Recommender Systems for Healthy Behavior Change

THÈSE N^{O} 7973 (2017)

PRÉSENTÉE LE 29 SEPTEMBRE 2017
À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS
GROUPE SCI IC PFP
PROGRAMME DOCTORAL EN INFORMATIQUE ET COMMUNICATIONS

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Onur YÜRÜTEN

acceptée sur proposition du jury:

Prof. D. Gatica-Perez, président du jury
Dr P. Pu Faltings, directrice de thèse
Prof. L. Chen, rapporteuse
Dr J. Zhang, rapporteur
Dr R. Boulic, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2017

...
I am the master of my fate,
I am the captain of my soul
— William Ernest Henley

To my beloved friends, family and country…

Acknowledgements

I am thankful to Dr. Pearl Pu, my thesis supervisor, for placing confidence in me for so many years. I thank the thesis jury; Dr. Jiyong Zhang, Prof. Li Chen, Dr. Ronan Boulic, and Prof. Daniel Gatica-Perez, for all of their insightful comments and suggestions. I thank Prof. Xiaojuan Ma for the time I have spent in the HCI initiative at HKUST.

In the course of my studies I have been blessed with the presence of countless friends from Turkey, Switzerland, Hong Kong, and all over the world. Thanks to them I persevered and thrived. Thank you Valentina, Gregoire, Marina, Ira, Ina, Andrew, Alina, Sergii, Igor, Aleksei, Florent, Goran, Fei, Julia, Georgeos, Artem, Renata, Radmila, Wenchang, Zhida, Ziming, Peng, Mingfei, Qing, Xuanwu, Anastasiia, Seva, Rysiek, Yasmin, Bart, Simon, Wolfgang, ... with you these years were full of inspiration and discovery.

Thank you Iraklis, Christos, Manolis, Stefanos, Eleni, Natassa, Pavlos, Nathalie, Matt, and my all other friends from the Greek community, for your ever-welcoming friendships. Thank you uncle Ioannis, aunt Ana, brothers Thanasis, Dimitris, and Alexandros.

Thanks to TURQUIA 1912, the association of Turkish students in Switzerland, for connecting me with brilliant, amazing people - Ertan, Fatih, Gun, Suat, Gokcen, Meric, Nariye, Can, Basak, Busra, Hilal, Gulperi, Kadir, Merve, Alp, and many, many more.

Thank you Ece, Okan, Yu, Adrian, Daniel, Sonia, Jean-Eudes and Andrea for all the inspirational conversations that dive into the world of philosophy, love, and happiness.

Thank you and farewell Burak - I miss you.

Thank you Love for the fun conversations. Thank you Ekin for the fantastic moments in our brief flatmateship. Thank you Nicolas for inspiring me with a joyful soul. Thank you Merve for all of our spontaneous dances. Thank you Agata and Konrad for our outdoor adventures and voyages. Thank you Ilke for running with me at great lengths and depths. Thank you Handan, Cem and Dilan for inspiring me with your compassion. Thank you Berker for being by my side in my cathartic moments. Special thanks to Yasar, Esat, Mehmet and Baran - our music, friendship, and laughters span the globe, and make it traversable within a heartbeat.

Thank you Nathalie! Hello, ocean heart...

Thank you Fusun and Birol, for inspiring me with your unconditional love and support.

And finally, dearest Ada.

This one is for you.

Lausanne, 24 August 2017 O. Y.

Abstract

Sedentary lifestyles and bad eating habits influence the onset of many serious health problems. Healthy behavior change is an arduous task, and requires a careful planning. In this thesis, we propose that behavior recommenders can help their users achieve healthy behavior change. Such a system should inspire its users with small, incremental and achievable goals. For this, it must resolve a trade-off between two opposing objectives: help the user achieve a steady improvement in target behavior, and avoid extreme goals that may injure or discourage the user. This is an unprecedented challenge in the recommender systems research.

If the system understands the impacts of past interventions for behavior change, it can determine its users' behavioral responses to its own recommendations. This implies a specific data curation, in which we not only measure people's behavior but also deliberately introduce an intervention to monitor its effect on people's patterns. In turn, the system can use these existing users' information to derive the right procedure for effective recommendations.

In this study we capitalize on this insight and develop InspiRE - our behavior recommender framework. Through InspiRE we propose the following contributions: 1) We design the data curation. 2) We develop the novel approaches for behavior profiling 3) We develop an evaluation process for this novel type of recommender system, and also compare it with traditional, similarity-based recommendation approach.

We curate a dataset that contains information of daily step counts and social intervention for 83 people. InspiRE successfully uses the observations from this dataset, and proposes recommendations that are both effective and feasible. We also show that InspiRE can generalize to other dimensions of well being: we demonstrate this through a dataset that contains the snacking patterns of 73 people, who receive message-based interventions. We observe that InspiRE's recommendation strategy is in line with theories of behavior change.

Keywords: Behavior recommenders, time series analysis, clustering, data curation

Résumé

Le mode de vie sédentaires et les mauvaises habitudes alimentaires peuvent influencer l'apparition de nombreux problèmes de santé importants.
Changer pour un comportement plus sain n'est pas une tâche aisée, mais les recommandations de comportement peuvent aider les utilisateurs à y parvenir. Un tel système a pour but de proposer à ses utilisateurs des objectifs modestes, progressifs et réalisables.
Pour cela, il doit réaliser un compromis entre deux objectifs opposés : aider l'utilisateur à atteindre une amélioration constante du comportement cible, et éviter de recommander des objectifs extrêmes qui pourraient être contreproductif ou décourager l'utilisateur. Ces systèmes représentent une branche encore méconnue des systèmes de recommandation.
Si le système peut évaluer l'impact des interventions passées sur le comportement de l'utilisateur, il peut déterminer les réponses comportementales de ses utilisateurs à ses propres recommandations. Cela implique un traitement de données spécifiques, dans lequel nous mesurons non seulement le comportement des utilisateurs mais aussi introduisons délibérément une intervention pour surveiller son effet sur leur profil. Ainsi, à son tour, le système peut utiliser les informations des utilisateurs existants pour établir la bonne procédure pour des recommandations efficaces.
Dans cette thèse, nous développons cette idée et proposons InspiRE - notre framework de recommandation de comportement. Grâce à InspiRE, nous proposons les contributions suivantes : 1) Nous établissons des lignes de conduites pour le traitement des données. 2) Nous développons de nouvelles approches pour la définition de profils de comportement 3) Nous développons un processus d'évaluation pour ce nouveau genre de système de recommandation, et aussi le comparons avec une approche basée sur les systèmes de recommandation traditionnelles.
Nous utilisons des informations sur le nombre d'enjambées effectuées quotidiennement et les interactions sociales de 83 personnes. InspiRE utilise avec succès les observations faites sur ces données et propose des recommandations à la fois efficaces et réalisables. Nous montrons également que InspiRE peut être plus généralement appliqué à d'autres dimensions du bien-être : nous démontrons cela à travers des données sur les habitudes alimentaires de 73 personnes pour lesquelles nous proposons une intervention basée sur un système de messages. Nous observons que InspiRE propose une stratégie de recommandation en adéquation avec les théories du changement de comportement.

Mots-clés : Recommandation de comportement, analyse des séries temporelles, clustering, traitement de données

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation and Challenges

Obesity, diabetes, and heart diseases are some of the most serious health problems of the modern societies. Many studies and reports identify daily habits as the primary factors of these health problems [OSH+10], and some even claim that being inactive is as dangerous as smoking.[1] In order to attain a healthier lifestyle, it is necessary for us to make changes in our daily habits and behaviors. This is, however, an arduous task. The difficulty of such a behavior change lies not in contemplating on the change itself, but in figuring out how to adopt the necessary behavior patterns over time. As many previous exercising-related studies show, people may lose motivation and relapse [CP14a], or injure themselves [CH02] because of exercising too much or arbitrarily. In other words, behavior change requires not only significant effort but also a careful planning.

Recent advances in wearable sensor technology and research on activities of daily living (ADLs) grant the possibility to track people's behaviors and develop data-driven systems to assist people in managing their personal well-being [Coo10b, DVD+02]. Successful realizations of such systems pave the way for many potential applications, e.g., early detection of health risks and sending alerts to caretakers and medical experts when necessary. Among these possibilities, we are most excited about developing *behavior recommender systems*. This novel type of recommender system can provide its users with recommendations, eventually helping them achieve a healthy behavior change.

Users of such systems can receive suggestions to pair up with an exercise partner, or to follow a day-by-day plan to increase their physical activeness. From a data analytical perspective, these suggestions are essentially temporal ADL patterns generated from other users (*interpersonal* recommendations) or from the past of the target user (*intra-personal* recommendations). Figure 1.1 conveys a conceptual case for such a recommendation: the system observes a user's data (purple line with no dashes), compares it to other patterns up to now, and decides the optimal

[1] See, for instance, http://www.bbc.co.uk/news/uk-wales-politics-18876880 and http://www.telegraph.co.uk/news/health/news/12044585/Obesity-has-become-a-national-threat-like-terrorism.html

recommendation.

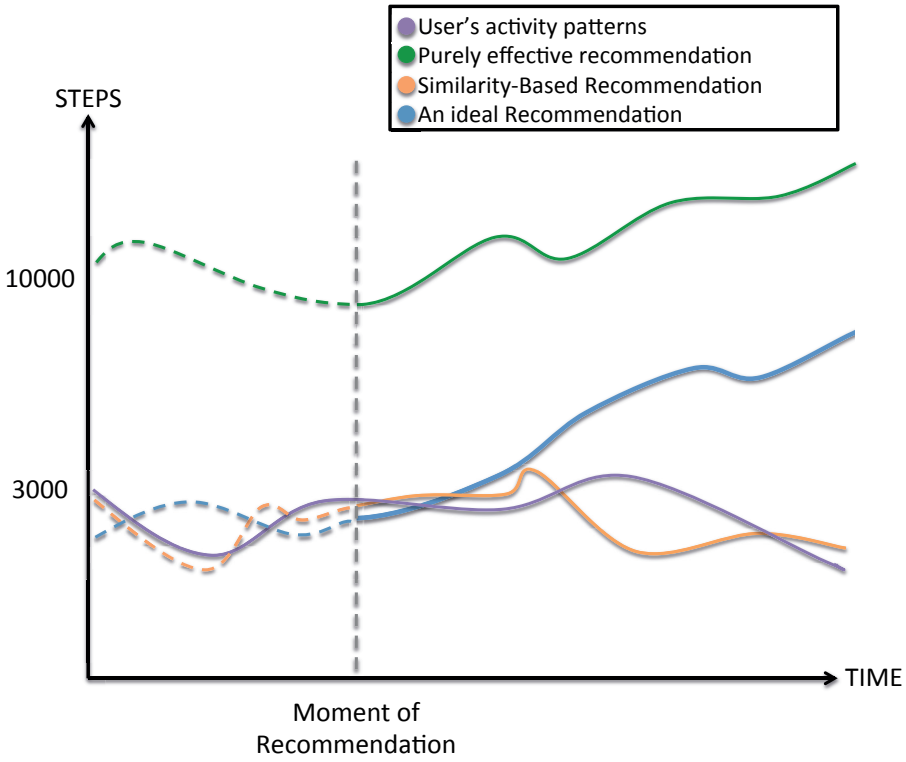

Figure 1.1 – The conceptual plot for recommendations. The recommender system analyzes the pre-recommendation patterns of existing users (dashed lines) to generate the recommendations.

How would the system compute such optimal suggestions? In order to further familiarize with this novel sort of recommendations, let us consider the first panel in Figure 1.2. Here we depict the experience of John, a typical user of a behavior recommender system called InspiRE:

> John is inactive and overweight. He wants to become more active so that he can avoid many health issues. He had received a wearable activity tracker as a gift from his wife, and found out that WHO suggests walking 10,000 steps per day for a healthy life.[2]. Unfortunately within three days, he injured himself by going from 500 steps per day to 6500 steps per day. He is naturally worried: he does not know how to become more active without getting injured.
>
> But then he discovers a system called InspiRE, which promises him safe behavior changes. He installs the system, which informs him to follow his current activity routines as naturally as possible. On the 7th day, he receives a pattern recommendation:

[2]The World Health Organization (WHO) states that 10,000 steps per day is a recommended level of activeness for adults aged 18-64 years: http://www.who.int/dietphysicalactivity/factsheet_recommendations/en/

> a plot which shows a series of daily step goals he should achieve from that point onward. The recommendation is sophisticated, as it gives details on which days to rest, maintain or increase his activeness. He takes this suggestion, which turns out to be achievable for him: Two weeks later, he finally manages to walk 10,000 steps per day. His success story will be an inspiration for others.

Parallel to this scenario, an optimal recommendation is the one that realizes the principles of well-established theories of behavior change. These theories suggest that John can best achieve the behavior change if he followed suggestions to pursue small, incremental and achievable goals [Ban86, NC02, PV97].

For such suggestions, it might be tempting to consider a traditional recommender system. These systems exploit similarities between users and recommend items based on their like-mindedness. They can simply track their users' activity patterns to employ this strategy. Had we adopted this approach, we'd use the behavior patterns of like-minded users to generate our recommendations to John. In that case, we would recommend the orange line (the half-dashed line with an average of 3000 steps) depicted in Figure 1.1. But, this suggestion obviously will not work for behavior change: a sedentary person like John can easily follow other sedentary people's activity patterns, but then he cannot achieve his goal of behavior change. The behavior recommender should deliver *effective* recommendations that help John achieve a steady improvement in physical activeness.

On the other hand, such recommendations should also be *feasible*, as otherwise they may set extreme goals that can injure or discourage its users. For example, in Figure 1.1 we convey John as a user who is not used to walking 10000 steps a day. Had the system suggested John the green line (with 10000+ steps) and he tried to follow it, he would have had to rapidly increase his activities by 7000 steps. Such a suggestion may lead to severe injuries.

From these two conceptual examples, we can understand that the system should make a critical trade-off between the effectiveness and feasibility of its recommendations to any given user. The behavior recommender system thus must model its users' innate capabilities for behavior change, but a mere collection of behavior patterns does not lend itself to this crucial information.

If the system understands the impacts of past interventions for behavior change, it can determine its users' behavioral responses to its own recommendations. The system can acquire this understanding if it observes successful and failed attempts for behavior change. The solution to our conflict is thus a specific data curation, in which we not only measure people's physical activeness but also deliberately introduce an intervention to monitor its effect on people's patterns. We can choose this intervention as pairing up people with an exercise partner, or sending them daily messages to motivate for exercise. Using this curated dataset, the system can exploit the behavior patterns that worked in the past, and avoid the ones that did not. To recommend activity patterns to a novel user, the system will first find people who used to follow similar activity patterns as this user. Then, among those people, it will choose those who responded to the intervention and improved their activity patterns. This would lead the system generate the ideal recommendation

depicted as the blue line in Figure 1.1, whose initial (dashed) values are similar to the user's, but nevertheless increases steadily.

Using this insight, the system will realize the flow of events in the second panel of Figure 1.2:

> Unbeknownst to John, InspiRE had already been curating data from other users and categorizing them based on their behavior patterns and their responses to its recommendations in the past.
>
> InspiRE considers Charlie and Mary, whose behavior patterns had been very similar to John's, and then both of them received recommendations. InspiRE identifies that Charlie relapsed to a less active lifestyle, but Mary managed to increase her activeness and switched to a more active behavior pattern without getting injured. Since Mary was similar to John in the beginning, the system decides that Mary's activity pattern will be an ideal candidate to guide John to safely increase his activeness.

The characters Charlie and Mary in this scenario actually represent groups of people who have distinct behavioral responses to recommendations or interventions. Charlie represents what we may call *non-responders* and *temporary responders*, whereas Mary represents *responders*. A useful behavior recommender analyses the differences between these profiles, particularly their temporal dynamics. As a result, the recommendation will be the aggregate of patterns which satisfies the trade-off between effectiveness and feasibility of the behavior change.

1.2 Research Agenda

In this thesis, we capitalize on these insights and respond with InspiRE, a novel behavior recommender system framework. Specifically, we design and build its analytical components in order to address the following key research challenges:

1. **Data Curation**. To generate the optimal recommendations, the system must use the examples of proven behavior change. There are generally no annotations in sensor-based data to indicate whether people are maintaining or improving their levels of activeness towards such suggestions. Many existing methods rely on such annotations [DBPV05, GER15, SJS05], and therefore are impractical in our case. Given these constraints, what information should the system collect besides pure observations of activities? And what should it use to measure the usefulness of the potential recommendations?

2. **Behavior Profiling**. Behavior recommenders must make sense of ADL routines of their users. The temporal characteristics of these routines are so diverse that merely comparing the average levels of activeness will fail to capture the distinctions between them [EP09, FDRK12]. Furthermore, given the trends in wearable technology, the system must also

leverage sensor-based data to extract useful behavior patterns. What is the suitable approach to obtain common behavior patterns from raw sensor data?

3. **Methods of Evaluation.** The recommendations must be both safe to perform and useful enough to improve the users' well-being. Some proven behavior changes will be too demanding for a new user. Furthermore, inactive users should never receive recommendations based on other like-minded, inactive people. What is the appropriate strategy to deliver these recommendations, and how can we evaluate these recommendations?

1.3 Main Contributions

In this thesis, we address these unmet research challenges with InspiRE - our behavior recommender framework. With InspiRE, we propose the following contributions:

- InspiRE's design is inspired from Social Cognitive Theory [Ban86], Trans-Theoretical Model [PV97], and Flow Concept [NC02]. We show that, with our data curation strategy, it is possible to generate suggestions for small, incremental, and achievable behavior changes.
- We have proposed a methodology to infer broad patterns from wearable sensor data via time series clustering [YZP14, YP16]. In this thesis, we extend this proof-of-concept method, validate its computational complexity, and obtain behavior profiles of users. These profiles capture the temporal dynamics of users' patterns (*Temporal Profiles*), as well as their behavioral responses to interventions (*Intervention Profiles*).
- Since this system is novel, it is yet another challenge to define the methods of evaluation. We propose and report three levels of validation that correspond to the methods in profiling and recommendation. We also test the system with varying granularity in data and additional contextual information such as user demographics.

This thesis covers the data curation, analytics and evaluation for InspiRE, while the sensor setup is out of scope. As a matter of fact, we designed the system to be able to work with activity data from any type of sensor, as long as it is in time series format. For the sake of clarity throughout the study, we demonstrate the capabilities of the system with an activity dataset of steps, and a nutrition dataset. Secondly, we leave persuasion strategies, the user evaluation and acceptance of recommendations as a future work.

1.4 Thesis Structure

We organize this thesis as follows:

- In Chapter 2, we review the definition of behaviors, theories of behavior change, persuasive

technology, recommender systems, and behavior profiling, followed by a summary of our contributions.

- In Chapter 3, we summarize the datasets used throughout the thesis, and a preliminary study that explores the relation between sensor data and well-being.
- In Chapter 4, we present FactorHabiTS, our framework to process wearable sensor data for profiling.
- In Chapter 5, we show how to curate the datasets so that InspiRE's methods can correctly function.
- In Chapter 6, we elaborate and demonstrate the analytical engine of InspiRE, including the behavior profiling, recommendation, and evaluation methods.
- In Chapter 7, we review the contributions of our thesis, as well as enumerating several directions for future studies.

Figure 1.2 – The scenario that illustrates how InspiRE helps John become more active.

2 Background

In this chapter, we review the background for behavior recommender systems. The background consists of a wide range of topics, from the concept of behavior and behavior change theories, to recommenders and activity analysis systems.

2.1 What is a Behavior?

In the introduction of this thesis, we revealed that a behavior recommender system must have its foundations on theories about behavior and behavior change. We refer to two of the most comprehensive definition of a behavior:

> Behavior is the the response of the system or organism to various stimuli or inputs, whether internal or external, conscious or subconscious, overt or covert, and voluntary or involuntary. [MK14]

> A behavior is a relation that consists of behavior actor, operation, interactions, and their properties. [Cao10]

Performing a behaviour in a repetitive, consistent manner leads to forming habits [And03, LJPW10]. Habits help us automate our behaviors in order to free the mental resources for other tasks. On the other hand, habit formation makes it more difficult to perform behavior change (see Section 2.3 for the extensive review of behavior change theories).

In this thesis we are interested in behaviors that can be quantified with sensors and have influence on well-being.

2.2 Quantifiable Information for Behavior Recommenders

A behavior recommender system can only deliver recommendations based on the types of behaviors it can track in a continuous manner. The tracking of health-related information can be roughly divided into two categories: External and internal body measures [Sma12b]. Both of these categories have their corresponding sensors.

The external measures can be grouped under the categories of *Nutrition*, *Exercise*, *Sleep* and *Stress* (i.e., *NESS*). The market is now being populated by many affordable products (sensors and software) that support the monitoring of NESS such as Fitbit, BodyMedia, EM Wave, etc [Sma12b]. These sensors typically collect temporal information on heart rate, galvanic skin response, number of steps, calorie expenditure, the distribution of active and sedentary periods, and so forth. The level of details provided by these sensors also have a rich variety, allowing both advanced users and perpetual intermediates to benefit. These kind of sensors are also generally non-invasive, and hence can be utilized throughout the day.

While it is known that the external sources of measurement have relations with the internal metabolism (such as stress having strong effects on biochemicals in blood and nervous systems [Sma12b]), detecting the underlying causes of some health problems (such as diabetes and cancer) may require a direct access to internal body measurements. The data from internal measurements can provide up to hundreds of variables about cholesterol, sugar and acid level, and hormonal measures. These variables are used to interpret the status in cell system, sugar system, hormone system, liver and kidneys, cardiovascular system, and inflammation [Sma12b]. For even further investigation, one can survey the information related with human genomics, and even a complete systems biology profiling. The measurements of these kind of information can be achieved via biomarkers in blood, saliva, and stool [Sma12b].

There are two main issues in internal body tracking and genomics. First, the utilization of related sensors requires a great effort from the user, and some users can be sensitive about the kind of information collected (at least, more than the external body measures). As such, it might be difficult to have continuous measurements of the data. Secondly, the collected information can require an expert in the field to interpret.

In summary, the survey on self-quantification suggests that a user-friendly behavior recommender system should be able to provide content that is mainly related with external body measures. Nevertheless, further studies may yield more accessible representations of internal body measures.

2.3 Theories for Behavior Change

The first and foremost task in designing a behavior recommender system is to identify the optimal strategies for healthy behavior change. Fortunately, studies on behavior change prove to be abundant sources of inspiration. These studies investigate many aspects of human behavior,

including the stages of habit formation [PV97], action possibilities [Gre94, SMCUU07], factors associated with successful change [KWM+97] and the conditions that maximize a person's engagement in specific tasks [WTR94]. Among them, two well-established theories are the most relevant to our study: Trans-Theoretical Model (TTM) [PV97] and Social Cognitive Theory (SCT) [Ban86].

Trans-Theoretical Model (TTM) [DPF+91, PV97] models the behavior change as a progression through six distinct stages. These stages are called *Precontemplation*, *Contemplation*, *Preparation*, *Action*, *Maintenance* and *Termination*. A person's involvement gradually increases at each successive stage until the *Termination* stage, where the person is certain that he will not relapse to the old habit. Each stage requires different strategies to support the behavior change. For instance, a person in *Preparation* stage may benefit from advices on how to start the first steps of action. An earlier user study called Fish'n Steps [LML+06] uses this theory to measure the success of their system.

Social Cognitive Theory (SCT) [Ban86] relates a person's learning and actions with social interactions and other environmental influences. In the context of behavior change, SCT states that a person can remember and apply the sequences of event that lead to the behavior change of other people. Consequently, each person is both an agent and a responder to change. Based on SCT, successful strategies for behavior change should involve people interacting within a social environment. Despite its useful guidelines, to our best knowledge, this theory has not been explicitly applied yet in computational studies.

From TTM and SCT, we derive the following guidelines towards designing recommenders for behavior change:

- Changes occur in distinct stages such as contemplation, planning, action, and maintenance. Every stage may require different strategies for intervention.
- Each stage of behavior change has different requirements, so a behavior recommender should avoid one-size-fits-all approaches in generating recommendations. Instead, system should categorize people with the available data, and aim to generate personalized recommendations.
- To maximize the sustainability of change, the system should make sure to set small, incremental, and achievable goals for its users.
- People are subject to *reciprocal determinism* in behavior change: they can be both agents and responders for change. Thus, the system should help its users receive positive influences from each other's achievements in behavior change.

Behavior recommenders could very well use these ideas to model the innate capability of a user to carry out and sustain behavior changes with physical activities. To our best knowledge, this has never been implemented before this study.

2.4 Engaging Users for Behavior Change

2.4.1 The Concept of Flow

Flow [JTMS01, MJ99, NC02] is a concept from positive psychology. With an insight that describes a good life as "*...one that is characterized by complete absorption in what one does*" [NC02], the concept of Flow investigates the relationship between the difficulty of tasks and the perceived capabilities of a person. The concept of flow is well-studied in many areas. Sample studies include the investigation on the dimensionality and correlates of flow in human-computer interactions [WTR94] and the assesment of flow in physical activities [JE02]. The flow theory states that a user's engagement is maximized when the difficulty of performing a task matches with a person's capabilities. In the realm of recommendations, this can be translated to the balance between a person's capability to perform behavior changes and the difficulty to perform the recommended activities (see figure 2.1).

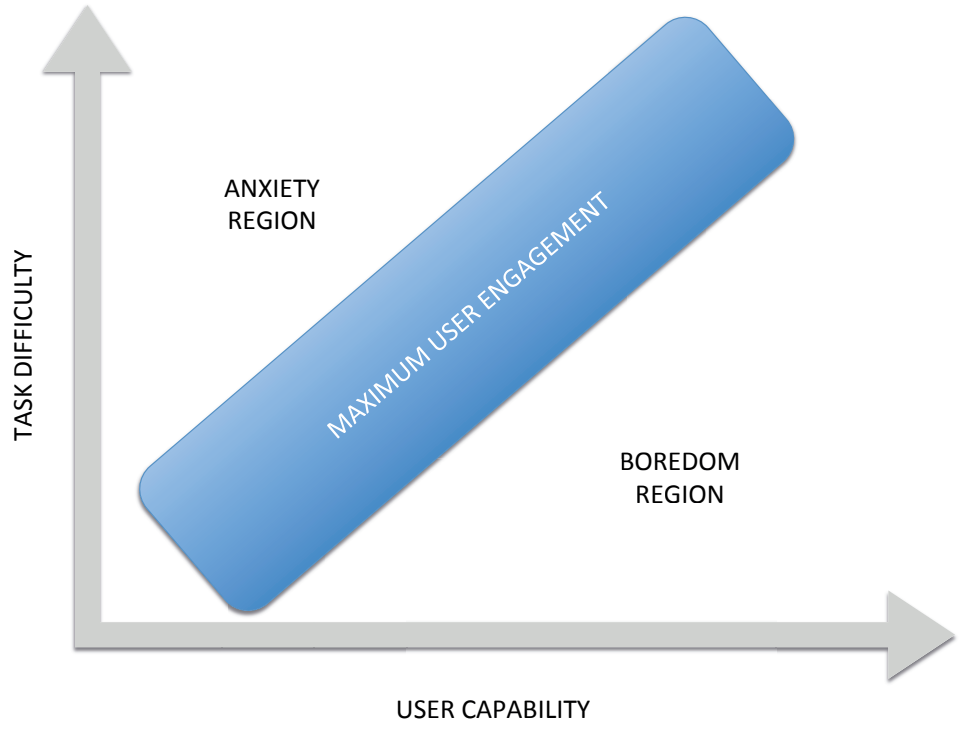

Figure 2.1 – Application of the Flow theory to behavior change tasks

In terms of action opportunities and capabilities, the concept of flow may share some parallels with the concept of affordances. The notion of affordances deals with modeling interfaces or data representations based on perceived action opportunities mediated by action capabilities, and it is well studied in various literature such as HCI and robotics [Gre94, KN12, SMCUU07]. On the

other hand, the flow concept deals more with the effects of the balance (or lack thereof) between opportunities and capabilities on the level of engagement, absorbtion, satisfaction, etc. of the users in performing a certain action, task, etc.

2.4.2 Persuasion and Persuasive Systems

Alternative or complementary to recommendations, some studies promote physical activities and healthy habits through persuasive technologies. Such approaches draw inspirations from the six principles of persuasion stated by Robert Cialdini [Cia01]: *reciprocity*, *commitment and consistency*, *social proof*, *authority*, *liking*, and *scarcity*. A particular example is the Fish'n Steps application [LML+06], which links a user's daily step count to the growth and activity of an animated virtual character (a fish in a tank).

1. Reciprocity - People tend to accept the requests of people who first do them a favor.
2. Commitment and Consistency – If people commit, orally or in writing, to an idea or goal, they are more likely to honor that commitment because of establishing that idea or goal as being congruent with their self-image. Persuasive systems may implement this principle by making their users pledge for certain goals (e.g., "I will walk 10,000 steps today!")
3. Social Proof - People will do things that they see other people are doing. In the case of physical activeness, this can be implemented with messages and visualizations which convey other people steadily increasing their level of activeness up to 10,000 steps.
4. Authority - People will tend to follow suggestions from authority figures. In the case of physical activeness, this can be implemented with messages like: "Experts in World Health Organization recommends to walk 10,000 steps per day for a healthy life"
5. Liking - People are easily persuaded by other people that they like. A recommendation from a loved one or an attractive person has a higher likelihood to persuade the target user.
6. Scarcity - Perceived scarcity will encourage people to take action. For example, offering gifts for a limited time in exchange of walking 10,000 steps.

The impact of the feedback ultimately depends on using the appropriate strategy of persuasion. Below are some strategies employed by prior studies:

1. Power of Praise [Fog02]: Praise with different framing makes it easier to persuade people. For instance, the message "You have done great! Walk 400 steps more to reach your goal" works better than "You lack 400 steps towards your goal".
2. Negativity Bias [KH87]: People pay more attention and give more weight to negative than positive information. So when delivering feedback, we should be careful and do not

over-emphasize the negative aspects of a person's current situation. Persuasive systems should put more concentration on positive sides.

3. Humor Effect [Sch02]: Humorous items are more easily remembered than non-humorous ones. In order to draw the users' attention to the relevant information, studies advise to add some humor, e.g., "Wow, with your lifetime number of steps, you can walk around the Niles river!"

4. Rhyme as Reason, or Eaton-Rosen phenomenon[MT00]: A statement is judged as more truthful when it is rewritten to rhyme. We see that, for instance, the public transportation system in Lausanne, Switzerland uses this technique very frequently: "Vélo à bord? Je lui prends son titre de transport" (meaning: "Bringing bicycle on board? I buy it's ticket for transport")

5. "Maybe Later" [Fog02] "Maybe Later" works better than "no". It provides psychological hints. "You should not eat that dessert" vs. "Maybe you should take that dessert later"

6. Using words to simplify complex relations [TBV12a]: In one of HCI studies, the researchers presented to the participants the correlations between their habits and their weights. This information made the participants more engaged in losing weight.

7. Identify the susceptibility towards persuasion [KLS10, KRMA12]: In one user study, researchers measured the susceptibility of the participants to different types of persuasion (reciprocity, scarcity, authority, commitment, consensus, and liking) to reduce their snacking. They have found out that when message wordings are tailored based on participants' most susceptible persuasion strategy, they are more likely to reduce snacking than using random persuasion strategies.

2.5 Recommender Systems

Recommender systems assist the difficult task of decision making in our everyday tasks [RV97]. They are ubiquitous technologies in product or service recommendations. In these systems, the profiling stage is essential to generate personalized product recommendations. Systems can infer profiles from user ratings, online behaviors, and product information. We can broadly categorize profiling methods as user-based and item-based. User-profiling approaches analyse user's past decisions as well as the behaviors and decisions from similar users [ABG+97]. Item-profiling instead analyses the properties of previously rated/bought items to recommend similar items for a given user. [LSY03]. One can also employ utility functions to represent users' preferences with varying levels of priority. Such preferences may originate from both items' and users' characteristics [AT05].

Besides user and item profiling, recommender systems can also be categorized based on the alternatives to obtain such profiles [AT05]. In cases where data sparsity is an important limitation, many studies prefer to treat this task as a matrix decomposition problem, i.e., collaborative

filtering. The variants of this approach, such as Singular Value Decomposition [GR70], aim to obtain item or user profiles as lower rank approximations of rating patterns. Extensions of such methods can accommodate temporal relations [DL05] and contextual information [VMO+12].

Many others use probabilistic modeling to capture complex relations between users and items. For instance, a content-based recommender system treats recommendation as a sequential decision problem, and solves it with a Markov Decision Process design [SBH02]. Probabilistic modeling is also popular in hybrid approaches where both collaborative filtering and content filtering are involved [Bur02]. One notable example is the three-way aspect model [PPL01], which analyzes co-occurrences among users, items, and item content.

Despite the differences of methods (i.e., matrix-based vs. probabilistic), the key characteristic in traditional recommenders is nevertheless the same: mine the past data to elicit preferences, and use these preferences to maximize the likelihood that a user prefers thus buys certain products from the system. This typically leads to the designs that generate better recommendations based on similarities, i.e. either from like-minded people and/or items that have similar characteristics to the ones the given user has already bought or rated. However, the objective of a behavior recommender is different: to maximize the likelihood that a given person becomes more active. The recommendations should be both safe to perform and useful enough to increase the users' activeness. A similarity-based recommendation strategy will be particularly ineffective for inactive users, since such recommendations would be generated from like-minded, inactive people.

Another topic of interest is the evaluation of recommender systems. One alternative is to use indices like F-1 score (see Appendix A.4) to compare users' actual item ratings and the rating predictions of the system. The other alternative is to measure or estimate the level of acceptance of the recommendations. Technology Acceptance Model (TAM) [Dav86] tells us the relation between perceived usefulness, usability, and the adoption rate of a certain technology. This model has been extensively studied and validated on recommender systems. For instance, Hu and Pu proposed a personality-based recommender system, and then investigated the acceptance issues of such a system [HP09, HP10]. TAM model has a significant implication on the evaluation of recommender systems: Increased levels of perceived usefulness and usability not only improves the adoption of a behavior recommender system, but also increases the chances that its users will consider applying its recommendations to their daily routines.

2.6 State of the Art for Behavior Profiling and Recommendation

2.6.1 Behavior Profiling

The analytical building blocks of a behavior recommender system has strong parallels with many other research topics in activity analysis. The most prominent of these topics can be categorized as *Activity Recognition*. The objectives of activity recognition are to model human behavior and

make predictions for the future activities of sensor users. Such systems have a wide range of applications, including smart homes, emergency detection systems, and location-oriented systems (e.g. traveling recommendations).

In activity recognition and analysis (AR/A) studies, we find behavior profiling as an analogy to the user and item profiling of traditional recommenders. While recommenders use profiling to represent common rating patterns and to deliver optimal recommendations, AR/A systems use profiling to represent common behavior patterns and to improve the intelligence of the environments in which people live and work [CK14]. However, similar to the traditional recommenders, we observe that AR/A systems typically treat profiling as either a probabilistic modeling or matrix factorization task.

Some typical probabilistic approaches for behavior profiling use Bayes classifiers, Markov Logic Networks [GER15], conditional Random Field Models and Hidden Markov Models [CPV14, Coo10b, NH12]. Topic modeling, a technique adapted from document-word analysis for mining semantic data, is another alternative for probabilistic modeling of behaviors [FGP14, CdIAK14]. Some studies even rely on more complex alternatives such as Hidden Semi-Markov Model [DBPV05, KEK10] to extract behavior profiles with variable time granularities.

Alternative to probabilistic modeling approaches, some studies adapt collaborative filtering methods to obtain user profiles in a model-free fashion. This straightforward adaptation showed some promise in location and proximity data such from mobile phone [ZLN13]. One notable example is the "Eigenbehaviors" study that employed Principal Component Analysis [EP09]. This strategy decomposes the data into eigen-vectors, and allows each user to be modeled as a weighted combination of such eigen-vectors. Computer vision studies [WY13] inspired activity analysis literature to apply robust matrix factorization methods [YZP14] and handle the noise in exercise datasets from wearable sensors accelerometers.

The design of AR/A methods depend greatly on the data collection procedure. Initially using manually logged information [SJS05], AR/A methods also exploited the advances in sensor technology to deliver systems that use environmental sensors [CdIAK14, Coo10b, DBPV05, RCHSE11a], and mobile sensor data such as location and bluetooth [ZLN13, FGP14]. A further challenge is the annotation of the collected data. Earlier approaches were typically tested on presence, location and duration information collected from indoor sensors [WA05]. Thus, it was easy to designate a set of probabilistic states, each of which would correspond to a distinct activity pattern. However, adapting the same approach to physical exercise or outdoor location patterns proved to be more difficult - some studies use clustering as a pre-processing step [SJS05], or employ voting to find the best-fit model [RCHSE11a].

One potential way to handle the diversity of sensor data is to employ a taxonomy of ADLs. One certain taxonomy divides activities into three broad categories [CCAY13, NHdlC15]: *Basic ADLs (BADL)* such as bathing, brushing teeth, dressing, using toilet, eating and drinking, sleeping; *Instrumental ADLs (IADL)* such as preparing meals, preparing drinks, resting, housekeeping,

using a telephone, taking medicine; and *ambulatory activities* such as walking, doing exercise, transitional activities, and stationary activities. Some studies determine data labels using a sample of activities from this taxonomy and other sources such as Barthel's Index [MB65]. Such methods then use setups with multiple sensors [FVN10] to recognize these activities. These approaches obtained high classification rates with a limited number of BADLs and IADLs. The taxonomy gives a coarse overview of the activities, but much more detailed work needs to be done to infer fine-grained information (such as calorie expenditure etc.), especially in the context of physical exercises. Lastly, taxonomy-based systems depend on the quality of annotation in the training datasets. Considering the vast amount of data required to train the state-of-the-art methods, this requires significant manual effort.

Some very recent studies exploit the potential of location and movement patterns for applications such as movie recommendations [CPCP13]. However, given its analogy with user-item profiling and the variety of existing methods, it is surprising that behavior profiling has never been considered in recommending behavioral changes. This might be because many of the methods, especially optimizations in probabilistic modeling approaches [Moo96] require expert knowledge to work in some well-controlled settings. For more realistic settings, there is a need of a method to identify frequently occurring behavior patterns without expert knowledge and ground truth. Another reason is that the research in behavior recommenders is in its early stages. We now proceed to discuss behavior recommenders in detail.

2.6.2 Behavior Recommendations

There is an ever-increasing demand for adaptive, preventive interventions in the medical domain [CMB04]. To our best knowledge, there are only few studies who responded to this demand and proposed designs for a behavior recommender [FDRK12, SJC15].

For instance, HealthAware system [SJC15] relies on manually specified rules, which were derived from some general guidelines developed by Centers for Disease Control and Prevention [1]. This approach offers no means for personalization or adaptability to changing conditions - such guidelines would only help the system identify what to improve.

The Intrapersonal Retrospective Recommender [FDRK12] follows a more data driven approach: it identifies stable patterns in users' personal histories, and selects the patterns that have the largest impact on users' goals as recommendations. Since this approach abandons information from other users, it resembles to content-filtering methods in traditional recommender systems. A significant drawback is that the recommended activities would never be novel. Thus the recommendations will never be helpful to those who never succeeded to improve his well-being in the past. This study mentions the high diversity of habits as a reason to discard the potentially useful patterns from other people. Research on AR/A has viable solutions to this diversity problem by identifying structures in behavior routines [EP09, FGP14, YZP14], but none of them were tested in behavior

[1] http://www.cdc.gov/

recommender systems prior to this study. Lastly, the existing studies evaluate their systems only with the perceived usefulness reported by the users [FDRK12, LML+06, SJC15]. It is also necessary to know if the recommendations have any significant influence on the daily routines of a person in a positive way.

Finally, it is noteworthy to consider the concept of Just-In-Time Adaptive Interventions (JITAI). JITAIs intend to intervene the patients and provide support with an accurate timing by detecting health states with potentially elevated vulnerability [NSST+14, PRA+09, ROJM08]. JITAI designs aim to provide many types of supports: mental illness mangement, smoking cessation, weight management, and so forth [BZKB+13]. These interventions can be specified by experts or initiated by the patients, and delivered through mobile applications and text messages [FKR+11]. Despite these enthusiastic design attempts, the research on JITAIs are still on their early stages, and thus need further evidences and theoretical grounding [DMCY13]. Many such designs still rely on manual inputs from users and fixed rules specified by medical experts. Thus there is a need to integrate sophisticated methods to JITAIs in order to generate automated interventions.

2.7 Summary

Building a behavior recommender requires a thorough analysis of existing studies from a variety of research fields. In this survey, we investigated theories of behavior change, profiling methods in traditional recommender systems, behavior profiling methods and some pioneer studies towards behavior and lifestyle recommendations.

First, by analyzing psychology studies, we found two major theories of behavior change to ground the inspirations for our analytical methods: TTM, SCT, and Flow Concept. These theories recognize the difficulty of changing habits, and help us accurately frame the task of behavior recommendation as delivering personalized, small but incremental, and socially influenced suggestions. Second, we have analysed the state-of-the-art advances in traditional recommender systems, particularly, the user and item profiling methods. We note that traditional systems for item recommendations typically employ strategies that optimize their recommendations based on like-minded people. But a behavior recommender should not recommend patterns of inactive people to inactive people.

Next, we surveyed behavior profiling studies, which employ methods with high levels of sophistication, but requires expert knowledge and annotations to predict future activity sequences of people. We also note it is rather a novel challenge to process sensor-based time series data in the context of behavior recommender systems: One of the early systems resorts to fixed rules to determine when and what to recommend [SJC15]. Another study implements the intrapersonal-retrospective strategy [FDRK12]: for a given person, only his own history of stable patterns are considered to generate recommendations. Without an access to sophisticated AR/A methods, these systems either compromise personalizability or recommendation novelty. Both alternatives hinder the usefulness of behavior recommendations. Lastly, there are studies that design JITAIs

[NSST+14], who aim to detect vulnerable health states of people in order to intervene at the right time and with the right measures. A careful analysis on these studies reveals that there is much to be done for a viable JITAI framework, especially towards their data analytical components. For all such efforts towards behavior recommenders, there is a further need to justify the designs with existing theories of behaviour change.

The research on behavior recommenders has the following impacts:

- Supporting Persuasive Systems and User Studies: Behavior recommendations can enhance the acceptance of the persuasive systems applications we reviewed in Section 2.4.2. Specifically, users' trust and perceived usefulness of the system will greatly improve when the system can generate accurate and useful recommendations. Based on the technology acceptance model we reviewed in Section 2.5, this eventually leads to an increased adoption of the technology.
- Predictors of Successful Behavior Change: Long term deployment of behavior recommenders will generate large amount of data with rich information about users' interactions with recommendations. This in turn can help researchers find new factors that influence the success of behavior change and overall well-being of wearable sensor users.
- Supporting Emergency Detection Systems with Proactive Measures: Recent years witnessed a surge of systems that detect health-related emergency conditions (such as heart attacks and falls [SZD+15]) and subsequently send alerts to caretakers. Such systems can easily be complemented with behavior recommenders, which help users adopt preventive measures against many potential diseases.

3 Preliminaries and Common Material

3.1 Used Datasets for Experiments

Throughout the thesis, we have used various datasets to demonstrate our contributions. In this section, we briefly mention each one of them, and refer the user to the relevant chapters. Notice that they are curated in very different manners. We postpone the justification of these differences in further chapters of the thesis.

- The Reality Mining Dataset [EPL09] is collected between 2004 and 2005 from a longitudinal study in MIT, Boston. It includes 94 users (students, professors and staff). Each participant was given a mobile phone with several pre-installed pieces of software to record various information, including call logs, Bluetooth devices in proximity, cell tower IDs, application usage, and phone status (such as charging and idle). The Reality Mining dataset has been used in many studies. For example, Zheng et al. [ZN12a] use probabilistic reasoning techniques to discover behavior patterns and group users. The dataset does not include annotations of user daily activity. We use this dataset in our prelimiary study in Section 3.2, where we show the link between sensor-based activities and actual well-being reported by sensor users.

- YELP Academic Dataset [1]: YELP dataset contains approximately 2.3 million ratings from 70,000 users for 15,000 businesses. In this study, we analyse the 2014 version of this ever-growing dataset, which includes the ratings within the time period from 01-02-2005 to 28-01-2014. Similar to other rating datasets, the YELP dataset is very sparse: there are 402 out of 3284 days with no records of ratings or reviews. On average, each user has 33.69 ratings (minimum 1, maximum 3286). We use this dataset to validate the scalability of our sensor data processing method as explained in Chapter 4

- YQZ Dataset: This dataset contains the daily steps counts of 1000 participants of a social exercising campaign, who wear different sensors to measure their level of activeness. The

[1] https://www.yelp.com/dataset_challenge

dataset also contains the gender, height, weight, exercise group id, company id, and age (all anonymized). We use this dataset to validate our sensor data processing method as explained in Chapter 4

- HealthyTogether Datasets: This dataset was curated to discover common physical activity routines of people and analyse the effects of social interventions on the behavior patterns. More specifically, it was curated via a longitudinal user study that involved a wearable sensor (Fitbit) and our custom mobile application called HealthyTogether. The dataset contains calorie expenditure and steps each participant to the longitudinal study.

  In this thesis, we used two versions of this dataset:

  - HT-48, the original one, curated to validate our sensor data processing method as explained in Chapter 4;
  - HT-83, where we expanded the dataset to 83 users to validate our recommender system as explained in Chapter 6

- SNACK dataset: This dataset contains daily, self-reported number of snacks of a set of people. During the data collection period, these people received messages to motivate them to cut their unhealthy snacks. We use this dataset in Chapter 6 to show that we can generalize our recommendations beyond physical activities to other dimensions in the four pillars of well-being, i.e., NESS (See Section 2.2 for a detailed review of these pillars.)

3.2 Preliminary Study: Defining the Relation Between Sensor-based Activity Data and Well-Being

Pervasive healthcare systems provide automated wellness monitoring [KPG03] and activity suggestions to improve the well-being of the user. The user is equipped with various sensors, which collect information on users' metabolism, activity, location, and so on. The ever-increasing number of diseases and deaths due to inactivity [2] strongly indicate that such systems should become an indispensable component of our lives. There are already various studies which investigate the goals that should and can be achieved through such systems, such as maintenance of physical health [DAC+09, TBV12a, TLR+07, ZPSB04], and providing the means for self-monitoring [LDF11, MKK+12, TBV12a, TLR+07].

We explored two issues in our preliminary studies. First is to find the activity patterns of users using the information collected from mobile devices. Second is to investigate how daily activities are correlated to people's satisfaction from life, measured through survey data. We adopt the well-known Reality Mining dataset [EPL09] in our study, through which we extract daily activities and apply Structural Equation Modeling (SEM) to find their relations with reported levels of satisfaction.

[2] see http://www.bbc.co.uk/news/uk-wales-politics-18876880

The Reality Mining Dataset [EPL09] is collected between 2004 and 2005 from a longitudinal study in MIT, Boston. It includes 94 users (students, professors and staff). Each participant was given a mobile phone with several pre-installed pieces of software to record various information, including call logs, Bluetooth devices in proximity, cell tower IDs, application usage, and phone status (such as charging and idle). The Reality Mining dataset has been used in many studies. For example, Zheng et al. [ZN12a] use probabilistic reasoning techniques to discover behavior patterns and group users.

	High-Social ($P(u,t) > \bar{P}_u$)	Low-Social ($P(u,t) \leq \bar{P}_u$)
Phone On, Location = *Home*	HomeSocial	HomeRest / Sleep*
Phone On, Location = *Work*	WorkSocial	Working
Phone On, Location = *Elsewhere*	LeisureSocial	LeisureRest
Phone Off from Mid-night to 8AM	Sleep	
Phone Off from 8AM to Mid-night	PrivateActivity	

Table 3.1 – The rules to estimate the activities for a given user u. (*): Being at home from 8PM to 8AM without any SMS or voice call action is labeled as *Sleep*)

The dataset do not include annotations of user daily activity. We use *communication* information (the number of SMS and voice calls every hour), *proximity information* (the number of devices discovered in bluetooth scans every 5 minutes) and *location* information (hourly recorded as *home, work, elsewhere*) in the dataset to *estimate* the activities. For this, we consider *intrapersonal* data, i.e., we compare the proximity information of a person in a given hour with his own average hourly proximity information. For a given time t, $P(u,t)$ denotes the number of bluetooth devices in the proximity of the user u, and $\bar{P}_u$ denotes the average number of bluetooth devices discovered hourly by the same person. If $P(u,t) > \bar{P}_u$, we label the user in *high-social* mode, otherwise he is labeled in *low-social* mode. We summarize these activities (8 in total) in Table 3.1. According to our estimation method, an average student spends 7.4 hours on sleep, 6.8 hours on work, 1 hours on break in work (*WorkSocial*), 7.3 hours on leisure outside (*LeisureRest* and *LeisureSocial*), and 1.5 hours on other activities (HomeSocial, HomeRest, PrivateActivity). These numbers are similar to the findings in 2012 Sodexo University Lifestyle Survey report [3].

We also include four *communication-related* features: the number of SMS, number of phone calls, *proximity* (proximity count and proximity time) information. We also notice that users' activities are very different between weekdays and weekends, so we calculate them separately. Thus we obtain 24 features for activities on weekdays and weekends for each user as shown in

[3] http://uk.sodexo.com/uken/media-centre/press-releases/university-lifestyle.asp

Table 3.2.

Id	Activity Features
1-2	HomeSocial {weekday, weekend}
3-4	HomeRest {weekday, weekend}
5-6	Sleep {weekday, weekend}
7-8	WorkSocial {weekday, weekend}
9-10	Working {weekday, weekend}
11-12	LeisureSocial {weekday, weekend}
13-14	LeisureRest {weekday, weekend}
15-16	PrivateActivity {weekday, weekend}
17-18	BluetoothDeviceCount {weekday, weekend}
19-20	BluetoothDeviceTime {weekday, weekend}
21-22	VoiceCallCount {weekday, weekend}
23-24	SMSCount {weekday, weekend}

Table 3.2 – Activity Features.

Lastly, we represent the regularity of user activities through their *entropies*. The entropy of a feature x can be calculated as:

$$\mathscr{H}(x) = -\sum_{t\in[1,24]}\sum_{c\in\mathscr{C}_x} p(c|t)log(p(c|t)), \tag{3.1}$$

where t denotes the time in hours. $\mathscr{C}_x$ is a generic notation for the set of available items of the given feature x. For instance, for activity entropy $\mathscr{H}(x = activity)$, $\mathscr{C}_x$ is the set of all possible activities. Then, $p(c|t)$ would denote the probability of having activity c at time t. A person with high activity entropy would have irregular amounts and distribution of activities while he/she participated to the longitudinal study. We compute entropies for *activity*, *social time*, *location* and *proximity* for each user for weekdays and weekends respectively, providing us with *8* regularity features as shown in Table 3.3.

Id	Regularity Features
25-26	ActivityEntropy{weekday, weekend}
27-28	LocationEntropy{weekday, weekend}
29-30	ProximityEntropy{weekday, weekend}
31-32	SocialEntropy{weekday, weekend}

Table 3.3 – Regularity Features.

There are *25* survey questions in the dataset, *10* of which are *self-reported* measures of happiness, health and travel frequency of the users (See Table 3.4). In this work we use such self-reported satisfaction information to represent users' happiness. These fields have different ranges: question with $Id = 42$ ranges between 1-5; question with $Id = 41$ ranges between 1-4; and the rest range between 1-7. To obtain a unified interpretation, we scaled their values to a common range, i.e., 1-5. We have performed a linear scaling, which retains the information of the original values.

Id	Survey Questions
33	I am satisfied with my experience at MIT thus far
34	I am satisfied with my current social circle
35	I feel I have learned a lot this semester
36	I am satisfied with the content and direction of my classes and research this semester
37	I am satisfied with the support I received from my circle of friends
38	I am satisfied with the level of support I have received from the other members in my group
39	I am satisfied with the quality of our group meetings
40	I am satisfied with how my research group interacts on a personal level
41	Have you been sick recently?
42	Have you travelled recently?

Table 3.4 – Satisfaction/wellness questions

We apply Principal Component Analysis (PCA [Pea01]) on the features to group them into *factors* (expressed in terms of eigenvalues and eigenvectors). We choose the eigenvalue threshold as *1* to determine the number of factors, and the factor loading threshold as *0.55* in order to include the features for further analysis. We use the IBM SPSS software and apply the varimax rotation (an orthogonal rotation) in the factor analysis. We identify 6 *factors* from the activity features, and 2 *factors* from the survey data. We name the factors with respect to the feature with the highest positive loading, as conveyed in Figures 3.1 and 3.2. More precisely, two factors are satisfaction-related (*Social Life Satisfaction* and *Research/Study Satisfaction*), and another three are regularity-related (*Social Entropy, Location Entropy* and *Proximity Entropy*) and the remaining three are related with the activity patterns (*Leisure and Sleep*, *Working Activities*, and *Communication Activities*). We discard some features (*WeekdayHomeRest*, *WeekendHomeRest*, *WeekendSocialEntropy*, *WeekdayLeisureSocial*, *WeekendLeisureSocial*, *WeekdayHomeSocial*, *Health*) since their loadings are below *0.55*.

To understand how these activity features affect self-reported satisfaction, we use Structural Equation Modeling [Pea00], which can be used both to explore and confirm hypotheses of causal assumptions between groups of features, and model noises in the data with latent (unobserved) variables. To our knowledge, there is only one study that uses SEM for daily activity analysis - specifically, for predicting sequence of activities based on commute data [KP12]. The dataset of that study includes solely self-reported activities and their durations. In contrast, the Reality Mining dataset was collected using modern sensor technology.

We have followed commonly accepted thresholds for factor analysis[4]: as shown in Figures 3.1

[4] see http://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/thresholds for a summary of thresholds

		Component					
		1	2	3	4	5	6
Proximity Entropy	**weekdayProximityEntropy**	**.920**	-.034	-.031	-.094	.145	-.085
	weekendsBluetoothTime	**.858**	-.270	.110	-.010	-.091	.013
	weekendProximityEntropy	**.858**	.048	-.077	-.089	-.025	-.201
	weekdaysBluetoothTime	**.805**	-.166	.001	.060	.093	.100
	weekendsBluetoothCount	**-.750**	-.071	.020	-.056	.072	.117
	weekdaysBluetoothCount	**-.750**	-.188	-.001	-.130	-.080	.140
Communication Activities	**weekendsVoiceCount**	-.006	**-.853**	.137	.052	.010	.127
	weekdaysSMSCount	-.015	**.841**	-.093	.172	.064	-.048
	weekdaysVoiceCount	.050	**-.819**	.156	.110	.114	.018
	weekendsSMS	.013	**.798**	.006	.148	.101	-.121
	weekendActivityEntropy	.325	**.567**	-.358	-.302	.276	.120
	weekendSocialEntropy	*-.372*	*-.455*	*.411*	*.217*	*-.017*	*-.265*
	weekdayHomeRest	*-.148*	*.291*	*.167*	*.021*	*-.145*	*.014*
Leisure and Sleep Activities	weekendHomeSleep	-.078	.123	**-.798**	-.044	.089	-.192
	weekdayLeisuring	.064	-.029	**.763**	-.376	.266	.091
	weekendLeisuring	.050	-.018	**.758**	-.394	.231	.207
	weekdayHomeSleep	-.040	.116	**.687**	.449	-.287	.187
	weekdayActivityEntropy	-.273	-.334	**.579**	.322	-.186	-.079
	weekendLeisureSocial	*-.065*	*.312*	*-.440*	*-.021*	*-.097*	*.271*
	weekdayLeisureSocial	*.327*	*.113*	*-.421*	*-.288*	*.082*	*.227*
Working Activities	**weekdayWorking**	.120	.072	-.113	**.865**	.259	.124
	weekdayWorkSocial	-.129	.108	.106	**.789**	-.015	.042
	weekendWorking	.272	-.017	.076	**.742**	.218	.078
	weekendWorkSocial	.099	.051	.032	**-.553**	-.059	.069
	weekendHomeRest	*.070*	*-.140*	*-.125*	*.178*	*-.162*	*-.134*
Social Entropy	weekendPrivateActivity	.016	.034	.016	-.039	**-.799**	.030
	weekdayPrivateActivity	-.126	.005	.012	-.334	**-.707**	-.036
	weekdaySocialEntropy	.461	.182	-.195	-.163	**.565**	.060
	weekendHomeSocial	-.078	.020	.159	.199	**.551**	-.373
Location Entropy	**weekdayLocationEntropy**	-.223	-.226	.186	.058	-.031	**.839**
	weekendLocationEntropy	-.309	-.189	.166	.221	-.011	**.808**
	weekdayHomeSocial	*.113*	*.239*	*-.065*	*-.156*	*-.400*	*.441*

Figure 3.1 – Principal Component Analysis for Activity Features.

and 3.2, each factor has at least 2 features with factor loading larger than 0.7.

The structural model fit for our hypothesis is shown in Figure 3.3. The model goodness-of-fit indices ($\chi^2 = 1160.5$, $df = 473$, $p < 0.05$, $RMR = 0.008$), and R^2 values ($R^2 > 0.1$ for all) surpass

		Component	
		1	2
Social Life Satisfaction	SAT.GroupInteractPersonal	**.848**	-.223
	SAT.SupportFromGroupMembers	**.786**	-.151
	SAT.GroupMeetings	**-.738**	.035
	SAT.SupportFromFriends	**.735**	-.056
	SAT.SocialCircle	*-.457*	*.033*
Research Satisfaction	SAT.Learning	-.399	**.707**
	SAT.ResearchContentAndDirection	-.277	**.682**
	SAT.Overall.MIT	-.270	**.634**
	Health	*-.043*	*-.513*
	TravelFrequency	*-.111*	*-.473*

Figure 3.2 – Principal Component Analysis for Survey Data.

their recommended values. In this model, we have drawn paths from the activity related factors (Entropies, leisure, sleep, work, and communication activities) to the satisfaction-related factors. The analysis conveys three interesting causal assumptions:

- *Social Entropy - Leisure and Sleep - Social Life Satisfaction:* The increase in social activity regularities (i.e., the decrease of *Social Entropy*) improves both *Leisure and Sleep*, and *Social Life Satisfaction*. Furthermore, *Leisure and Sleep* also has direct positive influence on *Social Life Satisfaction*. Thus, we say that *Social Entropy* has an amplifying effect. To illustrate this, we select top ten regular users and top ten irregular users with respect to the feature of *social entropy*, and compare their satisfaction levels. We observe that in average the regular users report 40.74% higher satisfaction score (significant, $p = 0.023$) in the survey question with $Id = 34$ than the irregular users.

- *Working Activities - Leisure and Sleep - Social Life Satisfaction: Working Activities* have an amplifying effect on *Leisure and Sleep* and *Social Life Satisfaction*, but with a different interpretation: While both *Working activities* and *Leisure and Sleep* positively influence *Social Life Satisfaction*, working activities have negative influence on leisure and sleep. This implies that spending more time at work lowers the time for sleep and other activities. However, this analysis does not exactly show how to compute an equilibrium between work and leisure and sleep activities.

- *Working Activities - Social Life Satisfaction - Research Satisfaction: Research Satisfaction* is positively influenced by both *Working Activities* and *Social Life Satisfaction*. Similar with the previous observation, the *Working Activities* factor has an amplifying effect.

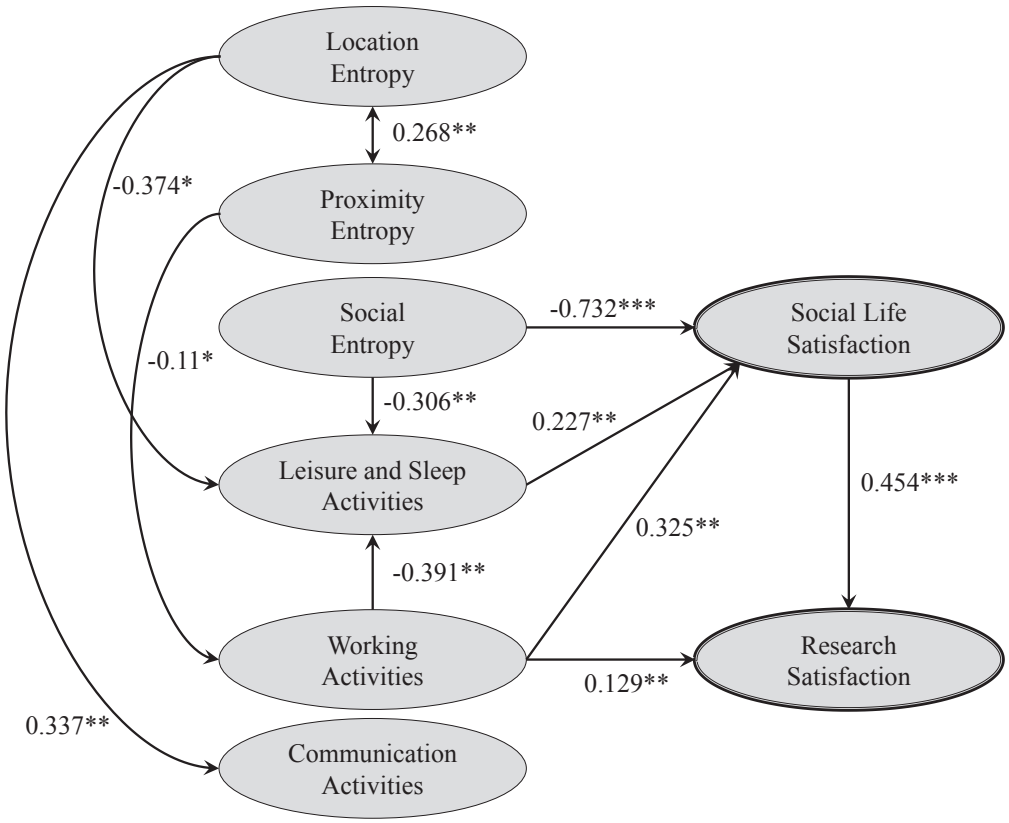

Figure 3.3 – The model fitted with SEM. The values on the directed paths denote the standardized regression weights of the model. For example, when *Leisure and Sleep Activities* goes up by one standard deviation, *Social Life Satisfaction* also goes up by 0.227 standard deviations. The paths with significance levels $p < 0.1$, $p < 0.05$, and $p < 0.001$ are marked with *, ** and ***, respectively. For brevity, we omitted the features of the factors and the paths that do not have statistical significance.

Other regularity-related factors (location, bluetooth proximity) are crucial to *Leisure and Sleep*, *Communication*, and *Workplace Activities*. Thus they indirectly regulate satisfaction: the lower the entropy is, the higher satisfaction with research and social life a user would have.

In summary, we have analysed Reality Mining dataset to identify the predictors of life satisfaction. Our results reveal meaningful relations between activities and satisfaction. More specifically, our analysis shows that work, leisure and sleep activities, and regularities in the daily activities have both direct and indirect influences over the reported levels of satisfaction. These findings can guide us toward better designs for behavior recommender systems. Specifically, given our observations on the activity entropy measurements, we designed our methods so that they capture the regularities. In the subsequent chapters we perform advanced time-series analysis to discover various behavior profiles and propose appropriate recommendations. Furthermore, the Reality Mining dataset provides a limited amount of information for our purpose. Thus, we also curated our own dataset with continuous sensor measurements on users' physical activities.

4 FactorHabiTS: Decomposing of Activities of Daily Living to Discover Routine Clusters

Our data curation strategy (see Chapter 5), with the help of modern sensor technology, helps us collect massive time series data for activities of daily living (ADLs). A behavior recommender system can use this data to infer broad patterns, such as common daily routines. In order to do so, the system must employ appropriate methods to process the ADL datasets. Most of the existing approaches either rely on a model trained by a preselected and manually labeled set of activities, or perform micro-pattern analysis with manually selected length and number of micro-patterns. Since real life ADL datasets are massive, such approaches would be too costly to apply. Thus, there is a need to formulate unsupervised methods that can be applied to different time scales.

We propose FactorHabiTS, a novel approach to discover clusters of daily activity routines. We use a matrix decomposition method to isolate routines and deviations to obtain two different sets of clusters. We obtain the final memberships via the cross product of these sets. We validate our approach using two real-life ADL datasets and a well-known artificial dataset. Based on average silhouette width scores, our approach can capture strong structures in the underlying data. Furthermore, results show that our approach improves on the accuracy of the baseline algorithms by 12% with a statistical significance ($p < 0.05$) using the Wilcoxon signed-rank comparison test.

4.1 Introduction

The abundance of wearable sensors helps people track their Activities of Daily Living (ADLs), and promises substantial opportunities in pervasive healthcare. For instance, existing medical studies depend on self-reported survey data [OSW+12, ROK+12], but they could be complemented with sensor-based measurement. Additionally, there is an ongoing effort to develop personalized lifestyle recommendations based on people's daily habits [FDRK12]. Using these tools, people can quantify their physical activities and internal metabolism over time [Sma12a]. Some systems also incorporate simple techniques to deliver correlation information for personal data [TBV12b]. However, researchers must employ even more sophisticated methods to understand what physical activity patterns people adopt, and whether these patterns cause variations in the level of physical

activeness within individuals (intrapersonal differences) or groups of people (interpersonal differences). A pattern analysis on activity routines can help identify such information, and thus enhance the usefulness of pervasive healthcare systems.

Such ambitious objectives require a reliable means to organize (e.g. clustering) massive, sensory ADL data into categories of temporal patterns followed by people. Each of these patterns would characterize the temporal dynamics of a general behavior trend - for instance, an increase of daily number of steps by 5000 steps over 30 days. People's data would be associated to these trends with additional temporal dynamics such as possible deviations (e.g. two exceptional days of inactiveness) and warps (achieving the same goal over 34 days instead of 30). Until recently, this need was addressed with activity recognition methods based on data labeling and probabilistic modeling. They would be subsequently evaluated by their recognition accuracy. The existing methods on ADL analysis either explicitly specify models for a preselected set of activities [WOCRM08], or analyse and extract features from repetitive micro-patterns (i.e., motifs). The first approach requires expert knowledge, thus it is costly and delivers a restricted understanding of the data. In the second approach, the appropriate granularity for micro-patterns must be exhaustively searched for any given dataset. As such, despite early successes [BI04, Coo10a], studies that adopt these approaches report on a limited amount of physical activities, likely monitored in laboratory conditions [PPO10, ZWGT13]. However, this requires expert knowledge and content-dependent modifications over standard modeling techniques - which are costly, and unlikely generalizable over different, ever enlarging datasets [YZK15, YZP14]. Thus, there is a need to formulate unsupervised methods that can be applied to different time scales.

We observe that people adopt some activity routines in their daily living, with some possible deviations every day. Based on this observation, we propose FactorHabiTS, a novel approach to analyse time series activity data. We pre-process the time series with a smoothing filter [HP97] and extract routines and deviations via a sparse and low rank matrix decomposition technique [LCM10]. We separately cluster the routines and deviations, and then perform a cross product between routine-clusters and deviation-clusters to find the final memberships for each entry.

FactorHabiTS's core methodology (Low Rank and Sparse Decomposition – LRSD) involves matrix factorization, which requires minimal external supervision. However, given the myriad of such methods [AT05], there is a need to assess their applicability to ADLs. Therefore, in this chapter, we also make a critical assessment of matrix factorization approaches on analyzing ADL datasets. More concretely, we perform a comparative evaluation of two state-of-the-art approaches (LRSD and Time-SVD++). Neither of these two state-of-the-art approaches were originally designed with ADLs in mind: LRSD [LCM10] is used to reduce the dimensionality in a possibly corrupted image data so as to capture regular and symmetric structures. However, this design could also permit it to characterize ADL data in terms of common trends and minor deviations, and identify the temporal patterns that people follow. Likewise, TimeSVD++ [Kor10] can model the changes of user product preferences over time very successfully - especially considering the fact that such datasets are notably sparse, i.e., they have many missing points to be fixed. This suggests that TimeSVD++ may, in a similar manner, model the changes in people's

ADLs over time, so that distinct temporal patterns can be easily identified by a subsequent clustering method.

We evaluate these approaches through two criteria: scalability and clustering quality. With our results we first confirm that these methods' superiority over basic clustering approaches, and also demonstrate notable differences between ADL datasets and customer ratings datasets.

Our contributions in this chapter are as follows:

- Our approach is different from prior work as it is model-free, and it uses the whole time series data as opposed to a subset of motifs or features.
- We propose a novel combination of low rank and sparse matrix decomposition and time warping techniques for activity analysis. To our knowledge, our approach is the first one in the activity analysis studies to incorporate this approach.
- We show, on two real-life datasets of accelerometer data (calorie expenditure and steps) and of different time scales, that our method can capture distinct structures in ADL time series that are associated with different levels of activeness. Furthermore, we show on a well-known synthetic dataset [KK03] that we can also obtain high accuracy scores on labelled time series data.
- We further demonstrate that FactorHabiTS is scalable to large datasets, and it performs better than other state-of-the-art approaches in processing dense behavior datasets.

4.2 Related Work

Activity analysis studies follow two general directions. The first approach constructs a model of some preselected activities, and establishes the fitness of this model through methods such as Bayesian Learning [ZN12b] and Hidden Markov Models [Coo10a]. The obtained models can serve to predict people's house activities [Coo10a], to group the users based on their activity routines [ZN12b], or to identify common activity routines [ZN12b]. Model-based methods are commonly applied on datasets of location and motion sensors. To obtain sound results in their models, researchers study incorporate domain expert knowledge (and perhaps manually annotate the dataset). This requires substantial effort, and constrains the quality of the analysis to the extent of the expert's knowledge ahead of the quality of the dataset.

As an alternative, studies from the second approach extract features from frequently occurring patterns (motifs in other words), and then construct classifiers based on these features. The bioinformatics field spearheads the research on discovering frequent patterns (we refer the readers to the paper of Sandve and Drablos [SD06] for an extensive review). Typically each pattern-based activity recognition study proposes a custom motif-detection algorithm [PPO10, PHL12, RCHSE11b], while some prefer to directly incorporate state-of-the-art pattern detection

algorithms such as random projection [VAS09] and Closet+ [AEA+08]. Subsequently, for classification, studies either apply state-of-the-art supervised learning techniques such as Support Vector Machines, Decision Trees [PHL12] or incorporate custom data structures (like graph-based clustering [VAS09], and routine-tree [AEA+08]). It is also possible to construct Hidden Markov Models based on the extracted patterns [RCHSE11b] or apply ensemble learning [ZWGT13]. Motif-based studies obtained empirical success on datasets from a large variety of sources: environmental motion sensors, wearable accelerometers, pressure sensors, and medical analysis data (such as blood tests and urinalysis).

Due to the computational complexity of finding motifs, some studies prefer a fixed length and number of motifs [VAS09]. Some other studies report that the accuracy (or other quality measures) of the classification and clustering consistently improves as the number of motifs increases [RCHSE11b]. On the other hand, some studies show that clustering the entire set of subsequences does not produce meaningful results [KL05]. Therefore, the scientists may have to exhaustively search for the optimal length, and the number of motifs in their studies. This, again, may limit the representation capabilities of the systems.

4.2.1 Probabilistic Modeling

Activity recognition approaches [CK14] typically propose probabilistic approaches like Bayes classifiers, Conditional Random Field Models and Hidden Markov Models, driven by the motivation to recognize and understand people's ADLs using wearable and environmental sensors. These approaches typically designate a set of probabilistic transitioning rules between some states (either of well-being or of distinct activity patterns), and try to validate these rules on a dataset. Topic modeling, a technique adapted from document-word analysis for mining semantic data, is another alternative for probabilistic modeling of behaviors [CdIAK14, FGP14].

One issue in the probabilistic modeling approach for activity recognition is that there is no standard probabilistic model that is scalable for every kind of dataset. Thus each study proposes a special modification (especially Hidden Markov Model variants), so that the model can handle the properties of some specific dataset. Consequently, where some methods employ simple clustering procedures for preprocessing [SJS05], other studies resort to more complex variations such as Hidden Semi-Markov Model [DBPV05], Markov Logic Network [GER15], or voting Multi-HMM model constructed with frequent pattern mining [RCHSE11a]. These variations imply the requirement of an immense amount of expert knowledge and additional computational complexity for each separate case. As a result, their solutions often lack generalizability.

4.2.2 Matrix Factorization

Matrix factorization is a well-known approach to reduce dimensionality in large datasets. Recommender systems [AT05] typically use this to obtain item and user profiles for many applications including product, music, and movie recommendations. Further applications for matrix factor-

ization include adaptive web searches based on user profiles [SHY04], and image and video processing [WY13].

In addition to its extensions to accommodate temporal relations [DL05, Kor10], contextual information [VMO+12] and probabilistic relations [PPL01, WY13], collaborative filtering requires minimal (if any) effort from users in constructing user preference profiles, making it a suitable approach for behavior profiling. Some studies empirically demonstrate this usefulness on sparse mobile phone data [ZLN13].

Another prominent example is Principal Component Analysis. Eagle and Pentland [EP09] employ Principal Component Analysis to decompose behavioral patterns from mobile phone data into eigenvectors, which they name as "eigenbehaviors". With this strategy, each user can be modeled as a weighted combination of eigenbehaviors, allowing further analysis to predict the missing patterns during a single day. On the other hand, Principal Component Analysis is known to be very sensitive to noise, and this study does not address this problem. A more recent study [YZP14] employs a noise-tolerant variant of PCA, called Linearized Alternating Direction Method [LCM10] on wearable accelerometer data. However, this PCA variant has a cubic computational complexity, rendering this proof-of-concept approach impractical to apply on large datasets.

4.3 Timeseries Clustering Method: FactorHabiTS

People adopt some activity routines in their daily living, with some possible deviations every day. We developed FactorHabiTS based on this insight: It isolates the regular trends from the deviations, processes them separately, and captures the activity routines as time series clusters.

We summarize the flow of data processing in Figure 4.1. We pre-process the ADL time series data with a smoothing filter [HP97] and apply a low rank and sparse decomposition [CLMW11] to isolate routines (L-Matrix) from the deviations (S-Matrix). We separately cluster L-Matrix and S-Matrix, using Dynamic Time Warping [KP99b] as the distance metric. We use the well-known Silhouette index [KR09b] to determine the optimal number of clusters. We then perform a cross product of the two separate cluster sets to find the final memberships for each day.

4.3.1 Smoothing Filter

The physical activity time series data may contain noise in the form of small fluctuations. Such characteristics of the raw data can deteriorate the quality of clustering. We address this issue by applying the Hodrick-Prescott filter [HP97]. This is a well-known trend analysis method in economics. The filter decomposes a given time series object $Y = (y_1, ..., y_m)$ into a summation

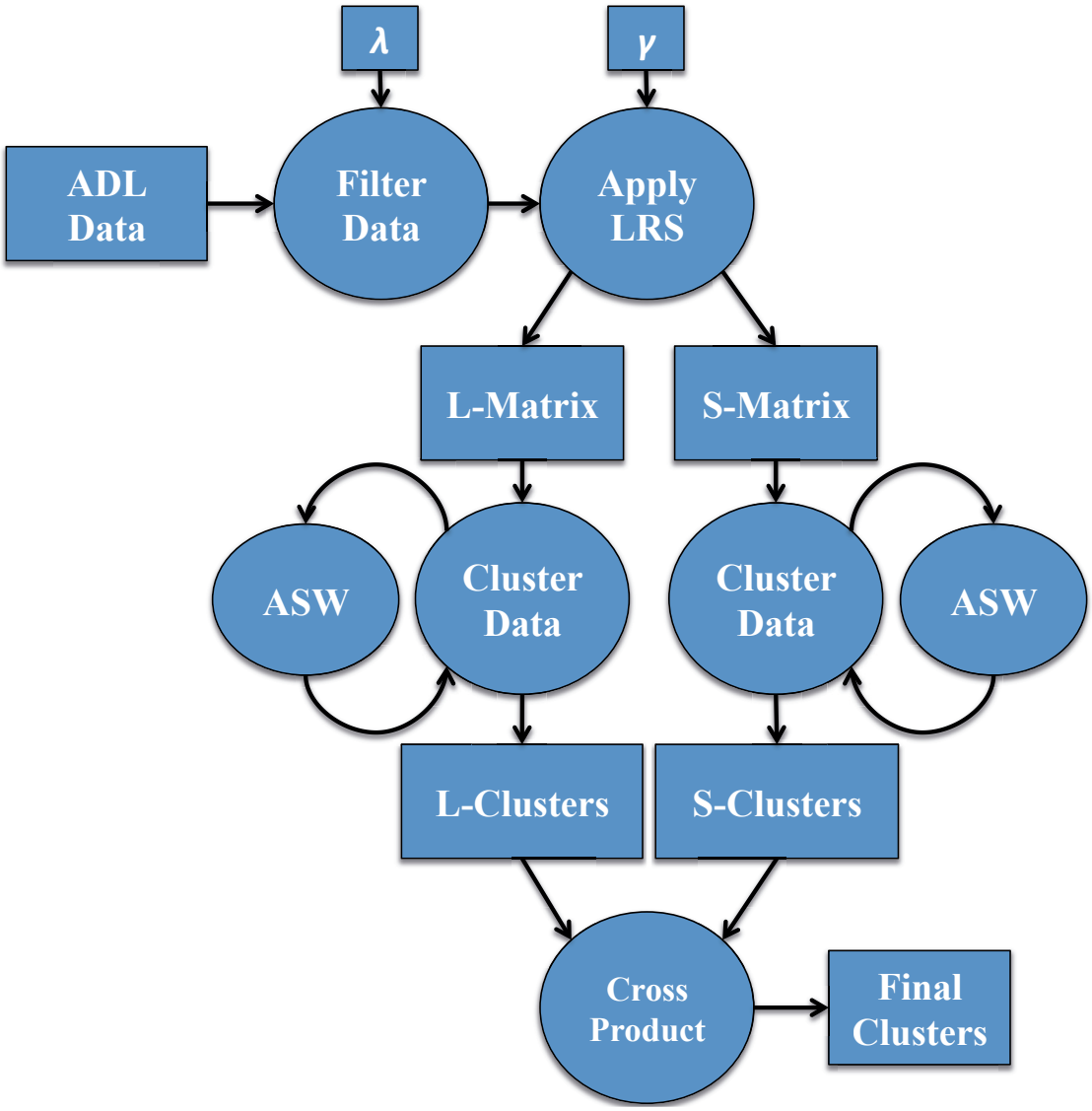

Figure 4.1 – The data flow in our approach. LRS stands for “Low rank and sparse decomposition", and ASW stands for “Average Silhouette Width"

$y_t = T_t + C_t$ such that the objective function

$$c = \sum_{t=1}^{m} {C_t}^2 + \lambda \sum_{t=2}^{m-1} ((T_{t+1} - T_t) - (T_t - T_{t-1}))^2, \tag{4.1}$$

is minimized over $(T_1, ..., T_m)$, where T_t represents the trend component (the desired output), and C_t represents the cyclical component. Increasing the smoothing parameter (λ) results in smoother trend components at a cost of more information loss. We discard the cyclical component and use

the trend component in the further steps.

4.3.2 Matrix Decomposition

The *low rank and sparse decomposition* (LRS) is a recently discovered approach that aims to capture regular and symmetric structures within a possibly corrupted data matrix [LRZM12]. While it is designed for image processing problems such as video surveillance and face recognition, it is also used other high-dimensional data mining tasks such as finding topic models in document analysis [MZWM10b].

Based on existing studies [CLMW11], we can formulate this decomposition problem as

$$L_{opt}, S_{opt} = min(\|L\|_* + \gamma\|S\|_0)$$
$$s.t.\ M = L + S,$$

where L is the low-rank matrix, and S is the sparse matrix. $\|L\|_*$ denotes the nuclear norm of L, which is the best approximation for the rank of L. $\|S\|_0$ is the number of non-zero entries in S. $\gamma > 0$ is the parameter to make a trade-off between the rank of L and the sparsity of S. Theoretical studies show that it is optimal to set γ as $1/\sqrt{max(n_1, n_2)}$, where n_1, n_2 are the number of rows and columns of M, respectively [CLMW11].

The interpretation of the L-matrix and S-matrix differs among the related studies. L is commonly regarded as the “true matrix", which is recovered from the errors and missing values denoted in S [ZT11]. In a related study, L contains linearly aligned images and S contains the rotational errors from the original matrix [PGW+10]. In some other image processing studies, L is considered to be the background and S the non-background objects in the given images [KC12b]. As such, depending on the application, the information in both of these matrices can be useful.

We use the Linearized Alternating Direction Method [LCM10] on the matrix of ADL time series data to identify common daily routines (in the form of low-rank matrix) and deviations (in the form of the sparse matrix). To our knowledge, our study is the first to apply the low rank and sparse decomposition approach to ADL analysis.

4.3.3 Distance Metric: Dynamic Time Warping

ADL routines are subject to nonlinear warps in the time dimensions (e.g. waking up 15 minutes late, having lunch for 30 minutes instead of 45, etc.). Dynamic Time Warping (DTW) is a dynamic programming-based distance metric to compensate these warps [BC94]. In contrast to Euclidean distance, DTW takes local misalignments into consideration, and reports the optimal warping path between the given two sequences. The DTW distance between the time series data

Q and P can be calculated as

$$DTW(Q,P) = min_W(\sum_{k=1}^{K} d(w_k)), \tag{4.2}$$

where $d(w_k) = (q_i - p_j)^2$ such that (q_i, p_j) is on the warping path w [Fu11]. This optimization problem can be solved by dynamic programming (longest common subsequence). Various studies with artificial datasets [KP99b], image data of letters in historical documents [RM03], speech data [SC78b], and kitchen tool usage data [PPO10] suggest that DTW improves the classification accuracy of the time series classification algorithms in comparison to Euclidean distance. DTW is sensitive to noise [Fu11]. This can be overcome by applying additional preprocessing [RM03]. We avoid this problem by applying Hodrick-Prescott filter before the matrix decomposition stage.

4.3.4 Clustering

We obtain pairwise distance matrices for the L-matrix and the S-matrix. Then we feed these distance matrices to agglomerative hierarchical clustering with complete linkage. As a result, for each row in the original data, there will be one cluster membership from L-matrix and one cluster membership from S-matrix. L-clusters represent the common trends and S-clusters represent the common deviations. To determine the final memberships, we perform a cross product of L-clusters and S-clusters, i.e., we explore all possible combinations of L-clusters and S-clusters. The maximum possible number of final clusters is (number of L-clusters) × (number of S-Clusters). We discard the clusters with no members. To guarantee the optimal number of clusters, we select the number of L-clusters and S-clusters that result in the highest Average Silhouette Width.

4.4 Experiments

4.4.1 Datasets

CBF Dataset.

This artificial dataset [KK03] contains time series objects that belong to one of three distinct shape characteristics (i.e., Cylinder $c(t)$, Bell $b(t)$ and Funnel $f(t)$, see Figure 4.2). The dataset can be generated with the following equations:

$$c(t) = (6+\eta)\chi_{[a,b]}(t) + \epsilon(t) \tag{4.3}$$

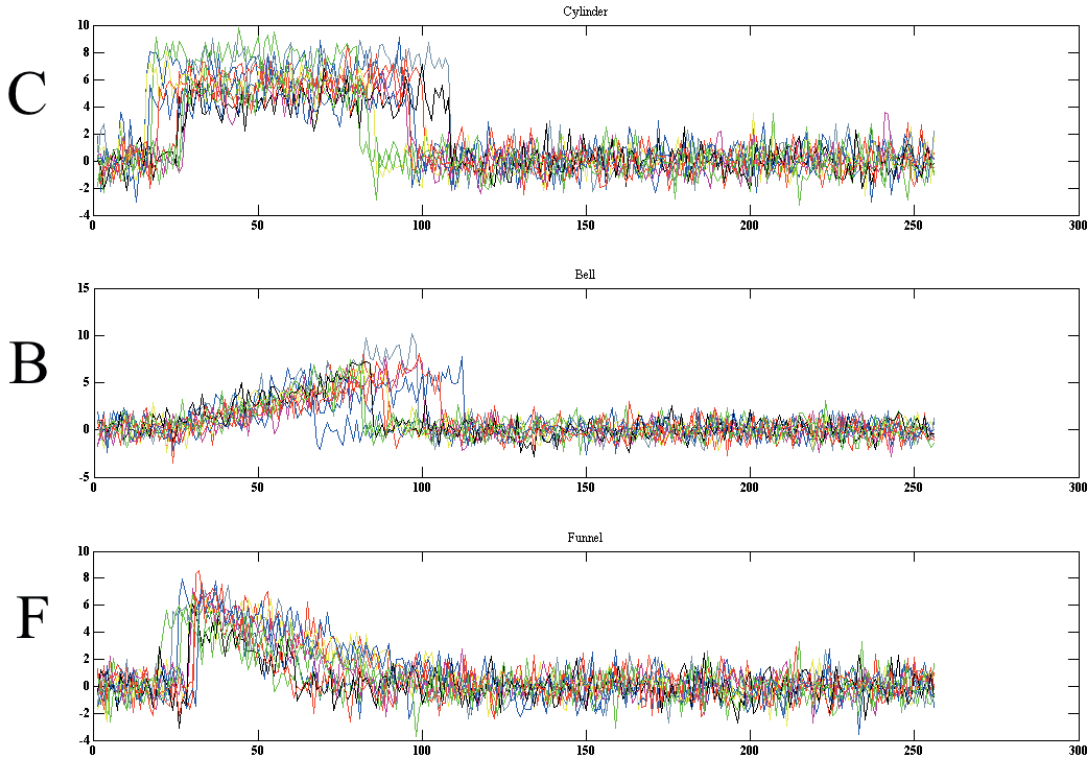

Figure 4.2 – Samples from the CBF dataset. The axes are unitless. Each class of objects (C:Cylinder, B: Bell, F: Funnel) is defined uniquely by its shape characteristics.

$$b(t) = (6+\eta)\chi_{[a,b]}(t)\frac{t-a}{b-a} + \epsilon(t) \tag{4.4}$$

$$f(t) = (6+\eta)\chi_{[a,b]}(t)\frac{b-t}{b-a} + \epsilon(t), \tag{4.5}$$

where η and $\epsilon(t)$ are drawn from a standard normal distribution, a is an integer drawn uniformly from [16, 32], and $b-a$ is an integer drawn uniformly from [32, 96]. We have generated 256 instances for each class (cylinder, bell, and funnel), each of which contains 256 data points.

E-Walk Dataset.

This dataset is the courtesy of the Yiqizou company, which provide a platform for people to form social groups and walk together. This dataset contains step counts of 236 people, who wore modern wearable accelerometers in October 2013 for a month. In its raw form, each data point represents activities during a single day. Due to some possible reasons (losing interest in the

program, forgetting to wear the sensors, sensor batteries running out, etc.), 3108 out of 7080 data points (approximately 44%) have the value 0. We represent steps time series data in a matrix where each row represents a person. There are a total of 236 time series objects, each of which has 30 data points, one for each day. Here, we can analyse the long-term usage of pedometers, and the patterns that differentiate the long-term physical performances.

HealthyTogether Dataset.

Previously collected for another study [CP14b], this dataset contains the calorie expenditure data of 48 users wearing Fitbit (a wearable accelerometer) for ten days in the period between April 2013 and June 2013. In its raw form, each data point represents activity during a single minute. This dataset do not have any missing values. We process the data in a matrix where each row represents a day. There are a total of 480 time series objects, each of which has 1440 data points. With this dataset, we can analyse the effects of daily routines on the daily physical performance.

4.5 Baseline Evaluations

4.5.1 Overall Comparison

We compare our method with some well-known baseline algorithms (namely, K-means, 1-nearest neighbor, and agglomerative hierarchical clustering). We employed Euclidean distance for K-means and DTW distance in 1-nearest neighbor and agglomerative hierarchical clustering.

Since the E-Walk and HealthyTogether datasets do not have labels, we evaluate our method via internal cluster evaluation. We specifically employ overall Average Silhouette Width [KR09b]. This value indicates the quality of the underlying structure of the clusters: values below 0.25 indicate no structure, values between 0.25 and 0.5 indicate a possibly strong structure, and values above 0.5 indicate a very strong structure [KR09b] (see Appendix A.3 for more details).

Cluster Id	Median of Daily Steps
E1	1842
E2	4194
E3	6461
E4	10357
E5	10646
E6	13782

Table 4.1 – The median of daily step counts for each cluster in E-Walk dataset, with ids matching with those in Figure 4.4.

Figure 4.3 conveys the average silhouette width scores for the three datasets. On average, our method outperforms baseline methods in ASW by 0.455, and it is able to capture clusters with high quality. We have applied Wilcoxon signed rank test with $p < 0.05$ to compare our method's

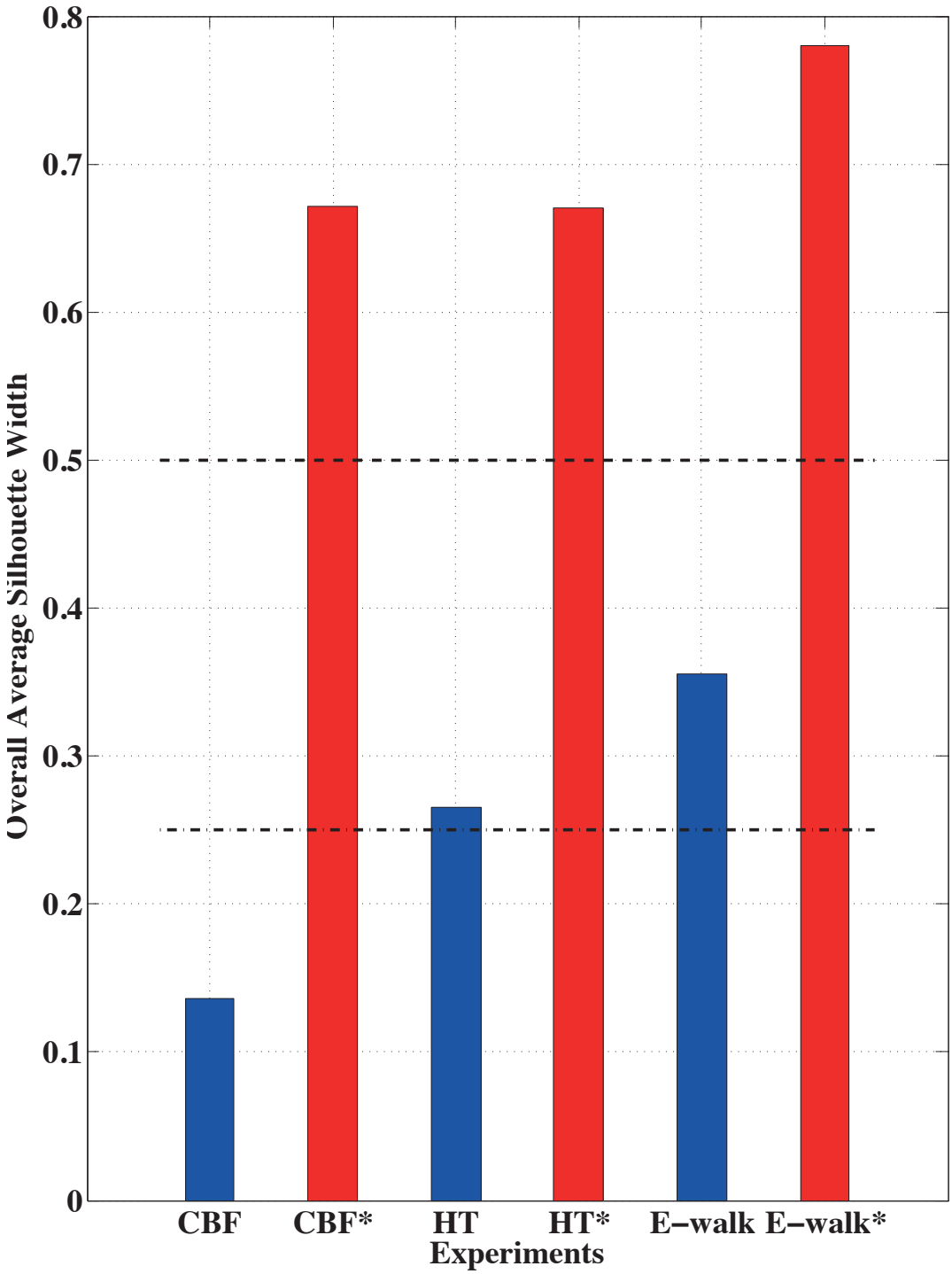

Figure 4.3 – The average silhouette width scores for clustering with (denoted by *) and without our method. “HealthyTogether" is abbreviated as “HT". The lines drawn on 0.25 and 0.5 denote boundary for acceptable and good values of ASW, respectively. We report the highest average score achieved with baseline methods.

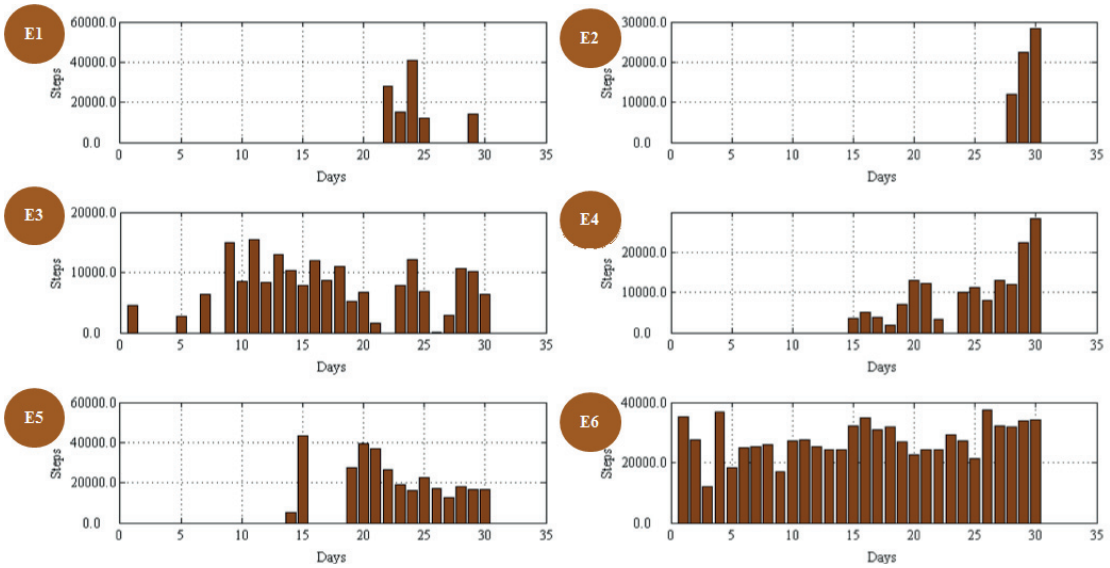

Figure 4.4 – The medians of the clusters for the E-Walk dataset ($\lambda = 100$ and $\gamma = 0.065$). Y axis represents the steps taken and X axis represents the days.

and baseline methods' ASW scores in each dataset, and validated the significance of these improvements.

4.5.2 CBF Results

Since CBF dataset contains labels, we also evaluated CBF dataset's output clusters with external evaluation indices (accuracy, F-1 score, normalized mutual information - NMI and Jaccard index) with 10-fold cross validation. Table 4.2 summarizes these scores in the CBF dataset. We refer the reader to Appendix sections A.4, A.5, and A.6 respectively on the formulas of F-1 score, NMI and Jaccard index.

Our approach outperforms baseline methods in terms of accuracy (by 12%), F-1 score (by 0.18), normalized mutual information (by 0.21), and cluster purity (by 0.25). For each of these indices, we compared our method against each of the baseline methods with Wilcoxon signed rank test with $p < 0.05$, and validated that these improvements are significant.

Experiment	Accuracy	F-1	NMI	Jaccard Index
K-means	0.75	0.62	0.51	0.46
Hierarchical	0.81	0.72	0.63	0.58
1-NN	0.93	0.87	0.78	0.77
Our method	0.95	0.92	0.85	0.86

Table 4.2 – The external index scores for the CBF dataset.

4.5.3 E-Walk Results

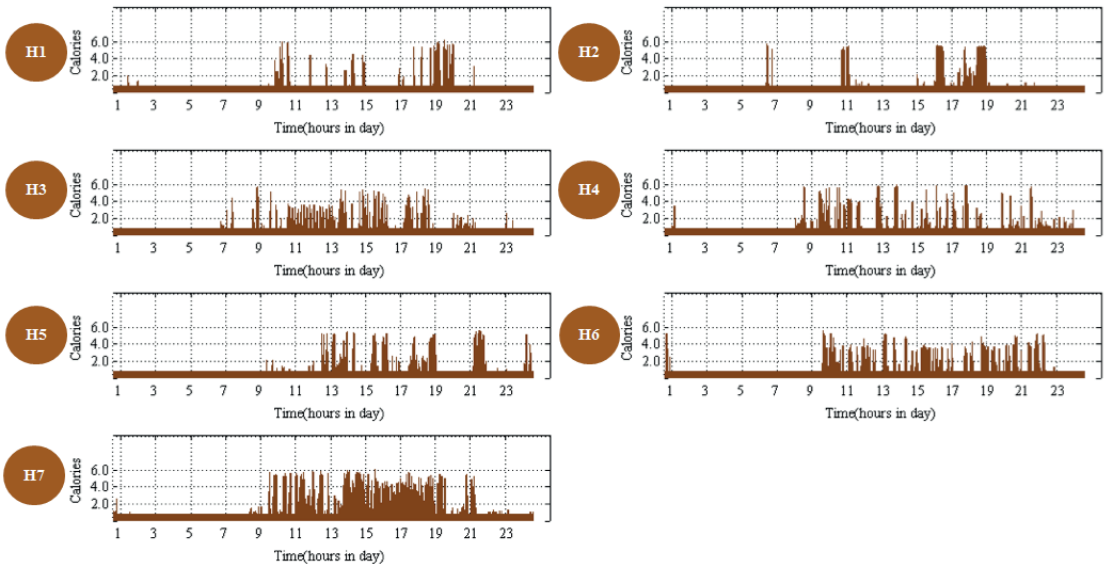

Figure 4.5 – The medians of the clusters for the HealthyTogether dataset ($\lambda = 100$ and $\gamma = 0.026$). Y axis represents the calorie expenditure and X-axis represents the hours in the day.

The representatives for each cluster (member with median number of average steps), and the selected values for the parameters λ and γ are shown in Figure 4.4. The median calorie expenditures for all clusters are shown in Table 4.1. Through the 6 clusters that we obtain from this dataset, we can observe the long-term usage patterns of pedometers. For instance, some people convey a novelty effect, i.e., they performed well in the early days of their pedometer usage, but then lost their engagement. Such people are generally grouped in the clusters with lowest average number of steps. We also observe that regularity of activeness has positive contribution towards higher average numbers of steps.

4.5.4 HealthyTogether Results

The representatives for each cluster (member with median calorie expenditure), and the selected values for the parameters λ and γ are shown in Figure 4.5. The median calorie expenditures for all clusters are shown in Table 4.3.

The results show that we can characterize 7 types (clusters) of daily activity routines. These routines can be associated to some persona, such as “Commuter" (H1), who has two main peaks in the morning and afternoon; “Afternoon Break-taker" (H2), who is more active in the afternoon with frequent “breaks"; “Early morning person" (H3), who is more active in the early times of the day; “The Frequent breaker" (H4), who takes frequent breaks through the day; “Night Person" (H5), whose is more active late at night; “Hyperactive" (H6), who has moderate, and continuous activeness through the day; and “Traveler" (H7), who has high and continuous activeness through the day.

Through these 7 clusters, we can observe how the intra-day patterns can contribute to the average daily activeness. The average step count increases from the “Commuter" type of daily routine to “Traveler" type of daily routine. Similar to the clustering results in the E-Walk dataset, we see that regular distribution of activeness contributes most to the level of activeness.

Cluster Id	Median of Daily Calories
H1	1412
H2	1519
H3	1587
H4	1640
H5	1660
H6	1862
H7	2353

Table 4.3 – The median of daily step counts for each cluster in HealthyTogether dataset, with ids matching with those in Figure 4.5.

4.6 Scaling up and State-of-the-Art Comparisons

Many probabilistic methods suffer from one or more of the following shortcomings: dependence on supervised labels or expert knowledge, or high computational complexity. As such, they are not scalable and efficient enough for large and unlabeled datasets for ADLs. Matrix Factorization approaches are however an exception, therefore they are more practical for our task. We now proceed to describe how to make FactorHabiTS scalable, and we present how we compare it with a state-of-the-art alternative in more detail.

4.6.1 Scalability issues in FactorHabiTS

The most common method to solve the Equation 4.3.2 has a time complexity of O(N^3) [LCM10], which would render this variant and the methods that employ it (e.g. [YZP14]) prohibitively costly for large datasets. For an optimized performance, we instead use Robust Grassmann Averages [HFB14]. This approach models the dimensionality reduction problem as the averages of subspaces spanned by the data. This modification helps the method discern the local deviations to a great extent. Given the input data $y_{1:N}$, each iteration in Robust Grassman Average proceeds with the following two equations:

$$\omega_n \longleftarrow sign(u_n^T q_{i-1}) \parallel y_n \parallel \tag{4.6}$$

$$q_i \longleftarrow \frac{\mu_{rob}(\omega_{1:N}, u_{1:N})}{\parallel \mu_{rob}(\omega_(1:N), u_(1:N)) \parallel} \tag{4.7}$$

where $\omega_{1:N}$ would denote weights, q_i is the weighted average of robust means computed at iteration i, $u_n = \frac{y_n}{\|y_n\|}$, and μ_{rob} denotes a robust average. In this study, we specify μ_{rob} as the trimmed mean.

4.6.2 TimeSVD++

TimeSVD++ [Kor10] is a variant of collaborative filtering that models temporal changes in customer preferences, i.e., concept drift. TimeSVD++ extends the existing factor model SVD++ [Kor08] and incorporates the temporal dynamics under three bias components (user bias $b_u(t)$, item bias $b_i(t)$, and global bias μ) at a given time t. TimeSVD++ uses these components, along with the user u's preference $p_u(t)$, item i's characteristics c_i , factor vector f, and the set of items already rated by user ($R(u)$) to predict the rating of user u for item i with the following prediction rule [Kor10] :

$$\widehat{r}_{ui}(t) = \mu + b_i(t) + b_u(t) + c_i^T(p_u(t) + |R(u)|^{\frac{-1}{2}} \sum_{j \epsilon R(u)} f_j) \tag{4.8}$$

Without disregarding its original function as a recommender system routine, we can also interpret TimeSVD++ as a procedure to reconstruct a time series dataset with missing values. The design of TimeSVD++ intends to capture long-term trends of temporal data, while avoiding the short-term patterns that would not have a predictive influence on future trends. Furthermore, as demonstrated in Netflix dataset, its predictive capabilities outperform existing state-of-the-art factor models (SVD, SVD++) while running very fast in rating datasets [Kor10].

4.6.3 Adoption of Matrix Factorization Methods

Adoption Approach: We use Low Rank and Sparse Decomposition in a flow of time series processing that captures common trends and deviations, followed by clustering based on Dynamic Time Warping (see Figure 4.6). We pre-process the ADL time series data with a simple moving average filter, and apply the decomposition to obtain two separate matrices for long-term patterns and short-term deviations. We separately cluster these two matrices, using Dynamic Time Warping (DTW) [BC94] with Keogh's lower bounding [KR05] as the distance metric. We then perform a cross product of the two separate cluster sets to find the final memberships for each time series object.

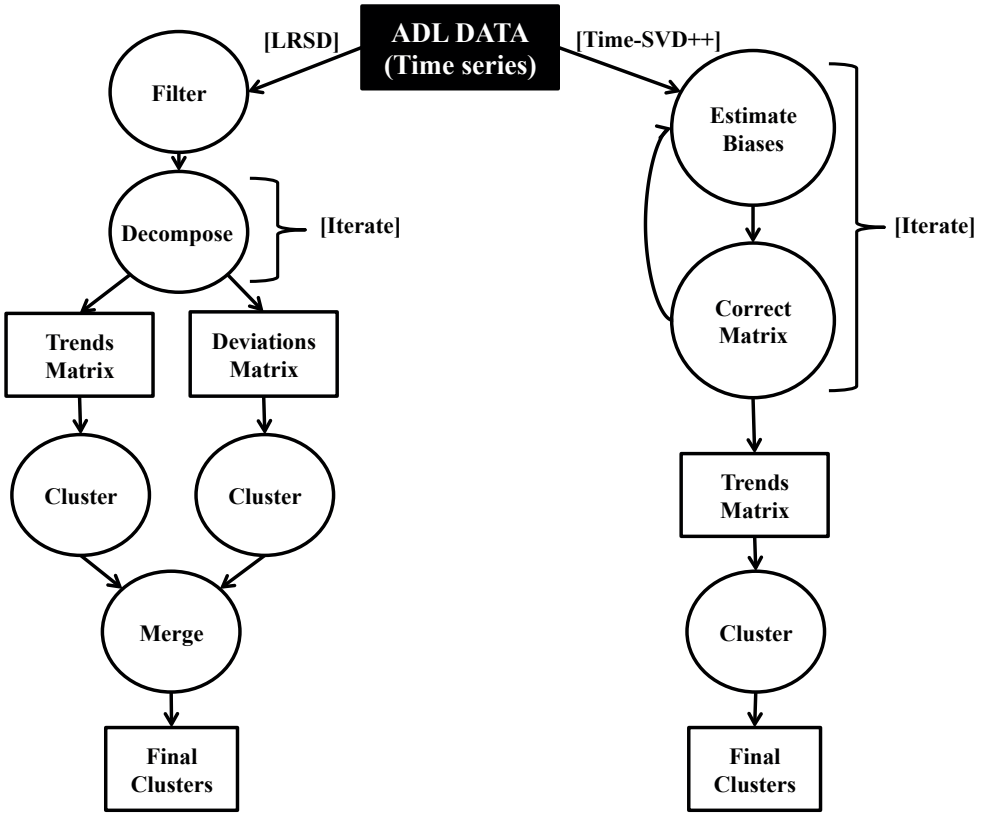

Figure 4.6 – The flow of data processing in FactorHabiTS- and TimeSVD++-based clustering. FactorHabiTS decomposes the data into two components (trends and deviations), while TimeSVD++ discards the deviations altogether

In its naïve implementation, the dynamic time warping costs $O(D^2)$ for comparing a single pair of time series objects. Various alternatives reduce this complexity [SC78a, KP99a]. For an optimized performance, we use Keogh's lower-bounded Dynamic Time Warping as the distance metric in the clustering phase [KR05]. Furthermore, we employ the simple moving average filter to circumvent DTW's sensitivity to noise.

We adopt TimeSVD++ as follows (Figure 4.6): we run the algorithm on the dataset with initializing the bias components based on the dataset's properties. This initialization maps the global bias μ to the global average of ADL levels (e.g. calorie expenditure) per unit of time (day, hour or minute), the user bias $b_u(t)$ to the average ADL level of the user up to time t, and the item bias $b_i(t)$ to the average ADL level of all users up to time t. $p_u(t)$ naturally maps to the current ADL level of u at time t, and f maps to other users' ADL level. With this setup, the algorithm corrects the matrix at time t based on past data and updates the bias parameters for future data. Thus, the warps and deviations are discarded. We then cluster the corrected matrix.

4.6.4 Complexity Evaluation

We denote U as the set of users, $|U| = N$, and D as the maximum possible number of observations (total number of businesses recorded in rating datasets and the length of sensor utilization in ADL datasets).

FactorHabiTS

The complexity of FactorHabiTS depends on the total complexity of filtering, decomposition and clustering phases (see Figure 4.6).

Filtering Phase: The simple moving average transforms a given time series object $X = (x_1, x_2, \ldots, x_D)$ to a rolling average $Y = (y_1, y_2, \ldots, y_D)$ by the following formula:

$y_k = y_{k-1} + \frac{(x_k - x_{k-l})}{l}$

where l is the length of the sliding window. This takes takes $O(1)$ operations to complete for each observation in the time series object, amounting the complexity for a single entry to be processed in $O(D)$ and the entire dataset to be processed in $O(ND)$.

Decomposition Phase: A single iteration in The Robust Grassmann Averages (TGA) has a computational complexity of $O(KND)$, where the parameter K denotes the number of components to be found in the dataset [HFB14]. It always holds that $K \leq D$, so the worst-case performance of TGA is bounded by $O(DND) = O(ND^2)$, raising the overall complexity of LRSD to $O(ND) + O(ND^2) = O(ND^2)$.

Time-SVD++ based clustering

A single iteration to train a model based on TimeSVD++ is determined as $O(\sum_{u \epsilon U} |R(u)|^2)$, where $R(u)$ denotes the set of items rated by the user u [Kor10]. This renders TimeSVD++ very efficient in processing highly sparse item rating datasets where $|R(u)| \ll D$. In datasets such as HealthyWalkers, however, the dataset is dense: $R(u)$ naturally converges to D, raising the complexity to $O(ND^2)$.

In summary, with the same number of iterations, matrix factorization steps in FactorHabiTS and TimeSVD++ have the asymptotically equivalent runtimes. Furthermore, since Euclidean distance and DTW with Keogh's lower bounding is the same, i.e., $O(D)$, both algorithms also have the same runtime for clustering.

4.6.5 Clustering Quality

Datasets

We evaluate the capabilities of the methods through two activity datasets:

- **The YELP Dataset**. YELP dataset contains approximately 2.3 million ratings from 70,000 users for 15,000 businesses. In this study, we analyse the 2014 version of this ever-growing dataset, which includes the ratings within the time period from 01-02-2005 to 28-01-2014. Similar to other rating datasets, the YELP dataset is very sparse: there are 402 out of 3284 days with no records of ratings or reviews. On average, each user has 33.69 ratings (minimum 1, maximum 3286).

  Processing: We process this dataset as a matrix where each row represents a business, and each data point is the cumulative rating of the business for a single day. In this manner, each cluster would represent a distinct rating pattern for businesses, thus help making post-hoc analysis on the common features of businesses that generally perform well or bad.

- **HealthyTogether**. In a previous study [CP14a], we have conducted a user study to investigate the effects of self-monitoring and social intervention on the overall activeness of users. The end product of this study is the HealthyTogether dataset, which contains the calorie expenditure data of 83 users wearing Fitbit (a wearable accelerometer) between September 2013 and September 2014. The length of user participation varied from 12 to 235 days ($\mu = 48.36$, $\sigma = 54.85$). Contrary to YELP dataset, this dataset do not have any missing values.

  Processing: We process the data in a matrix where each row represents a day, and each data point represents one minute. With this representation, there are a total of 4014 rows, each of which has 1440 data points, and users' data are processed altogether. In this manner, the clusters can be used to identify common daily routines adopted by the wearable sensor users. Such a segmentation of activity routines could help us develop specialized interventions to improve each of the daily routines.

Results

We use Average Silhouette Width (ASW) in comparing the clustering quality of the three approaches (see Appendix A.3). Kaufman and Rousseauw [KR09b] proposed this metric to measure in-cluster consistency and inter-cluster distinctiveness. ASW scores are bounded in the interval $[-1, 1]$. When comparing two methods, the one with a higher ASW score is said to be producing clusters with higher quality. Furthermore, ASW has suggested values for validation: any score below 0.25 would indicate a bad quality of clustering (comparable to random partitioning), scores within $[0.25, 0.5]$ would indicate an acceptable level of quality, and scores above 0.5 would indicate a high quality of clustering. Table 4.4 summarizes the clustering qualities of three alternative approaches: first approach is Hierarchical Agglomerative

Table 4.4 – The ASW scores for clustering algorithms on HealthyTogether and YELP datasets. Higher scores imply better clustering quality. (*) indicates an acceptable level, while (**) indicates a good level. The differences in the scores are statistically significant (Wilcoxon signed-rank test: $p < 0.05$).

Method	**HealthyTogether**	**YELP**
HAC without Matrix Factorization	0.26	0.46 (*)
TimeSVD++	0.4 (*)	0.59 (**)
FactorHabiTS	**0.69** (**)	**0.79** (**)

Clustering without any of the matrix factorization steps, and the other two are LRSD-based and Time-SVD++-based approaches. Wilcoxon signed-rank test on silhouette scores indicate statistically significant differences between their performances. ($p < 0.05$). Both TimeSVD++ and LRSD improve the clustering quality as opposed to baseline method. While all methods perform well in YELP dataset, the data density in HealthyTogether dataset takes its toll on their performances. The clustering qualities of baseline method and TimeSVD++ suffer particularly more (-43% and -32%, respectively) than FactorHabiTS (-12%). While both TimeSVD++ and FactorHabiTS can produce clusters with highly reliability in YELP dataset, only FactorHabiTS can do so in HealthyTogether dataset.

4.7 Discussion and Future Work

We proposed a novel approach to perform cluster analysis on ADL data. This approach is different from prior studies as it can process ADL time series without expert knowledge or micro-pattern extraction. Our approach is useful to reveal clusters with high external and internal evaluation scores, and it outperforms baseline algorithms (for instance, by 12% of accuracy and 0.455 points of average silhouette width) with statistical significance. The employed matrix decomposition technique makes our method suitable for high-dimensional data, paving the way for further possible applications such as analysing between-subject variabilities and multi-sensor data.

Our next step is to employ our understandings we obtained from this study to identify and elaborate on predictors or crucial behavior patterns that lend to activeness in daily physical activity routines. Such an analysis of clusters was shown to be useful in predicting illnesses based on behaviour patterns [MCLP10].

4.8 Chapter Summary

Future applications on Activities of Daily Living require a good level of understanding of ADL data clustering. In this chapter, we presented our efforts to improve this understanding via our novel method and a comparative study of alternative methods. We first briefly reviewed why methods based on matrix factorization are more suitable than other approaches for this

task. Then, we described two particular matrix factorization techniques used in different fields (computer vision and recommender systems, respectively). Following the discussion of their respective design strategies, we elaborated the modifications necessary to adapt these techniques for clustering ADL data. Then we presented their similarities and differences through theoretical and experimental analysis, with a particular emphasis on three key aspects: handling temporal dynamics, scalability, and sensitivity to data density. We have realized these analyses through runtime complexity analysis and measuring the clustering qualities through experiments on a physical activity dataset and a ratings dataset.

From a conceptual point of view, this chapter enhances our understanding of the properties of ADL datasets: they are similar to image datasets, in the sense that they can be dense and arbitrarily noisy. This became evident in the course of our comparisons as Time-SVD++, which is engineered for and works great in sparse ratings datasets, produces suboptimal results in clustering ADL datasets. We therefore validated the suggestion of Farrell et al. [FDRK12] that behavior profiling methods should not be sensitive to the level of sparsity in the datasets, and we add that such methods must also explicitly deal with noise and temporal dynamics (i.e., warps and deviations).

The clustering task we study is also closely related with the well-known collaborative filtering scheme in recommender systems. As such, from a practical point of view, this chapter solidifies the analytical building blocks of our behavior recommender system. With FactorHabiTS, our system can obtain temporal profiles, assign each user a temporal profile, and personalize its recommendations to each user. In Chapter 6, we show how to use temporal profiling with intervention data to obtain the full-fledged behavior profiles.

5 Data Curation

5.1 Introduction

The behavior recommender has to deliver personalized suggestions to its users. In order to do so, it would exploit users' past patterns of Activities of Daily Living (ADLs). However, just like traditional recommenders, the system's capabilities are limited for new users, as it will not have enough information for their past. This phenomena, *cold start problem*, is one of major challenge in recommender systems [SPUP02].

Traditional recommender systems attempt to solve this with eliciting the preferences of users, either by asking them to provide data or implicitly observing their behaviors. It is also possible to use a hybrid model of item-based and user-based similarities to further alleviate this issue [SPUP02]. While such methods do minimize user effort, it still holds that the recommenders should receive users' activities and behaviors (such as ratings).

As we established in the thesis introduction, the goals of behavior recommendations are more complicated than those of classical recommendations: they should inspire their recipients with small, incremental and achievable goals. This raises an additional challenge in collecting the required data. The data must allow the system to find a trade-off between two conflicting criteria, i.e., *effectiveness* and *feasibility* of recommendations: On one hand, the system must make sure that its users achieve a steady improvement in their ADL patterns. On the other hand, it should avoid setting extreme goals that can injure or discourage the users. The behavior recommender system thus must model its users' behavioral responses to potential recommendations, and make sure whether its recommendations would have a significant influence on its users' behavior patterns. Because of this requirement, the behavior recommender system cannot mimic the data collection from traditional recommender systems. A mere collection of behavior patterns does not lend itself to the crucial insight on why some recommendations or other external interventions may succeed or fail to help a user for behavior change.

In this chapter, we show how to perform the data curation in order to tackle these challenges. In

our data curation, we not only measure people's ADL patterns, but also deliberately introduce an intervention to monitor its effects on people's patterns. We base our solution on the following insight: if the system understands the impacts of past interventions for behavior change, it can predict its users' behavioral responses to its own recommendations. To further motivate this idea, consider the following scenario, which tells how InspiRE, a behavior recommender system, could generate the optimal recommendation for John:

> Unbeknownst to John, InspiRE had already been curating data from other users and categorizing them based on their behavior patterns and their responses to its recommendations in the past.
>
> InspiRE considers Charlie and Mary, whose behavior patterns had been very similar to John's, and then both of them received recommendations. InspiRE identifies that Charlie relapsed to a less active lifestyle, but Mary managed to increase her activeness and switched to a more active behavior pattern without getting injured. Since Mary was similar to John in the beginning, the system decides that Mary's activity pattern will be an ideal candidate to guide John to safely increase his activeness.

To demonstrate our data curation approach, we performed a longitudinal user study with 83 participants and investigated the influences of exercise partners on the overall activeness measured by wearable sensors. In this study we equipped participants with wearable sensors and asked them to form dyads, i.e., two-people exercise groups. While the wearable sensors provide us with the activity data, the formation of dyads provides us with the intervention data to bootstrap our recommendations. We recorded their pre-intervention and post-intervention patterns, which are then processed with our behavior profiling methods.

We continue this chapter by proposing how to generalize from this specific data curation study. In fact, our data curation approach can be applied to ADLs other than physical activities and interventions other than pairing up with exercise partners. In Chapter 3, we outlined a range of possible data sources to build the behavior recommender system [Sma12b]. In this chapter we complement this information with other possible interventions by investigating previous studies on behavior change. The resulting approach is a combination of single-case design and randomized controlled trials. We implement this data curation with a longitudinal user study, where we observe participants' behavior patterns and introduce an intervention to measure its potential impacts on the behavior patterns.

The contribution of this chapter is two-fold: first, we propose and implement a data curation strategy for physical activity recommendations. Second, we elaborate and provide suggestions for the generalized version of this data curation, linking with the widely accepted data collection approaches in medical experiments.

5.2 Related Work on Data Curation

Data curation implies a deliberate design to collect data. Such a design should guarantee a significant amount of data, all the while avoiding potential biases and noises. There exists a vast variety of data collection strategies. In behavior analysis studies, we observe four major styles of data curation: qualitative user studies, crowdsourcing, randomized controlled trials, and single case designs. In this section we review some examples of these strategies, and identify the relation between them and our approach.

5.2.1 Ethnography study

Ethnography is the branch of anthropology where the goal is to provide a detailed, in-depth description of people's everyday life and practice [1]. These studies are typically based on interviews, focus group, observations and field studies which can last days, months or years. These studies aim to capture the values, the attitudes, the motivation towards a particular service as well as the social context and the living settings involved [LSH+09a].

Depending on the desired outcome, researchers have used various ethnographic technics. Some ethnographic studies only focus on interviews [BJR10], while others may involve field trials and observations. Field observations are especially useful where the researchers would like to see the day-to-day interactions of the participants with a new technology. Example studies involve interactions with a conversational robot [SKH11] or wearable devices such as google glass [MVR+14] or physical activity trackers [MLO+16].

For the case of behavior recommenders, ethnography studies are invaluable as we can get insights on people's attitude on sensors and other means of data collection. These insights ultimately guide the design of a high quality data curation study. On the other hand, these studies cannot replace the actual data curation, as it is often not possible to attach a quantitative, statistical significance to the study results.

5.2.2 Crowdsourcing

Crowdsourcing provides a way to outsource the immense effort of data collection and annotation to a large group of people, usually through web. Typical designs for such a collection involve a set of micro-tasks to be completed by participants using a web-based questionnaire (such as Amazon Mechanical Turk [KCS08]) or mobile crowdsourcing platforms [YMH+09].

The participants are paid for each micro-task they accomplish. Alternatively, they can be motivated with various gamification strategies. Researchers should pay particular attention to motivate people to provide truthful and high-quality responses [KCS08].

[1] Brian A. Hoey. What is Ethnography ? http://brianhoey.com/research/ethnography/

Crowdsourcing is particularly useful in obtaining labels and annotations for the data. As such, it is becoming increasingly popular to apply crowdsourcing in fields of image analysis and text mining from social media. We refer the reader to the exemplary studies of Sintsova for further details [Sin16].

Our data curation design differs from crowdsourcing by two major aspects. First, the number of participants in a crowdsourcing study is undefined in the beginning, and the interaction with participants are terminated once they finish their micro task. In this aspect, our study is quite the opposite of a crowdsourcing, as we aimed to get detailed information about our participants, with the intention of delivering personalized recommendations. Second, contrary to typical crowdsourcing studies, we did not break down our study into repetitive micro-tasks. We only considered the data of the participants who have fully completed our data curation study. With these two major differences, we compromised from the total number of participants in favor of the length and the quality of the data.

5.2.3 Randomized Controlled Trials

Randomized controlled trial (RCT) is a term derived in clinical studies [CSB+81]. In this type of studies, participants are randomly chosen to receive a certain intervention. The strategy in this data collection procedure is to have a uniform distribution of various characteristics except the designated interventions or condition the researchers would like to test.

We can find different types of interventions and outcomes in randomized controlled trials. For instance, one study [ESBC+08] investigates the effects of a drug called raloxifene on fracture risk in postmenopausal women. The study divides participants into the raloxifene and placebo groups for a period of five years. The researchers found that there was no difference between these two groups in risk of nonvertebral fractures, but women treated with raloxifene reduced their risk of vertebral fractures.

Another RCT study [KBFK06] reports on the effect of multimedia education for children with asthma. A control group of pediatric patients with asthma was given standard asthma educational resources, while the experimental group of pediatric patients with asthma was given standard resources plus multimedia resources. The study found a reduction in daily symptoms, in emergency room visits, in school days missed, and in days of limited activity in the group given multimedia education resources.

RCT designs reduce statistical biases and help researchers determine the effect of the designed intervention. In this manner, randomized controlled trials make it particularly easy to evaluate the results with well-established statistical tools. On the other hand, RCT designs require data collection from hundreds of subjects, which makes it relatively difficult to apply RCT methods in fields such as eHealth and behavior change.

5.2.4 Single-Case Designs

Single-Case experimental designs (SCDs) offer an alternative approach to control groups in RCTs. In SCDs, every participant "serves as his/her own baseline" [Smi12, DCR13]. SCDs have two objectives:

1. To test the success of an intervention on a particular case (either a person or a community)
2. To provide evidence about its general effectiveness of the said intervention.

These approaches have been around for some decades, and are still being deployed in medical, educational and psychological studies. Such studies are also getting prominent in our field: we can name an earlier study with an exercise app that pairs its participants into exercise partners [CP14a], and a study involving adaptive persuasive messages for improving snacking behavior [KRMA12].

Providing a control group does have an advantage over the single-case designs (SCD) as it can take more external factors into account. For instance, suppose that we conducted an intervention study on students, which coincided with the exam periods. SCD results would be more difficult to justify, as every student goes through this period and is affected. On the other hand, SCDs achieve their goals with a relatively small sample size [Kaz82]. This is an advantage over fully randomized experiments, which may require hundreds of participants. Perhaps this is why, according to a study, SCDs are becoming more and more prominent applied research, particularly in eHealth, mHealth and behavior change [DCR13].

5.2.5 Our contributions

In this chapter we convey our implementation of a Single-Case Design. A close investigation confirms that we can associate our study with "Multiple Baseline Design" [DCR13], a subset of SCDs, which has a baseline period followed by an intervention period. Our design specifically concurs with "concurrent multiple baseline design", as our interventions happen at the same time (5th day) for the participants. Our design involves longitudinal sensor data collection, deliberate interventions, and interactions between participants in the course of the study. In this manner, the behavior recommender can have access to examples of proven behavior change. Our study also collected qualitative data, which leads us to derive suggestions for the sensor equipments for future data curation studies.

Upon a careful inspection, we also see that our curation strategy satisfy various heuristics proposed by prior SCD studies:

- A rule of thumb is that there must be at least five data points in the baseline period [HCH05]. Our studies satisfy this condition.

- There are some heuristics to evaluate the validity of the results [PB92]. These are: the a large change in the level (immediate effect), a large change in the mean (long-term effect), replication of effects across participants. Studies consider interrupted time-series analysis as valid means of evaluation, as it already includes these heuristics [BAW00]. We not only obtain these heuristics, but we also attached statistical significance.
- As suggested in the SCD studies, we support statistical results with visuals [PB92]. Specifically, we provide the aggregated time series plots of our participants. This further supports the validity of our statistical results throughout the thesis.

Our data curation does not re-introduce the baseline period after the intervention period. Such a study ("baseline-intervention-baseline") would be a Reversal study, where it is reasonable to assume that people can go back to their baseline conditions after the intervention. For instance, the researchers could stop administering a drug and see the changes in patients' metabolism. This is certainly not the case in HealthyTogether, our proposed study. Our intervention, our mobile app, cannot be removed like a drug prescription from our participants' lives. It induces a permanent effect on our participants' awareness of their partners' and their own activities. Also, some participants continued using the app even after the study.

5.3 Data Collection with the HealthyTogether study

5.3.1 Data Analysis Requirements

We take into consideration that analysis of sensor measurements can only work well if there is a clear specification of all the relevant context parameters of the sensing process. Algorithms usually need to be fed with accurate and well understood historic time series data.

We can further extend the requirements depending on what sort of analysis we wish to perform. For *supervised learning* approaches, the data need to be precisely labelled with the time points when interesting events happened. In the context of physical activity data, example to such events are: walking, cycling, running, rest, swimming, etc. If one can provide such data to the algorithms for training, the resulting classifiers will eventually be able to associate characteristic patterns of sensor measurements with the activity classes. To limited degrees it is possible to use another set of methods to make sense of unannotated data. We can categorize such settings as *unsupervised learning*. In this setting, it is more about discovering new patterns rather than validating some pre-defined categories of patterns.

From the perspective of data analysis, the data collection specifications can be broadly summarized as:

1. Sensor specifications: The collected information, the collection frequency, timestamps (down to the granularity of the sensor's collection frequency)

2. Scenario specifications: What is the context of usage, where are the sensors placed, any potential discrepancies?

3. Annotations: Careful recording of interventions, interesting observations, special events, the timestamp of such events

4. Data Formats and Standards: Keeping a consistent/integratable formats of the sensor recordings (JSON, CSV, SQL Tables), the database (MongoDB, MySQL, TinyDB)

5.3.2 Procedure

We take the approach from a prior study [CP14a] and design our user study with elements of social influence (see Figure 5.1). It is previously shown that following each other's activities serves as a social intervention, with an effect of an increase of people's activeness up to 15% on average [CP14a].

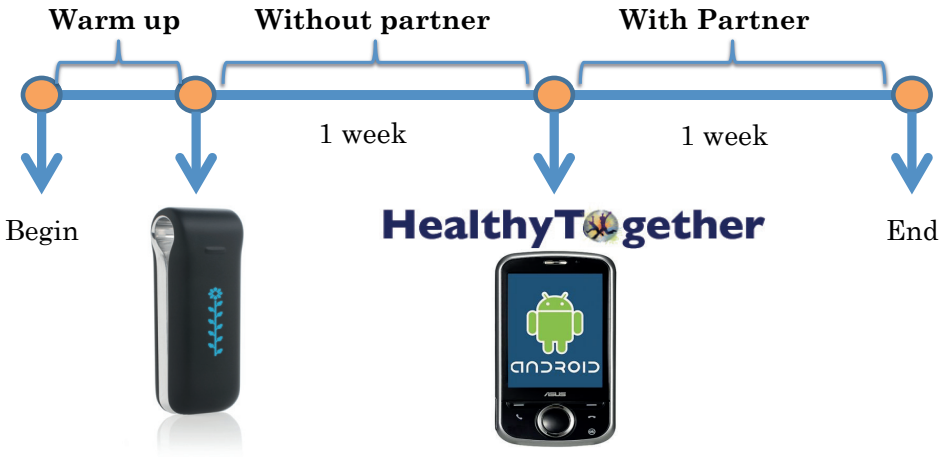

Figure 5.1 – We followed this timeline in our HealthyTogether study.

We organized the duration of the study as a warm-up session of 2 days, and two equally long phases of control and experiment (one week each). In the warm-up session, we distributed a wearable sensor to participants and let them familiarize themselves with it. We started our data collection from the beginning of the control phase, where the participants continued using the wearable sensor. We afterwards started the experiment phase: we gave the participants Android phones with an installation of our custom mobile application. Using this application, we asked the participants to form groups of two, i.e., *dyads*. The application's interface allowed the dyads to monitor each other's step patterns, as well as exchanging messages (see Figure 5.2).

In this study the participants were assigned to two conditions, which depend on the performances of their exercise partners:

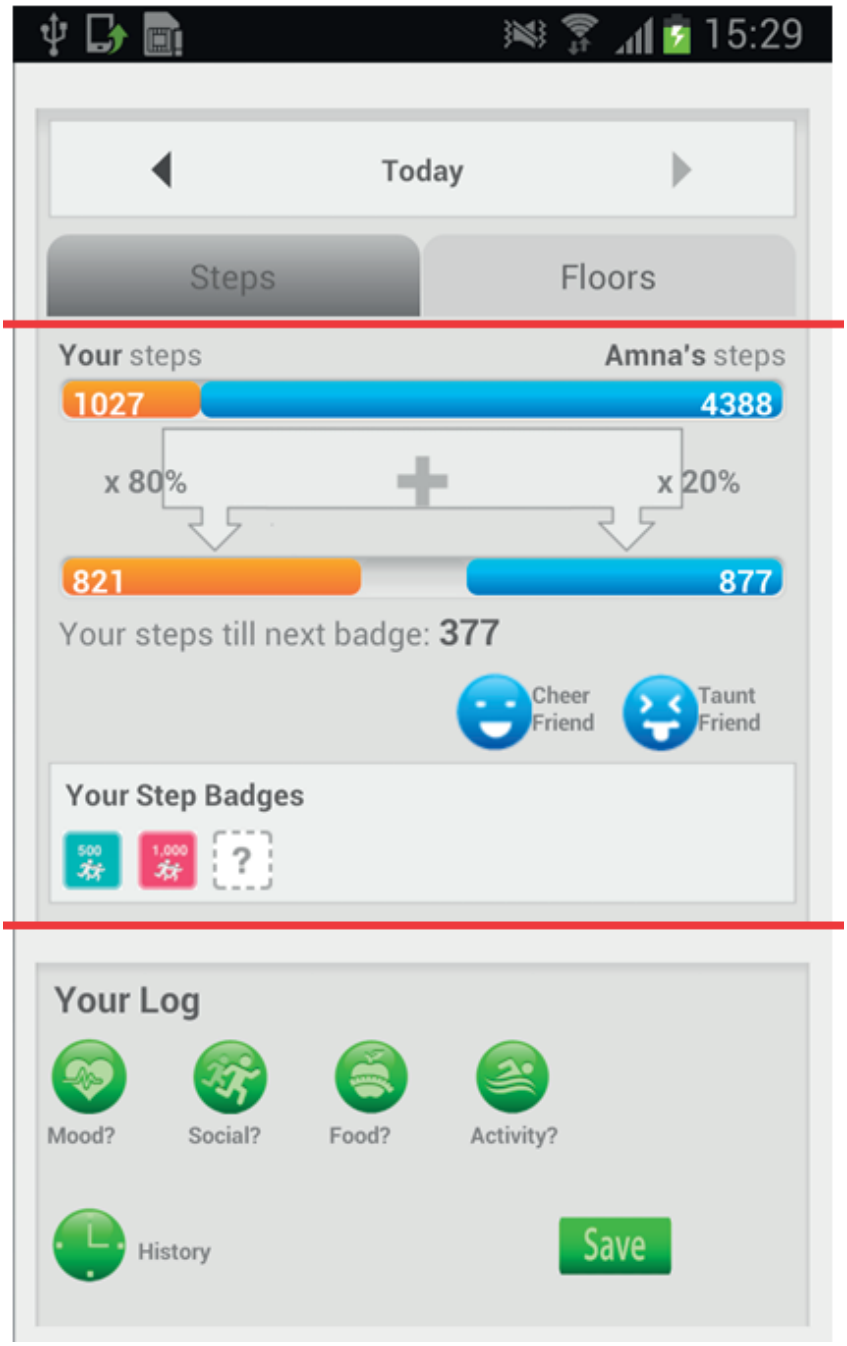

Figure 5.2 – The main screen of HealthyTogether, the mobile application used during the data curation study

- Harmonious Dyad: Each participant in this condition matches with a partner whose level of activeness is similar to him/her.
- Disparate Dyad: Each participant in this condition matches with a partner who is significantly more or significantly less active than him/her

We measure and validate the disparity between dyads with the disparity score, which we discuss in Chapter 6. Furthermore, we have conducted interviews with the participants in the beginning, middle, and the end of the study. The experiences of the participants have provided us with valuable insights about the experiment setup, which we elaborate in Section 5.5.2

5.3.3 Collected Data: HT-83

We recruited participants by on-campus advertisements and by our collaboration with a hospital (University Hospital of Geneva) in the region. 83 people joined this study, who are originally from 17 different countries. 15 of them had been diagnosed as Diabetes Type-II and 68 of them were non-patient students. We compensated participants with 50CHF gift cards.

While the original study lasted for two working weeks, we have extended the period of the study depending on the availability of the participants. The length of participation varied from 12 to 235 days ($\mu = 48.36$, $\sigma = 54.85$). Overall, we performed the curation between September 2013 and September 2014. Figure 5.3 summarizes the collected data. For the people who stopped participating after day 12, we filled missing values with the mean values of the continuing participants.

While it is possible to implement this procedure with any wearable sensor, we used Fitbit One, a wireless activity tracker. This sensor is unobtrusive and convenient to use. Furthermore, its API allows us to easily fetch data in time series. The sensing involves an accelerometer, and it tracks the steps, floors and calorie expenditure of the user over time. The resulting dataset contains time series of every user's steps with 1-minute precision, as well as the dates when the users were paired. In our main experiments, we aggregate the sensor data values so that each data point in a user's time series data corresponds to the total count of steps for a day. All of the data is stored in CSV format.

Thanks to the Fitbit API and our custom mobile application, our data is clearly annotated with timestamps. The labeling of physical activities (running, cycling, rest, etc.) proved to be particularly difficult, as the participants rarely entered logs to the diary available in the mobile app. The manual annotation of ADLs is one of the major bottlenecks in activity analysis studies. The implications of this bottleneck eventually leads us to develop unsupervised methods on calorie expenditure data to discover and analyse common behavior patterns, as we demonstrate in Chapter 4.

5.4 Data Curation with the SNACK study

As we outlined in Chapter 3, physical exercise is only one of the possible types of data to be used in behavior recommender system. In fact, thanks to the rapid advancement of wearable technology, digital food databases and computer vision techniques, it has been becoming easier than ever to track nutrition habits of people.

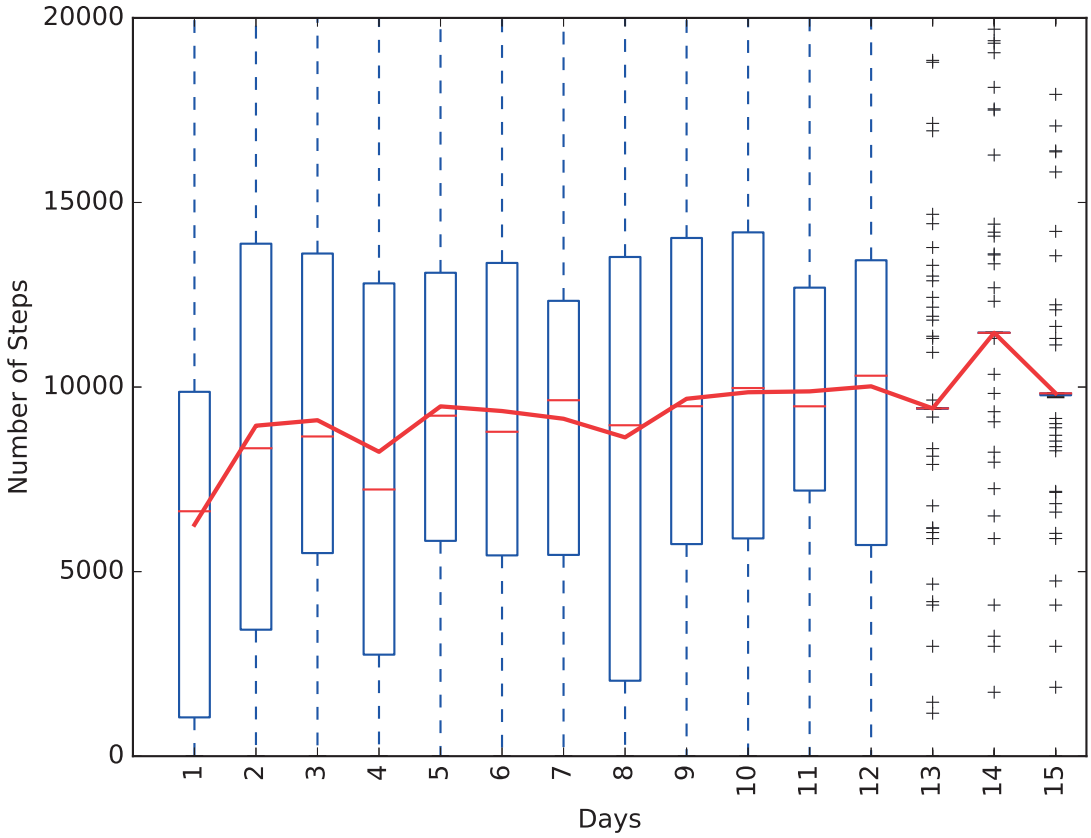

Figure 5.3 – The aggregated time series data of HealthyTogether Users. Our intervention starts at day 5. Original study lasts for 12 days in total, although there were participants that continued afterwards. For the sake of clarity of the figure, we limited the number of days to 15.

Surprisingly, the data curation strategy arrived earlier than the sensor technology for nutrition. An earlier study investigated the effects of persuasive message to the snacking habits over a period of two weeks [KRMA12].

This study, which we call SNACK, follows the timeline in Figure 5.4. The participants first fill in a questionnaire, which quantifies the participants' susceptibility to the 6 distinct styles of persuasion (see Chapter 2 for a list of these elements). Then, for 5 days, they regularly logged a diary consisting of the following quesitons:

- The number of snacks they had today
- The number of unhealthy snacks they had today

In the second half of the study, As the participants continued logging their snacking diaries, they started to receive daily persuasive messages to cut down their snacks. In this part, the participants

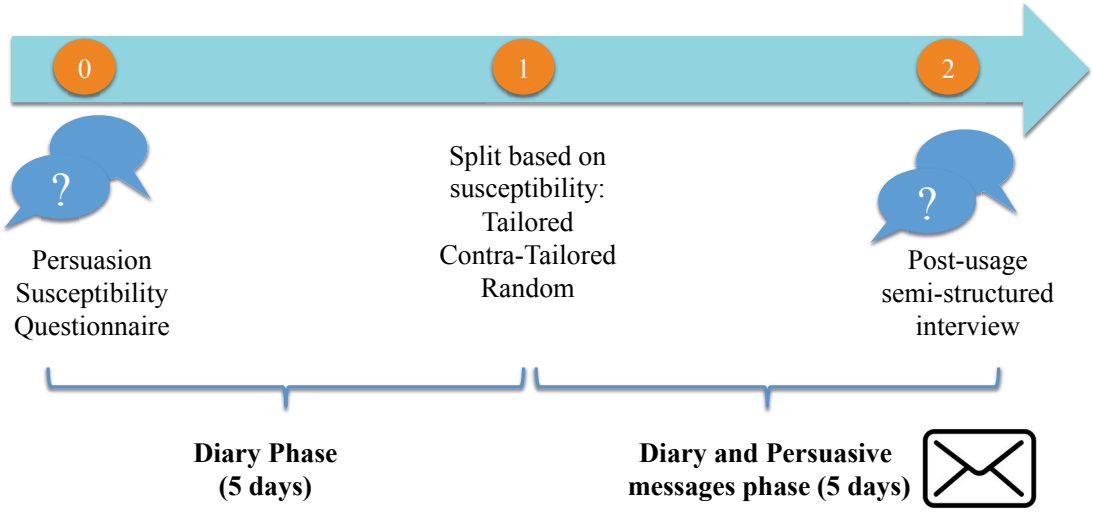

Figure 5.4 – This is the timeline for the SNACK study

were assigned to three conditions, which depend on their scores in the susceptibility questionnaire. These conditions determined the content of the message they received every day:

- Tailored Condition: Each participant in this condition receives messages whose contents are tailored based on the persuasion strategy that he/she is most susceptible to.
- Contra-Tailored Condition: Each participant in this condition receives messages whose contents are tailored based on the persuasion strategy that he/she is least susceptible to.
- Random Condition: Each participant in this condition receives randomly selected persuasive messages.

The resulting study consists of 73 people, all Dutch citizens, from various age groups and genders. Figure 5.5 summarizes the collected data. We filled missing values with the mean values of the continuing participants. We perform an analysis of this dataset in Chapter 6. This study proposes daily messages as an alternative form of intervention.

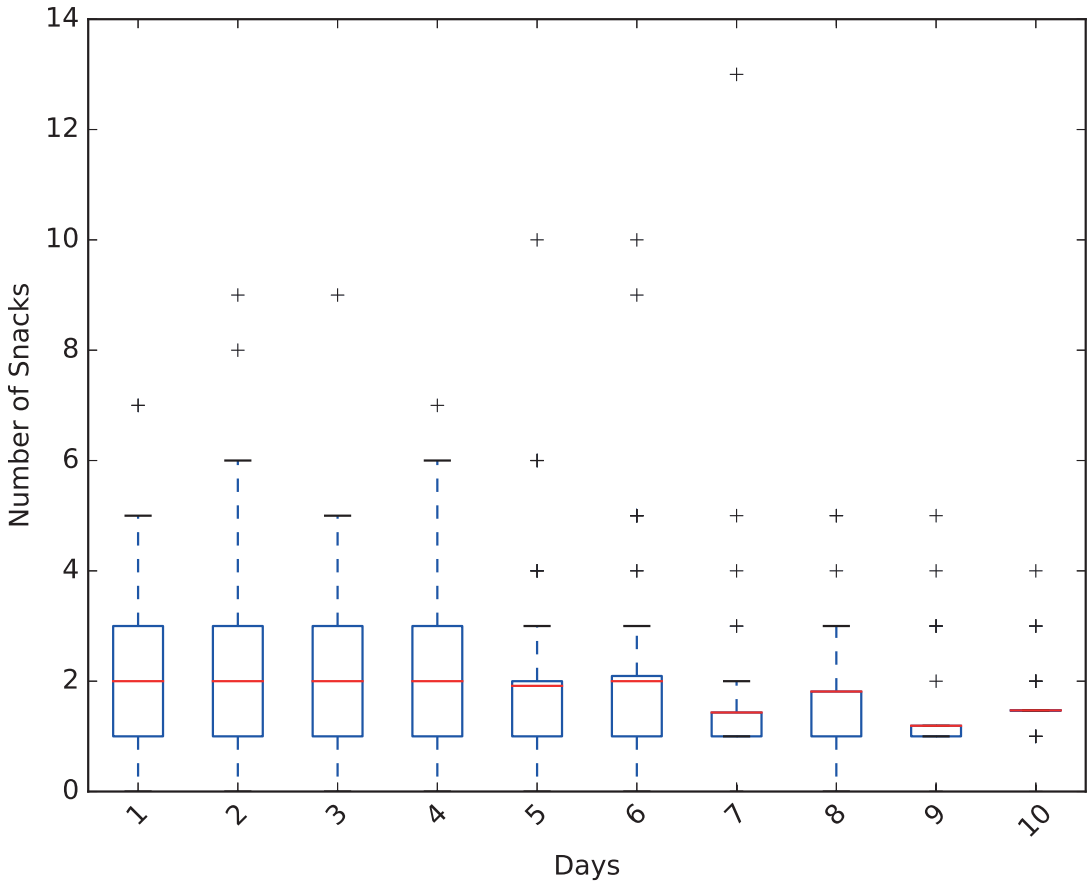

Figure 5.5 – The aggregated time series data of SNACK Users. The intervention starts at day 5. Study lasts for 10 days in total.

5.5 Results

5.5.1 Generalization

Based on the HealthyTogether and SNACK studies, we can layout the following key elements for a generalized form of our data curation:

- Technology-Mediated Interventions: As shown in the HealthyTogether and SNACK datasets, we can specify the interventions as social influence (exercise partners), persuasion (daily messages), and daily goal setting.
- Recording pre-intervention and post-intervention periods: This step ensures that we can measure the impact of the intervention on a given participant.

Given enough number of participants, we can introduce a third element, i.e., splitting the participants to experiment and control groups. In this manner, this curation study becomes an

extention of Randomized Controlled Trials.

In summary, our generalized data curation is a longitudinal study that involves users in an experiment over an extended period of time. We investigate how technology-mediated interventions can users achieve active lifestyles. We propose a randomized controlled trial, where we monitor two groups of patients. One group starts receiving our custom intervention 3 weeks after the start. The diagram in Figure 5.6 below shows how Agata, a typical participant, would undergo this user study.

5.5.2 Sensors for Future Curation Studies

We identified that since many ADLs are collected through sensors, it is critical to ensure that such studies will involve sensors that satisfy a number of criteria. With the contribution of qualitative feedback from the study participants, we identified these criteria as follows:

- Accuracy: The quality of the sensor dictates the usefulness of data for the recommendations.
- Safety: Proper safety measures ensure long-term usage of the wearable sensor.
- Data accessibility and seamless data transition: Sensors that support wireless data transmission and APIs for data fetching is more practical than those that require a USB connection for data upload.
- Ease of use and usefulness: Wrist-worn or pocket sensors are perceived as easy to use. Furthermore, it is preferable by participants to have a device that can sense multiple types of information (calories, heart rate, sleep, etc.)
- Affordability: Cheaper sensors are easier to deploy for a data curation study.

We leveraged these findings and outlined our comparisons between a number of physical activity sensors (see Figure 5.7) for future studies on data curation.

5.6 Chapter Summary

A functional behavior recommender requires a deliberately curated dataset. Such a curation should not only adress the classical cold start problem, but also help the behavior recommender find the optimal trade-off between effective and feasible recommendations.

In this chapter, we describe how to design this data curation. In our approach we measure people's behavior, but we also deliberately introduce an intervention to monitor its effect on people's patterns. We demonstrate our approach through HealthyTogether study, in which we curated physical activities of 83 people along with a social intervention (pairing up the users with exercise

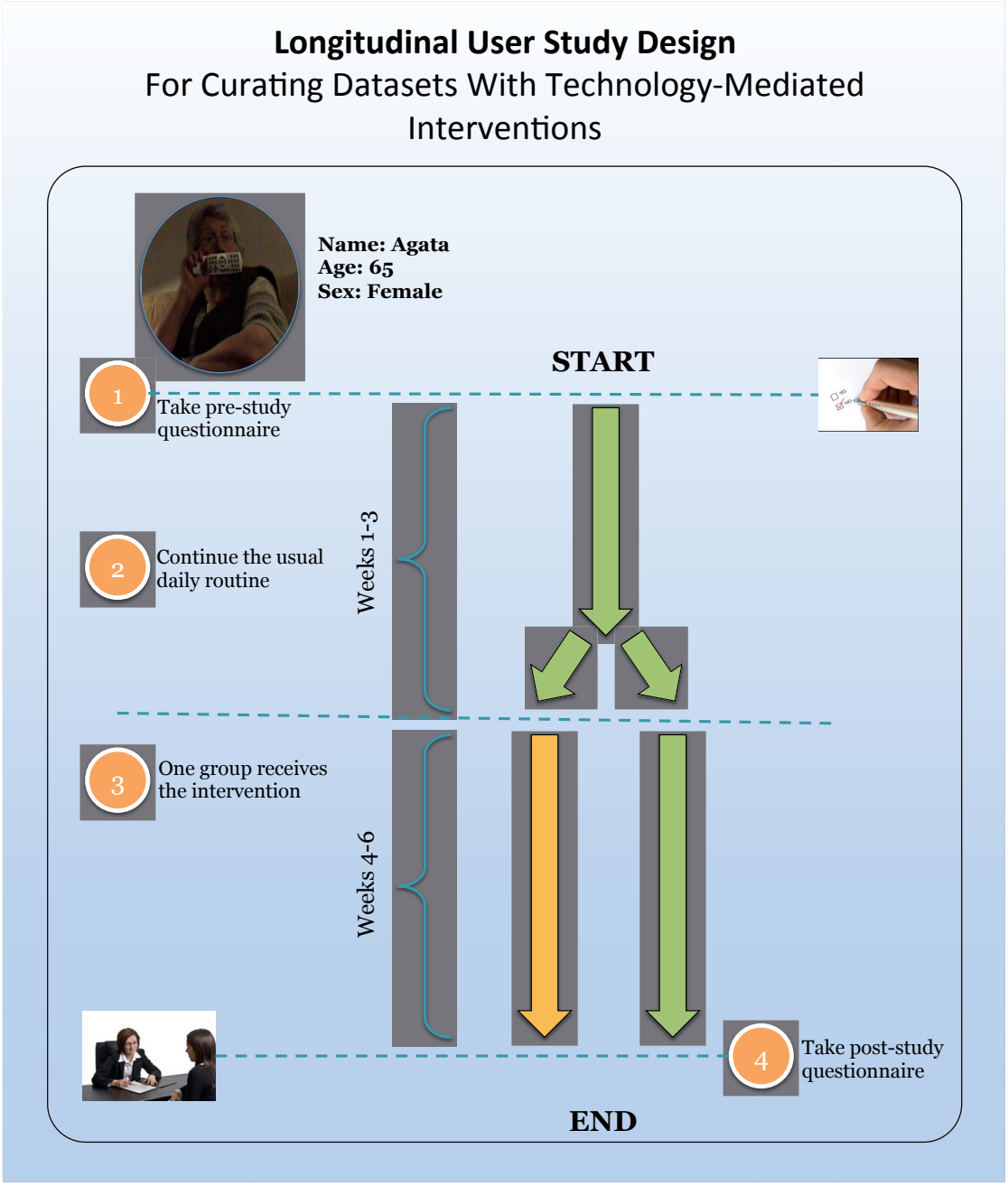

Figure 5.6 – Our proposed outline for future data curation studies

partners). We also summarize a previously collected dataset, which contains snacking logs of 73 along with a message-based information.

Our data curation approach is a special case of Single Case Design, specifically, the concurrent multiple baseline design. Given enough number of participants, our data curation can easily extend into a Randomized Controlled Trial, where we can split the participants to experiment and

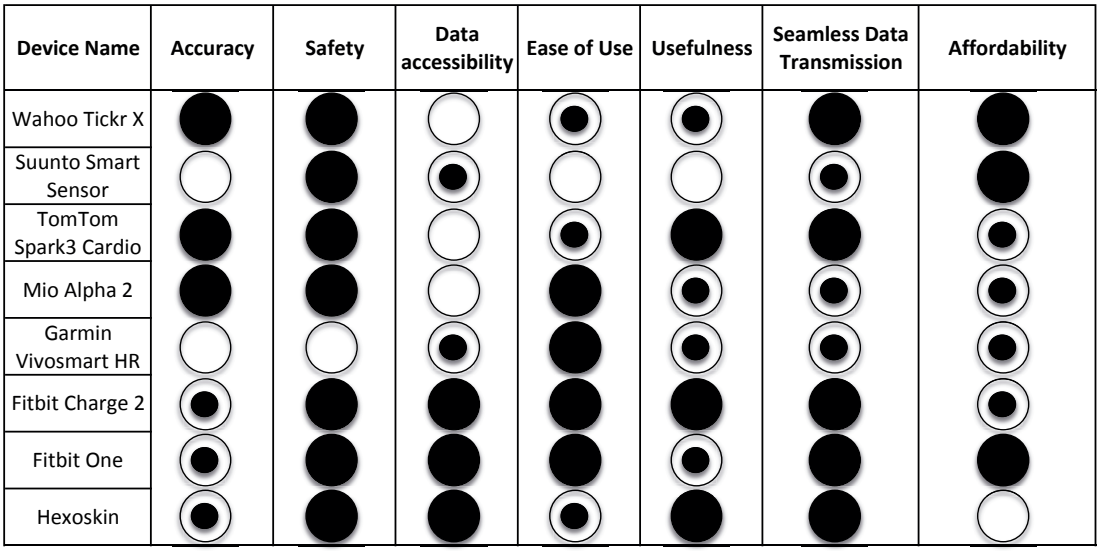

Figure 5.7 – A comparison of off-the-shelf sensors based on 7 criteria. A full black circle indicates that the given sensor fully satisfies the given criterion. A partially black circle indicates a partial fulfilment, whereas a white circle indicates that the sensor does not fulfil the given criterion

control groups. Furthermore, we identified that since many ADLs are collected through sensors, it is critical to ensure that such studies will involve sensors with high reliability and usability. We leveraged these findings and outlined our user study design and sensor comparisons for future studies on data curation.

6 Behavior Profiling and Evaluation for Recommendations

6.1 Challenges

In this chapter, we enlist the technical challenges for our proposed behavior recommender in Section 6.3, and outline our methods to adress them. Figure 6.1 depicts the flow of information between these methods.

- **Data Curation**: The system must use observations on behavior change attempts in addition to measurements of raw physical activities. We deliberately curated a dataset through a user study so that the system could have access to such information. In this user study, we equipped participants with wearable sensors and asked them to form two-people exercise groups. In this manner, we managed to obtain the activity and intervention data to bootstrap our recommendations. In order to preserve the flow of this chapter and elaborate the data curation more in detail, we describe it in Section 5.3.3 of Chapter 5.
- **Temporal Profiling**: The system must process rich, yet noisy and diverse information with temporal data to capture the common behavior patterns. We name the clusters of common behavior patterns as *temporal profiles*, and obtain them with a model-free combination of noise filtering, matrix decomposition and time series clustering. We describe this method in Section 6.4.
- **Intervention Profiling**: The sensor data lacks the manual annotations for proven behavior changes. Thus the system must itself discover the users who have responded to recommendations and improved their activeness. We call the distinct groups of responses as *Intervention Profiles*. The system uses a well known statistical method called Interrupted Time Series Analysis [WSZRD02] to compute the users' *Intervention Profile*, and consequently determines the response type of the Temporal Profile clusters in Section 6.4. We describe this method in Section 6.5. These profiles are closely related with our deliberate curation of the dataset described in Section 5.1.
- **Evaluation**: The recommendations should be both useful and safe for the system's users.

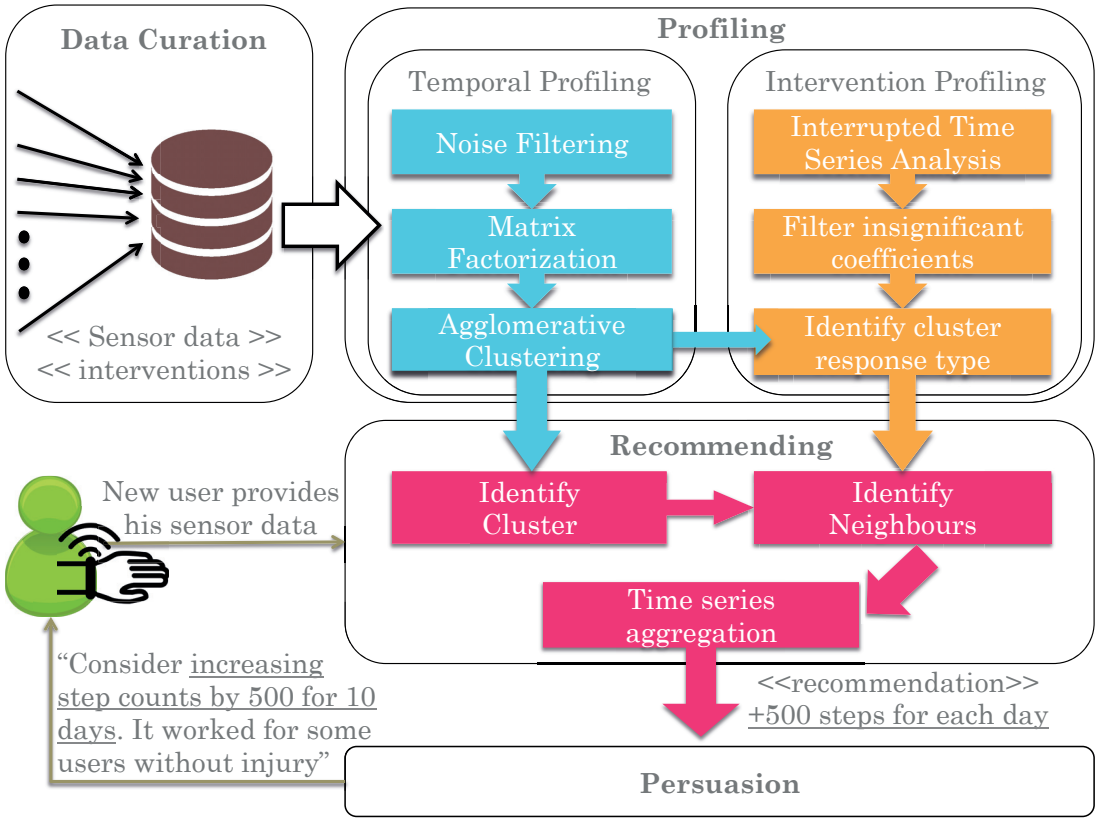

Figure 6.1 – InspiRE applies the processes illustrated in this figure to generate the recommendations

In relation to the users' existing patterns, the recommenders should either help the users improve or maintain their trends for behavior change. Since the research in behavior recommenders is in its early stages, it is yet another challenge to define the methods of evaluation. We propose and report three levels of validation that correspond to the respective methods. We also test the system with varying granularity in data and additional contextual information such as user demographics. We describe our evaluations in Section 6.3.

6.1.1 The Guidelines of Our Solutions

In summary, we generate behavior profiles as tuples $\langle TP, IP \rangle$, where TP represents the temporal profile, and IP represents the intervention profile. The recommendation, TS_{REC}, is a time series object which maximizes the likelihood that the recipient user will have a steady rate of

improvement, rendering him/her as a Responder:

$$TS_{REC} = argmax(p(user = Responder | userprofile = \langle TP, IP \rangle, recommendation = TS)) \tag{6.1}$$

To guide users for a successful behavior change, the system aims to recommend small and incremental patterns based on proven behavior change. As we argue in this thesis, this means that the recommendations should solve the critical trade-off between feasible and effective recommendations. This trade-off will quantitatively depend on a user's existing patterns. On the other hand, any such recommendation should adhere the following guidelines:

1. *Maintenance:* If the user's pre-recommendation pattern resembles to those of Responders, then can the recommendation help the user maintain this pattern?
2. If the user's pre-recommendation pattern resembles to those of Non-Responders or Temporary Responders:
    - *Effectiveness:* If the user may fail because of not performing enough, can the recommendation help the user speed up?
    - *Feasibility:* If the user may fail because of overdoing, can the recommendation help the user improve with a slower but safer rate of change?

We now discuss the related work and our implementations in more detail.

6.2 Related Work

6.2.1 Hidden Markov Models

In Chapter 2, we have provided an abstract overview of behavior profiling and recommender systems. In Section 2.6.2, we noted that very few studies have been performed in behavior recommender systems. Nevertheless we can still consider some state-of-the-art work in music recommendations. The items (i.e., songs) in such systems are similar to ADLs, as they can also be represented as time-series data with continuous values. Our survey shows that music recommender systems handle the temporal properties of songs predominantly with various probabilistic modeling approaches [AKS12, HMB12, PC09, YGK+08]. The most powerful of them, Hidden Markov Models, offer many advantages:

- HMMs are powerful graphical tools to model temporal dynamics. It can process users' ADLs to model micro-patterns and their occurrences as states and transitions. A careful

tuning of these states and transitions provides a good compression of data, i.e., it greatly simplifies the representation of broad patterns. This makes HMMs a candidate method to construct behavior profiles.

- The HMMs can be used to generate possible future ADL trends of users as a collection of micro-patterns, which are represented as states in HMM. One can use the Viterbi algorithm [Vit67] to identify the best sequence of micro-patterns that eventually leads the user to a state of improved ADLs and therefore well-being. Thus, this property of HMM also renders it a method to generate pattern recommendations.

Various prior music recommender studies employ such models to handle the temporal characteristics of songs [HMB12, PC09]. Nevertheless, Markov Modeling variants do have three critical weaknesses that render them unusable in the context of behavior recommendations:

1. As we demonstrate in Chapter 4.6, ADL datasets can rapidly become massive in size. To meet these trends, the system should incorporate scalable methods for data processing. Unfortunately, the Viterbi algorithm [Vit67], a core function in Hidden Markov Models, is prohibitively expensive: For a user's timeseries ADL of length D, the dynamic programming for finding the best path through a model with S states and E edges takes $O(S^D)$ space and $O(E^D)$ time.

2. The HMM needs immense amount of training data. The training involves repeated iterations of the Viterbi algorithm.

3. For a given training dataset, there are many possible HMMs. As a general rule of thumb, smaller models are easier to understand, but larger models can fit the data better. Without the expert knowledge on determining the states and/or prior probabilities for training, the choice of the ideal model remains rather arbitrary.

We can therefore state the advantage of our recommendation and profiling methods over the Markov Models in two perspectives: In the algorithmic perspective, our behavior profiling is more scalable and data-economic than Markov-Model based approaches. In the perspective of human efforts, our approach removes the need of expert knowledge and effort in behavior profiling, effectively allowing the researchers to concentrate on the time and the type of intervention to be delivered as recommendations.

6.2.2 Deep Learning

Deep learning [LBH15] is a family of approaches that have recently gained momentum in machine learning research. The design of such methods originally drew inspirations from neuroscientific studies, which modeled biological neural networks.

Deep learning methods are effective in learning multiple layers of abstraction. This renders them particularly useful in a wide range of computer vision tasks [HZRS16], from edge detection to face detection, object tracking, and annotation generation.

To our best knowledge, deep learning approaches have not yet been applied to behavior profiling and recommendations. This is mainly due to the following unmet prerequisites for deep learning models:

1. Deep learning variants require immense amount of training data. In some cases, including ours, the researchers can not meet this prerequisite.
2. The initalization and momentum is crucial to obtaining an acceptable quality in the network [SMDH13]. Typically the related parameters are manually tuned for each dataset, and this again requires expert knowledge. This is also in line with latest studies which compare sophisticated Deep Learning architectures against baseline methods in recommender systems. To our best knowledge, baseline methods like k-NN can still outperform deep learning architectures [JL17].

Despite the rapid advances in deep learning methods, we can argue that our current recommendation and profiling methods are more advantageous over deep learning methods in terms of a) handling small datasets and b) minimal human effort to determine the initialization of the models. However, near-future resarch on behavior recommenders will find it useful to test new approaches in deep learning.

6.3 Recommendation

As we explained in Chapter 1, a behavior recommender must find the optimal trade-off between safe and effective pattern for its users. With this constraint in mind, we can assess three possible ways to generate pattern suggestions: non-personalized, naive similarity-based, and balanced recommendations. A non-personalized approach (e.g. taking the global average of user patterns) discards a user's innate capabilities in generating suggestions. This clearly violates the well-established principles of Trans-Theoretical Model and Social Cognitive Theory (see Section 2.3). In this perspective, it is straightforward to deduce that a non-personalized approach cannot guarantee that the recommendations will satisfy the trade-off challenge.

A naive, similarity-based approach obtains recommendations as described in Algorithm 1. However, this approach will help the users maintain their existing patterns. For instance, in the case of activity recommendations, inactive users will receive pattern suggestions based on other inactive users. Therefore a similarity-based recommendation will not be able to engage the users according to the the well-established principles of the Flow Concept (see Chapter 2.4).

In this chapter, we propose a third alternative. In this alternative, the system obtains people's

ALGORITHM 1: The SIM Algorithm: Computing recommendations using only similarities

Input: TS_u - User u's data in time series format, $Clusters$ - the list of temporal profiles, TS_i - time series data of each User i in dataset
Output: TS_{REC} - the average trend based on the best neighbours for User u
$neighbor\_candidates$ = heap($maximum\_element\_limit$ = 10);
$closest\_cluster$ = argmin(DTW_LB_KEOGH (TS_u , $Clusters[c].centroid$));
for *each* TS_i *in* $closest\_cluster$ **do**
    $distance = DTW\_LB\_KEOGH(TS_u, TS_i)$;
    $neighbor\_candidates$.insert(item=TS_i, value=$distance$);
end
return temporal_average($neighbor\_candidates$)

behavior profiles, and generates recommendations with the following procedure:

1. Get the user's data thus far, and identify the best matching temporal pattern followed by the system's users. The system obtains these patterns as described in Section 6.4.

2. Find the users whose patterns were similar to user's patterns, but responded positively to recommendations. The system discovers the responses of users (and whether they improved their patterns) beforehand as described in Section 6.5.

3. Identify and output the average trends of the closest successful users (produced by averaging the values on each time point as in Appendix A.7)

This procedure in effect finds TS_{REC}, out of all time series objects (TS) that can be obtained from our dataset such that:

$$TS_{REC} = argmax(p(user = Responder | userprofile = \langle TP, IP \rangle, recommendation = TS)) \tag{6.2}$$

The search space of all possible TS objects is prohibitively large, but we can use the nearest-neighbour approach to approximate the optimal recommendation. Algorithm 2 describes our approach in more detail. We use the 10-nearest neighbors procedure with DTW_LB_KEOGH to obtain the most eligible time series patterns for recommendation (See Appendix A.2). The comparisons are based on pre-recommendation patterns: we assume that the new user has not received a recommendation before. With the temporal profiles obtained in Section 6.4, the algorithm determines the closest cluster to the target user faster than comparing his patterns with every other user. The intervention profiles obtained in Section 6.5 help the algorithm skip the clusters that are populated with non-responders and temporary responders. The resulting

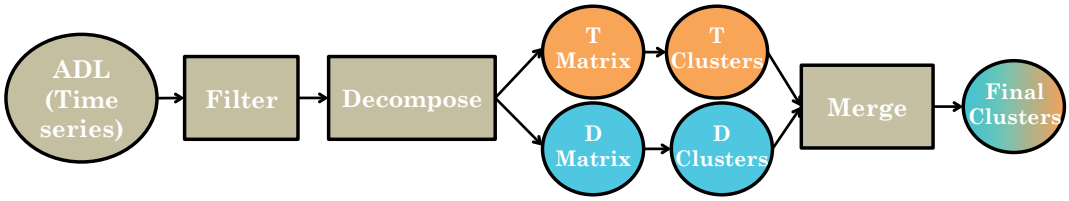

Figure 6.2 – This figure illustrates how InpiRE processes time series data in temporal profiling stage. $T-$ and $D-$ stands for Trends and Deviations respectively. Figure reproduced with permission [YP16] for the sake of clarity.

recommendation is a temporal pattern, whose daily rate of change depends on the new user's pattern and the existing attempts at behavior change.

ALGORITHM 2: Computing Recommendations with InspiRE

Input: TS_u - User u's data in time series format, $Clusters$ - the list of temporal profiles, TS_i - time series data of each User i in dataset
Output: TS_{REC} - the average trend based on the best neighbours for User u

```
neighbor_candidates = heap(maximum_element_limit = 10);
closest_cluster = argmin(DTW_LB_KEOGH (TS_u , Clusters[c].centroid ));
for each cluster C in {Clusters \ closest_cluster} do
    if (C.response_type == Responder ) then
        for each TS_i in C do
            distance = DTW_LB_KEOGH(TS_u, TS_i);
            neighbor_candidates.insert(item=TS_i, value=distance);
        end
    end
end
return temporal_average(neighbor_candidates)
```

6.4 Behavior Profiling: Temporal Profiles

Raw time series data typically consists of rich, yet noisy and diverse information. In order to capture this information, we employ an approach as depicted in Figure 6.2. We have developed this method in a prior work [YZP14], which is described more in detail in Chapter 4. In summary, we take the following steps: First, we apply filtering to remove excessive noise in our time series dataset. Then, we decompose the dataset into two matrices: one matrix contains common trends, and the other matrix contains deviations. As soon as we obtain these two matrices, we perform a clustering operations on them in parallel, and discover the broad patterns of trends and deviations. Finally, we merge the broad patterns from trends and deviations in order to obtain final clusters. The collection of these final clusters, i.e., the set of distinct behavior patterns, represents the *temporal profiles*.

Prior applications [KC12a, MZWM10a, YZP14] validate this technique's noise tolerance, minimal dependency on external labels, and superiority against baseline methods in terms of clustering quality in time series data [YZP14]. With it we obtain trends matrix and deviations matrix. Once this stage is completed, the final clusters are the *temporal profiles*, i.e., the set of distinct behavior patterns. The system can identify a user's temporal profile by comparing the patterns from her time series data with the clusters.

6.5 Behavior Profiling: Intervention Profiles

While sensor data generally lacks annotation, it is still possible to assess external effects, i.e. interventions on the time series data. We take this opportunity to develop *Intervention Profiling* for InspiRE. This method has two major stages: In the first stage, we compute each user's response to a recommendation. This stage represents the responses in two quantities: the immediate and the daily rate of the change a person's level of activeness after the recommendation. In the second stage, we identify the categories of the responses using the quantities we obtained from the first stage. These categories are the *intervention profiles*. We now describe these stages in more detail.

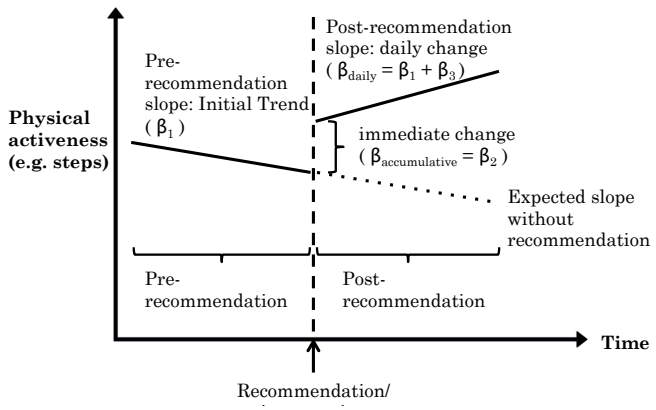

Figure 6.3 – A conceptual chart that depicts how Interrupted Time Series analysis models the intervention and change.

6.5.1 Computing Responses

In this stage, we use pre-recommendation data as baseline, and post-recommendation data to measure the impact of the recommendation on the time series data. We treat recommendations as interventions, and compute the responses through an interrupted time series (ITS) analysis, which is known to be the strongest quasi-experimental approach for evaluating the longitudinal effects

of interventions [WSZRD02]. We solve the following linear regression problem on a user's time series data:

$$TS_t = \beta_0 + \beta_1 * time_t + \beta_2 * intervention_t + \beta_3 * time\_after\_intervention_t \quad (6.3)$$

where TS_t denotes the user's number of steps at day t, where t and $time_t$ ranges from 1 to the length of the sensor usage of the user. $intervention_t$ is an indicator for time t occurring before ($intervention_t$=0) or after ($intervention_t$=1) the time of recommendation. $time\_after\_intervention_t$ is the number of days after the recommendation. It values as 0 before the recommendation and ($time_t$–$time\_of\_recommendation$) after the recommendation.

Next, as we depict in Figure 6.3, we obtain the value $\beta_1 + \beta_3$ as β_{daily}: this value indicates the daily rate of increase in steps after the intervention. At the same time, we obtain the value β_2 as $\beta_{accumulative}$: this value explains the accumulative increase of steps not included in β_{daily}, i.e., short-term surges in the given time series data. It is theoretically possible to model more than one intervention. In such cases, we can update the equation 6.3 with additional terms to calculate the daily rate of increase and the accumulative increase in response to each additional intervention.

6.5.2 Identifying the Response Categories

The interpretation of the coefficients depends on the data. The HT-83 dataset (see Section 5.1) consists of step data, thus positive β-values are more desirable: they indicate increased level of activeness. On the other hand, the SNACK dataset (see Section 5.4) consists of number of unhealthy snacks during the day, thus negative β-values are more desirable: they indicate decreased level of snacking.

Upon inspecting the averages of statistically significant $\beta_{accumulative}$ and β_{daily} values within the clusters we obtained from HT-83 (Section 6.4), we can identify three distinct intervention profiles of users based on their responses to the recommendation:

- **Responders (R)**: Those who adopt a steady increase of activeness after the social recommendation. A cluster is R when β_{daily} is positive and larger than β_1, and $\beta_{accumulative}$ is non-negative.
- **Temporary Responders (TR)**: Those who have an immediate increase when the recommendation takes place, but do not maintain a steady increase. A cluster is TR if $\beta_{accumulative} > 0$, but $\beta_{daily} = 0$.
- **Non-Responders (NR)**: Those who do not have a steady increase, and perhaps even continue losing the level of activeness after the recommendation. A cluster is NR if

$beta_{daily} < 0$, or $beta_{daily}$ is not significantly different than $beta_1$ while $\beta_{accumulative}$ is non-positive.

With this analysis, users are further differentiated in terms of their responses to recommendations in addition to their temporal profiles. This means, for instance, users from two different temporal profiles could be "Responders" (both have significantly increased their level of activeness after the recommendation).

It is possible to measure the impact of any type of intervention with this method. In our primary evaluations, we used the introduction of exercise partners as the intervention - just like receiving a pattern as recommendation, the participants in our data curation study were able to see their exercise partner's pattern and adjust their behavior accordingly.

6.6 Evaluations

In this section, we discuss the validation of InspiRE's data engine:

- To validate *temporal profiling* stage, we show that we obtain distinct and internally consistent profiles. We measure this by inspecting the internal validation indices (Average Silhouette Width) of the clusters we obtained.
- To validate *intervention profiling* stage, we show that intervention analysis helps the system identify people's capability for behavior change, and that recommendations influence people's behavior patterns to a significant extent. We measure these features by analyzing the regression coefficients and coefficients of determination.
- To validate *recommendation stage*, we investigate the disparities between the patterns from a user, his/her activity partner (if applicable), and from recommendations generated for this user. We measure these relations by computing disparity scores, and showed that our recommendation strategy is the best approach to obtain the trade-off between effective and feasible recommendations. (See Section 6.6.2 for details).

We have used the HT-83 Dataset (Section 5.3.3) for our primary validations, and the SNACK dataset (Section 5.4) to show that we can generalize our recommendations beyond physical activity patterns, to snacking patterns.

6.6.1 Validating Profiling Stages

Quality of Temporal Profiles

We validate temporal profiles with Average Silhouette Width [KR09a] - see its details in Appendix A.3. Figure 6.4 summarizes the ASW scores for clusters obtained when the data granularity

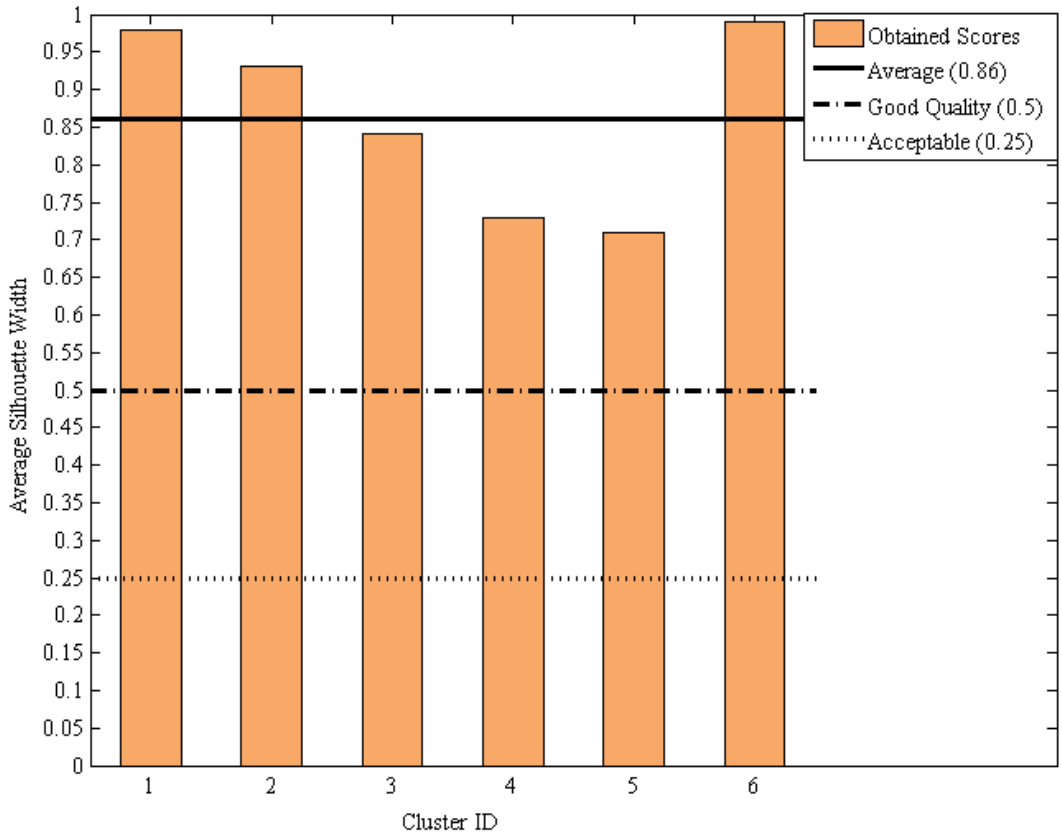

Figure 6.4 – Average Silhouette Widths of each cluster from the Temporal Profiling stage. ASW measures clustering quality, and higher values indicate better clustering quality. 0.5 is the boundary for high quality clustering, and 0.25 is the boundary for acceptable quality. Our method produces high quality temporal profiles.

consists of daily step counts for users. In this study, we obtained 6 clusters. The average silhouette width is 0.86, which indicates a strong level of clustering quality: thus every temporal profile represents a distinct pattern.

Significance of Intervention Profiles

We validate the significance of behavior change attempts with the statistical results concerning the models fitted on activity data with Interrupted Time Series analysis. This includes significance of the regression coefficients, and the R^2 statistic (i.e., coefficient of determination), as well as the differences between pre- and post-intervention slopes (as we elaborated in Section 6.5).

Table 6.1 summarizes the average of regression coefficients from the analysis of users that belong to each cluster from temporal profiles. The coefficients in the fitted models are statistically significant ($p < 0.05$). We observe that Intervention Profiles introduce further details on temporal profiles: Users in each profile have distinct starting trends and unique responses to recommenda-

tions. For instance, users in C1 have a high accumulative increase in their activeness, but fail to maintain a steadily increasing trend. However, users in C6 can do a similar accumulative increase and keep increasing their activeness over time. This further suggests that it is necessary to tailor different recommendations to each profile.

In addition to these evidences, the minimum adjusted coefficient of determination in models is 0.12 (in C1), therefore satisfying $adjusted - R^2 > 0.1$ for all solutions. As such, while there can be many unobserved, contextual factors that could influence the change of a person's activeness, our intervention profiling successfully explains the influence of recommendations to the post-recommendation activeness to a significant extent. This, in effect, removes the need to collect external ground truth information to validate recommendations.

By the end of our intervention profiling procedure, we identify that there are 8 TR, 29 NR, and 46 R users in the HT-83 dataset.

Table 6.1 – The parameters obtained via ITS on HT-83 clusters. β_0 to β_3 are coefficients of fitted linear model, with $\beta_{accumulative} = \beta_2$ and $\beta_{daily} = \beta_1 + \beta_3$ These coefficients are statistically significant in all cases ($p < 0.05$), thus our intervention profiling can detect the potential impacts of recommendations.

Cluster	Initial Trend (β_1)	Accumulative Change ($\beta_{accumulative}$)	Daily Change (β_{daily})	Minimum Adjusted-R^2	Interpretation
C1	-5.5	1044.35	0	0.12	Temporary Responder
C2	-2.43	-497.24	4.292	0.18	Non-Responder
C3	-53.3	527.46	40.56	0.22	Responder
C4	21.99	-822.34	-17.85	0.22	Non-Responder
C5	11.27	521.66	12.7	0.15	Responder
C6	-350.6	1466.5	638.98	0.32	Responder

6.6.2 Validating the Recommendation Stage

The implementation of the guidelines in Section 6.1.1 may vary based on specific cases. In the absence of the partner information, as we demonstrate in Section 6.7, we can examine the differences between patterns of the users and the recommendation. In cases like HT-83 dataset, where the users do have exercise partners, we can further enrich our examinations: we can compare the relations between the patterns of the user, the partner and the recommendation.

In all cases, it is necessary to quantify the differences between the time series patterns. Toward this end, we calculate disparity scores - a variant of mean absolute error. In the case of HT-83, for instance, we can calculate the disparity between a user and a partner or the recommended pattern

as:

$$disparity(user, partner) = \frac{\sum_{t=1}^{T} |d_t^{user} - d_t^{partner}|}{T} \tag{6.4}$$

where d represents the number of steps in a single day, and T is the length of the given user's data in days. Higher values of $disparity(user, partner)$ partners indicate more difference between the patterns of the user and his exercise partner. Consequently, a small difference between $disparity(user, partner)$ and $disparity(user, recommendation)$ indicates a stronger similarity between this user's partner's pattern and the system's recommendation.

Disparities Between Exercise Partners

Figure 6.5 summarizes the distribution of $disparity(user, partner)$ scores in R, NR, and TR users. The average disparity in R, NR and TR users are 931, 3685, and 3822 respectively. The ANOVA test between the three categories (R, TR, and NR) concludes that these differences are statistically significant ($F = 24.5, p < 0.001, df = 2$). It is rather predictable that the level of disparity within TR users is slightly higher than NR users: TR users start off with high levels of activeness but then drop down. NR users never achieve such surges of activeness.

These results suggest that for HT-83 participants, optimal recommendations would emphasize more on *feasibility*of the suggested number of steps per day. Thus we need to perform two tests. First, recommended patterns for NR and TR should be significantly easier for them to follow. In other words, $disparity(user, recommendation)$ should be significantly smaller than $disparity(user, partner)$ for NR and TR users. Second, these two quantities should be statistically similar for R users - this observation guarantees that recommendations help R users to maintain their pace of improvement.

Comparing Recommendations and Actual Exercise Partners

Table 6.2 summarizes the statistical comparisons between users' disparities with their actual partners, and similarity-based and InspiRE's recommendations. We see that both recommendation strategies will minimize the disparity between the recommended pattern and user's current pattern. However, we also observe that:

- A similarity-based recommendation strategy (i.e., finding 10-nearest neighbors for a user to recommend) always recommends patterns with significantly smaller disparity scores than InspiRE's approach (597.2 versus 856.5 on average, t=-3.72, p <0.001). Thus it is also more likely that the similarity-based strategy will recommend inactive people's patterns to users who are already inactive.

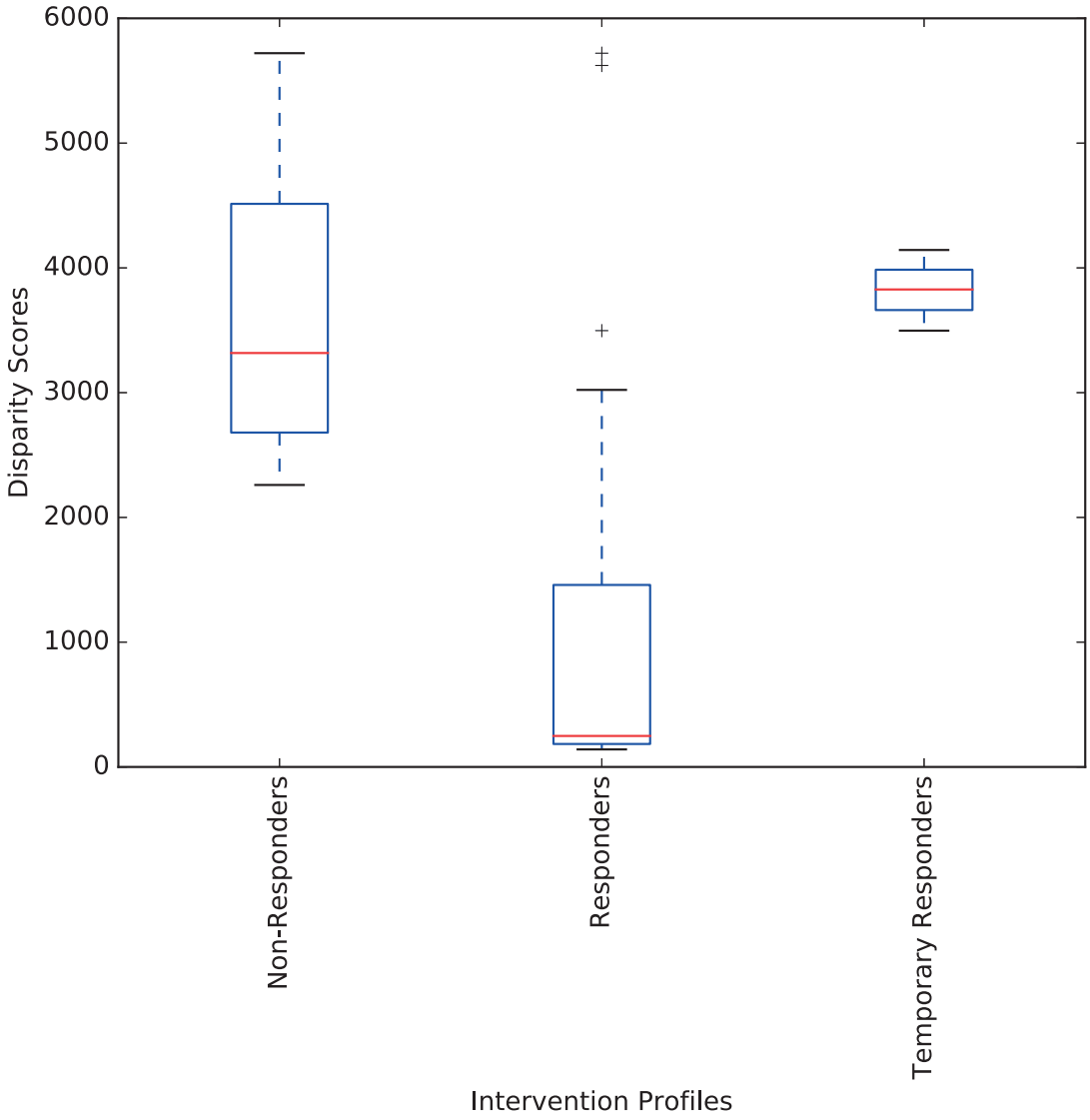

Figure 6.5 – The $disparity(user, partner)$ scores of Responders (R), Non-Responders (NR) and Temporary Responders (TR).

- InspiRE's recommendation strategy produces very similar recommendations for R users ($p = 0.26$), whereas a similarity-based recommendation strategy generates significantly different recommendations for R users ($p = 0.002$).

We conclude that our recommendation strategy overlaps with the strategies of users with proven behavior changes, and improves the chances of future users to have behavior change.

Table 6.2 – Comparing $disparity(user, partner)$ and $disparity(user, recommendation)$ scores. InspiRE's strategy produces more desirable recommendations.

Profile	Actual vs. Similarity-Based	Actual vs. InspiRE	ANOVA Test
Responders	t = 3.08, p = 0.002	t = 1.11, p = 0.26	F = 7.07, p = 0.001
Others	t = 5.87, p <0.001	t = 5.94, p <0.001	F = 30.5, p <0.01

6.6.3 Additional Observations

We further tested the system's capabilities and limitations by tuning various parameters involved in our procedures. We particularly focused on the data length and granularity, as well as using contextual information to guide the system.

Optimal Training Set Size for Predicting the Intervention Profile

Users' activity trends may change over time. This may cause shifts in their intervention profiles. In order to maintain the quality of recommendations from Algorithm 2, the system may need to reassess the profile assigned to its users. It is thus an important task to determine when this reassessment should take place. Toward this end, we measured the predictability of users' intervention profile (being an R, NR, or TR user) given a certain portion of their data.

We used each user's data in relative lengths, i.e., 1/6, 1/5, 1/4, 1/3 and 1/2 to set up the cases for the prediction. For each case, we implemented a 10-nearest neighbor with 10-fold cross validation. We measure the accuracy of predictions by unbiased F1-score (see Appendix A.4 for its calculation).

We summarize the F1 scores paired t-test comparison results in Table 6.3. We see that the prediction quality increases together with the relative length of training data - the more data the system has, the better the predictions. However, this improvement is no longer significant when we increase the relative length from 1/3 to 1/2 ($p = 0.28$). This hints that InspiRE should reassess the intervention profile when the length of pre-recommendation data of a user is less than a third of his entire data. For instance, assuming that InspiRE gives recommendations to a user 1 month after having started using the system, it should reassess this user's patterns no later than 2 months after the recommendation took place.

This result is particularly interesting, as it is supported by other behavior studies. For instance, a weight loss analysis study suggests that 1/3 of a user's weight loss trend is a significant predictor of the rest of his trend [WNW+11]. Furthermore, this analysis also suggests the maximum length of the recommendation pattern for a user should not surpass twice the length of his/her ADL data.

Table 6.3 – F1-scores based on the relative length of the training set. Results indicate that InspiRE should recalculate the profiles after twice the amount of time has passed since the last recommendation.

Relative length of training data	1/6	1/5	1/4	1/3	1/2
F1-score	0.63	0.67	0.7	0.75	0.76
Paired t-test significance	p = 0.01				
		p=0.04			
			p=0.03		
				p=0.28	

Incorporating Contextual Information for Profiling

The quality of recommendations may alter with external/contextual information to guide both of the profiling stages. For instance, as in our dataset, there could be two demographically distinct populations that would use the system. Using contextual information can help the system make better predictions for the intervention profiles, and consequently tailor more personalized recommendations.

There are sophisticated methods to automatically detect relevant contextual information in the domain of movie recommendations [OTTK13]. Such methods are yet to be developed for behavior recommenders. As such, for the sake of scope of this thesis, we have manually determined a context: we separately processed Students and Patients in the HT-83 dataset (see Section 5.3.3), and compared the prediction quality of this style of partitioning against the case when the system processes the mixture of these two populations. When we train our system separately for patients and students, we observe that it predicts the intervention profiles more accurately than when we train the system with mixed data - see Table 6.4.

A further inspection on the temporal profiles validates this finding: We observe that (a) clusters are statistically distinct, and their average steps are different and (b) Patients and Students have statistically different distribution across the clusters ($ANOVA: F = 10.42, p < 0.01$). Thus, students and patients in our datasets follow different daily activity routines, and thus they may require different (and sometimes conflicting) strategies to achieve more active lifestyles.

Table 6.4 – F1-scores we obtain when we partition the dataset based on population. Separating the populations result in better predictions for intervention profiles.

All Users	Students only	Patients only
0.62	0.68	0.82

Extracting Temporal Profiles in Varying Granularities in the Dataset

We had aggregated the sensor data into daily points ın our main experiments. In this setting, the Temporal Profiling procedure processed activity data for every minute for every user. The overall

ASW scores for this setting is 0.67, which is lower than original setting (1 data point for each day) but still above the 0.5 threshold for high quality clustering. The statistical distinctiveness of temporal patterns in these clusters (ANOVA with repeated measurements, $p < 0.01$), further demonstrates the robustness of the temporal profiling stage.

6.7 Generalizing InspiRE: The SNACK Dataset

We use the SNACK dataset (see Section 5.4) to show the generalizability of InspiRE. This dataset contains the daily unhealthy snack logs of 73 people, all Dutch citizens, from various age groups and genders. In the context of nutrition habits, the task of a behavior recommender is to inspire users to cut down the number of unhealthy snacks taken each day.

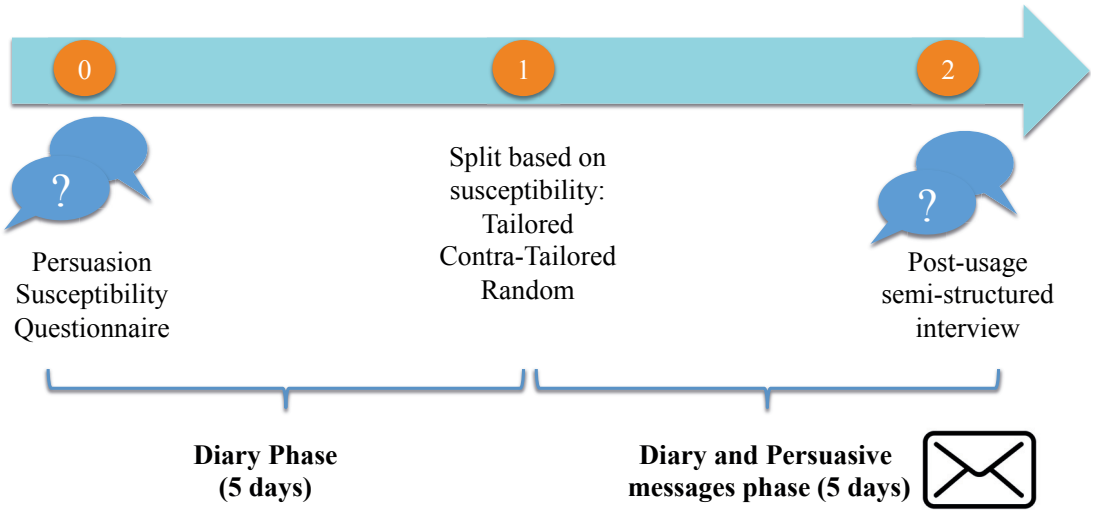

Figure 6.6 – The timeline for SNACK study. $N = 73$ participants joined the study.

Figure 6.7 conveys the behavior profiles in the SNACK datasets. InspiRE obtains 4 clusters for Temporal Profiles, with good clustering quality (ASW > 0.5). Through intervention profiling, InspiRE identifies these clusters as NR, R, R and TR, respectively. By the end of our intervention profiling procedure, we identify that there are 18 TR, 19 NR, and 36 R users in the SNACK dataset.

Table 6.5 summarizes the average of regression coefficients from the analysis of SNACK users that

Figure 6.7 – The Behavior Profiles and recommendations generated by InspiRE on SNACK dataset. The blue line represents the median of the users in each profile, whereas the green line represents the median of the users that InspiRE uses to generate recommendations for each user in the given profile.

belong to each cluster from temporal profiles. The coefficients in the fitted models are statistically significant ($p < 0.05$). In addition to these evidences, the minimum adjusted coefficient of determination in models is 0.22 (in C1), therefore satisfying $adjusted{-}R^2 > 0.1$ for all solutions. For the coefficients of SNACK Intervention profiles, we reverse our interpretations we made in HT-83 Intervention Profiles. In this case, it's the negative values that indicate a person improving his/her well-being (by decreasing the amount of unhealthy snacks)

The way we evaluate the recommendation quality on the SNACK dataset is slightly different than HT-83 dataset, as the SNACK users did not have a partner. We nevertheless follow the guidelines we proposed in Section 6.6.2: We instead compare the recommendations of the naive SIM Algorithm (see Algorithm 1) and InspiRE's Algorthim (see Algorithm 2) in three key tests:

1. We compare whether the two approaches recommend different patterns. We do so by comparing the quantities $disparity(user, SIM)$ and $disparity(user, InspiRE)$
2. We compare Responder users' post-intervention slopes and the slopes of the patterns

Table 6.5 – The parameters obtained via ITS on SNACK clusters. β_0 to β_3 are coefficients of fitted linear model, with $\beta_{accumulative} = \beta_2$ and $\beta_{daily} = \beta_1 + \beta_3$ These coefficients are statistically significant in all cases ($p < 0.05$), thus our intervention profiling can detect the potential impacts of recommendations.

Cluster	Initial Trend (β_1)	Accumulative Change ($\beta_{accumulative}$)	Daily Change (β_{daily})	Minimum Adjusted-R^2	Interpretation
C1	-0.31	-0.05	0.31	0.22	Non-Responder
C2	0.05	-0.34	-0.08	0.41	Responder
C3	-0.08	-0.34	-0.13	0.36	Responder
C4	-0.27	-0.56	0	0.35	Temporary Responder

generated by both algorithms. The latter quantity should be statistically similar to the former. In that case, we can state that the recommendation can help the responder users maintain their rate of improvement.

3. We compare Other users' (NR and TR combined) post-intervention slopes and the slopes of the patterns generated by both algorithms. The difference between these two quantities should be significant, so that we can state that the recommendations can help NR's and TR's to improve their behaviors.

For the first test, a t-test validates that InspiRE's recommended patterns have higher disparities from the users than the patterns generated by similarity-based recommendations (t=-2.70, p=0.007). Table 6.6 summarizes the tests 2 and 3.

In the case of SNACK dataset, we see that both similarity-based recommendations and InspiRE-based recommendations help R profiles maintain their patterns ($p > 0.05$). For NR and TR profiles, InspiRE's recommended patterns could help the users with a significantly higher rate of improvement ($p < 0.05$). On the other hand, similarity-based recommendations cannot deliver rates that are significantly better than users' existing patterns ($p > 0.05$).

In other words, InspiRE's recommendations can help Non-Responders and Temporary responders significantly improve their patterns, as well as help Responders maintain their patterns. A similarity-based recommendation strategy cannot achieve the same effect.

Table 6.6 – Comparing the average of post-intervention (PI) slopes of users, similarity-based recommendation patterns and InspiRE patterns.

	PI	Similarity-Based	InspiRE	ANOVA Test
Responders	-0.67	-1.09 (t = 1.43, p = 0.15)	-0.62 (t = -0.16, p = 0.87)	F = 1.97, p = 0.14
Others	-0.44	-0.77 (t = 1.09, p = 0.27)	**-1.09 (t = 2.25, p = 0.02)**	**F = 3.11, p = 0.04**

As a final step of validation in SNACK dataset, we investigated how well the behavior profiles

explain the real life conditions of the participants of the SNACK study. In this study, the participants were assigned to *Tailored Condition (TC)*, *Contra-Tailored (CTC)*, and *Random Condition (RC)* - see Section 5.4 for more details on these conditions. Their results had indicated that TC participants had the most stable rate of improvement. They were followed by RC participants, whose patterns shown little and unstable improvement. CTC participants were the least to improve, and they even showed signs of relapse [KRMA12].

We have observed a strong overlap between our behavior profiles, and the conditions assigned to study participants:

- The composition of SNACK Cluster $C1$, which contains Non-Responders, is 60% CTC, 20% TC, and 20%RC participants.
- The composition of SNACK Cluster $C2$, which contains Responders, is 60% TC, 30% RC, and 10%CTC participants.
- The composition of SNACK Cluster $C3$, which contains Responders, is 50% TC, 40% RC, and 10%CTC participants.
- The composition of SNACK Cluster $C3$, which contains Responders, is 100% RC participants.

Therefore, we further validated that our profiling methods realistically represent users' temporal patterns and their responses to interventions.

6.8 Chapter Summary

In this chapter, we have elaborated on the core components of InspiRE the behavior recommender. We have justified its recommendation strategy through the behavior change theories that we reviewed in Chapter 2.

We have performed a rigorous validation on the behavior profiling and the recommendation strategies of InspiRE. In our evaluations, we have demonstrated that InspiRE successfully obtains the trade-off between effectiveness and feasibility of recommendations. In more details, our findings show that:

- Our methods obtain temporal profiles that capture distinct behavior patterns among users. Thus we minimize human effort to categorize behavior patterns.
- Our intervention profiling successfully explains the influence of recommendations to the user's post-recommendation patterns to a significant extent. As such, the system can use baseline data instead of manual annotations.

- Our recommendation strategy coincides with the strategies followed by users with successful behaivor change.

We have also shown that similarity-based recommendations cannot guarantee the desirable suggestions to achieve behavior changet: such an approach ultimately recommends the users to maintain their existing pattern, whether they are improving or declining their well-being.

We demonstrated the capabilities of InspiRE through two datasets: while HT-83 dataset is a collection of physical activity habits, the SNACK dataset is a collection of nutrition habits. InspiRE's success on both datasets suggests that our strategy can generalize to multiple dimensions of well-being. We are excited by the future opportunities to test InspiRE on other dimensions of well-being such as sleep and stress.

InspiRE performs a high-quality behavior profiling before generating the recommendations. We composed this behavior profiling as temporal profiles and intervention profiles. The advances of sensor technology suggests that future deployments of InspiRE will perform the behavior profiling exclusively on sensor-based data. As such, the temporal profiling component includes various techniques to handle sensor-based data. We elaborate more on these techniques in Chapter 4.

InspiRE can only exhibit its capabilities if it processes a carefully curated dataset. In Chapter 5, we give the details of this data curation approach, including the key considerations in designing a longitudinal study for data collection.

7 Conclusions

7.1 Summary

An active lifestyle prevents the onset of many diseases associated with obesity, diabetes and heart-related issues. This requires a careful planning and conduct of behavior change. In this thesis, we proposed that a *behavior recommender system* can help its users with effective but feasible behavior recommendations. Design of such a system is an ambitious goal, and it calls for many important research challenges. We proposed a sophisticated set of methods as the analytical building blocks of InspiRE, our recommender system for healthy behavior change. We developed the appropriate approaches to address key challenges towards behavior recommenders, namely; data curation, behavior profiling, and evaluations.

We extended the state-of-the-art methods to meet these challenges, and described how the system generates recommendations as behavior patterns. We represented behavior profiles as tuples, which consist of *temporal* and *intervention* profiles.

1. **Identifying temporal profiles:** For this challenge, we demonstrated in Chapter 4 how to obtain distinct clusters from high-dimensional time series activity data as *Temporal Profiles*, which exploit common trends and deviations in ADL time series data.

2. **Identifying intervention profiles:** To measure the impact of recommendations, we implemented interrupted time series analysis and obtained statistical representations of proven behavior change. Our methods in Chapter 6 help the system identify the baseline (pre-recommendation) data, and detect significant differences in users' behavior patterns before and after recommendations.

3. **Data curation:** In Chapter 5, we proposed our data curation study with real users to obtain all the necessary information for our profiling methods.

We rigorously evaluated InspiRE's method with various datasets:

1. **The Reality Mining Dataset** to draw insights about the relations between well-being and activitiy patterns (see Chapter 3). Here, we discovered that regularity of behavior patterns are strong predictors of increased well-being.

2. **Cylinder-Bell-Funnel** (an artificial dataset), **YQZ** (a dataset of physical activities), and **HT-48** (our initially curated dataset) for proof-of-concept validation of our data processing method (see Chapter 4). We have shown that our methods outperform the baseline methods in terms of clustering quality.

3. **YELP Academic Dataset** (a massive dataset of ratings), **YQZ** (a dataset of physical activities) to evaluate the scalability of our temporal profiling methods (see Chapter 4). We have shown that our method has the same level of scalability with state-of-the-art matrix factorization methods, while outperforming them in physical activity datasets.

4. **HT-83** (a sensor-based physical activity dataset) to evaluate the quality of recommendation and behavior profiling. Here, we have developed and validated our recommendation strategy, which outperforms classical similarity-based approach.

5. **SNACK** (a dataset of snacking habits) to show the generalizability of our methods to other domains, and to investigate how well the behavior profiles explain the real life conditions of the participants of an earlier study. We have observed a strong overlap between our behavior profiles, and the conditions assigned to study participants.Therefore, we further validated that our profiling methods realistically represent users' temporal patterns and their responses to interventions.

We demonstrated that our design is in line with the suggestions from the theories of behavior change:

1. **Validating the temporal profiles:** We used Average Silhouette Width scores to measure to what extent can our methods obtain distinct temporal patterns from the given datasets. We rigorously showed that our method outperforms baseline and state-of-the-art methods for clustering ADL data.

2. **Validating the intervention profiles:** We ensured that our intervention profiles statistically fit our curated dataset and the temporal profiles. Thanks to this validation, InspiRE can predict with the users' most likely responses to various recommendations with high confidence.

3. **Validating the recommendation strategy:** We argued that as a recommendation proposes the most likely strategy to improve the recipient's patterns, it also captures the optimal trade-off between effectiveness and feasibility. For this challenge, we used the disparity scores to examine the differences between users, their exercise partners, and the recommendations. This examination helped us to see that some users needed a faster pace of improvement (more effective than their existing patterns), while the others needed to maintain or slow

down (more feasible than their existing patterns). Our analysis in HealthyTogether and SNACK datasets showed that InspiRE not only achieves these criteria, but also outperforms the classical similarity-based recommendation approach.

7.2 Implications of Our Studies

Our results have several practical and theoretical implications for the related resarch in the future. In a practical perspective, to our best knowledge, InspiRE is the first behavior recommender that follows a recommendation strategy backed by behavior change theories. This results in a timely contribution given the increasing demand on adaptive interventions in preventive medicine [CMB04, NSST+14]. Our system further contributes to the research on behavior recommenders as a first step to use sensor-based data to generate recommendations for safe behavioral changes. Lastly, the analytical methods of InspiRE requires minimal manual input. As such, the findings of this study allows us to divert the knowledge and efforts of experts towards other complicated challenges in wellness management systems, such as finding the right instruments of intervention (social influence, medications, etc.). Further enhancements of InspiRE will lead to a personal wellness management system for injury-free behavior change.

In a theoretical perspective, our study has improved our understanding on the connection between sensor-based data and well-being. It establishes a connection between wearable sensors and theories of behavior change: it is possible to use sensor-based data to identify small and incremental steps for safe behavior change.

7.3 Recommendations for Building Behavior Recommenders

In the preparation of this thesis, we have tackled several challenges under the categories of data curation, behavior profiling, and evaluation. Based on our findings, we have derived the following guidelines for researchers and practitioners in this field:

Data Curation:

- *Consider sensor data as the primary source of information*. We have shown that the wearable sensor technology is becoming more prominent every coming day. This current trend is not likely to reverse. For practical reasons, consider the sensors that collect external measures such as nutrition, exercise, sleep, and stress of their users.

- *Consider adopting Single-Case Designs and Randomized Controlled Trials.* This thesis shows that behavior recommender systems must calculate their users' potential behavioral responses to recommendations. When we delibrerately introduce interventions to the longitudinal study, the curated dataset contains the necessary information to make these calculations. We have shown that it is possible to implement Single-Case Designs,

which helps the system obtain pre-intervention and post-intervention patterns of the participants. The researchers should implement randomized controlled trials given that there is a sufficient number of participants (typically in the order of hundreds). Following these approaches guarantee the validity of the subsequent experiments on recommendations.

Behavior Profiling:

- *Consider the data processing approaches that minimizes manual annotation.* Our investigations showed that it is impractical to annotate the ever-growin sensor datasets. Thus, the behavior profiling should employ unsupervised methods to process low-level information.
- *Consider the statistical properties of behavior patterns.* We have noted that people adopt routines in their activities of daily living, with some possible deviations every day. Behavior profiling component of the system can exploit this insight by separately treating trends and deviations. Such a treatment results in clusters with high internal validity, i.e., they can accurately represent the common behavior patterns.
- *Consider adopting algorithms that can handle time series data from wearable sensors*. We have shown that when the system explicitly handle temporal dynamics, they can further increase the quality of behavior profiles. These systems should also incorporate methods to leverage the abundant but noisy sensor signals.
- *Consider modeling the behavioral responses for the recommendations.* We have shown that methods like interrupted time series analysis can measure the significance and the impact of the recommendations or interventions employed in the data curation. This further enriches the behavior profiles, and allows the system to generate recommendations based on previous users who significantly improved their behavior patterns.

Evaluations:

- *Consider evaluating the quality of behavior profiles.* The quality of recommendations ultimately depend on how well does the system forms the behavior profiles. We have used internal validity indices (such as Silhouette Width) and statistical assessments of the obtained profiles. Such assessments enable the researchers to perform a fair comparison between different recommendation strategies.
- *Consider the trade-off between effective and feasible recommendations.* In this thesis, we investigate the disparities between the patterns from a user, his/her activity partner (if applicable), and from recommendations generated for this user. We showed that our recommendation strategy is the best approach to obtain the trade-off between effective and feasible recommendations. Computing an MAE-like disparity score helps us measure these relations accurately.

7.4 Directions for the Near Future

7.4.1 Profiling the Persuadability of Users

We regard persuasion as an indispensable aspect in delivering behavior recommendations, as persuasive recommendations have higher chances of acceptance and adoption by the users of the system. This requires incorporating persuasion strategies to the system, along with a qualitative study to validate the acceptance of recommendations. Earlier studies [LML+06] explored various strategies to encourage physical activeness, and validated the success of their approach with theories of behavior change such as Trans-Theoretical Model [PV97]. We are further motivated by well-established theories in technology acceptance [Dav86], user engagement [NC02], and principles of persuasion [Cia01]. Prior studies successfully applied these theories in exercising, pervasive systems, Human-Computer Interaction and eCommerce [BTS+13, HP09, HP10, KLS10, KRMA12, JTMS01, TBV12b, WTR94, OVM14]. Lastly, in our analysis of the SNACK dataset (see Section 6.7), we observed that there are some associations between the behavioral responses of users and the style of persuasion they have received. While we could not observe a statistical significance for this association, it further reinforces our motivation to incorporate elements of persuasion to InspiRE.

In our future studies, we will extend our previous behavior profiling that comprised of temporal profiles and intervention profiles. This will result in a new component, the *persuadability profiles*. This extension inspires from the six principles of persuasion stated by Robert Cialdini [Cia01]: *reciprocity*, *scarcity*, *authority*, *commitment*, *consistency*, and *liking*. The research agenda for this extension includes:

1. Designing features for accurate descriptions of persuadability profiles
2. The information required to obtain the features
3. Conducting longitudinal user studies along with interviews and field observations
4. Validating the consistency and relevance of the collected information.

Besides the qualitative aspects, such a study can further enrich our offline validations with online observations. Moreover, it will help us devise approaches that render our behavior profiling methods even more adaptive to novel behavior patterns.

7.4.2 Predictors of Successful Behavior Change

In recent years much research work has been dedicated to detecting user activity patterns from sensor data such as location, movement and proximity. The researchers can exploit the relations between multiple types of behavior patterns and improve the design for healthcare systems and

behavior recommenders. For instance, a system like InspiRE could integrate water intake trends of a user, and use it to enrich its snacking recommendations.

While there exists several studies that employ different techniques to recognize activity patterns [MKK+12, SQM12, ZN12a]; little has been done to understand the relations between them [TBV12a]. Furthermore, how sensor-based daily activities influence people's well-being (such as their satisfaction from work and social lives) is not well explored. Thus, a detailed analytical procedure is required to identify the structural relations between activities and well-being.

In this thesis, we performed a preliminary investigation on the relationship between users' daily activity patterns and their life satisfaction level. Our results show that our analytical procedure can identify meaningful assumptions of causality between activities and satisfaction. Particularly, keeping regularity in daily activities can significantly improve the life satisfaction.

As one future step, we will further investigate the relationship between ADL patterns and indicators of physical well-being.

7.5 Further Directions for Future studies

7.5.1 Intra-Personal Recommendations

A behavior recommender system may as well operate on *intrapersonal retrospective* data, i.e., providing recommendations to a user based only on his/her own history. The initial ideas for such an intrapersonal recommender system is established by Farrell et al. [FDRK12]. However, to our best knowledge, it has not yet been tested on sensor data.

The basic benefit of such a design is a guarantee on personalized recommendations - in case where people's physical capabilities are significantly different from one another, this type of recommender minimizes any risk of injury caused by trying interpersonal behavior patterns. However, it also introduces two fundamental challenges: First, an intrapersonal retrospective recommender will require much more data than interpersonal recommenders. Second, a pure intrapersonal retrospective recommendation can only come from past data. As such, there must be additional mechanisms to guarantee that recommendation has novel elements that help the user improve his activeness.

In this study, we have not explicitly tested intrapersonal retrospective recommendations with InspiRE. However, our system does not require additional modifications to work in that setting. In our future studies, we intend to report a comparison of intrapersonal and interpersonal recommendations, as well as the optimal balance between these two strategies.

7.5.2 Emergency Detection

Our behavior recommender design has a pro-active outlook to managing well-being. In the future, this system can also incorporate solutions to react in cases of emergency (e.g. detecting heart attacks and falls [SZD+15]). As opposed to our analysis on long-term patterns, such cases require an analysis on short-term data.

The research on activity recognition produced many potential methods to address such an opportunity. For instance, well-established methods [LKLC03] can detect unexpected behaviors from time-series data. Our survey also reveals well-studied principles for fall detection [NFR+07], along with sample implementations based on gyroscope and accelerometer sensors [LSH+09b] and vision systems [NCM04]. In our future studies, we will assess the opportunity to merge InspiRE with such methods for a more extensive system for pervasive well-being management.

7.5.3 Detecting Context in Behavior Datasets

In Section 6.6.3, we revealed that contextual information can significantly improve the quality of recommendations. Some subdomains of recommender systems have incorporated methods to recognize useful contextual information [OTTK13]. This is yet to be implemented for behavior recommender systems. In our future studies, we will experiment the extension of such methods on InspiRE.

A Appendix

In this chapter, elaborate on the formulas and algorithms omitted from the main text for the sake of the flow.

Further instructions and potential extensions in each of the components, as well as their implementations will be maintained in Bitbucket [1] and GitHub [2]. The components of the system was implemented in different programming languages and environments: Java, C++, MATLAB, Python, R.

A.1 Low Rank and Sparse Decomposition

Given a matrix M, the decomposition is achieved by solving:

$$min \parallel A \parallel_* + \lambda \parallel E \parallel_1 \text{ such that } M = A + E \tag{A.1}$$

Here, $\parallel . \parallel_*$ denotes the nuclear norm of a matrix, $\parallel . \parallel_1$ is the number of non-zero entries in a matrix, and λ is a positive parameter. The most common method to solve the equation above has a time complexity of $O(N^3)$ [MZWM10a], which is very costly for large datasets. We instead use Robust Grassmann Averages [HFB14]. Its worst-case performance is $O(ND^2)$, where N is the number of rows and D is the number of columns.

[1] https://bitbucket.org/onuryuruten/behaviorrecommender
[2] https://github.com/onuryuruten

ALGORITHM 3: DTW_ LB_ KEOGH

Input: Time series objects Y,Z and window length r
Output: The DTW Distance
$LB\_sum = 0$;
for *each (timestamp, value) in Y* **do**
    $lower\_bound$ = min(Z[(timestamp-r):(timestamp+r)]);
    $upper\_bound$ = max(Z[(timestamp-r):(timestamp+r)]);
    if $value > upper\_bound$ **then**
        $LB\_sum = LB\_sum + (value - upper\_bound)^2$;
    else
        $LB\_sum = LB\_sum + (value - lower\_bound)^2$;
    end
end
return $\sqrt{LB\_SUM}$

A.2 Dynamic Time Warping

Dynamic Time Warping between the time series data Y and Z can be calculated as:

$$DTW(Y,Z) = min_w \left[\sum_{k=1}^{K} d(w_k) \right] \tag{A.2}$$

where $d(w_k) = (y_i - z_j)^2$ such that (y_i, z_j) is on the warping path w. In this naïve form the dynamic time warping costs $O(D^2)$ for comparing a single pair of time series objects given $D = max(length(Y), length(Z))$. Various extensions of this method reduce this complexity [SC78a, KP99a]. For an optimized performance, we use Dynamic Time Warping with Keogh's lower bound[KR05], which runs in linear time. Algorithm 3 describes this modified version of Dynamic Time Warping.

A.3 Average Silhouette Width

The ASW for a given cluster C is calculated as:

$$ASW_C = \frac{\sum_{i \epsilon C} s(i)}{|C|} \text{ such that } s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{A.3}$$

where $a(i)$ is the average dissimilarity of i with all other data within its cluster, and $b(i)$ is the average dissimilarity of i to the closest cluster that i does not belong. The scores are bounded in the interval $[-1, 1]$. Any score below 0.25 would indicate a bad quality of clustering, scores within

$[0.25, 0.5]$ would indicate an acceptable level of quality, and scores above 0.5 would indicate a high quality of clustering [KR09a].

A.4 The Unbiased F1 Score

The unbiased F1 score avoids biased measurements in cross-validation experiments, especially under high class imbalance in the datasets. It is calculated as follows [FS10]:

$$F_{unbiased} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{A.4}$$

In this equation, TP represents the number of "True Positive" classifications (behavioural changes correctly predicted as Responder), FP represents the number of "False Positive" classifications (Non-Responders/Temporary Responders misclassified as Responder), and FN represents the number of "False Negative" classifications (Responders misclassified as Non-Responders/Temporary Responders). F1-scores are bounded within the interval [0,1].

A.5 Normalized Mutual Information

Normalized Mutual Information (NMI) measures the correlation between the true labels and the labels predicted by the system. This measure scales the results between 0 (no mutual information) and 1 (perfect correlation). Normalized Mutual Information is obtained by:

$$NMI = \sqrt{H(labels_{actual}) * H(labels_{predicted})} \tag{A.5}$$

where H stands for the marginal entropy.

A.6 Jaccard Index

Jaccard index measures the similarities between two sets. In our setting, these sets represent actual and predicted labels of the datapoints. Jaccard index is bounded between 0 (total dissimilarity) and 1 (perfect similarity)

$$J(A = labels_{actual}, B = labels_{predicted}) = \frac{|\, A \bigcap B \,|}{|\, A \,| + |\, B \,| - |\, A \bigcap B \,|} \tag{A.6}$$

A.7 Temporal Average

We calculate the temporal average as in Algorithm 4:

ALGORITHM 4: temporal_ average

Input: TS_i - User i's data in time series format ($1 \leq i \leq N$), D - the maximum length of time series in the given subset of data
Output: TS_{avg} - the temporal average
for *l in (1,D)* **do**
    $currentSum = 0$;
    $currentCount = 0$;
    for *i in (1,N)* **do**
        if *(length*$(TS_i) \leq l$ *)* **then**
            $currentCount$++;
            $currentSum = currentSum + TS_i[l]$;
        end
    end
    $TS_{avg}[l] = currentSum \div currentCount$;
end
return TS_{avg}

Bibliography

[ABG+97] Mark Ackerman, Daniel Billsus, Scott Gaffney, Gordon Khoo, Seth Hettich, Dong Joon Kim, and Ray Klefstad. Learning probabilistic user profiles: Applications for finding interesting web sites, notifying users of relevant changes to web pages, and locating grant opportunities. *AI magazine*, 18(2), 1997.

[AEA+08] Raza Ali, Mohamed ElHelw, Louis Atallah, Benny Lo, and Guang-Zhong Yang. Pattern mining for routine behaviour discovery in pervasive healthcare environments. In *Information Technology and Applications in Biomedicine, 2008. ITAB 2008. International Conference on*, pages 241–244. IEEE, 2008.

[AKS12] Natalie Aizenberg, Yehuda Koren, and Oren Somekh. Build your own music recommender by modeling internet radio streams. In *In Proceedings of the 21st international conference on World Wide Web*, pages 1–10. ACM, 2012.

[And03] B.R. Andrews. Habit. *he American Journal of Psychology. University of Illinois Press.*, 14(2):121–149, 1903.

[AT05] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 2005.

[Ban86] A. Bandura. *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall, Inc, 1986.

[BAW00] A Biglan, D Ary, and AC Wagenaar. The value of interrupted time-series experiments for community intervention research. *Prev Sci*, 1(1):31–49, Mar 2000.

[BC94] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.

[BI04] Ling Bao and Stephen S Intille. Activity recognition from user-annotated acceleration data. *Pervasive computing*, pages 1–17, 2004.

[BJR10] Jeremy Birnholtz and McKenzie Jones-Rounds. Independence and interaction: Understanding seniors' privacy and awareness needs for aging in place. In

Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI, pages 143–152. ACM, ACM New York, NY, USA, 2010.

[BTS+13] Frank Bentley, Konrad Tollmar, Peter Stephenson, Laura Levy, Brian Jones, Scott Robertson, Ed Price, Richard Catrambone, and Jeff Wilson. Health mashups: Presenting statistical patterns between wellbeing data and context in natural language to promote behavior change. *ACM Transactions on Computer-Human Interaction*, 20(5):30, 2013.

[Bur02] Robin Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370, 2002.

[BZKB+13] D. Ben-Zeev, S.M. Kaiser, C.J. Brenner, M. Begale, J. Duffecy, and D.C. Mohr. Development and usability testing of focus: A smartphone system for self-management of schizophrenia. *Psychiatric Rehabilitation Journal*, 36(4):289–296, 2013.

[Cao10] Longbing Cao. In-depth behavior understanding and use: the behavior informatics approach. *Information Science*, 180(17):3067—-3085, 2010.

[CCAY13] Saisakul Chernbumroong, Shuang Cang, Anthony Atkins, and Hongnian Yu. Elderly activities recognition and classification for applications in assisted living. *Expert Systems with Applications*, 40(5):1662–1674, 2013.

[CdIAK14] F. Castanedo, D.L. de Ipina, H.K. Aghajan, and R. Kleihorst. Learning routines over long-term sensor data using topic models. *Expert Systems*, 31(4):365–377, 2014.

[CH02] Priscilla M. Clarkson and Monica J. Hubal. Exercise-induced muscle damage in humans. *American journal of physical medicine and rehabilitation*, 81(11):52–69, 2002.

[Cia01] Robert Cialdini. The science of persuasion. *Scientific American*, 284(2):76–81, Feb 2001.

[CK14] Diane J. Cook and Narayanan Krishnan. Mining the home environment. *Journal of intelligent information systems*, 43(3):503–519, 2014.

[CLMW11] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[CMB04] Linda M. Collins, Susan A. Murphy, and Karen L. Bierman. A conceptual framework for adaptive preventive interventions. *Prevention science*, 5:185–196, 2004.

[Coo10a] Diane J Cook. Learning setting-generalized activity models for smart spaces. *IEEE intelligent systems*, 2010(99):1, 2010.

[Coo10b] D.J. Cook. Learning setting-generalized activity models for smart spaces. *IEEE intelligent systems*, 99:1, 2010.

[CP14a] Y. Chen and P. Pu. Healthytogether: exploring social incentives for mobile fitness applications. In *In Proceedings of the Second ACM International Symposium of Chinese CHI*, pages 25–34. ACM, April 2014.

[CP14b] Yu Chen and P Pu. Healthytogether: Exploring social incentives for mobile fitness applications. In *The second International Symposium of Chinese CHI 2014, Toronto, Canada, April 26 - 27*. ACM, 2014.

[CPCP13] Alberto Huertas Celdran, Manuel Gil Perez, Felix J. Garcia Clemente, and Gregorio Martinez Perez. Design of a recommender system based on users' behavior and collaborative location and tracking. *Journal of Computational Science*, 12:83–94, 2013.

[CPV14] P. Chahuara, F Portet, and M Vacher. Making context aware decision from uncertain information in a smart home: A markov logic network approach. *Ambient Intelligence*, 8309:78–93, 2014.

[CSB+81] TC Chalmers, H Smith, 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, May 1981.

[DAC+09] Tamara Denning, Adrienne Andrew, Rohit Chaudhri, Carl Hartung, Jonathan Lester, Gaetano Borriello, and Glen Duncan. Balance: towards a usable pervasive wellness application with accurate activity inference. In *Proceedings of the 10th workshop on Mobile Computing Systems and Applications*, number 5, 2009.

[Dav86] FD Davis. *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. PhD diss. Massachusetts Institute of Technology, 1986.

[DBPV05] T.V. Duong, H.H. Bui, D.Q. Phung, and S. Venkatesh. Activity recognition and abnormality detection with the switching hidden semi-markov model. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1 of *CVPR 2005*, pages 838–845. IEEE, June 2005.

[DCR13] Jesse Dallery, Rachel Cassidy, and Bethany Raiff. Single-case experimental designs to evaluate novel technology-based health interventions. *J Med Internet Res*, 15(2):1–17, 2013.

[DL05] Y. Ding and X. Li. Time weight collaborative filtering. In *In Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 485–492. ACM, October 2005.

[DMCY13] L. Dennison, L. Morrison, G. Conway, and L. Yardley. Opportunities and challenges for smartphone applications in supporting health behavior change: Qualitative study. *Journal of Medical Internet Research*, 15(4), 2013.

[DPF+91] C. C. DiClemente, J. O. Prochaska, S.K. Fairhurst, W.F. Velicer, M. M. Velasquez, and J.S. Rossi. The process of smoking cessation: an analysis of precontemplation, contemplation, and preparation stages of change. *Journal of consulting and clinical psychology*, 59(2):295, 1991.

[DVD+02] Jacques Demongeot, Gilles Virone, Florence Duchene, Gila Benchetrit, Thierry Herve, Norbert Noury, and Vincent Rialle. Multi-sensors acquisition, data fusion, knowledge mining and alarm triggering in health smart homes for elderly people. *Comptes Rendus Biologies*, 325(6):673–682, 2002.

[EP09] Nathan Eagle and Alex Sandy Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.

[EPL09] Nathan Eagle, Alex Pentland, and David Lazer. Inferring social network structure using mobile phone data. In *PNAS*, volume 106, pages 15724–15278, 2009.

[ESBC+08] K. E. Ensrud, J.L. Stock, E. Barrett-Connor, D. Grady, L. Mosca, and K. Khaw. Effects of raloxifene on fracture risk in postmenopausal women: The raloxifene use for the heart trial. *Journal of Bone and Mineral Research*, 23(1):112–120, 2008.

[FDRK12] R.G. Farrell, C. M. Danis, S. Ramakrishnan, and W.A. Kellogg. Intrapersonal retrospective recommendation: lifestyle change recommendations using stable patterns of personal behavior. In *In Proceedings of the First International Workshop on Recommendation Technologies for Lifestyle Change*, LIFESTYLE 2012, page 24, Dublin, Ireland, September 2012.

[FGP14] K. Farrahi and D. Gatica-Perez. A probabilistic approach to mining mobile phone data sequences. *Personal and ubiquitous computing*, 18(1):223–238, 2014.

[FKR+11] C. Free, R. Knight, S. Robertson, R. Whittaker, P. Edwards, W. Zhou, and I. Roberts. Smoking cessation support delivered via mobile phone text messaging (txt2stop): A single - blind, randomised trial. *The Lancet*, 378(9785):49–55, 2011.

[Fog02] B.J. Fogg. *Persuasive technology: using computers to change what we think and do*. Ubiquity, 2002(December), 5, 2002.

[FS10] G. Forman and M. Scholz. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. In *ACM SIGKDD Explorations Newsletter*, volume 12, pages 49–57, Dublin, Ireland, 2010. ACM.

[Fu11] Tak-chung Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.

[FVN10] Anthony Fleury, Michel Vacher, and Norbert Noury. Svm-based multimodal classification of activities of daily living in health smart homes: sensors, algorithms, and first experimental results. In *Transactions on information technology in biomedicine*, volume 14, pages 274–283, March 2010.

[GER15] K.S. Gayathri, Susan Elias, and Balamaran Ravindran. Hierarchical activity recognition for dementia care using markov logic network. *Personal and ubiquitous computing*, 19(2):271–285, 2015.

[GR70] Gene Golub and Christian Reinsch. Singular value decomposition and least squares solutions. *Numerische mathematik*, 5, 1970.

[Gre94] J.G. Greeno. *Gibson's Affordances*. 1994.

[HCH05] RH Horner, EG Carr, and J. Halle. The user of single-subject research to identify evidence-based practice in special education. *Except Child*, 71(2):165–179, 2005.

[HFB14] Soren Hauberg, Aasa Feragen, and Michael J. Black. Grassmann averages for scalable robust pca. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014.

[HMB12] Negar Hariri, Bamshad Mobasher, and Robin Burke. Context-aware music recommendation based on latent topic sequential patterns. In *In Proceedings of the sixth ACM conference on Recommender systems*, RECSYS, pages 131–138. ACM, 2012.

[HP97] Robert J Hodrick and Edward C Prescott. Postwar us business cycles: an empirical investigation. *Journal of Money, credit, and Banking*, pages 1–16, 1997.

[HP09] Rong Hu and Pearl Pu. Potential acceptance issues of personality-based recommender systems. In *In proceedings of the 3rd ACM Conference on Recommender Systems*, RECSYS, pages 221–224, New-York City, NY, USA, October 2009. ACM.

[HP10] Rong Hu and Pearl Pu. A study on user perception of personality-based recommender systems. In *UMAP*, UMAP, pages 221–224, Hawaii, USA, June 2010. Lecture Notes in Computer Science.

[HZRS16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. IEEE, 2016.

[JE02] Susan A. Jackson and Robert C. Eklund. Assessing flow in physical activity: The flow state scale–2 and dispositional flow scale–2. *Journal of Sport and Exercise Psychology*, 24(2):133–150, 2002.

[JL17] Dietmar Jannach and Malte Ludewig. When recurrent neural networks meet the neighborhood for session-based recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, RecSys '17, pages 306–310, New York, NY, USA, 2017. ACM.

[JTMS01] S.A. Jackson, P.R. Thomas, H.W. Marsh, and C. J. Smethurst. Relationships between flow, self-concept, psychological skills, and performance. *Journal of applied sport psychology*, 13(2):129–153, 2001.

[Kaz82] Alan Kazdin. *Single-Case Research Designs*, volume 1. New York: Oxford University Press. ISBN 0-19-503021-4, 1982.

[KBFK06] S. Krishna, E.A. Balas, B.D. Francisco, and P. Konig. Effective and sustainable multimedia education for children with asthma: A randomized controlled trial. *Childrens Health Care*, 35(1):75–90, 2006.

[KC12a] A. Kyrillidis and V. Cevher. Matrix alps: Accelerated low rank and sparse matrix reconstruction. In *Statistical Signal Processing Workshop*, SSP, pages 185–188. IEEE, August 2012.

[KC12b] Anastasios Kyrillidis and Volkan Cevher. Matrix alps: Accelerated low rank and sparse matrix reconstruction. In *Statistical Signal Processing Workshop (SSP), 2012 IEEE*, pages 185–188. IEEE, 2012.

[KCS08] Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 453–456. ACM, 2008.

[KEK10] T.L.M. Van Kasteren, G. Englebienne, and B.J.A Krose. Activity recognition using semi-markov models on real world smart home datasets. *Ambient Intell. Smart Environ.*, 2:311–325, 2010.

[KH87] D.E. Kanouse and L.R. Jr Hanson. Negativity in evaluations. *Preparation of this paper grew out of a workshop on attribution theory held at University of California. Lawrence Erlbaum Associates, Inc.*, 1987.

[KK03] Eamonn Keogh and Shruti Kasetty. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 7(4):349–371, 2003.

[KL05] Eamonn Keogh and Jessica Lin. Clustering of time-series subsequences is meaningless: implications for previous and future research. *Knowledge and information systems*, 8(2):154–177, 2005.

[KLS10] Maurits Kaptein, Joyca Lacroix, and Privender Saini. Individual differences in persuadability in the health promotion domain. In *International Conference on Persuasive Technology*, pages 94–105, 2010.

[KN12] V. Kaptelinin and B. Nardi. Affordances in hci: toward a mediated action perspective. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 967–976. ACM, May 2012.

[Kor08] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD, pages 426–/434, Las Vegas, Nevada, USA, August 2008. ACM.

[Kor10] Yehuda Koren. Collaborative filtering with temporal dynamics. *Communications of the ACM*, 53(4):89–97, 2010.

[KP99a] Eamonn Keogh and M.J. Pazzani. Scaling up dynamic time warping to massive datasets. *Principles of Data Mining and Knowledge Discovery*, pages 1–11, 1999.

[KP99b] Eamonn J Keogh and Michael J Pazzani. Scaling up dynamic time warping to massive datasets. In *Principles of Data Mining and Knowledge Discovery*, pages 1–11. Springer, 1999.

[KP12] A.R. Kuppam and R.M. Pendyala. A structural equations analysis of commuters' activity and travel patterns. *Journal of Transportation*, 28(1):33–54, 2012.

[KPG03] Ilkka Korhonen, Juha Pärkkä, and Mark Van Gils. Health monitoring in the home of the future. *IEEE Engineering in Medicine and Biology Magazine*, 22(3):66–73, 2003.

[KR05] Eamonn Keogh and Chotirat Ann Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and information systems*, 7(3):358–386, 2005.

[KR09a] L. Kaufman and P.J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. 2009.

[KR09b] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. Wiley. com, 2009.

[KRMA12] Maurits Kaptein, Boris De Ruyter, Panos Markopoulos, and Emile Aarts. Adaptive persuasive systems: a study of tailored persuasive text messages to reduce snacking. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2:10, 2012.

[KWM+97] M.L. Klem, R. R. Wing, T.M. McGuire, H.M. Seagle, and J.O. Hill. A descriptive study of individuals successful at long-term maintenance of substantial weight loss. *The American journal of clinical nutrition*, 66(2):239–246, 1997.

[LBH15] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[LCM10] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[LDF11] Ian Li, Anind K. Dey, and Jodi Forlizzi. Understanding my data, myself: Supporting self-reflection with ubicomp technologies. In *Proc. 13th international conference on Ubiquitous computing*, pages 405–414, 2011.

[LJPW10] P Lally, CH Van Jaarsveld, HW Potts, and J. Wardle. How are habits formed: Modelling habit formation in the real world. *European journal of social psychology*, 40(6):998–1009, 2010.

[LKLC03] J. Lin, E. Keogh, S. Lonardi, and B. Chui. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11. ACM, June 2003.

[LML+06] J. J. Lin, L. Mamykina, S. Lindtner, G. Delajoux, and H.B. Strub. Fish'n'steps: Encouraging physical activity with an interactive computer game. In *Ubiquitous Computing*, UBICOMP, pages 261–278, Berlin Heidelberg, 2006. Springer.

[LRZM12] Xiao Liang, Xiang Ren, Zhengdong Zhang, and Yi Ma. Repairing sparse low-rank texture. In *Computer Vision–ECCV 2012*, pages 482–495. Springer, 2012.

[LSH+09a] Qiang Li, John Stankovic, Mark Hanson, Adam Barth, John Lach, and Gang Zhou. Accurate, fast fall detection using gyroscopes and accelerometer-derived posture information. In *Sixth International Workshop on Wearable and Implantable Body Sensor Networks*, pages 138–143. IEEE, June 2009.

[LSH+09b] Qiang Li, John Stankovic, Mark Hanson, Adam Barth, John Lach, and Gang Zhou. Accurate, fast fall detection using gyroscopes and accelerometer-derived posture information. In *Sixth International Workshop on Wearable and Implantable Body Sensor Networks*, pages 138–143. IEEE, June 2009.

[LSY03] Greg Linden, Brent Smith, and Jeremy York. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1), 2003.

[MB65] F. Mahoney and D. Barthel. Functional evaluation: The barthel index. *Maryland State Medical Journal*, 14:56–61, 1965.

[MCLP10] Anmol Madan, Manuel Cebrian, David Lazer, and Alex Pentland. Social sensing for epidemiological behavior change. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 291–300. ACM, 2010.

[MJ99] H.W. Marsh and S.A. Jackson. Flow experience in sport: Construct validation of multidimensional, hierarchical state and trait responses. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(4):343–371, 1999.

[MK14] Elizabeth A. Minton and Lynn R. Khale. *Belief Systems, Religion, and Behavioral Economics*. New York: Business Expert Press LLC, 2014.

[MKK+12] Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. Affectaura: An intelligent system for emotional memory. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 849–858, 2012.

[MLO+16] Siobhan K. McMahon, Beth Lewis, Michael Oakes, Weihua Guan, Jean F. Wyman, and Alexander J. Rothman. Older adults' experiences using a commercially available monitor to self-track their physical activity. *JMIR mHealth and uHealth*, 4(2):2, 2016.

[Moo96] Todd Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996.

[MT00] M.S. McGlone and J. Tofighbakhsh. Birds of a feather flock conjointly (?): rhyme as reason in aphorisms. *Psychological Science*, 11(5):424–428, 2000.

[MVR+14] Roisin McNaney, John Vines, Daniel Roggen, Madeline Balaam, Pengfei Zhang, Ivan Poliakov, and Patrick Olivier. Exploring the acceptability of google glass as an everyday assistive device for people with parkinson's. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2551–2554. ACM, 2014.

[MZWM10a] K. Min, Z. Zhang, J. Wright, and Y. Ma. Decomposing background topics from keywords by principal component pursuit. *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 269–278, October 2010.

[MZWM10b] Kerui Min, Zhengdong Zhang, John Wright, and Yi Ma. Decomposing background topics from keywords by principal component pursuit. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 269–278. ACM, 2010.

[NC02] J. Nakamura and M. Csikszenmihalyi. *Handbook of positive psychology*, pages 89–105. 2002.

[NCM04] Hammadi Nait-Charif and Stephen McKenna. Activity summarisation and fall detection in a supportive home environment. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 4 of *ICPR*, pages 323–326. IEEE, August 2004.

[NFR+07] Norbert Noury, Anthony Fleury, Pierre Rumeau, A.K. Bourke, G.O. Laighin, Vincent Rialle, and J.E. Lundy. Fall detection-principles and methods. In *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1663–1666. IEEE, August 2007.

[NH12] Norbert Noury and Tareq Hadidi. Computer simulation of the activity of the elderly person living independently in a health smart home. *Computer methods and programs in biomedicine*, 108(3):1216–1228, 2012.

[NHdlC15] Qin Ni, Ana Belen Garcia Hernando, and Ivan Pau de la Cruz. The elderly's independent living in smart homes: A characterization of activities and sensing infrastructure survey to facilitate services development. *Sensors*, 15(5):11312–11362, 2015.

[NSST+14] S. Nahum-Shani, S.N. Smith, A. Tewari, K. Witkiewitz, L.M. Collins, B. Spring, and S.A. Murphy. Just-in-time adaptive interventions (jitais): An organizing framework for ongoing health behavior support. *(Technical Report No. 14-126). University Park, PA: The Methodology Center, Penn State.*, 2014.

[OSH+10] N Owen, PB Sparling, GN Healy, DW Dunstan, and CE Matthews. Sedentary behavior: Emerging evidence for a new health risk. In *Mayo Clinic Proceedings.*, pages 1138–1141, 2010.

[OSW+12] L.G. Ogden, N. Stroebele, H.R. Wyatt, V.A. Catenacci, J.C. Peters, J. Stuht, and J.O. Hill. Cluster analysis of the national weight control registry to identify distinct subgroups maintaining successful weight loss. *Obesity*, 20(10):2039–2047, 2012.

[OTTK13] Ante Odic, Marko Tkalcic, Jurij F. Tasic, and Andrej Kosir. Predicting and detecting the relevant contextual information in a movie-recommender system. *Interacting with Computers*, 25(1):74–90, 2013.

[OVM14] Rita Orji, Julita Vassileva, and Regan L. Mandryk. Modeling the efficacy of persuasive strategies for different gamer types in serious games for health. *User Modeling and User–Adapted Interaction*, 24(5):453–498, 2014.

[PB92] BS Parsons and DM Baer. The visual analysis of data and current research into the stimuli controlling it. *Single-case research design analysis: New directions for psychology and education*, 1(1):15–40, 1992.

[PC09] John Paisley and Lawrence Carin. Hidden markov models with stick-breaking priors. In *Transactions on Signal Processing*, volume 57, pages 3905–3917. IEEE, 2009.

[Pea01] K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559—-572, 1901.

[Pea00] Judea Pearl. *Causality: Models, Reasoning, and Inference*, volume 29. Cambridge: MIT Press. ISBN 0-521-77362-8, 2000.

[PGW+10] Yigang Peng, Arvind Ganesh, John Wright, Wenli Xu, and Yi Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images.

In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 763–770. IEEE, 2010.

[PHL12] Dhaval Patel, Wynne Hsu, and Mong Li Lee. Integrating frequent pattern mining from multiple data domains for classification. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pages 1001–1012. IEEE, 2012.

[PPL01] A. Popescul, D.M. Pennock, and S. Lawrence. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 437–444. Morgan Kaufmann Publishers Inc., August 2001.

[PPO10] Cuong Pham, Thomas Plötz, and Patrick Olivier. A dynamic time warping approach to real-time activity recognition for food preparation. In *Ambient Intelligence*, pages 21–30. Springer, 2010.

[PRA+09] K. Patrick, F. Raab, M.A. Adams, L. Dillon, M. Zabinski, C.L. Rock, and G.J. Norman. A text message–basedintervention for weight loss: Randomized controlled trial. *Journal of Medical Internet Research*, 11(1), 2009.

[PV97] J.O. Prochaska and W.F. Velicer. The transtheoretical model of health behavior change. american journal of health promotion. *American journal of health promotion*, 12(1):38–48, 1997.

[RCHSE11a] P. Rashidi, D.J. Cook, L.B. Holder, and M. Schmitter-Edgecombe. Discovering activities to recognize and track in a smart environment. *IEEE Transactions on Knowledge and Data Engineering*, 23(4):527–539, 2011.

[RCHSE11b] Parisa Rashidi, Diane J Cook, Lawrence B Holder, and Maureen Schmitter-Edgecombe. Discovering activities to recognize and track in a smart environment. *Knowledge and Data Engineering, IEEE Transactions on*, 23(4):527–539, 2011.

[RM03] Toni M Rath and Raghavan Manmatha. Word image matching using dynamic time warping. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–521. IEEE, 2003.

[ROJM08] W. T. Riley, J. Obermayer, and J. Jean-Mary. Internet and mobile phone text messaging intervention for college smokers. *Journal of American College Health*, 57(2):245–248, 2008.

[ROK+12] N.R. Reyes, T.L. Oliver, A.A. Klotz, C.A. LaGrotte, S.S. Vander Veur, A. Virus, and G.D. Foster. Similarities and differences between weight loss maintainers and regainers: a qualitative analysis. *Journal of the Academy of Nutrition and Dietetics*, 112(4):499–505, 2012.

[RV97] Paul Resnick and Hal Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

[SBH02] G. Shani, R.I. Brafman, and D. Heckerman. An mdp-based recommender system. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 453–460. Morgan Kaufmann Publishers Inc., August 2002.

[SC78a] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49, 1978.

[SC78b] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43–49, 1978.

[Sch02] S.R. Schmidt. The humour effect: Differential processing and privileged retrieval. *Memory*, 10(2):127–138, 2002.

[SD06] Geir Kjetil Sandve and Finn Drablos. A survey of motif discovery methods in an integrated framework. *Biol Direct*, 1(11), 2006.

[SHY04] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of the 13th international conference on World Wide Web*, pages 675–684. ACM, May 2004.

[Sin16] Valentina Sintsova. *Advancing Fine-Grained Emotion Recognition in Short Text: A PhD Dissertation*. ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE, 2016.

[SJC15] A. Sano, P. Johns, and M. Czerwinski. Healthaware: An advice system for stress, sleep, diet and exercise. In *International Conference on Affective Computing and Intelligent Interaction*, ACII, pages 546–552. IEEE, September 2015.

[SJS05] Steven L. Scott, Gareth M. James, and Catherine A. Sugar. Hidden markov models for longitudinal comparisons. *Journal of the American Statistical Association*, 100(470), 2005.

[SKH11] Alessandra Maria Sabelli, Takayuki Kanda, and Norihiro Hagita. A conversational robot in an elderly care center: an ethnographic study. In *Proceedings of the 6th international conference on Human-robot interaction*, HRI, pages 37–44, New York, NY, USA, 2011. ACM.

[Sma12a] Larry Smarr. Quantifying your body: A how–to guide from a systems biology perspective. *Biotechnology Journal*, 7(8):980–991, 2012.

[Sma12b] Larry Smarr. Quantifying your body: A howto guide from a systems biology perspective. *Biotechnology Journal*, 7(8):980–991, 2012.

[SMCUU07] E Sahin, M.R. Dogar M. Cakmak, E. Ugur, and G. Ucoluk. To afford or not to afford: A new formalization of affordances toward affordance-based robot control. *Adaptive Behavior*, 15(4):447–472, 2007.

[SMDH13] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, volume 28, pages 1139–1147, Feb 2013.

[Smi12] Justin Smith. Single-case experimental designs: A systematic review of published research and current standards. *Psychol Methods*, 17(4), 2012.

[SPUP02] AI Schein, A Popescul, LH Ungar, and DM Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM conference on Research and development in information retrieval*, SIGIR, pages 253–260. ACM, August 2002.

[SQM12] N. K. Suryadevara, T. Quazi, and Subhas C. Mukhopadhyay. Smart sensing system for human emotion and behaviour recognition. *Perception and Machine Intelligence*, 7143:11–22, 2012.

[SZD+15] G. Shi, J. Zhang, C. Dong, P. Han, Y. Jin, and J. Wang. Fall detection system based on inertial mems sensors: Analysis design and realization. In *International Conference on Cyber Technology in Automation, Control, and Intelligent Systems*, CYBER, pages 1834–1839, Shenyang, 2015. IEEE.

[TBV12a] Konrad Tollmar, Frank Bentley, and Cristobal Viedma. Mobile health mashups: Making sense of multiple streams of wellbeing and contextual data for presentation on a mobile device. In *6th International Conference on Pervasive Computing Technologies for Healthcare*, pages 65–72, 2012.

[TBV12b] Konrad Tollmar, Frank Bentley, and Cristobal Viedma. Mobile health mashups: Making sense of multiple streams of wellbeing and contextual data for presentation on a mobile device. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2012 6th International Conference on*, pages 65–72. IEEE, 2012.

[TLR+07] Christopher C. Tsai, Gunny Lee, Fred Raab, Gregory J. Norman, Timothy Sohn, William G. Griswold, and Kevin Patrick. Usability and feasibility of pmeb: a mobile phone application for monitoring real time caloric balance. *Mobile Networks and Applications archive*, 12:173–184, 2007.

[VAS09] Alireza Vahdatpour, Navid Amini, and Majid Sarrafzadeh. Toward unsupervised activity discovery using multi-dimensional motif detection in time series. In *IJCAI*, volume 9, pages 1261–1266, 2009.

[Vit67] Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967.

[VMO+12] K. Verbert, N. Manouselis, X. Ochoa, M. Wolpers, H. Drachsler, I. Bosnic, and E. Duval. Context-aware recommender systems for learning: a survey and future challenges. *IEEE Transactions on Learning Technologies*, 5(4):318–335, 2012.

[WA05] Daniel H. Wilson and Chris Atkeson. Simultaneous tracking and activity recognition (star) using many anonymous, binary sensors. In *International Conference on Pervasive Computing*, pages 62–79. Springer, 2005.

[WNW+11] T.A. Wadden, R.H. Neiberg, R.R. Wing, J.M. Clark, L.M. Delahanty, J.O. Hill, and M.Z. Vitolins. Four-year weight losses in the look ahead study: Factors associated with long-term success. *Obesity*, 19(10):1987–1998, 2011.

[WOCRM08] Virginia G Wadley, Ozioma Okonkwo, Michael Crowe, and Lesley A Ross-Meadows. Mild cognitive impairment and everyday function: evidence of reduced speed in performing instrumental activities of daily living. *American Journal of Geriatric Psych*, 16(5):416–424, 2008.

[WSZRD02] A.K. Wagner, S.B. Soumerai, F. Zhang, and D. Ross-Degnan. Segmented regression analysis of interrupted time series studies in medication use research. *Journal of clinical pharmacy and therapeutics*, 27(4):299–309, 2002.

[WTR94] J. Webster, L. K. Trevino, and L. Ryan. The dimensionality and correlates of flow in human-computer interactions. *Computers in human behavior*, 9(4):411–426, 1994.

[WY13] Naiyan Wang and Dit-Yan Yeung. Bayesian robust matrix factorization for image and video processing. In *Computer Vision (ICCV)*. IEEE, 2013.

[YGK+08] Kazuyoshi Yoshii, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model. In *Transactions on Audio, Speech, and Language Processing*, volume 16, pages 435–447. IEEE, 2008.

[YMH+09] T. Yan, M. Marzilli, R. Holmes, D. Ganesan, and M. Corner. Bayesian robust matrix factorization for image and video processing. In *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, pages 347–348. ACM, November 2009.

[YP16] Onur Yuruten and Pearl Pu. Factoring the habits: Comparing methods for discovering behavior patterns form large scale activity datasets. In *International Conference on Big Data Analytics, Data Mining and Computational Intelligence (BIGDACI). Part of Multi Conference on Computer Science and Information Systems (MCCSIS)*. IADIS, July 2016.

[YZK15] Jie Yin, Qing Zhang, and Mohan Karunanithi. Unsupervised daily routine and activity discovery in smart homes. In *37th Annual International Conference in*

Engineering in Medicine and Biology Society (EMBC), pages 5497–5500. IEEE, 2015.

[YZP14] Onur Yuruten, Jiyong Zhang, and Pearl Pu. Decomposing activities of daily living to discover routine clusters. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*. AAAI, June 2014.

[ZLN13] J. Zheng, S. Liu, and L.M. Ni. Effective routine behavior pattern discovery from sparse mobile phone data via collaborative filtering. In *Pervasive Computing and Communications*, PerCom. IEEE, March 2013.

[ZN12a] Jiangchuan Zheng and Lionel M. Ni. An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 153–162. ACM New York, 2012.

[ZN12b] Jiangchuan Zheng and Lionel M Ni. An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 153–162. ACM, 2012.

[ZPSB04] K. Zhang, F. Pi-Sunyer, and C. Boozer. Improving energy expenditure estimation for physical activity. *Medicine and Science in Sports and Exercise*, 36(5):883–889, 2004.

[ZT11] Tianyi Zhou and Dacheng Tao. Godec: Randomized low-rank & sparse matrix decomposition in noisy case. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 33–40, 2011.

[ZWGT13] Yonglei Zheng, Weng-Keen Wong, Xinze Guan, and Stewart Trost. Physical activity recognition from accelerometer data using a multi-scale ensemble method. In *Twenty-Fifth IAAI Conference*, 2013.

PERSONAL

Surname / First name	**YÜRÜTEN, ONUR**
Address(es)	**RESIDENCE: 6B, CHEMIN DE L'OCHETTAZ, 1025, ST. SULPICE, SWITZERLAND**
E-mail	**onuryuruten@gmail.com**
Phone number	**+41 78 679 70 31**
Skype id	**onuryuruten**
LinkedIn	**https://ch.linkedin.com/in/onuryuruten**
Current Website	**http://hci.epfl.ch/members/onur-yuruten/**
Nationality	Turkish (with Residence Permit B in Switzerland)
Date of birth	MAY 3,1989

WORKING EXPERIENCE

Period	September 2012 – present day.
Position	**Doctoral Assistant** at EDIC Doctoral Student school in Computer Communication and Information Sciences at **EPFL** (Ecole Polytechnique Fédérale de Lausanne); Switzerland.
Project	Development of **InspiRE,** a recommender system for safe and healthy behavior change. Worked on data science aspects: wearable sensor data analysis, time series analysis, machine learning, statistical modeling, data visualization and recommendation. Technologies: **Android (Java), C++, Python, Matlab, Javascript, Time series analysis, PCA, SVD, SEM, Clustering, SVM**. Supervisors: Pearl Pu (pearl.pu@epfl.ch) and Boi Faltings (boi.faltings@epfl.ch)
Period	October 2016 – January 2017
Position	**Research Intern** HCI Group, HKUST; Hong Kong
Project	Developing social media crawling, text and image analysis methods and performing qualitative studies for the HCI Initiative to analyse the relations between eating habits, social media habits and emotional well-being. Technologies: **Topic modeling (LDA/HCA), Python, C++, Javascript, Redis, TensorFlow, MongoDB, CSS, JotForm, D3, UX qualitative study design**. Supervisor: Xiaojuan Ma (mxj@cse.ust.hk)
Period	June 2010 – August 2012.
Position	**Research Assistant** at KOVAN Research Laboratory; Ankara, Turkey
Project	Development of the intelligence of **iCub** (*www.icub.org*), an open source cognitive humanoid robotic platform. Teaching language concepts with machine learning techniques. Technologies: **C++, Matlab, ROS, Visualeyez, OpenCV, WEKA, SVM, Neural Networks**. Supervisors: Sinan Kalkan (skalkan@ceng.metu.edu.tr) and Erol Şahin (erol@ceng.metu.edu.tr)
Period	June 2009 – September 2009.
Position	**Software Engineering Intern** at University of Ghent and DAIKIN; Belgium
Project	Development of a simulation software (front end and back-end). Technologies: **C#**, **SQL Server**.
Period	June 2008 – August 2008.
Position	**Software Engineering Intern** at Altay Corporation, Ankara, Turkey
Project	Developed back-end components for the Ministry of National Security information system. Technologies: **Java, J2EE, Hibernate, Spring, JSF, Jasper Reports**

EDUCATION

Period	September 2012- September 2017 (expected).
Title	**Ph.D. in Computer Science** at Ecole Polytechnique Fédérale de Lausanne
Principal subjects	**Machine Learning for Recommender Systems**, time series analysis and statistics.
Degree	Ph.D. In Computer Science.

Period	June 2010 - August 2012
Title	**Master's degree in Computer Engineering** at Middle East Technical University, Turkey
Principal subjects	**Machine Learning for Robotics,** computer vision, neural networks
Degree	M.S. In Computer Science. Final full marks with honor: 3.46/4.00
Period	September 2006 – June 2010.
Title	**Bachelor's Degree** at Bilkent University, Department of Computer Science
Degree	Bachelor of Science, Final full marks with high honor: 3.75/4.00

AWARDS

EDIC Department Fellowship (2012-2013) Bachelors: High honor in every semester, granted with success scholarship every semester (2006-2010).

RESEARCH ACTIVITY

Period	June 2010 - current
Positions	Research Assistant at Middle East Technical University (2010-2012), Research Assistant in EPFL (2012-current), Research intern in HKUST (2016-2017)
Publications	In 2 top journals, 5 top conferences, 2 workshops. 1 journal submission in progress (see http://hci.epfl.ch/members/onur-yuruten/ for a full list, publications available upon request)

LINGUISTIC SKILLS

MOTHER TONGUE	**Turkish**
OTHER LANGUAGES - 1	**English**
Level	Excellent.
Certificates	**TOEFL** certificate of English (111 on iBT - 2012).
Period Abroad	I worked 4.5 years (ongoing) in Lausanne Switzerland, two months (ongoing) in Hong Kong, two months in Belgium in environments where English is the everyday language for communication. The official language in all of my higher education schools is English.
OTHER LANGUAGES - 2	**French**
Level	Intermediate-Fluent.
Certificates	**Centre de Langues, EPFL,** certificate of French (B2 - 2016).
Period Abroad	I spent 4.5 years (ongoing) in Lausanne Switzerland, living in an international environment where French is the everyday language for communication.

COMPUTER SKILLS

OS used	Windows, OSX, Ubuntu-Linux, Android.
Programming Languages	**Back-End**: Java, C++, Python, C#, C **Front-End**: Javascript, CSS. **DB**: MySQL, SQL Server, MongoDB.
Version Control Software	Git, Svn.
Software Engineering	Object Oriented, Design Patterns
Scientific Computing Software	Matlab, R
Certificates	Turkish Driver's Licence
PERSONAL INTERESTS	Playing piano & guitar, dance, President of Turkish Students Association in Switzerland, hiking