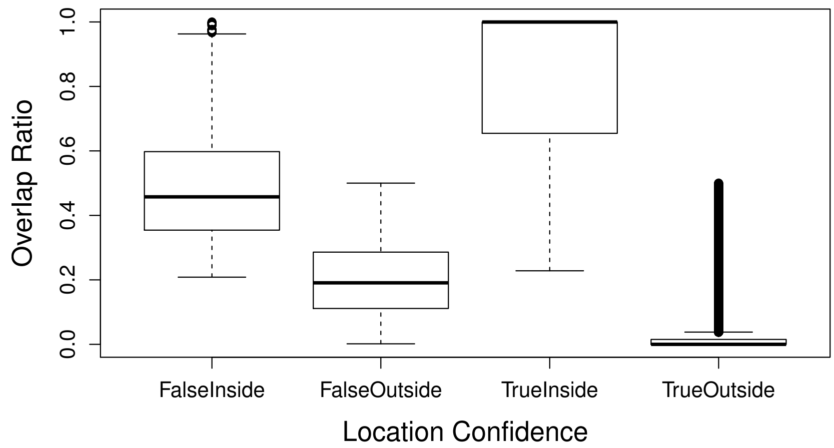


Figure 5.5: Box Plot capturing the distribution of overlap ratio across the four location classes of participants.

tutes the *true location* or *ground truth* of that participant. A given error distribution is then applied across all participants shifting them from their true location. The resulting position of each participant now constitutes the *system location* of that participant. Figure 5.4 shows a screenshot of the simulator. The arrow captures the shift of each participant from their true location to their system defined location. The length of the arrow represents the magnitude of the location error. Given the (true location, system location) pair, we can evaluate the accuracy of the algorithm in estimating the number of false positives in a given event as well as measuring the reliability of using overlap ratio to represent location confidence.

5.2.2 Results: Using overlap ratio to represent location confidence

To evaluate the accuracy of using overlap ratio as our location confidence metric, participants are divided into four classes: 1) *TrueInside*-Participants whose true location and system location coordinates are inside the event location, 2) *TrueOutside*-Participants whose true location and system location coordinates are outside the event location, 3) *FalseInside*-Participants whose true location is outside the event location but the system location is within the region and 4) *FalseOutside*-Participants whose true location is within the event location but the system location is outside the region.

Note, we are truly only interested in two classes of participants, *TrueInside* and *FalseInside*, as these are the set of participants considered to have satisfied the event location condition. We however include all four classes in our observation of whether overlap ratio - area of intersection with the event location by the area of the minimum bounding rectangle as defined by the system location- is a good indicator of location confidence.

The overlap ratio was captured for ten different scenarios. The location error radius (0 to R) was uniformly distributed (discrete) across the participants, with the maximum location error radius R, ranging from 3 to 12 meters for each scenario. Each scenario was repeated 100 times, randomly generating the number of participants (min. 20, max. 100) during each run. The event location dimensions remained constant for the complete experiment, while the dimensions of the region outside the event location varied based on the maximum error radius for the given location error distribution scenario.

Figure 5.5 captures the distribution of the overlap ratio across all four location classes. We observe that the box plot for each class does not overlap significantly, suggesting the use of the overlap ratio to differentiate between the different classes of participants. Thus participants with a higher overlap ratio are more likely to be within the event location than participants with a lower ratio. I further evaluate its classification capability by building a Naive Bayes model in Weka using overlap ratio as the feature . The resulting model provides an accuracy of 84.6% in classifying a participant's true location based on the overlap ratio. Unfortunately, classification errors do still exist - Table 5.1 captures the confusion matrix. Of interest are those participants classified as *TrueInside* when in reality they should be classified as *FalseInside* - which is the reason why we attempt to estimate the number of false positives. However, despite these errors, we still consider overlap ratio to be a good representation of location confidence. Table 5.2 summarizes the mean and standard deviation of each class observed during the simulation run.

	TrueInside	FalseOutside	TrueOutside	FalseInside
TrueInside	5054	191	2	956
FalseOutside	0	991	2370	436
TrueOutside	0	1558	34172	456
FalseInside	811	717	62	1324

Table 5.1: Confusion Matrix: Using overlap ratio to classify a participants true location.

Class	Mean	SD	SE
FalseInside	0.4848	0.1724	0.0032
FalseOutside	0.2049	0.1177	0.0019
TrueInside	0.8315	0.2171	0.0027
TrueOutside	0.0391	0.0857	0.0004

All differences are significant (using student's t-test with $p = 0.05$).

Table 5.2: Mean and Standard Deviation of overlap ratio across the different location classes.

5.2.3 Results: Accuracy in estimating the number of False Positives

Given that the simulator captures both the true location and system location of each participant, we can compare the *true* number of false positives with the value *estimated* by the algorithm. We run the simulator for ten different uniform location error distribution scenarios (R=3 to 12 meters), with each scenario repeated 100 times. The number of participants (min. 20, max. 100) were randomly generated during each run. For each run, the number of false positives was then computed using the algorithm described in Section 5.1. Table 5.3 captures the output of the simulator for a single run.

Figure 5.6 shows a cdf plot of the percentage error in estimating the number of false positives by the algorithm. We observe that 80% of the time the algorithm can accurately estimate the number of false positives in an event with an error less than $\pm 25\%$.

P(N)	P(E)	FP(True)	FP(Estimate	Estimate Error (%)
20	8	0	1	12.5

Where *P(N)* is the total number of participants during the simulation run, *P(E)* is the number of participants that satisfied the Event Location condition and *FP()* is the number of false positives.

Table 5.3: Simulator Output for a single run with maximum error radius *R=3m*.

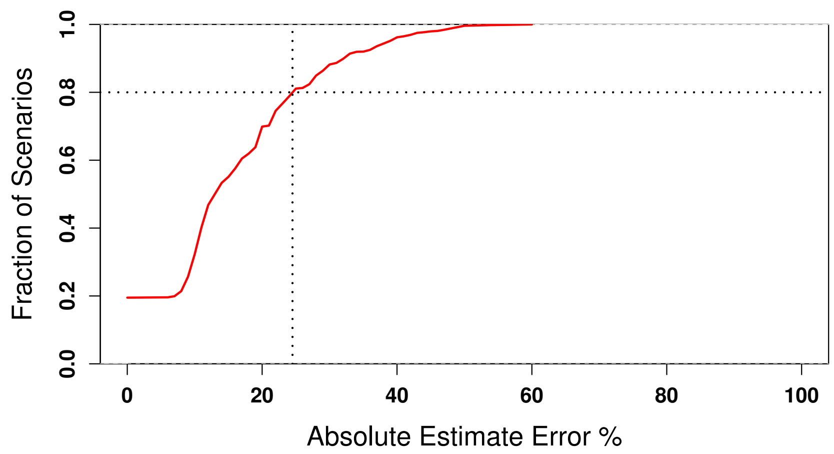


Figure 5.6: CDF of the % error in estimating the number of participants that did not satisfy the event conditions (false positives).

5.2.4 Discussion

In order to represent the location confidence metric, with respect to the given event location, the above algorithm requires as input the error radius detected at that location. This value is then used to construct the minimum bounding rectangle. However, several indoor localization systems [10, 11, 29], including the one used by LiveLabs does not directly provide the location error. RADAR [11], for example, creates a radio map of the indoor environment to estimate the users location. Here, the system instead, provides a probability vector of the *likely* location across different (nearest) landmark points. Using this information we can then derive the location error. For example, the error radius can be computed as the euclidean distance between the user's most likely and second most likely location, as estimated by the system.

This need to derive the necessary information raises an important requirement of handling uncertainty - the need to provide an interface that translates information across different context providers to that required by the system. Further, the current algorithm explores only two pieces of information, location confidence and estimating the number of false positives, required in handling event-based uncertainty. It would be wortwhile to understand other such metrics that can aid reasoning under context uncertainty.

Also, while this work explores handling uncertainty of the location attribute, similar effort is needed in representing the uncertainty of other context attributes supported by the experimentation system.

5.3 Handling Sample Size under Uncertainty

In Section 5.1, I introduced the Uncertainty-Handling module within Jarvis. I showed how the module defines a confidence metric for the location predicate as well as how it stochastically estimates additional information such as the number of false positives. However, while the module represented location uncertainty it failed to act upon it. Instead, the system relied on the experimenter to evaluate the impact of uncertainty on the experiment hypothesis.

I now extend the Uncertainty-Handling module within Jarvis to act on context uncertainty in addition to providing context confidence (uncertainty) information. In particular, I look into how we can dynamically compute the sample size for a given experiment under context uncertainty. Traditionally, sample size is computed under the assumption that the samples obtained are not tainted in any way. For example, a survey of 300 students assumes that all participants surveyed are indeed students. However, when automating subject selection for experiments, for example using mechanical turk [113], subjects can often be misrepresented, having an impact on the validity of the experiment [76]. As a result of such tainted samples, the sample size specified for the experiment may not be sufficient. Hence it becomes imperative to handle sample size selection under uncertainty and continue running the experiment till valid samples are obtained.

5.3.1 Why is Experiment Sample Size important?

Sample size determination is the act of choosing the number of observations or replicates to include in a statistical sample. The sample size is an important feature of any empirical study in which the goal is to make inferences about a population from a sample. In practice, the sample size used in a study is determined based on the expense of data collection, and the need to have sufficient statistical power. In complicated studies there may be several different sample sizes involved in the study: for example, in a survey sampling involving stratified sampling there would be different sample sizes for each population. In a census, data are collected on the entire population, hence the sample size is equal to the population size. In experimental design, where a study may be divided into different treatment groups, there may be different sample sizes for each group.

Sample sizes may be chosen in several different ways:

1. expedience - For example, include those subjects readily available or convenient to collect. A choice of small sample sizes, though sometimes necessary, can result in wide confidence intervals or risks of errors in statistical hypothesis testing.

2. target variance - standard deviation is the measure of dispersion or variability in the data. While calculating the sample size an investigator needs to anticipate the required variation in the measures that are being studied. For example, a smaller sample is sufficient if the population is more homogenous and therefore has a smaller variance or standard deviation.

3. target power - statistical power is the likelihood that a study will detect an effect when there is an effect there to be detected. Sample size can therefore be calculated so that enough subjects can be recruited to give the results

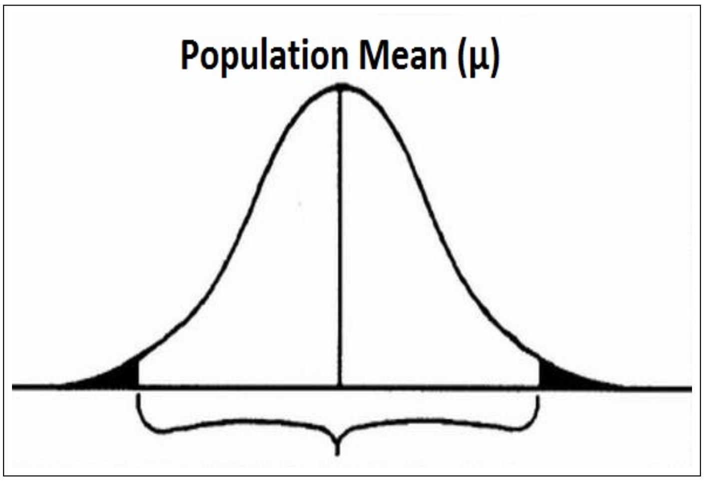
adequate power. However this may require making assumptions about the desired effect size and variance within the data.

5.3.2 Computing Sample Size

In addition to the purpose of the study and population size, three criteria usually will need to be specified to determine the appropriate sample size: the level of precision, the level of confidence or risk, and the degree of variability in the attributes being measured [97].

- **The Level of Precision:** The level of precision or sampling error, is the range in which the true value of the population is estimated to be. Thus, if a researcher finds that 60% of students in the sample have adopted a recommended practice with a precision rate of 5%, then he or she can conclude that between 55% and 65% of students in the population have adopted the practice.

- **The Confidence Level:** The confidence or risk level is based on concepts captured in the Central Limit Theorem. The key idea in the Central Limit Theorem is that when a population is repeatedly sampled, the average value of the attribute obtained by those samples is equal to the true population value. Furthermore, the values obtained by these samples are distributed normally about the true value, with some samples having a higher value and some obtaining a lower score than the true population value. In a normal distribution, approximately 95% of the sample values are within two standard deviations of the true population value [33]. In other words, this means that if a 95% confidence level is selected, 95 out of 100 samples will have the true population value within the range of precision specified earlier (Figure 5.7). There is always a chance that the sample you obtain does not represent the true population value. Such samples with extreme values are represented by the shaded areas in Figure 5.7. This risk is reduced for 99% confidence levels and



95% of sample means within two standardized deviations.

Figure 5.7: Distribution of Means for Repeated Samples

increased for 90% (or lower) confidence levels.

- **Degree of Variability:** The degree of variability refers to the distribution of attributes in the population. The more heterogeneous a population, the larger the sample size required to obtain a given level of precision. The less variable (more homogeneous) a population, the smaller the sample size. Note that a proportion of 50% indicates a greater level of variability than either 20%or 80%. This is because 20% and 80% indicate that a large majority do not or do, respectively, have the attribute of interest. Because a proportion of 0.5 indicates the maximum variability in a population, it is often used in determining a more conservative sample size, that is, the sample size may be larger than if the true variability of the population attribute were used [94].

There are several approaches to determining the sample size. These include using a census for small populations, imitating a sample size of similar studies, using published tables, and applying formulas to calculate a sample size [66]. For populations that are large, Cochran [32] developed an equation to give a representative sample for proportions.

$$n_0 = \frac{Z^2pq}{e^2} \tag{5.1}$$

Where n_0 is the sample size, Z^2 is the abscissa of the normal curve that cuts off an area α at the tails (1 - α equals the desired confidence level, e.g., 95%), e is the desired level of precision, p is the estimated proportion of an attribute that is present in the population, and q is 1-p. The value for Z is found in statistical tables which contain the area under the normal curve.

For example, suppose we wish to evaluate a program in which students were encouraged to adopt a new practice. Assuming we do not know the variability in the proportion that will adopt the practice: therefore, p=0.5 (maximum variability). Furthermore, suppose we want a 95% confidence level and 5% precision. The resulting sample size is computed as:

$$n_0 = \frac{Z^2pq}{e^2} = \frac{1.96^2(0.5)(0.5)}{0.5^2} = 385 Students \tag{5.2}$$

5.3.3 Computing Sample Size *under Uncertainty*

Note that handling sample size selection under uncertainty becomes important when considering experiments with more than one group. For experiments with a single population (100% treatment) using a certainty threshold, such that any event with a certainty smaller than this threshold will be discarded, or otherwise treated as certain is sufficient. For two sample experiments the degree of uncertainty within each population should be comparable in order to test the experiment hypothesis. This rules out any confounding effects between the two groups due to context confidence.

To do this we compute the confidence interval on the difference between means. The difference of means is traditionally used for hypothesis testing. A confidence interval for the difference between two means specifies a range of values within which the difference between the means of the two populations may lie. These intervals may be calculated by, for example, a producer who wishes to estimate the difference in mean daily output from two machines; a medical researcher who wishes to estimate the difference in mean response by patients who are receiving two different drugs; etc. The confidence interval for the difference between two means contains all the values of (μ_1-μ_2) (the difference between the two population means) which would not be rejected in the two-sided hypothesis test of:

H_0: μ_1=μ_2 against H_a: $\mu_1 \neq \mu_2$ i.e, H_0: μ_1-μ_2 = 0 against H_a: μ_1-$\mu_2 \neq 0$

If the confidence interval contains zero (more precisely, the parameter value specified in the null hypothesis), then we can say that there is no significant difference between the means of the two populations at the given level of confidence. Whenever an effect is significant, all values in the confidence interval will be on the same side of zero (either all positive or all negative). Therefore, a significant finding allows the researcher to specify the direction of the effect. We use this value to dynamically verify whether the current samples of the two populations are comparable and if not continue sampling till it is. Note the verification process will start when the initial target sample size is reached and will thereafter be re-computed after a participant has been added to each of the population groups with a context confidence metric below a given threshold. For example, if one group has a mean context confidence of 80% and the second group has a mean context confidence of 60%, participants are added *only* to the latter group (if the threshold set is 80%). As a result, the groups may not be of equal size.

Difference between means

Suppose we have two populations with means equal to μ_1 and μ_2. Suppose further that we take all possible samples of size n_1 and n_2. And finally, suppose that the following assumptions are valid:

- The size of each population is large relative to the sample drawn from the population. That is, N_1 is large relative to n_1, and N_2 is large relative to n_2. (In this context, populations are considered to be large if they are at least 10 times bigger than their sample.)
- The samples are independent; that is, observations in population 1 are not affected by observations in population 2, and vice versa.

- The set of differences between sample means is normally distributed. This will be true if each population is normal or if the sample sizes are large (n>30).

Given these assumptions, we know the following. The expected value of the difference between all possible sample means is equal to the difference between population means. Thus, E(x_1 - x_2) = μ_d = μ_1 - μ_2. The standard deviation of the difference between sample means (σ_d) is approximately equal to:

$$\sigma_d = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{5.3}$$

Confidence Interval on the Difference Between Means

When the population variances are known, the 100(1-α) percent confidence interval for μ_1-μ_2 is given by

$$(x_1 - x_2) \pm z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{5.4}$$

where z is the z-statistic and x_1 and x_2 are the sample means.

Statistical Power

There are two main reasons why a study may not show a significant difference between groups being studied (e.g. in a randomized trial of a new drug, or a case-control study testing the effect of location proximity on coupon redemption).

1. There really was no significant difference (hence a true negative result).

2. There was a difference but the study failed to detect it (false negative result). This may arise because the study was poorly designed (e.g. used imprecise measurements) or because the study was too small (in statistical jargon, it "lacked power").

The power of a study is therefore its ability to detect a difference, *if the difference in reality exists*.

Statistical power is affected chiefly by the size of the effect and the size of the sample used to detect it. Bigger effects are easier to detect than smaller effects, while large samples offer greater test sensitivity than small samples. Thus in addition to ensuring that the experiment groups are comparable in terms of context confidence, we must also ensure that this process does not impact the power of the study. This means adding additional participants to groups with a lower sample size. A tradeoff for increasing the statistical power of the study is the possibility of decreasing the mean confidence of a group. To illustrate, in Section 5.3.3, we stated that participants will be added to each of the population groups with a context confidence metric below a given threshold. This step can result in uneven group sizes which in turn will impact the power of the study. However, to counteract this effect, adding participants to a group with a current high mean context confidence, but low sample size, can decrease the overall confidence mean - a tradeoff we need to accept. Note, we chose a power of 80%; hence a true difference will be missed 20% of the time.

5.4 Experimental Results

5.4.1 Experiment Setup

I evaluated the algorithm for computing sample size, under the effect of context uncertainty as follows. The algorithm takes as input the number of experiment groups and the target sample size. Participants, with a context confidence following a normal sampling distribution with μ=0.5, are added to each group. One of the experiment groups include participants with a low context confidence (<30% confidence). For the evaluation, the starting sample size is varied between 10 and 100 (in increments of 10) and the percentage of participants with low context confidence (uncertainty factor) is varied between 10% and 90% (in increments of 10%) thereby increasing the group uncertainty. Thus a trial with a sample size of 50 and an uncertainty factor of 70% will include 35 participants with a context confidence

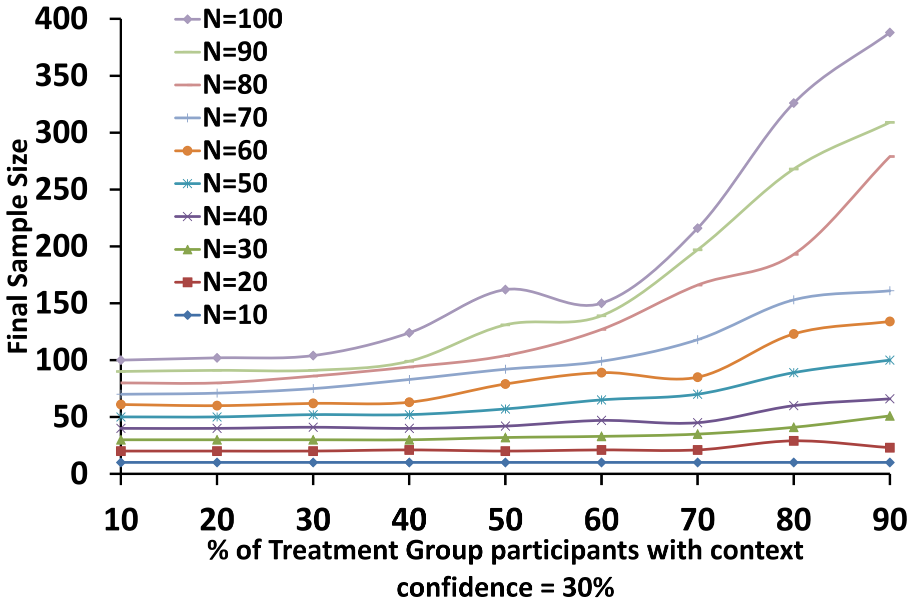


Figure 5.8: Computing sample size under context uncertainty for two groups. Original target sample size range N=10 to 100.

of 30%. For a given starting sample size and uncertainty factor, the final sample size is computed as an average of 100 trials.

5.4.2 Results: Using difference between means to compute sample size under Context Uncertainty

Figures 5.8 and 5.9 show the final sample size computed for an experiment with two and three groups respectively. As expected, when the uncertainty factor is high additional participants need to be included in the experimental group to ensure the different groups are comparable in terms of the context confidence.

5.5 Drawing an Experiment Conclusion under Uncertainty

An important goal of our system is to provide a platform that facilitates getting a better understanding of user behaviour towards context-based systems, through a process of experimentation. To answer questions such as "Does sending a coupon to consumers standing outside a coffee shop improve sales?" would require sending

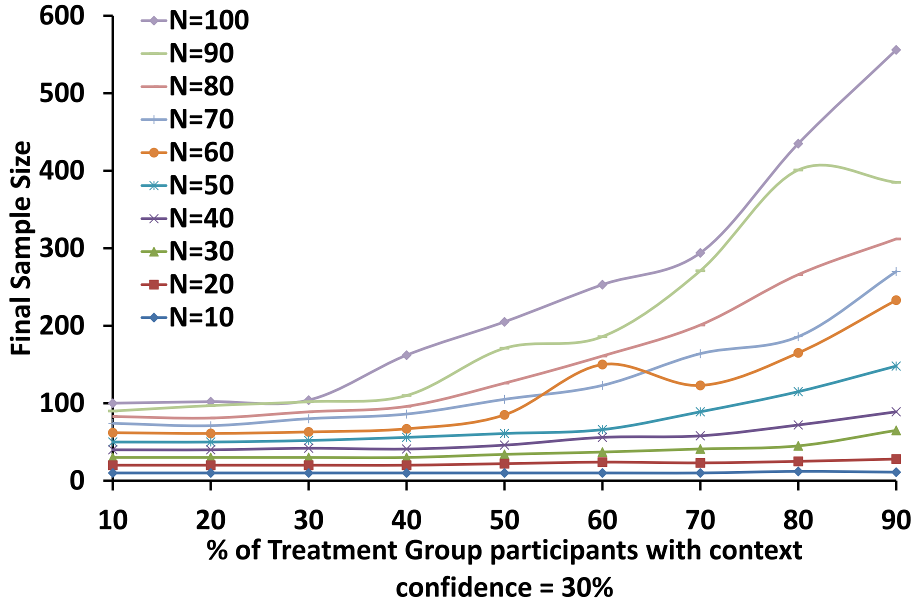


Figure 5.9: Computing sample size under context uncertainty for three groups. Original target sample size range N=10 to 100.

coupons to participants outside a coffee shop as well as other locations. In order to validate the outcome of that experiment, knowing the location confidence of the participants involved as well as the confidence of other event attributes are important. A high location confidence among the set of participants who were in front of the coffee shop *and* adopt the coupon can suggest a strong correlation between coupon adoption and location. It therefore becomes pertinent to provide such confidence information to experimenters. But the question arises as to how the experimenter should interpret this context confidence data?

5.5.1 Non-manipulated independent variable

In many experiment designs, one of the independent variables is a non-manipulated independent variable. The researcher measures it but does not manipulate it. A study by Schnall and colleagues [141] is a good example. One independent variable was disgust, which the researchers manipulated by testing participants in a clean room or a messy room. The other was private body consciousness, which the researchers simply measured.

Such studies are extremely common, and there are several points worth making about them. First, non-manipulated independent variables are usually participant variables (private body consciousness, hypochondriasis, self-esteem, and so on), and as such they are by definition between-subjects factors. For example, people are either low in hypochondriasis or high in hypochondriasis; they cannot be tested in both of these conditions. Second, such studies are generally considered to be experiments as long as at least one independent variable is manipulated, regardless of how many non-manipulated independent variables are included. Third, it is important to remember that causal conclusions can only be drawn about the manipulated independent variable.

For example, Schnall and her colleagues were justified in concluding that disgust affected the harshness of their participants moral judgments because they manipulated that variable and randomly assigned participants to the clean or messy room. But they would not have been justified in concluding that participants private body consciousness affected the harshness of their participants moral judgments because they did not manipulate that variable. It could be, for example, that having a strict moral code and a heightened awareness of ones body are both caused by some third variable (e.g., neuroticism). Thus it is important to be aware of which variables in a study are manipulated and which are not.

5.5.2 Context Confidence as a non-manipulated independent variable

Thus in running experiments where there exists a certain degree of uncertainty in the input contextual triggers, the context confidence metric can be considered as a non-manipulated independent variable. For example, consider the data gathered for an experiment capturing the relationship between proximity of the consumer to the store offering discount coupons and coupon adoption. Traditional methods involving handing the coupons physically, will ensure no ambiguity as to the location of the consumer. The relationship between variables would then be captured as shown in Table 5.4. However, when relying on the experimentation system to send the coupon based on the consumer's location, we will need to consider the location confidence as an additional independent variable (Table 5.5).

Location (IV)	**Coupon Adoption (DV)**
In front of Store	Adopted
In front of Store	Adopted
In front of Store	Adopted
In front of Store	Adopted
In front of Store	Adopted
Away from Store	Not Adopted
Away from Store	Not Adopted
Away from Store	Not Adopted
Away from Store	Not Adopted
Away from Store	Not Adopted

Table 5.4: Single Predictor Variable

Location (IV)	**Location Confidence (Non Manipulated IV)**	**Coupon Adoption (DV)**
In front of Store	0.7	Adopted
In front of Store	0.8	Adopted
In front of Store	0.7	Adopted
In front of Store	0.7	Adopted
In front of Store	0.9	Adopted
Away from Store	0.9	Not Adopted
Away from Store	0.7	Not Adopted
Away from Store	0.8	Not Adopted
Away from Store	0.7	Not Adopted
Away from Store	0.7	Not Adopted

The confidence metric captures how sure the system is that the triggering context condition was satisfied by the participant. For example, row 6 in this table indicates that the system was 90% confident that the participant was at a location away from the store.

Table 5.5: Context Confidence as an additional Predictor Variable

Drawing the experiment conclusion under uncertainty will then follow the normal evaluation process [105]:

1. Assess each variable separately first (obtain measures of central tendency and dispersion; frequency distributions; graphs); is the variable normally distributed?

2. Assess the relationship of each independent variable, one at a time, with the dependent variable (calculate the correlation coefficient; obtain a scatter plot); are the two variables linearly related?

3. Assess the relationships between all of the independent variables with each other (obtain a correlation coefficient matrix for all the independent variables); are the independent variables too highly correlated with one another?

4. Calculate the regression equation from the data.

5. Calculate and examine appropriate measures of association and tests of statistical significance for each coefficient and for the equation as a whole.

6. Accept or reject the null hypothesis.

Example

Consider an experiment attempting to verify the following hypothesis: *"Does location proximity impact coupon adoption?"*. Validating such an experiment would involve sending coupons to participants both near and away from a given establishment (e.g., Subway) and verifying whether a statistical difference exists in terms of coupon adoption between the two experiment groups.

Let the 2x2 frequency distribution matrix of coupon adoption between the two groups be as shown in Table 5.6.

	Take the Coupon	Do not take the Coupon
Near Subway	80	20
Away from Subway	20	80

Chi-square statistic is 72. The P value is 0.(α=0.05)

Table 5.6: H_0: Location proximity does not impact coupon adoption.

The chi-square test statistic shows that the result is significant, indicating that location proximity does indeed play a role in coupon adoption - in this case participants are more likely to take the coupon when they are close to the establishment offering the coupon. However, to verify that this significance holds, we need to factor in the relationship between participant location-confidence and coupon adoption. This is done through the following steps below (Figure 5.10 captures the pseudo code). The goal of these steps is to understand whether the context confidence supports (or not) the observed effect of the main hypothesis.

Step 1. Assess the relationship between the location confidence and coupon adoption. This can be done using logistic regression (since the dependent variable is categorical).

If the relationship is insignificant and the mean of the context confidence of all participants is *above* a set threshold we can safely <u>reject the null hypothesis</u>. However, if the relationship is insignificant and the mean of the context confidence of all participants is *below* a set threshold we need to <u>redo the experiment</u>. For example, if the the mean of the context confidence is 80% we can conclude that the observed coupon adoption effect is likely to be the ground truth. On the other hand, if the mean of the context confidence was 20%, the observed coupon adoption effect may be due to pure chance.

If the relationship is significant then we proceed to Step 2.

Step 2. Study the association between context confidence and coupon adoption within each experiment group. This will help answer questions such as: Is there a difference in coupon adoption across high and low location confidence?

To do this we first need to define the range for high and low location confidence. For example, high confidence can be defined as to be within the range 51% to 100% while low confidence is in the range of 0% to 50%. Once this range is defined we can construct the contingency table (for each experiment group) to capture the association between context confidence and coupon adoption.

For the above example let the contingency table for each experiment group be as shown in Table 5.7 and Table 5.8.

	Take the Coupon	Do not take the Coupon
High Location Confidence	70	10
Low Location Confidence	10	10

The Chi-square statistic is 14.0625. The P value is 0.000177. This result is significant at $p < 0.05$

Table 5.7: Experiment Group 1: Near Subway.

	Take the Coupon	Do not take the Coupon
High Location Confidence	20	60
Low Location Confidence	0	20

The Chi-square statistic is 6.25. The P value is 0.012419. This result is significant at $p < 0.05$.

Table 5.8: Experiment Group 2: Away from Subway.

For each experiment group, if the relationship between the confidence range and coupon adoption is significant, we then identify the component contributing the most to the chi-square statistic. For the experiment group *Near Subway* we are concerned with participants that take up the coupon while for the experiment group *Away from Subway* we are more concerned with the participants that do not take the coupon. If the contributing component is (High Location Confidence, Take the Coupon) for the *Near Subway* group *AND* if the contributing component is (High Location Confidence, Do not take the Coupon) for the *Away from Subway* group we can safely reject the null hypothesis. For all other conditions we need to redo the experiment. For example, from the above tables the contributing components of the chi-square statistic is 70 for the *Near Subway* group and 60 for the *Away from Subway* group. As both these components have a high location confidence it is very likely that the observed effect of coupon adoption was not due to chance.

Similarly, if the observed effect of the main hypothesis was insignificant, the context confidence can be used to determine whether it supports the failure to reject the null hypothesis or whether we need to redo the experiment.

5.6 Summary

The Behavioural Experimentation Platform, attempts to gain insight into consumer behaviour towards context based advertising, through event based experimentation. As these experiments rely on uncertain context, there is a need to identify as well as quantify this uncertainty. In this Chapter I describe how the platform defines a confidence metric for the location attribute as well as how it derives information, such as the number of false positives. The evaluation shows using overlap ratio to represent location confidence is reliable and that the algorithm to estimate the number of false positives has minimal errors. Both these values are important in understanding the outcome of an experiment and in turn defining it's success criteria. I also describe how context uncertainty impacts the experiment sample size as well as how verify the experiment hypothesis with additional context confidence information.

```
If χ^2(IV, Coupon Adoption)^$ //Relationship between the independent and dependent variable is significant
        If ¬ (logit(p) = β0 + β1 * Context Confidence)^$ then
                If x̄(Context Confidence)>Threshold then
                        Reject H0
                Else
                        Redo Experiment
        Else
        Measure Association (Context Confidence, Coupon Adoption)
        •   Define High/Low Context Confidence Range
        •   For each Independent Variable create contingency table (High/Low Context Confidence, Coupon
            Adoption)

        If χIV_A^2(High/Low Context Confidence, Coupon Adoption)^$
                If χIV_B^2(High/Low Context Confidence, Coupon Adoption)^$
                        If C (χIV_A^2) = (High Context Confidence, Take the Coupon)  &&
                          C (χIV_B^2) = (High Context Confidence, Do not take the Coupon)
                                Reject H0
                        Else
                                Redo Experiment
                Else
                        Redo Experiment
        Else
                Redo Experiment

Else //Relationship between the independent and dependent variable is *not* significant
        If ¬ (logit(p) = β0 + β1 * Context Confidence)^$ then
                If x̄(Context Confidence)>Threshold then
                        Accept H0
                Else
                        Redo Experiment
        Else
        Measure Association (Context Confidence, Coupon Adoption)
        •   Define High/Low Context Confidence Range
        •   For each Independent Variable create contingency table (High/Low Context Confidence, Coupon
            Adoption)

        If χIV_A^2(High/Low Context Confidence, Coupon Adoption)^$
                If χIV_B^2(High/Low Context Confidence, Coupon Adoption)^$
                        If C (χIV_A^2) = (High Context Confidence, Take the Coupon)  &&
                          C (χIV_B^2) = (High Context Confidence, Take the Coupon)
                                Accept H0
                        Else
                                Redo Experiment
                Else
                        Redo Experiment
        Else
                Redo Experiment
```

logit-logistic regression.
p is the probability of the binary outcome variable (coupon adoption) indicating failure/success to be 1.
\$-significant at $p<0.05$.
χ^2-chisquare test.
C(χ^2)-Contributing cell.
IV-independent variable. Near Subway(A), Away from Subway(B).

Figure 5.10: Pseudo code for drawing an experiment conclusion under uncertainty

Chapter 6

Related Work

In this chapter, I present the related work relevant to this dissertation. My work spans context-aware computing, behavioural science and human-computer interaction (HCI). At their juncture lies the opportunity to create a suitable tool for conducting social experiments in an unobtrusive way. To the best of my knowledge, this is the first work that enables real-time experimentation using context-based triggers to select participants for sending interventions to their mobile device.

Section 6.1 presents the work related to incorporating context in system design to make them more responsive to the different social settings in which they might be used. I show that even though user context has been used in many different scenarios before, very little knowledge exists on how the different contextual factors influence user experience in a certain situation, motivating the need for a behavioural experimentation platform.

One of the key challenges of ubiquitous advertising is reaching the right people with the right ads. In Section 6.2, I describe prior research that uses context to provide consumers with relevant content and explain how I used those ideas in targeting the right participant sample for experiments involving promotions as the intervention.

Due to the complex nature of information sources, input events almost always carry a certain degree of uncertainty and/or ambiguity. Until recently, most systems did not make any effort to handle uncertainty in information sources. In Section 6.3 I present an overview of existing approaches for handling context uncertainty.

Finally, in Section 6.4, I present the work relevant to understanding mobile usage. In particular, I discuss previous research that employ simple features extracted from the phone, such as user interaction with the notification center and screen activity as predictors of user behaviour.

6.1 Context and User Experience

User experience (UX) is a concept established during the past decade to describe the holistic nature of users' interaction with information technology products and the associated services. Many user experience models and frameworks have been proposed which explain aspects that affect the overall UX of a product [52, 60, 72]. One common trait in these models and frameworks is that they recognize the importance of context as a central aspect influencing user experience.

There are multiple definitions of what context is and how it influences the use of interactive products [25, 38, 39, 42]. Especially in user experience studies involving mobile products, understanding contextual factors is important because they are used in greatly varying contexts.

Although the importance of context has been acknowledged in UX research, very little knowledge exists on how the different contextual factors influence user experience in a certain situation. One of the reasons is that context is often defined as a single influencing factor, such as car context or mobile context [71]. Social and physical contexts have been the most commonly studied factors (66% and 61% of the studies, respectively) with other contextual factors less frequently involved [78]. Wigelius et al. showed that multiple contextual factors are usually present in the mobile context [156], and therefore, it would be important to study context holistically to determine which factors are influencing the user experience in a particular situation.

Understanding *how* context affects the interaction or user experiences of a mobile system is very important for system development. There are countless amounts of studies which report how mobile devices, applications and services have been evaluated in different contexts of use. The studies have ranged from specific domains (e.g., museums, hospitals) and different user groups (e.g., teenagers, visually impaired) to studying various tasks (e.g., writing messages while driving, navigating in an urban environment) that users perform with these systems.

The objective of these studies has usually been to understand individual contextual factors and overcome challenges that the contextual factors set for interaction design. For example, Barnard et al. have studied how changes in contextual factors (task type, motion and lighting level) affect the performance of mobile device users [16]. In their study, the users were operating a PDA and performed reading comprehension and word search tasks. The results indicate that changes in the contextual factors will also affect the interaction with a mobile device.

Jones et al. argue that the evaluation of a mobile system should take place in an actual context [69]. However, it is sometimes difficult to access the context of use or to include all relevant contextual factors to the evaluation. For example, Kjeldskov et al. reported a study in which they developed a mobile system to support safety-critical collaboration tasks on board a container ship [149]. In this study, the evaluations were arranged in a laboratory setting, because it was not possible to use the actual context, that is, a real container ship. For these kinds of evaluations, it would be beneficial to recognize which contextual factors are essential for the evaluation and might have significant influence of user experience.

Although it has been acknowledged that context affects the user experience of a product, which contextual factors among the all possible factors are actually affecting the user experience of mobile products has not been studied extensively. Mallat et al. have studied the acceptance of a mobile ticketing service and concluded that context was a significant determinant for consumer intention of using a service [91]. They reported that the service was found useful when users were in a hurry or the need for a ticket was unexpected, other alternatives for purchasing tickets were not available, or there was a queue at the sales point. All these reasons imply that there are different contextual factors which drive the use of the service, but the contextual factors were not identified in the study.

While previous studies provide valuable information on what kinds of contextual factors are involved in the mobile context or the user experience of mobile systems, they do not describe in detail the relationship between these two aspects. After reviewing an extensive amount of previous studies of mobile devices, Jumisko-Pyykk stated that, in most of the studies, context of use is mainly understood as a relatively static phenomenon and the studies consider only few contextual factors, such as the social and physical context. Other relevant factors are less explored [71].

By building a context-based experimentation system that target real people on their devices in real-time using a multitude of triggering contexts we overcome the limitations of prior work. Through this process of experimentation we seek out the context that affects user experience in a specific situation. The triggering context facilitates identifying the most meaningful experience, i.e. the core experience, for a user in an episode of usage. Understanding the influence of context is key for designing services that take into account how changes in the context strengthen or distract positive user experience.

6.2 Context-based Advertising

At one end we have stores with promotions for their products and services. At the other end are consumers with their needs. Ubiquitous advertising [80] attempts to match the two to create the greatest possible impact on the customers. However, such a service is riddled with many challenges [45, 127] the foremost being ad relevancy.

While a number of researchers [27, 96, 109, 134, 159, 162] have suggested the use of user context to provide a higher degree of ad personalization, most research has focused on using only location to increase the relevancy of advertisements delivered. Alto et al. proposed a location-based advertising (LBA) system that proactively sends ads to mobile phones when a user passes by a certain store, using Bluetooth for localization [2]. In the AdNext system [75], the authors propose to use mobility patterns to predict which store the user is likely to visit next, and show him advertisements related to that store. Hardt et al., consider using various physical contexts such as location and user activities to serve ads [59]. Commercial LBAs such as Shopkick [144] and PROMO [114] also exploit current user location to push relevant advertisements. More recently SmartAds [110] deliver contextual ads to mobile apps by taking into account the content of the page the ads are displayed on.

Other competitors fall into two broad categories. In the first category, applications like Froogle [54], LiveCompare [36], CompareEverywhere [67] and ShopSavvy [122] cater towards on-the-fly price comparisons for mobile users. In the second category are deal information services, like Citibank's Card Information Service [31] and mobile applications like Mobiqpons [101], that attempt to inform consumers about the best deals available. However, these applications fail in two main areas; first, they are usually targeted to just certain malls and/or certain payment cards (only Citibank cards for example), and second, they do not take into consideration user preferences - factors that are important in getting the best deal! In all the cases we observe that the burden of finding the best deal is left to the consumers — the applications just provide different ways to find lists of deals matching certain broad criteria (such as all restaurant deals etc.).

Collaborative filtering(CF) techniques have also been used for targeted advertising [30, 128, 136]. CF requires users to actively participate and express his or her preference by rating items in the system. These inputs are used to build user-item (or user-rating) matrices and recommend similar promotions for users with similar taste. We believe however, that user interests are highly situational and hence require users to explicitly state their preference every time the system is used. However, we could incorporate such additional signals to further increase promotion relevance.

Delivering relevant ads to web pages has been well studied. While the usage scenarios are inherently different from mine, I will discuss representative papers for sake of completeness. Ribeiro et al. [130], propose a series of strategies to match ads with web pages, based on keywords extracted from both the ad body and the web page content. Beyond keywords several other features have been employed for delivering relevant ads, including semantics mined from web pages (e.g, topics) [19] and user interactions in browsing [35].

Natural language processing(NLP) techniques are gaining prominence in mobile applications. NLP helps to greatly facilitate access and use of applications, encouraging greater adoption. Siri [148], a personal assistant mobile application uses NLP to answer questions and make recommendations. The application adapts to a user's individual preferences over time and personalizes results, and performs tasks such as making dinner reservations while trying to catch a cab. Maluuba designed an API based on NLP and other machine learning techniques that aids in processing the natural human language [92]. In our model, we applied NLP techniques for estimating the semantic similarity of deal items.

6.3 Context Uncertainty

Context-aware systems can't always identify the current context precisely, hence they need support for handling uncertainty. Various mechanisms such as probabilistic logic, fuzzy logic and Bayesian networks are used for reasoning about uncertainty [23, 58, 126]. MiddleWhere [126] uses probabilistic reasoning techniques to deduce a person's location confidence. However, this technique assumes the availability of precise information associated with the location sensing technology, such as the probability of a false positive. In contrast, we assume no such information is readily available and instead attempt to estimate the number of false positives for a given scenario.

There is also considerable work in reducing context uncertainty using sensor fusion [77, 146] as well as through user mediation [37, 53]. These efforts are orthogonal to our work, where we focus on representing location uncertainty, as opposed to reducing it, and associating this uncertainty with the outcome of an event.

Finally, while uncertainty is a significant problem for many other ubiquitous computing applications, it is not *as* problematic for advertising. From the advertiser's perspective, any reduction in uncertainty is welcome. Taking this viewpoint, current context based advertising applications (commercial and research driven) [144, 75, 158] do not handle or represent uncertainty in their context stream. As a result, any visibility as to why a consumer did not react to a given stimulus is ignored - something which our platform intends on correcting.

6.4 User Feedback

Current approaches to capturing mobile usage can be categorized into four classes: direct observation, lab-based evaluation, self- report, and automatic loggingeach offering different, limited visibility into human behaviour and user experience.

Automatic tracing typically records usage information passively without requiring user intervention (from the infrastructure [161] or directly from the device [124]). This technique scales well across users and collects large amounts of data; however, important information such as user intention and perception is lost. In contrast, in situ self-reports such as the Diary Method [112, 131] and the Experience Sampling Method (ESM) [34, 65] offer insight in to these otherwise imperceptible details, but at a cost of user involvement. Thus, the sampling rate is much lower than in automatic logging and the method does not scale as well (e.g., participant compliance diminishes over time). Direct observation methods like shadowing [73] can provide rich qualitative accounts of device usage and human behaviour; however, the method can only be applied to a small number of participants at a time and not all contexts are conducive to being studied (e.g., a formal business meeting) [131]. In addition, it is subject to observer bias, and the small form factors of mobile devices make it difficult to observe both the participant and their device screens. Finally, laboratory methods offer an environment to rigidly control device and context parameters for experimentation; however, usage is artificial and removed from its natural setting.

One common approach in combining logging and qualitative feedback is through the use of interviews. Logs can be used to cue a participant's memory during interviews, thereby reducing recall biases [40, 116, 163]. However, interviews do not scale well across large numbers of participants. In addition, participants may still suffer from some form of recall bias or memory lapse even with cueing.

SenseCam [55] offers an entirely different approach in collecting qualitative and quantitative data; digital photographs are automatically captured and annotated with sensor data via a pendant worn around the neck. The photographs allow for qualitative assessments of ground truth (e.g., user appears to be indoors) and provide good cue points for interviews; however, they do not collect user feedback in situ and the continuous photography raises privacy concerns. Moreover, participants are required to wear an additional device.

Chapter 7

Conclusion

In this chapter, I summarize, in Section 7.1, the research contributions made by this dissertation. I then describe possible future work in Section 7.2 . Finally, Section 7.3 places this dissertation in perspective with the current state of mobile sensing research.

7.1 Contribution

Imagine that you receive an advertisement on your mobile phone while you are busy with some important work. How will you react if it is an attractive offer promoting one of your favourite products at a nearby store? Will you take advantage of this opportunity as soon as possible? What if the advertisement had appeared while you were relaxing at home with family and friends? Answers to these questions are important for mobile marketing academicians and researchers, marketing managers who are likely to reach customers through mobile phones for promoting their products and curious to know the impact of mobile advertising. With the advent of smartphones the reach of advertisements and promotions have been extended even further. The question remains however, if there is a right time and right place for consumers to receive and respond such offers. *Can we develop a system for helping experimenters answer such questions about user behaviour?*

This thesis shows that it is possible to build *Jarvis*, an experimentation system that can assist in running in situ realtime experiments, targeting real participants on their smart phones based on multiple context-specific events. I show that through this process of context-based experimentation, it is possible to observe the contextual factors that have an influence on user behaviour. I also show that the system supports a diverse range of experiment predicates and controlled experiment designs and that the experiment specification can be done with ease.

At one end we have advertisers (experimenters) with ads for their products and services while at the other end are consumers (participants) with their needs. In running context-based advertising experiments there is (often) a need for the experimentation system to match the two to create the greatest possible impact on the customers and in turn get the right experiment sample. In this thesis I demonstrate a matching and scoring algorithm that accurately factors user's preferences when matching deals and is capable of combining structured and unstructured deal information into a single score. Doing so, will allow the experimentation system to target the right set of participants.

Furthermore, though sensing has becoming more cost-effective and ubiquitous, the interpretation of sensed data as context is still imperfect and will likely remain so for some time. A challenge facing the development of realistic and deployable context-aware services, therefore, is the ability to handle imperfect, or *ambiguous*, context. However, unlike most applications that ignore any uncertainty in the sensed data and its interpretations in this thesis I show how Jarvis handles context uncertainty. More specifically, I show how the module defines a confidence metric for the location predicate as well as how it stochastically estimates additional information such as the number of false positives. In doing so, I provide adequate information to the experimenter to process the results of an experiment.

Finally, the true power of the experimentation system is not only in running *in situ* experiments but also collecting both qualitative and quantitative data that provide a feedback of that experiment. Jarvis monitors the selected participants (for a given experiment) for a set period of time and records what they did in response to the experiment stimulus. In my thesis I show how user mobile interaction can act as a potential source of user reaction to a context-based experiment. Such data could perhaps provide insight into user motivation, perception, and satisfaction towards the context-driven intervention.

7.2 Future Work

In this section, I present avenues for future work on Jarvis.

7.2.1 More Context Categories, Experiment Designs & Participant Selection Techniques

Context Categories

Location was one of the first contextual factors to be studied in the development of context-aware systems [139]. In addition to location, researchers have also recognized that there are other factors that influence how users experience mobile systems. Schmidt et al. [140] presented that context could be divided into two main categories: human factors and physical environment. Under these main categories, there would be several subcategories. Subcategories of the Human Factors category include user, social environment and task. Subcategories of the Physical Environment category include condition, infrastructure, and location.

Further, Forlizzi et al. [52] noted that interaction with a product is a constant stream of interaction sequences and emotions, and it causes emotional and behavioural changes in a user. Contextual factors play an important role in these changes, and they can amplify previous negative or positive experience, alter positive experience into negative experience, or vice versa. It therefore becomes necessary to extend Jarvis to support these context categories not only as experiment triggering conditions but also capturing context in terms of how it changes user experience during the experiment life-cycle.

Experiment Designs

Experience sampling method (ESM) is a systematic way of having participants provide samples of their ongoing behaviour [61]. Participants' reports are dependent upon either a signal, pre-established intervals, or the occurrence of some event.

1. Signal contingent: The participant is signaled with a beeper, cell phone call, or programmed watch at random times within a fixed time period (e.g., between 8 AM and 9 PM). At the signal, the participant records the behaviour of interest (e.g., activity, location, mood, thoughts).

2. Interval contingent: The participant is assigned pre-set intervals for reporting events. For example, before going to bed at night, the participant fills out a checklist of the day's activities.

3. Event contingent: The event is determined by the research project, for example, watching a movie, or phoning a friend. The participant makes a record whenever the key event occurs. The recording of the event depends upon (is contingent) on its occurrence.

Although currently not supported, ESM is one type of experiment design that could easily be supported by Jarvis. It is important that other such treatment and observational study designs are identified to increase the experiment-design handling capabilities of the system.

Participant Selection: Support for Stratified Randomization

Stratified randomization [150] refers to the situation in which strata are constructed based on values of prognostic variables and a randomization scheme is performed separately within each stratum. For example, suppose that there are two prognostic variables, age and gender, such that four strata are constructed (Table 7.1:

	Treatment A	Treatment B
male, age <18	12	12
male, age ≥ 18	36	37
female, age <18	13	12
female, age ≥ 18	40	40

Table 7.1: Stratified Randomization

The strata size usually vary (maybe there are relatively fewer young males and young females within the sample population). From the purview of the experimentation system, the objective of stratified randomization is to ensure balance of the treatment groups with respect to the various combinations of the context variables. Simple randomization will not ensure that these groups are balanced within these strata so permuted blocks are used within each stratum are used to achieve balance.

However, if there are too many strata in relation to the target sample size, then some of the strata will be empty or sparse. This can be taken to the extreme such that each stratum consists of only one participant each, which in effect would yield a similar result as simple randomization. Support for stratified randomization will need to ensure the number of strata used to a minimum for good effect.

7.2.2 Infer Context Rules

Advertising in pervasive computing environments presents some unique challenges. One of the challenges is figuring out the best possible time for and the best possible way of delivering the ad. Here is where user context information can help us figure out the best time and the best delivery mechanism.

Pervasive environments can monitor the user's activities using a variety of techniques like looking at his schedule, looking at what applications he is running, seeing whether any other people are in the same room as him, etc. Also, certain types of ads may have greater impact when the user is in a particular context. For example, a beer ad or a pizza delivery ad *may* gain better mileage when it is played to a group of men who are watching a football game. At the same time, some other ads may be inappropriate when a person is with a certain company.

The decision of when and how to deliver the ad (intervention) could be made by the experimentation system, the experimenter (advertiser) or both. The experimenter could define some rules as to how and when it wants certain types of ads to be presented. It could also give the system some flexibility to choose the best time and mechanism depending on context. Jarvis would then need to have mechanisms that allow experimenters to define rules for delivering advertisements. An experimenter could also indicate that it would like it's ads to be delivered in a serendipitous manner too [45]. The system could also have it's own rules on how and when advertisements can be delivered. For example, ads should definitely not be delivered when the participant is in a meeting or is driving. However, identifying such rules is not always trivial. For example, it is not known whether offering discount coupons to people who have been standing outside a coffee shop for 10 minutes will improve sales as opposed to sending it when they pass by? One possibility of inferring such rules is by analysing outcomes of similar previous experiments.

7.2.3 Longer Term Research

The previous sections discussed shorter term research relevant to Jarvis. In this section, I discuss longer term research in the general area of context-aware computing.

The falling cost of adding sensing and communications to consumer products will mean that a typical family home, in a mature affluent market, could contain several hundred smart objects by 2022, according to Gartner [70]. This additional context information will not only aid a better understanding of the user, but will also lead to additional business opportunities and challenges in collecting and managing the context information.

Business Opportunities

Events emanating from mobile personal devices are only one source of event-driven context. If context is the information about the current situation, events generated from relevant applications, websites, cloud services, social platforms, databases and devices can all supply context to a context-aware application. The future of mobile computing will therefore be context-aware computing with mobile applications adjusting to the user's location, identity and past behaviours [132]. Contextual mobile applications will lead to new user experiences that will be simple, visually attractive, compelling and interactive, such as:

- Exploring: Providing information in an exploratory form, for example when a user wants to know what's potentially interesting and nearby.

- Refining: A user is seeking more information about something which might be a place, and object, a physical product or even another person.
- Suggesting: Context will be proactive, but suggestions must be delivered in ways that are themselves contextual and appropriate. For example, receiving an audio message through a Bluetooth headset might be best for a user who is driving.
- Controlling: Sometimes a user will need to specify their requirements. This can be done by typing text but, increasingly, will use new experiences such as pointing at something to indicate "I want one like that."
- Directing: Contextual systems will frequently involve location and will need to direct a user to a destination. This might be done through maps or by unobtrusive directions.

Information will be at the core of enabling such personalized used of devices. The information will include sensor-based data available and collected from mobile devices such as temperature, speed or location. Other sources of information will come from both corporate and external sites such as social media or public data. Enabling such use cases will require new ways of collecting, managing, delivering and rendering information to users that takes into consideration personalization but also the variety of devices.

Impact on Data Collection

Given the potential volume of data produced by mobile devices, organizations need to build a strategy to deliver value from the information. The need to collect this information in a cost-effective way, and the use cases that it will support will be very diverse. This will affect the following aspects of big data:

1. Volume of data generated by the devices.
2. Velocity of the data. Data is collected at very different rates. It can be sensor-

generated data that is constantly being collected (such as location information) or it can be user-generated data (such as making a call) that is infrequent.

3. Variety. Mobile devices produce different data types. Data can be images, audio or sensor data.

Managing Data and Adding Value

For sensor data, the major challenge resides in the cost-benefit analysis of managing the data. The collection of sensor data is not free and unless value is delivered from it, the value of collecting this data should be carefully considered. Managing the data collected from sensors is not just about potential storage challenges. A number of other requirements need to be addressed, including:

1. Privacy protection

2. Perishability The data collected might only be valid for a limited amount of time. Some of it may be worth storing for further pattern analysis. For example, wi-fi data collected for location analysis can be stored to perform offline pattern analysis.

3. Fidelity This applies to the level of trust required for the use case. How complete and accurate the data collection needs to be will be depend on the use case. For example, incomplete location data only affects the precision and calculation of the best route, whereas incomplete call detail records collection affects directly the ability to process accurately the bills.

4. Linking Enabling the data collection is often only valid when the data collected can be related to other data. This means that the application can recognize pertinent events from within the data stream and quickly analyse all of the available information assets to determine the best course of action. This is about more than just data collection. For example, for retailers to perform promotional offerings for clients while in the store, they must combine

customer-relationship management data, social data to refine their preferences and in-store location information.

7.3 Closing Remarks

Behavioural sciences have for long provided a systematic analysis and investigation of human behaviour through controlled and naturalistic observation, and disciplined scientific experimentation. However, these investigations are often subjected to the limitations of laboratory settings and the cost overhead of observational studies.

In 2012 a consumer insight report quipped *"The smartphone is perceived as a second brain or a best friend, and apps are becoming an emotionally important and integral part of people's daily lives"* [1]. They weren't too far off! Today smart devices can infer your mood [88], tell whether you liked the movie or not [14], keep a check of your fitness and eating patterns [28] and much more. These devices give us unprecedented access to everyday life helping us better understand the individual.

An obvious partnership is to therefore employ mobile sensing technology to provide unobtrusive access to human social behaviour. What is needed then is a platform that provides experimenters access to this deep, fine-grained, near-real time human context while exposing an experimentation service that frees experimenters from many experimental chores. This dissertation aims to provide just that.

Jarvis, is based on the well-known randomized controlled trial (RCT) experimental process. By representing context predicates of an experiment as events, I create an event-driven system that enables a real-time in situ participant enrolment phase. The underlying behavioural experimentation system also provides the components necessary to automate the additional phases of the RCT - intervention allocation, follow-up, and data analysis - with the goals of minimizing selection bias and allocation bias while maximizing the statistical power.

Bibliography

[1] Emerging App Culture. Technical report, Ericsson Consumer Lab, April 2012.

[2] L. Aalto, N. Göthlin, J. Korhonen, and T. Ojala. Bluetooth and wap push based location-aware mobile advertising system. In *Proceedings of the 2nd international conference on Mobile systems, applications, and services*, MobiSys '04, pages 49–58, New York, NY, USA, 2004. ACM.

[3] T. Abdelzaher, Y. Anokwa, P. Boda, J. Burke, D. Estrin, L. Guibas, A. Kansal, S. Madden, and J. Reich. Mobiscopes for human spaces. *IEEE Pervasive Computing*, 6(2):20–29, Apr. 2007.

[4] A. Adi and O. Etzion. Amit - the situation manager. *The VLDB Journal*, 13(2):177–203, May 2004.

[5] A. Ansari, S. Essegaier, and R. Kohli. Internet recommendation systems. *Journal of Marketing Research*, XXXVII:363–375, Aug. 2000.

[6] A. Artikis, O. Etzion, Z. Feldman, and F. Fournier. Event processing under uncertainty. In *Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems*, DEBS '12, pages 32–43, New York, NY, USA, 2012. ACM.

[7] A. Artikis, M. Sergot, and G. Paliouras. A logic programming approach to activity recognition. In *Proceedings of the 2Nd ACM International Workshop on Events in Multimedia*, EiMM '10, pages 3–8, New York, NY, USA, 2010. ACM.

[8] A. Artikis, A. Skarlatidis, F. Portet, and G. Paliouras. Logic-based event recognition. *The Knowledge Engineering Review*, 27:469–506, 12 2012.

[9] L. Ashworth, P. R. Darke, and M. Schaller. No one wants to look cheap: Trade-offs between social disincentives and the economic and psychological incentives to redeem coupons. *Journal of Consumer Psychology*, 15(4):295 – 306, 2005.

[10] P. Bahl and V. N. Padmanabhan. Enhancements to the radar user location and tracking system. Technical Report MSR-TR-2000-12, Microsoft Research, February 2000.

[11] P. Bahl and V. N. Padmanabhan. Radar: an in-building rf-based user location and tracking system. Institute of Electrical and Electronics Engineers, Inc., March 2000.

[12] J. Y. Bakos. Reducing buyer search costs: Implications for electronic marketplaces. *Management Science*, 43(12):1676–1692, Dec. 1997.

[13] R. K. Balan, A. Misra, and Y. Lee. Livelabs: Building an in-situ real-time mobile experimentation testbed. In *Proceedings of the 15th Workshop on Mobile Computing Systems and Applications*, HotMobile '14, pages 14:1–14:6, New York, NY, USA, 2014. ACM.

[14] X. Bao, S. Fan, A. Varshavsky, K. Li, and R. Roy Choudhury. Your reactions suggest you liked the movie: Automatic content rating via reaction sensing. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, pages 197–206, New York, NY, USA, 2013. ACM.

[15] O. Barak, G. Cohen, A. Gazit, and E. Toch. The price is right?: Economic value of location sharing. In *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, UbiComp '13 Adjunct, pages 891–900, New York, NY, USA, 2013. ACM.

[16] L. Barnard, J. S. Yi, J. A. Jacko, and A. Sears. Capturing the effects of context on human performance in mobile computing systems. *Personal Ubiquitous Comput.*, 11(2):81–96, Jan. 2007.

[17] P. Basiliere. Customer communications management software providers must meet the demands of the nexus of forces. Technical report, Gartner Report, Nov 2012.

[18] N. Bolger, A. Davis, and E. Rafaeli. Diary Methods: Capturing Life as it is Lived. *Annual Review of Psychology*, 54(1):579–616, 2003.

[19] A. Broder, M. Fontoura, V. Josifovski, and L. Riedel. A semantic approach to contextual advertising. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 559–566, New York, NY, USA, 2007. ACM.

[20] J. R. Bult and T. Wansbeek. Optimal selection for direct mail. *Marketing Science*, 14(4):378–394, 1995.

[21] E. Carrol, M. Czerwinski, A. Roseway, A. Kapoor, and m.c. Schraefel. Food and mood: An exploration in emotional eating intervention. ACII 2013, 2013.

[22] S. Carter, J. Mankoff, and J. Heer. Momento: Support for situated ubicomp experimentation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 125–134, New York, NY, USA, 2007. ACM.

[23] P. Castro, P. Chiu, T. Kremenek, and R. R. Muntz. A probabilistic room location service for wireless networked environments. In *Proceedings of the 3rd International Conference on Ubiquitous Computing*, UbiComp '01, pages 18–34, London, UK, UK, 2001. Springer-Verlag.

[24] Center For Sprogteknologi. *POS-tagger*. `http://ida.hum.ku.dk/tools/index.php/?lang=en`.

[25] M. Chalmers. A historical view of context. *Computer Supported Cooperative Work (CSCW)*, 13(3-4):223–247, 2004.

[26] T. C. Chalmers, H. S. Jr., B. Blackburn, B. Silverman, B. Schroeder, D. Reitman, and A. Ambroz. A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials*, 2(1):31 – 49, 1981.

[27] P.-T. Chen and H.-P. Hsieh. Personalized mobile advertising: Its key attributes, trends, and social impact. *Technological Forecasting and Social Change*, 79(3):543–557, Mar. 2012.

[28] H.-T. Cheng, F.-T. Sun, M. Griss, P. Davis, J. Li, and D. You. Nuactiv: Recognizing unseen new activities using semantic attribute-based learning. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '13, pages 361–374, New York, NY, USA, 2013. ACM.

[29] K. Chintalapudi, A. Padmanabha Iyer, and V. N. Padmanabhan. Indoor localization without the pain. In *Proceedings of the Sixteenth Annual International Conference on Mobile Computing and Networking*, MobiCom '10, pages 173–184, New York, NY, USA, 2010. ACM.

[30] E. N. Cinicioglu, P. P. Shenoy, and C. Kocabasoglu. Use of radio frequency identification for targeted advertising: A collaborative filtering approach using bayesian networks. In *Proceedings of the 9th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, ECSQARU '07, pages 889–900, Berlin, Heidelberg, 2007. Springer-Verlag.

[31] Citibank Singpore Ltd. *Citi Shopper (SM)*. `http://citishopper.citi.com/`.

[32] W. Cochran. *Sampling techniques*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1977.

[33] Confidence Interval. `http://en.wikipedia.org/wiki/Confidence_interval`.

[34] S. Consolvo and M. Walker. Using the experience sampling method to evaluate ubicomp applications. *IEEE Pervasive Computing*, 2(2):24–31, Apr. 2003.

[35] K. S. Dave. Computational advertising: leveraging user interaction contextual factors for improved ad relevance targeting. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 757–758, New York, NY, USA, 2012. ACM.

[36] L. Deng and L. P. Cox. LiveCompare: grocery bargain hunting through participatory sensing. In *Proceedings of the 10th Workshop on Mobile Computing Systems and Applications (HotMobile)*, Santa Cruz, CA, Feb. 2009.

[37] A. Dey, J. Mankoff, G. Abowd, and S. Carter. Distributed mediation of ambiguous context in aware environments. In *Proceedings of the 15th annual ACM symposium on User interface software and technology*, pages 121–130. ACM, 2002.

[38] A. K. Dey. Understanding and using context. *Personal Ubiquitous Comput.*, 5(1):4–7, Jan. 2001.

[39] A. K. Dey, G. D. Abowd, and D. Salber. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Hum.-Comput. Interact.*, 16(2):97–166, Dec. 2001.

[40] A. K. Dey, K. Wac, D. Ferreira, K. Tassini, J.-H. Hong, and J. Ramos. Getting closer: an empirical investigation of the proximity of user to their smart phones. In *Proceedings of the 13th international conference on Ubiquitous computing*, UbiComp '11. ACM, 2011.

[41] A. Dickinger and M. Kleijnen. Coupons going wireless: Determinants of consumer intentions to redeem mobile coupons. *Journal of Interactive Marketing*, 22(3):23 – 39, 2008.

[42] P. Dourish. What we talk about when we talk about context. *Personal Ubiquitous Comput.*, 8(1):19–30, Feb. 2004.

[43] P. Edwards, I. Roberts, M. Clarke, C. DiGuiseppi, S. Pratap, R. Wentz, and I. Kwan. Increasing response rates to postal questionnaires: systematic review. *BMJ*, 324(7347):1183, 2002.

[44] J. Elslander and K. Tanaka. A notification-centric mobile interaction survey and framework. In A. Jatowt, E.-P. Lim, Y. Ding, A. Miura, T. Tezuka, G. Dias, K. Tanaka, A. Flanagin, and B. Dai, editors, *Social Informatics*, volume 8238 of *Lecture Notes in Computer Science*, pages 443–456. Springer International Publishing, 2013.

[45] C. I. Eriksson and M. Akesson. Ubiquitous advertising challenges. In *Proceedings of the 2008 7th International Conference on Mobile Business*, ICMB '08, pages 9–18, Washington, DC, USA, 2008. IEEE Computer Society.

[46] O. Etzion and P. Niblett. *Event Processing in Action*. Manning Publications Co., Greenwich, CT, USA, 1st edition, 2010.

[47] S. K. F and G. D. A. Generation of allocation sequences in randomised trials: chance, not choice. *The Lancet*, 359(9305):515519, Feb 2002.

[48] H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin. Diversity in smartphone usage. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*, MobiSys '10. ACM, 2010.

[49] J. E. Fischer, C. Greenhalgh, and S. Benford. Investigating episodes of mobile phone activity as indicators of opportune moments to deliver notifications. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '11, pages 181–190, New York, NY, USA, 2011. ACM.

[50] J. Fogarty, S. E. Hudson, C. G. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. C. Lee, and J. Yang. Predicting human interruptibility with sensors. *ACM Trans. Comput.-Hum. Interact.*, 12(1):119–146, Mar. 2005.

[51] J. Fogarty, S. E. Hudson, and J. Lai. Examining the robustness of sensor-based statistical models of human interruptibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 207–214, New York, NY, USA, 2004. ACM.

[52] J. Forlizzi and S. Ford. The building blocks of experience: An early framework for interaction designers. In *Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, DIS '00, pages 419–423, New York, NY, USA, 2000. ACM.

[53] J. Froehlich, M. Y. Chen, S. Consolvo, B. Harrison, and J. A. Landay. My-experience: a system for in situ tracing and capturing of user feedback on mobile phones. In *Proceedings of the 5th international conference on Mobile systems, applications and services*, MobiSys '07. ACM, 2007.

[54] Froogle. `http://www.google.com/products`.

[55] J. Gemmell, L. Williams, K. Wood, R. Lueder, and G. Bell. Passive capture and ensuing issues for a personal lifetime store. In *Proceedings of the the 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, CARPE'04, pages 48–55, New York, NY, USA, 2004. ACM.

[56] B. George. The relationship between lottery ticket and scratch-card buying behaviour, personality and other compulsive behaviours. *Journal of Consumer Behaviour*, 2(1):7–22, 2002.

[57] Google now. `http://www.google.com/landing/now/`.

[58] D. Guan, W. Yuan, A. Gavrilov, S. Lee, Y. Lee, and S. Han. Using fuzzy decision tree to handle uncertainty in context deduction. In D.-S. Huang, K. Li, and G. Irwin, editors, *Computational Intelligence*, volume 4114 of *Lecture Notes in Computer Science*, pages 63–72. Springer Berlin Heidelberg, 2006.

[59] M. Hardt and S. Nath. Privacy-aware personalization for mobile advertising. In *Proceedings of the 2012 ACM conference on Computer and communications security*, CCS '12, pages 662–673, New York, NY, USA, 2012. ACM.

[60] M. Hassenzahl and N. Tractinsky. User experience a research agenda. In *March-April 2006*, pages 91–97.

[61] J. M. Hektner, J. A. Schmidt, and M. Csikszentmihalyi. *Experience sampling method: measuring the quality of everyday life*. SAGE Publications, Thousand Oaks, CA, USA, 2007.

[62] J. Hicks, N. Ramanathan, D. Kim, M. Monibi, J. Selsky, M. Hansen, and D. Estrin. Andwellness: An open mobile system for activity and experience sampling. In *Wireless Health 2010*, WH '10, pages 34–43, New York, NY, USA, 2010. ACM.

[63] C.-K. Hsieh, H. Tangmunarunkit, F. Alquaddoomi, J. Jenkins, J. Kang, C. Ketcham, B. Longstaff, J. Selsky, B. Dawson, D. Swendeman, D. Estrin, and N. Ramanathan. Lifestreams: A modular sense-making toolset for identifying important patterns from everyday life. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, SenSys '13, pages 5:1–5:13, New York, NY, USA, 2013. ACM.

[64] R. Hughes. James lind and the cure of scurvy: an experimental approach. *Medical history*, 19(04):342–351, 1975.

[65] S. S. Intille, J. Rondoni, C. Kukla, I. Ancona, and L. Bao. A context-aware experience sampling tool. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '03, pages 972–973, New York, NY, USA, 2003. ACM.

[66] G. D. Israel. Determining sample size. `http://edis.ifas.ufl.edu/pd006`, November 1992. [Online; accessed 10-April-2015].

[67] J. Sharkey. *CompareEverywhere*. `http://compare-everywhere.com`.

[68] J. Jaworska and M. Sydow. Behavioural targeting in on-line advertising: An empirical study. In J. Bailey, D. Maier, K.-D. Schewe, B. Thalheim, and X. Wang, editors, *Web Information Systems Engineering - WISE 2008*, volume 5175 of *Lecture Notes in Computer Science*, pages 62–76. Springer Berlin Heidelberg, 2008.

[69] M. Jones and G. Marsden. Mobile interaction design. 2006.

[70] N. Jones. The future smart home: 500 smart objects will enable new business opportunities. Technical report, Gartner Report, July 2014.

[71] S. Jumisko-Pyykkö and T. Vainio. Framing the context of use for mobile hci. *Int. J. Mob. Hum. Comput. Interact.*, 2(4):1–28, Oct. 2010.

[72] A. Kankainen. Ucpcd: user-centered product concept design. In *Proceedings of the 2003 conference on Designing for user experiences*, pages 1–13. ACM, 2003.

[73] M. Kellar, D. Reilly, K. Hawkey, M. Rodgers, B. MacKay, D. Dearman, V. Ha, W. J. MacInnes, M. Nunes, K. Parker, T. Whalen, and K. M. Inkpen. It's a jungle out there: Practical considerations for evaluation in the city. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '05, pages 1533–1536, New York, NY, USA, 2005. ACM.

[74] A. Khan, V. Ranjan, T.-T. Luong, R. Balan, and A. Misra. Experiences with performance tradeoffs in practical, continuous indoor localization. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2013 IEEE 14th International Symposium and Workshops on a*, pages 1–9, June 2013.

[75] B. Kim, J.-Y. Ha, S. Lee, S. Kang, Y. Lee, Y. Rhee, L. Nachman, and J. Song. Adnext: A visit-pattern-aware mobile advertising system for urban commercial complexes. In *Proceedings of the 12th Workshop on Mobile Computing Systems and Applications*, HotMobile '11, pages 7–12, New York, NY, USA, 2011. ACM.

[76] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 453–456, New York, NY, USA, 2008. ACM.

[77] I. König, B. N. Klein, and K. David. On the stability of context prediction. In *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, UbiComp '13 Adjunct, pages 471–480, New York, NY, USA, 2013. ACM.

[78] H. Korhonen, J. Arrasvuori, and K. Väänänen-Vainio-Mattila. Analysing user experience of personal mobile products through contextual factors. In *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, MUM '10, pages 11:1–11:10, New York, NY, USA, 2010. ACM.

[79] P. Korpipaa, J. Mantyjarvi, J. Kela, H. Keranen, and E.-J. Malm. Managing context information in mobile devices. *IEEE Pervasive Computing*, 2(3), July 2003.

[80] J. Krumm. Ubiquitous advertising: The killer application for the 21st century. *IEEE Pervasive Computing*, 10(1):66–73, Jan. 2011.

[81] J. Kukkonen, E. Lagerspetz, P. Nurmi, and M. Andersson. Betelgeuse: A platform for gathering and processing situational data. *Pervasive Computing, IEEE*, 8(2):49–56, April 2009.

[82] P. Lavrakas. *Encyclopedia of Survey Research Methods: A-M*. Number v. 1. Sage, 2008.

[83] K. Lee, J. Flinn, and B. Noble. The case for operating system management of user attention. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, HotMobile '15, pages 111–116, New York, NY, USA, 2015. ACM.

[84] Y. Lee, Y. Ju, C. Min, S. Kang, I. Hwang, and J. Song. Comon: cooperative ambience monitoring platform with continuity and benefit awareness. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*, MobiSys '12. ACM, 2012.

[85] Y. Lee, C. Min, C. Hwang, J. Lee, I. Hwang, Y. Ju, C. Yoo, M. Moon, U. Lee, and J. Song. Sociophone: Everyday face-to-face interaction monitoring platform using multi-phone sensor fusion. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '13, pages 375–388, New York, NY, USA, 2013. ACM.

[86] M. Leppaniemi and H. Karjaluoto. Factors influencing consumers willingness to accept mobile advertising: a conceptual model. *Int. J. Mob. Commun.*, 3(3):197–213, Mar. 2005.

[87] D. R. Lichtenstein, N. M. Ridgway, and R. G. Netemeyer. Price perceptions and consumer shopping behavior: A field study. *Journal of Marketing Research*, 30(2):pp. 234–245, 1993.

[88] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '13, pages 389–402, New York, NY, USA, 2013. ACM.

[89] D. Lin. An information-theoretic definition of similarity. In *ICML*. Morgan Kaufmann, 1998.

[90] G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80, Jan 2003.

[91] N. Mallat, M. Rossi, V. K. Tuunainen, and A. Öörni. The impact of use context on mobile services acceptance: The case of mobile ticketing. *Inf. Manage.*, 46(3):190–195, Apr. 2009.

[92] Maluuba Personal Assistant App. `http://techcrunch.com/2012/11/14/maluuba-launches-natural-language-processing-api-brings-siri-like-powers-to-any-app/`.

[93] G. Mark, D. Gudith, and U. Klocke. The cost of interrupted work: More speed and stress. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 107–110, New York, NY, USA, 2008. ACM.

[94] Measures of Variability. `http://onlinestatbook.com/2/summarizing_distributions/variability.html`.

[95] M. Mehl, S. Gosling, and J. Pennebaker. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90(5):862–877, 2006.

[96] M. Merisavo, S. Kajalo, H. Karjaluoto, V. Virtanen, S. Salmenkivi, M. Raulas, and M. Leppäniemi. An empirical study of the drivers of consumer acceptance of mobile advertising. *Journal of Interactive Advertising*, 7(2):41–50, 2007.

[97] G. Miaoulis and R. Michener. *An Introduction to Sampling*. Kendall/Hunt Publishing Company, 1976.

[98] E. Miluzzo, N. D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. B. Eisenman, X. Zheng, and A. T. Campbell. Sensing meets mobile social networks: The design, implementation and evaluation of the cenceme application. In *Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems*, SenSys '08, pages 337–350, New York, NY, USA, 2008. ACM.

[99] A. Misra and R. K. Balan. Livelabs: Initial reflections on building a large-scale mobile behavioral experimentation testbed. *SIGMOBILE Mob. Comput. Commun. Rev.*, 17(4):47–59, Dec. 2013.

[100] B. Mittal. An integrated framework for relating diverse consumer characteristics to supermarket coupon redemption. *Journal of Marketing Research*, 31(4):pp. 533–544, 1994.

[101] Mobiqpons. *Mobile Coupons*. `http://www.mobiqpons.com/`.

[102] S. Moriarty, N. Mitchell, and W. Wells. *Advertising: Principles and Practices*. Upper Saddle, New Jersey: Pearson Education, Canada, Ltd, 2009.

[103] C. Moxey, M. Edwards, O. Etzion, M. Ibrahim, S. Iyer, H. Lalanne, M. Monze, M. Peters, Y. Rabinovich, G. Sharon, and K. Stewart. IBM Redbooks | A Conceptual Model for Event Processing Systems, Mar. 2010.

[104] E. Mullen. Can different incentives influence participants choice? 2007.

[105] Multiple Regression. `http://web.csulb.edu/~msaintg/ppa696/696regmx.htm`.

[106] S. Munson and S. Consolvo. Exploring goal-setting, rewards, self-monitoring, and sharing to motivate physical activity. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2012 6th International Conference on*, pages 25–32, May 2012.

[107] K. Muralidharan, S. Gottipati, and R. Balan. Deal or no deal: Catering to user preferences. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, pages 199–202, March 2014.

[108] K. Muralidharan, S. Seshan, N. Ramasubbu, and R. K. Balan. Handling location uncertainty in event driven experimentation. In *Proceedings of the 8th ACM International Conference on Distributed Event-Based Systems*, DEBS '14, pages 206–212, New York, NY, USA, 2014. ACM.

[109] C. Narayanaswami, D. Coffman, M. C. Lee, Y. S. Moon, J. H. Han, H. K. Jang, S. McFaddin, Y. S. Paik, J. H. Kim, J. K. Lee, J. W. Park, and D. Soroker. Pervasive symbiotic advertising. In *Proceedings of the 9th workshop on Mobile computing systems and applications*, HotMobile '08, pages 80–85, New York, NY, USA, 2008. ACM.

[110] S. Nath, F. X. Lin, L. Ravindranath, and J. Padhye. Smartads: bringing contextual ads to mobile apps. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, MobiSys '13, pages 111–124, New York, NY, USA, 2013. ACM.

[111] D. Pacey, E. Dempster, M. Williams, A. Cawsey, D. Marwick, and L. MacKinnon. A toolkit for creating personalized presentations. In *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on*, pages 550–553, Oct 2003.

[112] L. Palen and M. Salzman. Voice-mail diary studies for naturalistic data capture under mobile conditions. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, CSCW '02, pages 87–95, New York, NY, USA, 2002. ACM.

[113] G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419, 2010.

[114] M. Papandrea, S. Giordano, S. Vanini, and P. Cremonese. Proximity marketing solution tailored to user needs. *2012 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, 0:1–3, 2010.

[115] K. Partridge and B. Begole. Activity-based advertising. In *Pervasive Advertising*, Human-Computer Interaction Series, pages 83–101. Springer, 2011.

[116] S. Patel, J. Kientz, G. Hayes, S. Bhat, and G. Abowd. Farther than you may think: An empirical investigation of the proximity of users to their mobile phones. In P. Dourish and A. Friday, editors, *UbiComp 2006: Ubiquitous Computing*, volume 4206 of *Lecture Notes in Computer Science*, pages 123–140. Springer Berlin Heidelberg, 2006.

[117] T. Pedersen. Information content measures of semantic similarity perform better without sense-tagged text. In *NAACL*, HLT, 2010.

[118] V. Pejovic and M. Musolesi. Interruptme: Designing intelligent prompting mechanisms for pervasive applications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, pages 897–908, New York, NY, USA, 2014. ACM.

[119] M. Pielot, R. de Oliveira, H. Kwak, and N. Oliver. Didn't you see my message?: Predicting attentiveness to mobile instant messages. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 3319–3328, New York, NY, USA, 2014. ACM.

[120] L. Pina, K. Rowan, A. Roseway, P. Johns, G. R. Hayes, and M. Czerwinski. In situ cues for adhd parenting strategies using mobile technology. In *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '14, pages 17–24, ICST, Brussels, Belgium, Belgium, 2014. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[121] B. Poppinga, W. Heuten, and S. Boll. Sensor-based identification of opportune moments for triggering notifications. *Pervasive Computing, IEEE*, 13(1):22–29, Jan 2014.

[122] R. Barnes. *ShopSavvy*. `http://www.biggu.com`.

[123] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas. Emotionsense: A mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, Ubicomp '10, pages 281–290, New York, NY, USA, 2010. ACM.

[124] M. Raento, A. Oulasvirta, R. Petit, and H. Toivonen. Contextphone: A prototyping platform for context-aware mobile applications. *IEEE Pervasive Computing*, 4(2):51–59, Apr. 2005.

[125] A. Ranganathan, J. Al-Muhtadi, and R. Campbell. Reasoning about uncertain contexts in pervasive computing environments. *Pervasive Computing, IEEE*, 3(2):62–70, April 2004.

[126] A. Ranganathan, J. Al-Muhtadi, S. Chetan, R. Campbell, and M. D. Mickunas. MiddleWhere: A middleware for location awareness in ubiquitous computing applications. In *Proceedings of the 5th ACM/IFIP/USENIX International Conference on Middleware*, Middleware '04, pages 397–416, New York, NY, USA, 2004. Springer-Verlag New York, Inc.

[127] A. Ranganathan and R. H. Campbell. Advertising in a pervasive computing environment. In *Proceedings of the 2nd international workshop on Mobile commerce*, WMC '02, pages 10–14, New York, NY, USA, 2002. ACM.

[128] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *CSCW*, pages 175–186, 1994.

[129] P. Resnick and H. R. Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, Mar. 1997.

[130] B. Ribeiro-Neto, M. Cristo, P. B. Golgher, and E. Silva de Moura. Impedance coupling in content-targeted advertising. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 496–503, New York, NY, USA, 2005. ACM.

[131] J. Rieman. The diary study: A workplace-oriented research tool to guide laboratory efforts. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, CHI '93, pages 321–326, New York, NY, USA, 1993. ACM.

[132] J. T. Roxane Edjlali, Nick Jones. Effects of mobility on information management. Technical report, Gartner Report, Aug 2012.

[133] A. Sahami Shirazi, N. Henze, T. Dingler, M. Pielot, D. Weber, and A. Schmidt. Large-scale assessment of mobile notifications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 3055–3064, New York, NY, USA, 2014. ACM.

[134] M. C. Sala, K. Partridge, L. Jacobson, and J. B. Begole. An exploration into activity-informed physical advertising using pest. In *Proceedings of the 5th international conference on Pervasive computing*, PERVASIVE'07, pages 73–90, Berlin, Heidelberg, 2007. Springer-Verlag.

[135] J. Salo. Retailer use of permission-based mobile advertising. *Advances in electronic marketing*, pages 139–156, 2005.

[136] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Itembased collaborative filtering recommendation algorithms. In *WWW*, pages 285–295, 2001.

[137] M. Satyanarayanan. Mobile computing: the next decade. In *Proceedings of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond*, MCS '10. ACM, 2010.

[138] A. Scharl, A. Dickinger, and J. Murphy. Diffusion and success factors of mobile marketing. *Electron. Commer. Rec. Appl.*, 4(2):159–173, July 2005.

[139] B. Schilit, N. Adams, and R. Want. Context-aware computing applications. In *Proceedings of the 1994 First Workshop on Mobile Computing Systems and Applications*, WMCSA '94, pages 85–90, Washington, DC, USA, 1994. IEEE Computer Society.

[140] A. Schmidt, M. Beigl, and H.-W. Gellersen. There is more to context than location. *Computers & Graphics*, 23(6):893 – 901, 1999.

[141] S. Schnall, J. Haidt, G. L. Clore, and A. H. Jordan. Disgust as Embodied Moral Judgment. *Personality and Social Psychology Bulletin*, 34(8):1096–1109, May 2008.

[142] K. F. Schulz, D. G. Altman, and D. Moher. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340, 2010.

[143] R. Sen, Y. Lee, K. Jayarajah, A. Misra, and R. K. Balan. Grumon: Fast and accurate group monitoring for heterogeneous urban spaces. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*, SenSys '14, pages 46–60, New York, NY, USA, 2014. ACM.

[144] Shopkick. `http://www.shopkick.com/`.

[145] B. Sibbald and M. Roland. Understanding controlled trials: Why are randomised controlled trials important? *BMJ*, 316(7126):201, 1998.

[146] S. Sigg, S. Haseloff, and K. David. An alignment approach for context prediction tasks in ubicomp environments. *Pervasive Computing, IEEE*, 9(4):90–97, October 2010.

[147] N. Singh and U. Varshney. Patterns of effective medication adherence: The role of wireless interventions. In *Wireless Telecommunications Symposium (WTS), 2014*, pages 1–10, April 2014.

[148] Siri. `http://www.apple.com/ios/siri/`.

[149] M. B. Skov and J. Kjeldskov. Creating Realistic Laboratory Settings: Comparative Studies of Three Think-Aloud Usability Evaluations of a Mobile System. In M. Rauterberg, M. Menozzi, and J. Wesson, editors, *INTERACT*. IOS Press, 2003.

[150] K. Suresh. An overview of randomization techniques: an unbiased assessment of outcome in clinical research. *Journal of human reproductive sciences*, 4(1):8, 2011.

[151] S. Swaminathan and K. Bawa. Category-specific coupon proneness: The impact of individual characteristics and category-specific variables. *Journal of Retailing*, 81(3):205 – 214, 2005.

[152] B. N. Taylor. *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*. DIANE Publishing, 2009.

[153] H. Taylor, A. Yochem, L. Phillips, and F. Martinez. *Event-Driven Architecture: How SOA Enables the Real-Time Enterprise*. Addison-Wesley Professional, 1st edition, 2009.

[154] X. Wang, D. Q. Zhang, T. Gu, and H. Pung. Ontology based context modeling and reasoning using owl. In *Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second IEEE Annual Conference on*, pages 18–22, March 2004.

[155] E. White, B. K. Armstrong, and R. Saracci. *Principles of exposure measurement in epidemiology: collecting, evaluating and improving measures of disease risk factors*. Oxford University Press, 2008.

[156] H. Wigelius and H. Vtj. Dimensions of context affecting user experience in mobile work. In T. Gross, J. Gulliksen, P. Kotz, L. Oestreicher, P. Palanque, R. Prates, and M. Winckler, editors, *Human-Computer Interaction INTER-ACT 2009*, volume 5727 of *Lecture Notes in Computer Science*, pages 604–617. Springer Berlin Heidelberg, 2009.

[157] Q. Xu, J. Erman, A. Gerber, Z. Mao, J. Pang, and S. Venkataraman. Identifying diverse usage behaviors of smartphone apps. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, IMC '11. ACM, 2011.

[158] B. Yan and G. Chen. Appjoy: personalized mobile application discovery. In *Proceedings of the 9th international conference on Mobile systems, applications, and services*, MobiSys '11. ACM, 2011.

[159] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen. How much can behavioral targeting help online advertising? In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 261–270, New York, NY, USA, 2009. ACM.

[160] Z. Yan, V. Subbaraju, D. Chakraborty, A. Misra, and K. Aberer. Energy-efficient continuous activity recognition on mobile phones: An activity-adaptive approach. In *Proceedings of the 2012 16th Annual International Symposium on Wearable Computers (ISWC)*, ISWC '12, pages 17–24, Washington, DC, USA, 2012. IEEE Computer Society.

[161] J. Yoon, B. D. Noble, M. Liu, and M. Kim. Building realistic mobility models from coarse-grained traces. In *Proceedings of the 4th International Conference on Mobile Systems, Applications and Services*, MobiSys '06, pages 177–190, New York, NY, USA, 2006. ACM.

[162] K. Zhang and Z. Katona. Contextual advertising. *Marketing Science*, 31(6):980–994, Nov. 2012.

[163] C. Zhou, P. Ludford, D. Frankowski, and L. Terveen. An experiment in discovering personally meaningful places from location data. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '05, pages 2029–2032, New York, NY, USA, 2005. ACM.

Appendix A

Database Schema

In this Appendix, I describe some of the key tables from the database schema supporting in situ experimentation. The database consists of 24 tables, as shown in Figure A.1, created using the postgresql DBMS. The database design meets the requirements of second normal form.

Details of a new experiment are stored in the following tables:

- experiment: The master table containing the key design specifications of the experiment such as the sample size and experiment duration.

- experiment_chain: In the case of a 'chained' experiment, details of the main experiment are stored in the experiment table while subsequent experiments in the chain are stored here.

- participant_spec: Stores triggering context information as well as demographic constraints of participants that need to be targeted.

  Information from these two tables are combined when registering with the Event Processing Agent (EPA).

The experiment intervention details are stored in the following tables:

- promotion, event: Interventions of type promotion and event are created through the content management portal and stored in the promotion and event table respectively.

- notices: Stores information on interventions of type survey, link and general. Details of the survey associated with a promotion or event (if created) are also stored in this table.
- system_message: Stores details of the reminder message associated with a promotion or event.

The following tables are used to store state information regarding selected participants of an experiment:

- exp_users_list: Contains IDs of participants selected for an experiment. These IDs map to details of the registered LiveLabs participants stored in the user_info table.
- promotion_view, event_view, system_message_view, promotion_like: Captures participant interaction with the experiment intervention.
- livelabs_tracking: Captures participant interaction with the LiveLabs mobile apps (clicking the notification, opening the application etc.)

Other tables such as bep_user, vendor_details, survey_sent are used for administrative and housekeeping purposes.

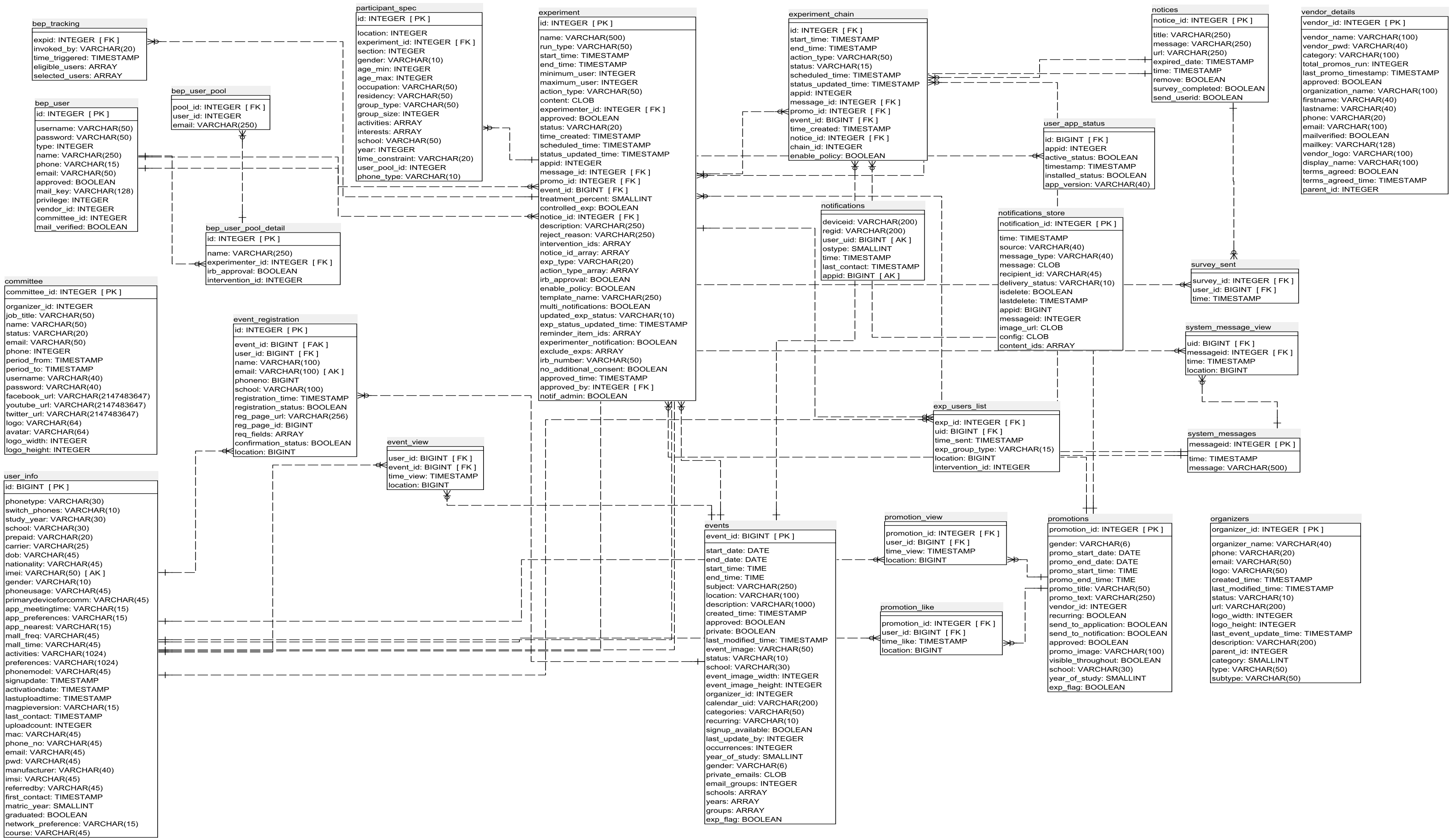


Figure A.1: Database Schema supporting the Behavioural Experimentation System

Appendix B

LiveLabs Context Collector

Context Type	Android	iOS
Location	GPS & Wi-Fi scans	Location API
Phone States	Attached celltowers, battery levels, network statistics, website visited,calendar, settings, contacts, media state	Battery level,calendar, contacts, settings, media states
Phone Events	Apps installed, Apps in use, Screen interactions, call & texting events	App state
Activity	accelerometer	accelerometer (only in foreground)
Raw Sensors	accel, compass, gyro, barometer, light	accel, compass, gyro, light (only in foreground)

Table B.1: Context Collected

The Context Collector run as a background service in the user space on both Android and iOS platforms. For Android, the Collector can be configured to collect sensor data, phone events, client-side indoor location coordinates, etc. Table B.1 details the various key events that are collected. For iOS, a background application gets a foreground processing timeslice once every 10 minutes; at that point, our collector can collect sensor data, limited location information and a much smaller set of phone events (see Table B.1). The collection policies of the Context Collector (what subset of sensors and events to collect, and frequently) are configurable by the LiveLabs server infrastructure, through a set of XML-based policies. Such policy-based collection allows us the flexibility of continuously adapting the granularity and fidelity of the sensor collection (with corresponding impact on the power drain) for each individual sensor. The current version of the Context Collector does not upload the collected data continuously; while the upload interval is configurable, it is currently set to 3 hours by default (implying a total of 8 collection episodes a day). Moreover, uploading happens only if the participant is on the SMU campus and connected to the SMU Wi-Fi network, so as to alleviate concerns about cellular data usage costs.[1]

[1] The text and table are an excerpt from a paper about LiveLabs [99]. Additional details of the LiveLabs platform can be found in the same paper.

Appendix C

User Study Documentation

In this Appendix, I list details of the artifacts used in validating this thesis. These primarily include details of the promotions and other interventions sent as part of an experiment as well as questionnaires assigned to participants to collect self-reporting information.

C.1 Experiment 1

Refer to Section 3.7.1 for details of the experiment.

C.1.1 Intervention Details

Figure C.1 shows the promotion used as an intervention for the experiment.

Figure C.1: Promotion used for Experiment 1

C.2 Experiment 2

Refer to Section 3.7.2 for details of the experiment.

C.2.1 Intervention Details

Figure C.2 shows the three promotions used as interventions for the experiment.

a) b) c)

Figure C.2: Promotions used for Experiment 2

C.2.2 Self-Report Questionnaire

Promotion Feedback Survey

User ID:*Require
Do not modify.

I liked the promotion that was sent.
Ref: Exclusive offer from MegaFash

- Strongly agree
- Somewhat agree
- Neither agree or disagree
- Somewhat disagree
- Strongly disagree

C.3 Experiment 4

Refer to Section 3.7.4 for details of the experiment.

C.3.1 ABIS Questionnaire

DIRECTIONS: People differ in the ways they act and think in different situations. This is a test to measure some of the ways in which you act and think. Read each statement and put an X on the appropriate circle on the right side of this page. Do not spend too much time on any statement. Answer quickly and honestly.

①	②	③	④
Rarely/Never	Occasionally	Often	Almost Always/Always

I am a careful thinker.	①	②	③	④
I plan trips well ahead of time.	①	②	③	④
I do things without thinking.	①	②	③	④
I concentrate easily.	①	②	③	④
I plan for job security.	①	②	③	④
I act “on impulse.”	①	②	③	④
I am self controlled.	①	②	③	④
I say things without thinking.	①	②	③	④
I don’t “pay attention.”	①	②	③	④
I act on the spur of the moment.	①	②	③	④
I plan tasks carefully.	①	②	③	④
I am a steady thinker.	①	②	③	④
I am future oriented.	①	②	③	④
I am the life of the party.	①	②	③	④
I feel comfortable around people.	①	②	③	④
I start conversations.	①	②	③	④
I talk to a lot of different people at parties.	①	②	③	④
I don't mind being the center of attention.	①	②	③	④
I don't talk a lot.	①	②	③	④
I keep in the background.	①	②	③	④
I have little to say.	①	②	③	④
I don't like to draw attention to myself.	①	②	③	④
I am quiet around strangers.	①	②	③	④

ABIS (ABbreviated Impulsiveness Scale) Administration and Scoring

The Abbreviated Impulsiveness Scale (ABIS) provides an efficient, reliable, valid, and generalizable measure of attentional, motor, and non-planning impulsiveness. The ABIS can be used as a brief alternative to the BIS-11 or as a model for reanalyzing previously collected BIS-11 questionnaire responses.

12	I am a careful thinker. (Reverse Scored)	①	②	③	④
7	I plan trips well ahead of time. (Reverse Scored)	①	②	③	④
2	I do things without thinking.	①	②	③	④
9	I concentrate easily. (Reverse Scored)	①	②	③	④
13	I plan for job security. (Reverse Scored)	①	②	③	④
17	I act "on impulse."	①	②	③	④
8	I am self controlled. (Reverse Scored)	①	②	③	④
14	I say things without thinking.	①	②	③	④
5	I don't "pay attention."	①	②	③	④
19	I act on the spur of the moment.	①	②	③	④
1	I plan tasks carefully. (Reverse Scored)	①	②	③	④
20	I am a steady thinker. (Reverse Scored)	①	②	③	④
30	I am future oriented. (Reverse Scored)	①	②	③	④

ABIS item order (using BIS-11 item numbering): 12, 7, 2, 9, 13, 17, 8, 14, 5, 19, 1, 20, 30

ABIS Scales:
Attention (5 items): 12, 9, 8, 5, 20
Motor (4 items): 2, 17, 14, 19
Nonplanning (4 items): 7, 13, 1, 30

Reverse-scored items (4, 3, 2, 1): 12, 7, 9, 13, 8, 1, 20, 30
Standard-scored items (1, 2, 3, 4): 2, 17, 14, 5, 19

To score each scale, take the average of the scores for each item on that scale (after reverse-scoring the specified items). Do not average across separate scales to produce combined scores.

References:

Coutlee, C.G., Politzer, C.S., Hoyle, R.H., Huettel, S.A. (Under Review). An Abbreviated Impulsiveness Scale Constructed Through Confirmatory Factor Analysis of the BIS-11.

The BIS-11 items used as the basis for the ABIS can be found at: http://www.impulsivity.org/

C.4 Assessment of Mobile Notifications

Refer to Section 3.7.5 for details of the study.

C.4.1 User Study Questionnaire

Research Participant Information and Consent/Authorization form for assessment of mobile notifications with minimal risk

a.) This user study has been commissioned for surveying amongst smartphone owners into the use and perception of mobile notifications.

b.) You will be asked to: Fill up a questionnaire on your usage patterns and perception on mobile notifications. The questionnaire should not take more than 5 minutes to complete.

c.) Your participation in the study is voluntary. You will be paid an amount of 5 SGD as a token of appreciation of completing the survey. You will need to provide your Name and NRIC/FIN number to receive the payment.

d.) There is no anticipated risk of injury or discomfort from participating in the study.

e.) You will not incur any monetary costs in participating in this study

Confidentiality

By participating in the study, you understand and agree that Singapore Management University may be required to disclose your consent form, data and other personally identifiable information as required by law, regulation, subpoena or court order. Otherwise, your confidentiality will be maintained in the following manner:

Your data, consent form and payment receipt will be kept separate. Your consent form will be stored in a secure location and will not be disclosed to third parties – in this manner, the data/information is stripped of any personally identifiable elements and becomes "anonymised". By participating, you understand and agree that the anonymised data and information gathered during this study may be used by Singapore Management University and published and/or disclosed by Singapore Management University to others outside of Singapore Management University. However, your name in your consent form, and contact information obtained during registration will not be mentioned in any such publication or dissemination of the anonymised research data and/or results by Singapore Management University.

Your confidentiality will be maintained during data analysis and publication/presentation of results using the following means: (1) You will be assigned a number, and your name will not be recorded with the number. The identity of the participants cannot be traced from the assigned numbers. (2) The researchers will save the data file by your number, not by name. (3) Only members of the research group will view any data unless you give them the optional permission to release them. (4) Any files will be stored in a secured location accessed only by authorized researchers.

Voluntary Participation

I understand that participation is voluntary. Refusal to participate will involve no penalty. I understand that I may discontinue participation at any time without penalty or loss of accrued benefits (Benefits are accrued in proportion to the amount of study completed or as otherwise stated by the researcher) to which I am otherwise entitled.

Age

Please note that you must be at least 18 years of age in order to participate in this study. If you agree to participate in this study you shall be taken to have declared that you are at least 18 years of age.

Right to Ask Questions & Contact Information

If you have questions about the study, desire additional information, or wish to withdraw your participation please contact:

Kartik Muralidharan
School of Information Systems
80 Stamford Road Singapore 178902
Email: kartikm.2010@smu.edu.sg
Phone: + (65) 83368909

If you have questions pertaining to your rights as a research participant; or to report objections to this study, you should contact:

IRB Secretariat
Singapore Management University
81 Victoria Street Singapore 188065
Email: irb@smu.edu.sg
Phone: 65-6828-1925

A G R E E M E N T

I have read the above terms and understand the nature of this study and give my consent to participate in it. I consent to the collection and use of my personal data in respect of the abovementioned research and agree that my consent has been reasonably obtained and is given in the manner required by the Personal Data Protection Act ("PDPA"). I further approve of the use and disclosure of the anonymised data and information that I give in this study for SMU's research purposes, including the sharing and dissemination of the anonymised data and information with external research collaborators, and also any other use or disclosure permitted under the PDPA and give SMU permission to reproduce the same in written or oral form where appropriate. I understand that the confidentiality of my personal information will be preserved with regard to the above.

Name *

Date *

Please click on the box below if you consent to participate in this research study. Otherwise, please close the browser now to exit. *

☐ I certify that I accept the above terms and that I am above the age of 18

Part A: Mobile Notifications

Today’s smartphones often use notifications to attract a user’s attention— when there’s an incoming text message or upcoming event, for example. We wish to understand the kind of notifications you receive on your mobile device and how you choose to address the alerts. Read each statement and select the appropriate response. Please answer honestly.

What is the OS of your current mobile device? *

- ○ Android
- ○ iOS

How important is your phone to you? *

- ○ Not very important. I can go for a day or more without it.
- ○ Somewhat important. A few hours without it probably won’t do any harm.
- ○ Very important. I have to have it with me all the time.

How would you rate your level of phone expertise? *

- ○ Novice - I use only the basic settings on my phone
- ○ Intermediate - I use some of the advanced setting on my phone
- ○ Expert - I know how to root the phone if I wanted to.

Do you know how to turn off notifications for selected apps on your phone? *

- ○ Yes
- ○ No
- ○ I didn't know I could do that!
- ○ Not sure what you are talking about.

How many notifications do you receive on your phone in a day? *

- ○ < 10
- ○ 10 - 20
- ○ 20 - 30
- ○ 30 - 40
- ○ > 40

How frequently do you check your phone for notifications? *

- ○ Only when my phone indicates that there is a notification (Blinking Light / Sound)
- ○ Once every few minutes.
- ○ Once every 10 minutes.
- ○ Once every 30 minutues
- ○ Once every 60 minutes
- ○ At most few times a day

What are the reasons for checking your phone? *

- I specifically check for missed notifications.
- Checking the time.
- Out of boredom.
- Force of habit.
- Other:

What are the reasons for not taking action on receiving a notification? *
For example when you receive an email notification, for what reason would you ignore the message.

- The notification isnt important enough.
- The notification looks like spam.
- Not enough time.
- I do not know the sender of the notification.
- Other:

Part B: Influence of Application Category

For each of the below application categories please give your notification preference.

Importance *
How relatively important are notifications related to the following application categories to you?

	Very Important	Important	Somewhat Important	Not Important
Calls & SMS				
Social applications (WhatsApp, Facebook)				
Reminders (Calendar)				
Promotions				
News				
App Updates				
Other app notifications (games, weather etc)				

Action *
What action do you generally take upon receiving a notification related to the following application categories?

	Click through immediately.	Click through later.	Ignore
Calls & SMS			
Social applications (WhatsApp, Facebook)			
Reminders (Calendar)			
Promotions			
News			
App Updates			
Other app notifications (games, weather etc)			

Timing *
Do you prefer receiving real-time or periodical notifications given the following application categories?

	Real-time notifications.	Periodical notifications.	Not sure.
Calls & SMS			
Social applications (WhatsApp, Facebook)			
Reminders (Calendar)			
Promotions			
News			
App Updates			
Other app notifications (games, weather etc)			

Selection *
Do you elect receiving all or only important notifications for the following application categories?

	All notifications.	Only important notifications.	Not sure.
Calls & SMS	○	○	○
Social applications (WhatsApp, Facebook)	○	○	○
Reminders (Calendar)	○	○	○
Promotions	○	○	○
News	○	○	○
App Updates	○	○	○
Other app notifications (games, weather etc)	○	○	○

Vibrations *
How often do you prefer notifications in the following applications categories to be accompanied by vibrations?

	Always.	Only when I switch vibrations for notifications on.	Never
Calls & SMS	○	○	○
Social applications (WhatsApp, Facebook)	○	○	○
Reminders (Calendar)	○	○	○
Promotions	○	○	○
News	○	○	○
App Updates	○	○	○
Other app notifications (games, weather etc)	○	○	○

Sound *
How often do you elect notifications in the following categories to be accompanied by sound of any kind?

	Always.	Only when I switch the sound for notifications on.	Never
Calls & SMS	○	○	○
Social applications (WhatsApp, Facebook)	○	○	○
Reminders (Calendar)	○	○	○
Promotions	○	○	○
News	○	○	○
App Updates	○	○	○
Other app notifications (games, weather etc)	○	○	○

Part C: LiveLabs App Notifications

Which of the following applications do you have installed on your mobile device? *

- ☐ LiveLabs
- ☐ Smuddy
- ☐ Eva

What are the reasons for not having the above apps on your phone? *

- ○ I did not find them useful
- ○ I was receiving too many notifications
- ○ I needed some space on my smartphone.
- ○ Other: ____

Have you turned off notifications for any of the following applications *

	Turned On but I dismiss them	Turned Off	Not sure	I do not have the app installed on my phone	I always click on the notifications from this app.
LiveLabs	○	○	○	○	○
Smuddy	○	○	○	○	○
Eva	○	○	○	○	○

How useful are the promotions you receive from LiveLabs? *

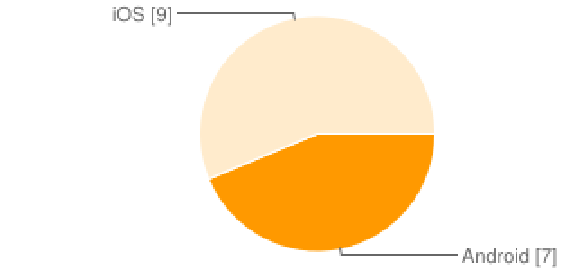
- Very Useful
- Somewhat Useful
- Not Useful at all.
- I consider it as spam
- I never knew promotions were sent.

Would you like to give us any feedback about our mobile applications or any additional opinion on mobile notifications? *

C.4.2 Summary of Responses

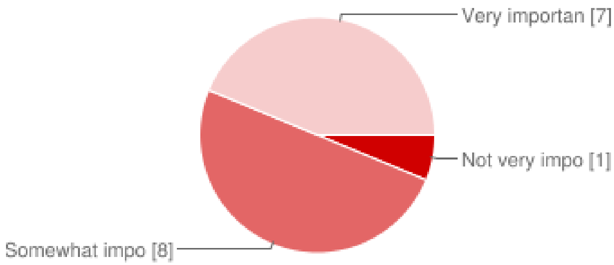
Part A: Mobile Notifications

What is the OS of your current mobile device?



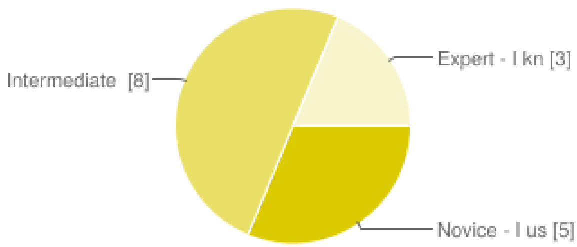
Android	**7**	43.8%
iOS	**9**	56.3%

How important is your phone to you?



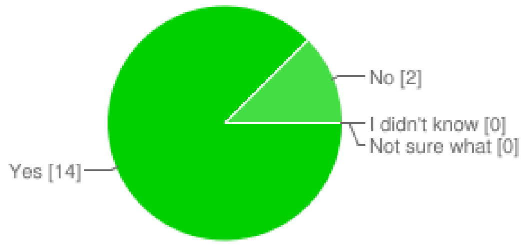
Not very important. I can go for a day or more without it.	**1**	6.3%
Somewhat important. A few hours without it probably won't do any harm.	**8**	50%
Very important. I have to have it with me all the time.	**7**	43.8%

How would you rate your level of phone expertise?



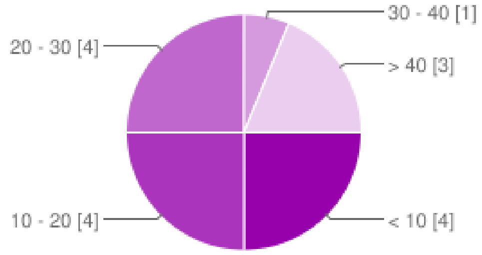
Novice - I use only the basic settings on my phone	**5**	31.3%
Intermediate - I use some of the advanced setting on my phone	**8**	50%
Expert - I know how to root the phone if I wanted to.	**3**	18.8%

Do you know how to turn off notifications for selected apps on your phone?



Yes	**14**	87.5%
No	**2**	12.5%
I didn't know I could do that!	**0**	0%
Not sure what you are talking about.	**0**	0%

How many notifications do you receive on your phone in a day?

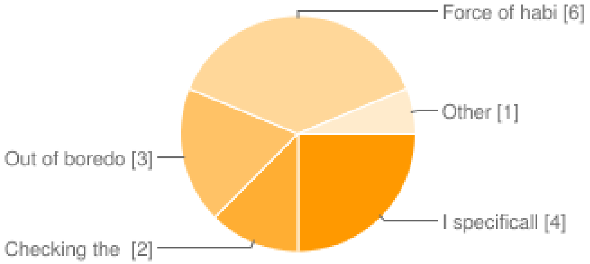


< 10	**4**	25%
10 - 20	**4**	25%
20 - 30	**4**	25%
30 - 40	**1**	6.3%
> 40	**3**	18.8%

How frequently do you check your phone for notifications?

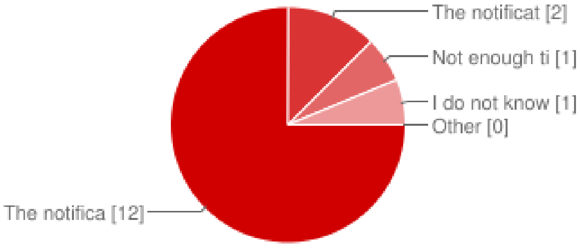
Only when my phone indicates that there is a notification (Blinking Light / Sound)	**7**	43.8%
Once every few minutes.	**1**	6.3%
Once every 10 minutes.	**4**	25%
Once every 30 minutues	**2**	12.5%
Once every 60 minutes	**0**	0%
At most few times a day	**2**	12.5%

What are the reasons for checking your phone?



I specifically check for missed notifications.	**4**	25%
Checking the time.	**2**	12.5%
Out of boredom.	**3**	18.8%
Force of habit.	**6**	37.5%
Other	**1**	6.3%

What are the reasons for not taking action on receiving a notification?



The notification isnt important enough.	**12**	75%
The notification looks like spam.	**2**	12.5%
Not enough time.	**1**	6.3%
I do not know the sender of the notification.	**1**	6.3%
Other	**0**	0%

Part B: Influence of Application Category

Calls & SMS [Importance]

Very Important	**7**	43.8%
Important	**9**	56.3%
Somewhat Important	**0**	0%
Not Important	**0**	0%

Social applications (WhatsApp, Facebook) [Importance]

Very Important	**3**	18.8%
Important	**11**	68.8%
Somewhat Important	**2**	12.5%
Not Important	**0**	0%

Reminders (Calendar) [Importance]

Very Important	**4**	25%
Important	**8**	50%
Somewhat Important	**3**	18.8%
Not Important	**1**	6.3%

Promotions [Importance]

Very Important	**1**	6.3%
Important	**1**	6.3%
Somewhat Important	**4**	25%
Not Important	**10**	62.5%

News [Importance]

Very Important	**0**	0%
Important	**0**	0%
Somewhat Important	**8**	50%
Not Important	**8**	50%

App Updates [Importance]

Very Important	**0**	0%
Important	**2**	12.5%
Somewhat Important	**0**	0%
Not Important	**14**	87.5%

Other app notifications (games, weather etc) [Importance]

Very Important	**0**	0%
Important	**1**	6.3%
Somewhat Important	**1**	6.3%
Not Important	**14**	87.5%

Calls & SMS [Action]

Click through immediately.	**15**	93.8%
Click through later.	**1**	6.3%
Ignore	**0**	0%

Social applications (WhatsApp, Facebook) [Action]

Click through immediately.	**12**	75%
Click through later.	**4**	25%
Ignore	**0**	0%

Reminders (Calendar) [Action]

Click through immediately.	**11**	68.8%
Click through later.	**4**	25%
Ignore	**1**	6.3%

Promotions [Action]

Click through immediately.	**1**	6.3%
Click through later.	**8**	50%
Ignore	**7**	43.8%

News [Action]

Click through immediately.	**1**	6.3%
Click through later.	**7**	43.8%
Ignore	**8**	50%

App Updates [Action]

Click through immediately.	**1**	6.3%
Click through later.	**6**	37.5%
Ignore	**9**	56.3%

Other app notifications (games, weather etc) [Action]

Click through immediately.	**0**	0%
Click through later.	**5**	31.3%
Ignore	**11**	68.8%

Calls & SMS [Timing]

Real-time notifications.	**15**	93.8%
Periodical notifications.	**1**	6.3%
Not sure.	**0**	0%

Social applications (WhatsApp, Facebook) [Timing]

Real-time notifications.	**16**	100%
Periodical notifications.	**0**	0%
Not sure.	**0**	0%

Reminders (Calendar) [Timing]

Real-time notifications.	**11**	68.8%
Periodical notifications.	**4**	25%
Not sure.	**1**	6.3%

Promotions [Timing]

Real-time notifications.	**4**	25%
Periodical notifications.	**10**	62.5%
Not sure.	**2**	12.5%

News [Timing]

Real-time notifications.	**2**	12.5%
Periodical notifications.	**12**	75%
Not sure.	**2**	12.5%

App Updates [Timing]

Real-time notifications.	**2**	12.5%
Periodical notifications.	**11**	68.8%
Not sure.	**3**	18.8%

Other app notifications (games, weather etc) [Timing]

Real-time notifications.	**1**	6.3%
Periodical notifications.	**10**	62.5%
Not sure.	**5**	31.3%

Calls & SMS [Selection]

All notifications.	**14**	87.5%
Only important notifications.	**2**	12.5%
Not sure.	**0**	0%

Social applications (WhatsApp, Facebook) [Selection]

All notifications.	**8**	50%
Only important notifications.	**8**	50%
Not sure.	**0**	0%

Reminders (Calendar) [Selection]

All notifications.	**6**	37.5%
Only important notifications.	**9**	56.3%
Not sure.	**1**	6.3%

Promotions [Selection]

All notifications.	**0**	0%
Only important notifications.	**13**	81.3%
Not sure.	**3**	18.8%

News [Selection]

All notifications.	**2**	12.5%
Only important notifications.	**10**	62.5%
Not sure.	**4**	25%

App Updates [Selection]

All notifications.	**2**	12.5%
Only important notifications.	**10**	62.5%
Not sure.	**4**	25%

Other app notifications (games, weather etc) [Selection]

All notifications.	**0**	0%
Only important notifications.	**10**	62.5%
Not sure.	**6**	37.5%

Calls & SMS [Vibrations]

Always.	**8**	50%
Only when I switch vibrations for notifications on.	**8**	50%
Never	**0**	0%

Social applications (WhatsApp, Facebook) [Vibrations]

Always.	**4**	25%
Only when I switch vibrations for notifications on.	**11**	68.8%
Never	**1**	6.3%

Reminders (Calendar) [Vibrations]

Always.	**4**	25%
Only when I switch vibrations for notifications on.	**11**	68.8%
Never	**1**	6.3%

Promotions [Vibrations]

Always.	**1**	6.3%
Only when I switch vibrations for notifications on.	**7**	43.8%
Never	**8**	50%

News [Vibrations]

Always.	**1**	6.3%
Only when I switch vibrations for notifications on.	**7**	43.8%
Never	**8**	50%

App Updates [Vibrations]

Always.	**1**	6.3%
Only when I switch vibrations for notifications on.	**7**	43.8%
Never	**8**	50%

Other app notifications (games, weather etc) [Vibrations]

Always.	**1**	6.3%
Only when I switch vibrations for notifications on.	**6**	37.5%
Never	**9**	56.3%

Calls & SMS [Sound]

Always.	**5**	31.3%
Only when I switch the sound for notifications on.	**7**	43.8%
Never	**4**	25%

Social applications (WhatsApp, Facebook) [Sound]

Always.	**1**	6.3%
Only when I switch the sound for notifications on.	**10**	62.5%
Never	**5**	31.3%

Reminders (Calendar) [Sound]

Always.	**2**	12.5%
Only when I switch the sound for notifications on.	**7**	43.8%
Never	**7**	43.8%

Promotions [Sound]

Always.	**0**	0%
Only when I switch the sound for notifications on.	**6**	37.5%
Never	**10**	62.5%

News [Sound]

Always.	**0**	0%
Only when I switch the sound for notifications on.	**6**	37.5%
Never	**10**	62.5%

App Updates [Sound]

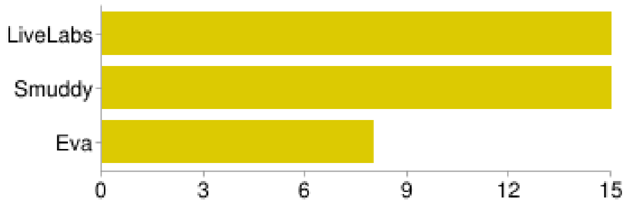
Always.	**0**	0%
Only when I switch the sound for notifications on.	**6**	37.5%
Never	**10**	62.5%

Other app notifications (games, weather etc) [Sound]

Always.	**0**	0%
Only when I switch the sound for notifications on.	**6**	37.5%
Never	**10**	62.5%

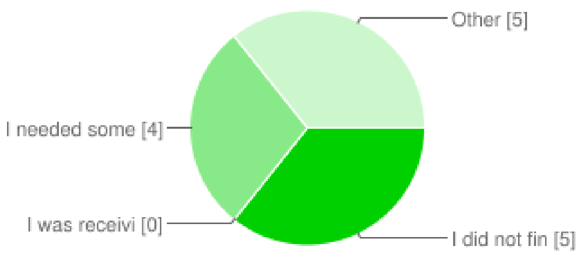
Part C: LiveLabs App Notifications

Which of the following applications do you have installed on your mobile device?



LiveLabs	**15**	93.8%
Smuddy	**15**	93.8%
Eva	**8**	50%

What are the reasons for not having the above apps on your phone?



I did not find them useful	**5**	35.7%
I was receiving too many notifications	**0**	0%
I needed some space on my smartphone.	**4**	28.6%
Other	**5**	35.7%

LiveLabs [Have you turned off notifications for any of the following applications]

Turned On but I dismiss them	**6**	37.5%
Turned Off	**2**	12.5%
Not sure	**4**	25%
I do not have the app installed on my phone	**0**	0%
I always click on the notifications from this app.	**2**	12.5%

Smuddy [Have you turned off notifications for any of the following applications]

Turned On but I dismiss them	**6**	37.5%
Turned Off	**3**	18.8%
Not sure	**4**	25%
I do not have the app installed on my phone	**0**	0%
I always click on the notifications from this app.	**2**	12.5%

Eva [Have you turned off notifications for any of the following applications]

Turned On but I dismiss them	**3**	18.8%
Turned Off	**3**	18.8%
Not sure	**4**	25%
I do not have the app installed on my phone	**4**	25%
I always click on the notifications from this app.	**1**	6.3%

How useful are the promotions you receive from LiveLabs?

Very Useful	**0**	0%
Somewhat Useful	**8**	50%
Not Useful at all.	**3**	18.8%
I consider it as spam	**3**	18.8%
I never knew promotions were sent.	**2**	12.5%

C.5 myDeal

Refer to Section 4.2.5 for details of the study.

C.5.1 User Study Questionnaire

Research Participant Information and Consent/Authorization form for minimal risk studies

a.) This user study has been commissioned for testing a novel mobile shopping assistant application.

b.) You will be asked to:
1. Take part in a lab study and perform several tasks (at most 30) in our simulated environment using our new application.

2. Fill out questionnaires about the shopping assistant application.

c.) The total time taken to engage in the lab-study part of this experiment, inclusive of training, will not exceed 1 hour.

d.) Your participation in the study is voluntary

e.) There will be no significant risk of injury or discomfort from participating in the study.

f.) You will not incur any monetary costs in participating in this study

g.) Your actions and opinions may be recorded for academic purposes at any time during the study. Any personal information divulged will be kept strictly confidential

h.) Your participation in this study may not result in any personal benefit to the initiators, but will contribute to ongoing research on the case for a digital wallet

i.) This study is completely funded by the Singapore Management University

j.) You will receive S$ 20 for successfully completing the study as a payment for your participation time and related efforts. The amount must be collected as directed from the Office of Finance in the university

k.) I understand that participation is voluntary. Refusal to participate will involve no penalty. I understand that I may discontinue participation at any time without penalty or loss of accrued benefits (Benefits are accrued in proportion to the amount of study completed or as otherwise stated by the researcher) to which I am otherwise entitled. I declare that I am at least 18 years of age.

Confidentiality

By participating in the study, you understand and agree that Singapore Management University may be required to disclose your consent form, data and other personally identifiable information as required by law, regulation, subpoena or court order. Otherwise, your confidentiality will be maintained in the following manner:

Your data and consent form will be kept separate. Your consent form will be stored in a secure location and will not be disclosed to third parties. By participating, you understand and agree that the data and information gathered during this study may be used by Singapore Management University and published and/or disclosed by Singapore Management University to others outside of Singapore Management University. However, your name, address, contact information and other direct personal identifiers in your consent form will not be mentioned in any such publication or dissemination of the research data and/or results by Singapore Management University.

Your confidentiality will be maintained during data analysis and publication/presentation of results using the following means: (1) You will be assigned a number, and your name will not be recorded with the number. The identity of the participants cannot be traced from the assigned numbers. (2) The researchers will save the data file by your number, not by name. (3) Only members of the research group will view the collected data. (4) All data will be stored in a secured location accessed only by authorized researchers.

Right to Ask Questions & Contact Information

If you have any questions about this study, you should feel free to ask them now. If you have questions later, desire additional information, or wish to withdraw your participation please contact:

Assistant Professor Rajesh Krishna BALAN
School of Information Systems
80 Stamford Road
Singapore 178902
Email: rajesh@smu.edu.sg
Phone: + (65) 6828 0879
Fax: + (65) 6828 0919

If you have questions pertaining to your rights as a research participant; or to report objections to this study, you should contact:

IRB Secretariat, Ms Stephanie Tan
Singapore Management University
81 Victoria Street
Singapore 188065
Email: irb@smu.edu.sg
Phone: 65-6828-1925

A G R E EM E N T

I have read the above terms and understood the nature of this study and give my consent to participate in it. I approve of the usage if the information that I give in this study for research purposes, and give the supervisors and their associates' permission to reproduce the same in written or oral form where appropriate. I understand that the confidentiality of my personal information will be preserved with regard to the above.

_________________________ _________________

SIGNATURE OF PARTICIPANT DATE

____

FULL NAME OF PARTICIPANT

I have explained and defined in detail the research procedures in which the subject (legal representative has given consent) has consented to participate.

_________________________________ __ _________________

SIGNATURE OF PRINCIPAL INVESTIGATOR DATE

____

FULL NAME OF PRINCIPAL INVESTIGATOR

Pre-Testing User Questionnaire

Please circle one option for the following questions:

a.) Have you used a touch phone? Yes No

b.) To what extent do you use the following features on your mobile phone (if any)

Calendar	Very Often	Sometimes	Rarely	Not at All
Maps	Very Often	Sometimes	Rarely	Not at All
Email	Very Often	Sometimes	Rarely	Not at All
Video Recording	Very Often	Sometimes	Rarely	Not at All
Taking Pictures	Very Often	Sometimes	Rarely	Not at All
Playing Games	Very Often	Sometimes	Rarely	Not at All
Music	Very Often	Sometimes	Rarely	Not at All
Data Synchronization	Very Often	Sometimes	Rarely	Not at All
Bluetooth Pair	Very Often	Sometimes	Rarely	Not at All
Installing Apps	Very Often	Sometimes	Rarely	Not at All
Rooting / Jail-break	Very Often	Sometimes	Rarely	Not at All
Browsing Internet	Very Often	Sometimes	Rarely	Not at All
Watch TV	Very Often	Sometimes	Rarely	Not at All

c.) To what extent do you browse for promotions/deals on the Internet from your phone?

Very Often Sometimes Rarely Not at All

d.) To what extent do you use applications that show you promotions/deals near you on the phone?

Very Often Sometimes Rarely Not at All

e.) How **important** is your phone to you? (**Please tick one option)**

1. Not very important. I can go for a day or more without it. [ ]
2. Somewhat important. A few hours without it probably won't do any harm. [ ]
3. Very important. I have to have it with me all the time. [ ]

PARTICIPANT CODE:

End of Task Questionnaire

a.) This task was **quick** to complete. **[Please circle one of the options below]**

Strongly Agree	Somewhat Agree	Neutral	Somewhat Disagree	Strongly Disagree

b.) This task was **easy** to perform. **[Please circle one of the options below]**

Strongly Agree	Somewhat Agree	Neutral	Somewhat Disagree	Strongly Disagree

c.) I am **confident** that the deal I chose was the best. **[Please circle one of the options below]**

Strongly Agree	Somewhat Agree	Neutral	Somewhat Disagree	Strongly Disagree

End of Study Questionnaire

1. **According to you which system rank ordered the deals in the best possible way? [Please circle one of the options below]**

   System 1 System 2 System 3 No preference I don't know

2. **Which system was the easiest to perform the task set on?**

   System 1 System 2 System 3

3. **Overall which system did you prefer**? **[Please circle one of the options below]**

   System 1 System 2 System 3

   Why?

   ____

4. What type of deals do you generally prefer? **[Please circle one of the options below]**

   Cash Back Discount Vouchers Free Stuff

5. (a)How important is location of the deal as opposed to savings? For example would you travel further if it meant getting a better offer?

   4. Very important. I rather go to someplace close to my current location. [ ]
   5. Somewhat important. An hour of travel won't do any harm. [ ]
   6. Not very important. I do not mind traveling longer distances if the saving is more. [ ]

   (b) Is this true for all types of products/activities?
   Yes No

   (c) If your answer was No, which type of products/activities are you referring to?

   ____

6. Would you be willing to pay for a system that provides a ranking of deals?
   Yes No

7. Would you prefer if the system showed deals based on your personal preferences, calendar, and time of year and so on?

   Yes No

8. What other category of deals, besides dining, would you like to see?

   ____

9. Your comments/suggestions on the myDeal application, or this study on a whole. Please **highlight** aspects of the tasks that you found challenging or areas that could be improved further.

____

____