How Does a Nation Walk? Interpreting Large-Scale Step Count Activity with Weekly Streak Patterns

BRIAN Y. LIM, National University of Singapore
JUDY KAY, University of Sydney
WEILONG LIU, National University of Singapore

Activity trackers are being deployed in large-scale physical activity intervention programs, but analyzing their data is difficult due to the large data size and complexity. As such large datasets of steps become more available, it is paramount to develop analysis methods to deeply interpret them to understand the variety and changing nature of human steps behavior. In this work, we explored ways to analyze the heterogeneous steps activity data and propose a framework of dimensions and time aggregations to interpret how providing a city-wide population with activity trackers, and monetary incentives influences their wearing and steps behavior. We analyzed the daily step counts of 140,000 individuals, walking a combined 74 billion steps in 305 days of a city-wide public health campaign. We performed data mining clustering to identify 16 user segments, each with distinctive weekly streaks in patterns of device wear and recorded steps. We demonstrate that these clusters enable us to interpret how some users increased their steps level. Our key contributions are: a new analytic method to scalably interpret large steps data; the insights of our analysis about key user segments in our large intervention; demonstrating the power to predictive user outcomes from their first few days of tracking.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing → Empirical studies in ubiquitous and mobile computing; Ubiquitous and mobile devices;** *Empirical studies in HCI;*

Additional Key Words and Phrases: Activity tracker, Usage patterns, Quantitative Analysis, Longitudinal Use

ACM Reference Format:

Brian Y. Lim, Judy Kay, and Weilong Liu. 2019. How Does a Nation Walk? Interpreting Large-Scale Step Count Activity with Weekly Streak Patterns. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 57 (June 2019), 25 pages. DOI: 10.1145/3328928

1 INTRODUCTION

The lack of physical activity has led to a global pandemic and is a major contributor to obesity, heart disease, diabetes, stroke, and depression [50]. The emergence of low-cost, wearable activity trackers has provided fertile ground for over a decade of Ubicomp research (e.g., [8–10,33]) and it has enabled clinical trials and interventions (e.g., [7,19,27,36,44]). Our work builds from one such intervention, a nation-wide public health campaign in Singapore, the National Steps Challenge ™ (NSC). The program is the first in the world to deploy fitness trackers at such a large scale (140,000 people walking 74 billion steps) over a 7-month period. A key feature of the program is its staged set of incentives. Participants could win three prizes over a minimum of 60 days and after

Author's addresses: Brian Y. Lim (brianlim@comp.nus.edu.sg) and Weilong Liu (e0046736@u.nus.edu), School of Computing, National University of Singapore, Singapore; Judy Kay (judy.kay@sydney.edu.au), School of Computer Science, University of Sydney.

2474-9567/2019/6-ART57 $15.00
https://doi.org/10.1145/3328928

that, they earned chances in a lucky draw. The resultant dataset, with each participant's daily step count and demographics, offers two very promising possibilities. First, it may provide new insights into the way people *participated in the program*, particularly how their *behavior was affected by the incentives*. Second, the large number of participants offers the tantalizing promise for finding *population segments with distinctive behaviors* – each segment representing people who had different responses to the opportunity to track their steps and to the incentives.

Current methods for analysing step data cannot realise these possibilities. Traditionally, public health data has been analysed using statistical methods with hypothesis testing and factor analysis (e.g., [18]). Some sophisticated methods can reveal associations between step counts and demographic factors, such as gender, age, body mass index, and income [48,58,59]. In contrast, HCI/Ubicomp research has reported *qualitative studies* of tracker adoption, use, and abandonment (such as [6,12,14]). Since people may not wear their tracker every day, the data is not a complete record of people's actual activity. Rather, it is a combination of two aspects of behavior: tracker wear and actual steps of activity, a topic that has had recent attention [29,39–41,56]. We particularly build upon the notions of streaks and breaks, defined by [39–41] as a way to describe people's behavior based on the temporal patterns in tracker wear and activity level.

We go beyond previous work to create an *analysis framework* based on two important innovations. First, we analyze the large dataset to establish a new set of *definitions of tracker wearing and step count patterns* to use in data pre-processing. Second, we define a *multi-step processing and analysis procedure* that results in clusters that enable us to identify population segments. These segments are distinctive in terms of the four core classes of information about them: tracker wearing behavior, daily step counts, the impact of the incentive prizes and demographics. These provide a new descriptive level for population segments, each with thousands of people. We further show the predictive power of our clustering approach. Our contributions are:

1. We propose an analytical and data organization framework to formalize the interpretation of steps activity in terms of arbitrary-length patterns and sequences. This builds upon prior definitions that take account of wear time with step counts [29,39,41,56], but we now establish the new definitions that emerged from analysis of our large dataset and that are needed to analyze it.
2. We refine concepts from qualitative descriptions into mathematical or engineered features, such that they can be computed and used for scalable data analysis. Specifically, our framework models steps activity in terms of patterns associated with the key dimensions of interest - tracking wearing, daily step counts, the incentives, demographics - at different time aggregation levels to generate clusters of steps behaviors and user segments.
3. We demonstrate that the framework can be applied to answering policy questions. Specifically, we applied the resulting behavioral and user clusters to explain user steps behavior regarding three questions about the efficacy of the physical activity intervention, particularly the incentives, and predict user physical activity outcomes from the first few days based on our clustered findings.

We organized our paper as follows: having introduced the (1) research objective, we next (2) discuss *related work* to explain the research gap, (3) provide an overview of the *National Steps Challenge*, (4) report the initial data analyses to clarify the nature of the data, the *challenges and requirements*, (5) explain our new core definitions of the ways to describe tracker wearing patterns, (6) describe the data analysis *technical procedure*, (7) present results of the multi-level clustering with patterns identified over time and for population segments, (8) demonstrate the application of our framework and ways to use clusters discovered to answer policy questions, (9) discuss generalizing and extending the framework along with its limitations, and (10) conclude with a summary of our contributions.

2 RELATED WORK

Since the introduction of accelerometers in mobile phones and wearables, public health, HCI and Ubicomp researchers have been interested in providing interventions with activity trackers, persuasive designs and behavioral principles to promote physical activity [8–11,18,22,27,33,36,49]. Beyond investigating how to design new interventions, this work focuses on understanding the impact of an intervention on user steps behavior.

2.1 Understanding the Usage of Activity Trackers

Randomized clinical trials with users have shown increases in step count with a computer-tailored pedometer-based intervention [7] and activity trackers [44]. While these studies provide strong evidence for the efficacy of interventions with activity trackers, it is important to understand how and why activity trackers are effective or not. Much research in HCI and Ubicomp has studied end-user adoption, engagement and abandonment of trackers primarily using qualitative methods, such as surveys and content analysis [4,15,23,24,26,30,52,54,66]. To help data owners, caregivers or health planners understand user behavior for longitudinal interventions, several models and tools have been developed. Li et al defined a five-stage model of personal informatics which supports questions that users ask for self-reflection on life log data [32]. Epstein explored the use of visual cuts to make sense of complex life logs in smaller subsets [17]. Morrison et al. developed SilverCloud to understand engagement and patterns of use of a web-based medical intervention [43]. SilverCloud shows an application usage navigation graph, a stripe time-series graph of usage for all users, a start-stop (time series) graph and a next action heat map (matrix) visualization. Tang and Kay developed iStuckWithIt [56] to help users understand their long-term levels of physical activity and daily patterns of device wearing patterns. While these tools and approaches are important to leverage visual analytics to understand step count activity, we focus on data analytic methods with visualized charts to draw insights from more complex data features.

Focusing on retrospective data analysis, other research has logged and analyzed tracker and smartwatch data to understand longitudinal use based on more quantitative or statistical analyses [29,40,41,46]. Using data from 50 to 500 participants, such work has found varying and inconsistent results regarding the engagement with, and abandonment of, activity trackers. Our work also analyzes retrospective tracker data use to understand different usage behaviors. Quisel et al. trained a convolutional neural network to predict health outcomes based on time series behavioral data, including physical activity of 1,996 users over 147 days of step count [47]. Instead, our analysis uses unsupervised machine learning to learn population segments based on behavior and demographics. Jeong et al. studied 50 smartwatch users over 203 days and performed spectral clustering on hourly steps data to identify three types of wearers (work-hour, active-hour, all-day) [29]. Similarly, Howie et al. performed latent class analysis on 628 users of a hip-worn accelerometer over 7 days to identify 5 activity types for males and females, which varied by physical activity vigorousness or sendentariness [28].

We focus on identifying patterns in daily step count (rather than intra-day step count) and how they vary across weeks and months. The work of Meyer and colleagues shares similar objectives [41]. They analyzed 104 individuals with steps data from 12 weeks to at least 9 months. Their analysis uses the notions of *streaks* of wear-time, *breaks* from wearing and *phases* of the combination of them. This enabled them to identify usage patterns in terms of phase types and user types based on an expert-driven, qualitative approach to analyse their quantitative data. We also use the time notions of streaks and phases to interpret usage behavior, although we needed to define these differently. Our work is also quite different in terms of scope and ecological validity as we investigated a much larger sample size (over 140 thousand users[1]), spanning diverse socio-economic groups (individuals were recruited citywide from multiple sites), and then we employed data mining techniques.

Analyzing temporal patterns in such data about daily steps presents several challenges which we highlight in Section 4. Our work relates to identifying different kinds of users, user segments, from time series activity data. This has been done for data types other than fitness trackers, particularly, clickstreams or mobile app usage logs. Zhao et al. analyzed smartphone app usage of 106 thousand users in China and identified 382 distinct types of users [67]. They identified 29 categories and used a k-means-MeanShift hybrid method to cluster app category usage sequences. Wang et al. analyzed over 135M clickstream events of 100k users and clustered user behaviors by partitioning on a similarity graph with an iterative feature pruning method [64]. To develop a visualization tool to explore clickstreams, Liu et al. investigated three techniques (bag-of-events, bag-of-motifs, mixture of Markov chains) to aggregate atomic events into higher levels of granularity [34]. We too performed

[1] The closest public health campaign, in terms of scale, is the Step It Up! Challenge [55] organized by Fitbit, Inc. with 61 billion steps over two weeks. However, no user segmentation analysis was reported.

data clustering, but used hierarchical clustering in a multi-scale manner. Furthermore, while clickstreams and app usage logs can describe events in a semantically meaningful way, we had to carefully engineer our data features from raw time series data.

3 APPLICATION AND DATA: NATIONAL STEPS CHALLENGE™

Our dataset comes from the National Steps Challenge ™ (NSC), a national-scale physical activity intervention program organized by the Singapore Health Promotion Board. This ran from 1 October 2015 to 8 May 2016. Members of the public were recruited throughout the country. All users were offered a free activity tracker and mobile app to track their daily step count. The program was designed to promote greater physical activity using two key features: (i) wearable activity tracker with mobile app and (ii) gamification with small rewards (which we refer to as the incentives). We analyzed data for 139,885 users, representing about 5% of the population in Singapore in 2015 [13]. The users were aged 12 to 90 years (median age 34), 59.7% female (40.3% male).

3.1 Incentive Reward System

Users could earn points each day by meeting targets:

- 5k steps for 15 pts,
- 7.5k steps for 45 points, and
- 10k steps for 60 points.

Accumulating points across multiple days, users earned "sure-win" prizes as reward vouchers:

- 600 points for the 1st voucher ($5 SGD ≈ $3.60 USD) — taking a minimum of 10 days,
- 1800 more points for the 2nd voucher ($15 SGD) — taking at least 30 more days and
- 1200 more points for the 3rd voucher ($10 SGD) — sustaining step counts for at least 20 more days.

This meant that winning all three prizes took a minimum of 60 days, and this could only be done if the person walked at least 10,000 on each of those days. After winning all three prizes, users were further incentivized to accumulate points for a lucky draw for substantial prizes. After 8 May 2016, users could not accumulate points for the "sure-win" rewards, but could still earn chances for the lucky draw. In this paper, we analyzed data from 1 Oct 2015 to 31 July 2016. Our study thus covered tracking beyond the official NSC campaign (i.e., beyond the "sure-win" period), allowing us to investigate post-incentive, sustained behavior changes.

3.2 Application Questions and Technical Challenges

To interpret the success of the campaign, several important application questions need to be evaluated. Here, we highlight three important questions about behavior changes, as well as the technical challenges associated with large-data analysis to address these key questions.

1. Did the incentives influence and drive steps activity?
2. Was the steps behavior sustained beyond the incentive period?
3. Did users increase their steps activity?

For all of these questions, we want to both measure the effects of the incentives on step counts (i.e., outcomes), and also to determine user behavior changes (i.e., *who* and *how* did they change) as these changes give insights into the long-term implications of the intervention. For example, for users who increased their steps behavior, what were their tracked activity patterns that may have facilitated their improvement? On the other hand, it is also important to understand patterns for other users who fail to improve their step counts.

4 INITIAL DATA ANALYSIS, REQUIREMENTS AND CHALLENGES

Now, we present the results of our first analyses of the dataset. This provides important new information since it is the first work to report analyses for such a large daily step count corpus. But more importantly, it established requirements for the design of the pre-processing of the data. Our lens on this analysis comes from the Ubicomp

perspective that has recognized the importance of the combination of people's tracker *wearing* behavior when interpreting daily step counts [29,39,41,56]. From this analysis we identified the challenges posed by the data. Table 1 summarizes these challenges and the technical approach our framework uses to tackle them, as explained in the next two sections.

Table 1. Technical challenges and approaches to understand temporal patterns from step count big data.

Challenge	Our Technical Approach
Imbalanced data	
Non-Gaussian steps level	Besides steps mean and standard deviation, analyze median and histogram.
Long-tail track duration	Analyze time durations with histogram bins and log transform.
Cyclic temporal steps pattern	Use the week as the foundation and smallest period of time to identify cyclic patterns.
Short and long breaks	Consider short breaks as within a week (i.e., <7 days) and long breaks exceed a week.
Routine habits and changes	Model routine as repeated weekly patterns. Changes in routine in terms of weekly patterns *or* separation by long breaks.
Diversity of users	Include demographic factors and user steps behavior in segmentation.

4.1 Handling Imbalanced Data

While many intervention studies report summary statistics of step count (e.g., mean, median, standard deviation), our initial analysis showed that many key variables are non-Gaussian and skewed. For example, the distribution of daily step count histogram in Fig. 1 has notable spikes that correspond to the incentive steps levels (5k, 7.5k and 10k steps) and non-wearing 0 steps (25.2% of days). Had people been freely setting their own targets, we would have expected a Gaussian distribution. Many widely used trackers also have default targets at 10k steps so we may have expected a peak at that level. In this histogram, we see a striking peak for days with step counts between 10k and 10.5k, just above the main incentives target.

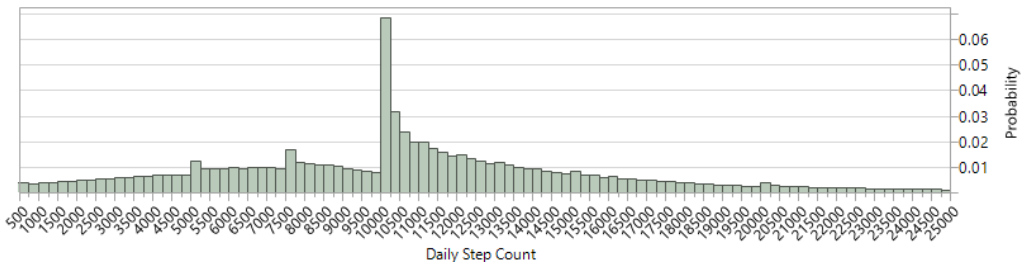


Fig. 1. Distribution of user daily step count (N=9.05 million total days); truncated at 100 days (max 305 days). There is a clear peak at 10-10.5k steps just over the 10k steps goal that gives the maximum daily points towards incentives.

The duration of tracking (see Fig. 2) also follows a long-tail distribution with a slight spike at 7 days. Most users have short participation lifespans <10 days, but some have very long lifespans (max 305, not shown). There were 8.2% of users who tracked for 1 day before abandoning; 6.1% of users for 7-8 days. It is also useful to see the total numbers of people:

- 74.2% (103,786) were tracking at 1 week from their start date
- 54% (75,588) at 1 month
- 39,504 (28.2%) at 3 months
- 12,345 (8.8%) at 6 months

These results are new and valuable information from a growing body of Ubciomp research into adoption, use, and abandonment (such as [6,12,14]), now for a large and diverse population within the context of a public health intervention. Given these distinctive and meaningful patterns, we propose that histograms should be used in analysis and reporting studies of such data.

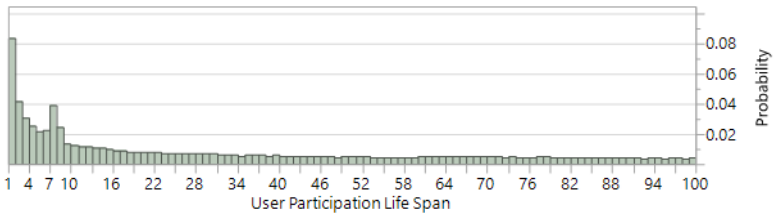


Fig. 2. Long-tail distribution of how long users participated in the program (life span). Life span mostly follows an exponential decay except for a peculiar spike at around one week.

4.2 Handling Cyclic Patterns: Weekly Time Aggregation

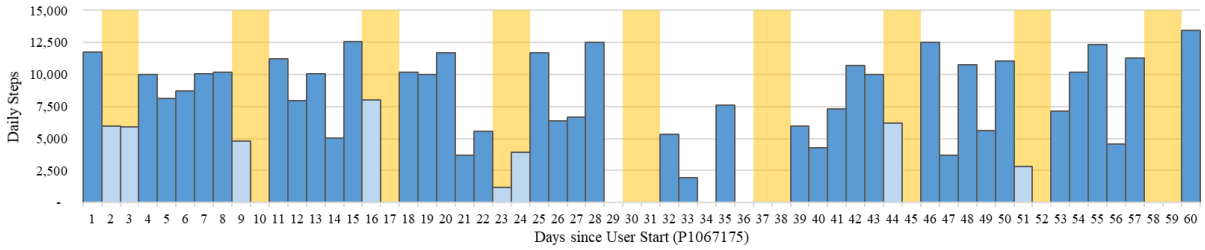


Fig. 3. Example time series of user daily steps to highlight cyclic weekly patterns. Steps level is not consistent: the user walks fewer steps or is non-active on the weekends (days with light blue bars, highlighted in yellow).

Although each user steps pattern may appear messy, by partitioning the daily steps into *weeks*, one can see a cyclical pattern emerge, as in the example user shown in Fig. 3. For example, public health and Ubicomp research report that it is common that people walk less on weekends than on weekdays [40,42,61]. We also observed other patterns within weeks. For example, some people's daily step counts ramped up from Monday to Friday for multiple weeks, or they spiked only on a single day of the week (perhaps that person's main exercise day). Some recent Ubicomp research has reported analysis of weekly tracker wear-time in terms of weekly adherence, the proportion of days that people *wear* their tracker [57]. But no work has reported analysis that uses the combination of both weekly wearing and step count patterns. Our analyses indicate the need for a greater awareness of the importance of the week for understanding activity tracker data. It also points to the need for new methods to take this into account when analyzing tracker data.

4.3 Detecting Routine Habits and Changes

Even with cyclical weekly patterns, we wanted to determine if these patterns recur and how long the same pattern continues. Recurring patterns may indicate a routine habit: it is a core goal of the intervention to help people build new habits of consistent healthy levels of physical activity. If a person consistently has >10k steps daily step counts, then this suggests a healthy habit that the initiators of this program would like to see in a larger proportion of the population. More complex patterns are also important. For example, the physical activity recommendations are for 30 minutes of moderate activity on *most* days of the week. So, a person who

makes the 10k target on 5 or 6 days would be doing well and we would like our analyses to be able to model this. We note that this analysis must account for *both tracker wear and actual daily step counts*. This means that a person whose week has 1-2 missing days may be modelled as doing well if they meet the 10k target on those days. This points to the complex mix of patterns of wear and actual daily step counts that could be meaningful. It points to the need for clustering methods that can discover such patterns.

4.4 Describing Wearing Behavior and Breaks

Following [41,57], we define active days as those with >= 500 steps and consider any day below this as a break. Our users took a total of 498k (498,047) breaks and 2.3M (2,315,088) break days, giving an average break length of 4.6 days. Most breaks were short. There were 87.5% <= 1-week long (7.3% lower than that the 94.8% reported by Meyer et al. [41]. These short breaks were dominated by 1-day breaks, which made up 49.2% of all breaks. The number of breaks per user follows a long-tail distribution with 25.6% of users taking no breaks (dropping out before a first break), 14.7% with one break only, 10.2% with two breaks. Before the end of the NSC, only 9,228 users remained, i.e., 130,657 users (out of 139,885) had abandoned tracking.

There are many reasons for breaking behavior including *forgetting* to wear the tracker for the day, failing to *upkeep* the tracker (e.g., battery not charged or device breakage), *skipping* wearing the device for diverse reasons (e.g., already aware of typical daily steps, having a sport activity where the tracker could not be worn, or expecting a low step count), or *suspending* tracking over a period of time (e.g., misplacing the device, taking a holiday, injury, or pregnancy) [6]. Importantly for our incentivized application, many users suspended tracking after having won a prize. Those who suspended after the 3rd final prize are similar to some instrumental trackers (users) identified by Epstein et al. [16] who stop tracking after the benefits fade or are withdrawn. However, we also found users who suspended after just the 1st or 2nd prize, perhaps suggesting that they felt that the incentives were insufficient to sustain their tracking activity or maintain their steps level. Intrinsic to our analysis of treating *weeks* as the atomic unit of analysis, we chose the cut-off duration of 7 days to define short and long breaks.

4.5 Handling Traits: Accounting for the Influence of Demographics

In HCI / Ubicomp research, with some notable exceptions [10,15], there is little consideration for how demographics influences user steps or wearing behavior. However, research in public health has shown the influence of demographics on user steps behavior [48,58,59]. A 2-way ANOVA analysis found that women walked *slightly* fewer steps/day than men (~500 steps), but this is more pronounced for younger men and women (~1k steps), and that older users had higher daily steps and participated for longer than younger users. The gender effect agrees with public health studies about gender differences [1,27] and the age effect agrees with work by Tang et al. [57]. Given the influence of demographic factors, we argue that it is important to consider these user traits to model their steps activity.

5 FRAMEWORK OF DISCOVERING INCENTIVE-DRIVEN PHYSICAL ACTIVITY POPULATION SEGMENTS

Building on previous research and our initial data analyses, we now explain how we defined the dimensions of analysis for our framework's feature engineering. The definitions are summarized in Table 2.

5.1 Dimensions of Incentive-Driven Physical Activity Intervention

We modeled steps activity using several **dimensions** to characterize various aspects of user behavior. We further refine them into **themes** to account for more aspects of interest. This allows us to identify more data features and paint a richer picture of user steps behavior with our data mining technique.

Table 2. Definitions in our framework to identify key behavioral patterns and user characteristics based on steps activity.

Daily Step Pattern	**Definition**
Steps level	Step count for each day in bins: 0, 500, 5000, 10000, 10500, 15000, 20000+ steps.
Non-wearing day	A day with no steps, i.e., step count = 0.
Non-active day	A day where the step count is <500 steps.
Active day	A day where the step count is ≥500 steps.
Temporal Patterns	**Definition**
Break	Period of one or more consecutive non-active days.
Short break	A break shorter than a week (1-6 days). e.g., one-day break on Monday.
Long break	A break lasting at least one week, i.e., ≥7 days.
Weekly pattern	Distinctive steps level based on the day of the week.
Streak	Note that a weekly pattern can have short breaks, e.g., where a user regularly does not wear the tracker on Mondays. Our clustering method explicitly defines weekly patterns as week clusters.
Phase	Period of one or more consecutive weeks with the same weekly pattern.
Phenotyping	**Definition**
User segment	A cluster of similar users in terms of temporal wear and daily steps patterns and demographics.

5.1.1 Engagement (steps level, adherence, tracking duration). Our primary measure of physical activity is the **daily step level**. Steps levels such as 10,000 steps/day are commonly used for goal setting to promote physical activity (e.g., [8,10,33,56]). After considering several methods [57] and in discussion with public health experts, we chose 500 steps as the threshold to eliminate non-purposeful use of the fitness tracker and 5k steps intervals for our histogram bins for the initial analysis, giving activity level bins bounded at steps level of 0, 500, 5k, 10k, 15k, and 20k. The histogram is helpful to represent another important metric, **adherence**, which is how often she adheres to a target steps-level in a time period. Finally, an important metric of engagement is the **tracking duration**, which can distinguish a dedicated user from one who has abandoned tracking.

5.1.2 Breaks (frequency, duration). We model how often users take breaks (**frequency**) and their **duration**. Based on the initial analysis, we define **Short Breaks** as <1 week and **Long Breaks** as at least a week.

5.1.3 Habit & Routine (pattern, frequency, diversity). A key goal in physical activity intervention programs is to get people to form healthy habits (e.g., 10k steps/day). We seek to discover week-long **patterns** and their **duration** which suggests a habit. Finally, users may not be consistent in their behavior or change their behavior. Hence, we model the **diversity** in terms of how many different types of behaviors each user may have.

5.1.4 Incentives (sufficiency, proximity, sustainability, achievement). Given their central role in our application for intervention, incentives are key for our analysis. To account for this we refined the step-level bins with one more to model the narrow band of 10-10.5k steps – matching the Fig. 1 spike in steps at 10k and just above (within ~10 minutes at purposeful steps cadence). We did not find strong patterns for the 5k- and 7.5k steps goals and, so, did not include these levels for special analysis. Other factors to model how steps behaviors may be influenced include **proximity** to winning the next prize (e.g., users may ramp up their steps), the **sustainability** of maintaining daily step level after winning a prize and for how long, and whether a reward was **achieved** at all.

5.1.5 Behavior change (steps level difference, adherence difference). The core goal of the intervention was to help people build healthy habits for physical activity. We particularly wanted to model the data to discover patterns characterising user segments with an **increase in steps level** (higher median step count) or at least an **improvement in adherence** (more active days in a time period) between the pre- and post-intervention periods. While some studies used longer periods of comparison (e.g., 4-week [33]), we constrained this length to be able to analyze user data with short life spans (e.g., ~2 weeks).

5.2 Temporal Modeling and Time Aggregate "Chunks"

To help to make sense of the complex patterns of daily step counts, we aggregated the data into time chunks so that we can reason about the behaviors in terms of fewer, but higher-level, semantically meaningful patterns, instead of in terms of numeric variables. Unlike conventional techniques that only consider step count (or adherence) level and changes in steps level at fixed time intervals (e.g., daily or weekly average), our approach allows arbitrary time lengths that span a few days to many weeks or even months.

Before explaining the rationale for our definitions of time aggregate chunks, we help the reader build a high-level understanding of the core ideas with an example. Fig. 4's shows how our definitions translated into clusters that we can use to describe the data for a person. The top bar chart shows daily step counts and the Days row of squares shows their color coding. This user's first week had just 4 active days, with the first day well above 10k steps (green day cells) and the 3 other active days had lower step counts, the last very low. Then she had a break of more than a week, followed by a single very active day in Week 3. In the 4th week, she had 5 active days, two in the narrow band of 10 to 10.5k (blue). After this, the tracking and activity levels take off (with lots of green and blue), albeit with some drops and a break at Day 42. In the rest of this section we will use the remaining four color-coded rows at the bottom of the figure to illustrate our definitions. (The full details of the color codes are in Fig. 8 which also shows more data for this person. In addition, Sections 6 and 7 explain the precise meanings of the particular clustered weeks, streaks, phases and prizes in this figure.)

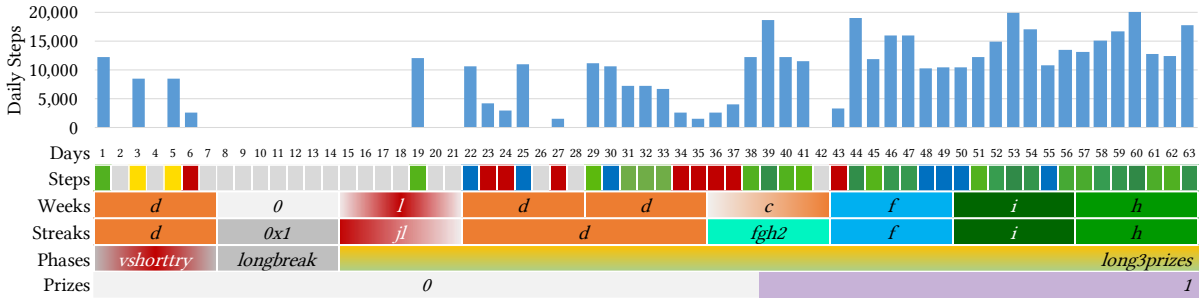


Fig. 4. Example partial time series a user's steps data with time chunks at different aggregate levels: days, weeks, streaks, and phases; different colors represent different types for each level (see Fig. 8 for the legend). The prizes won are also shown in the last row. The similarity of chunks (e.g., weeks) for each higher-level time aggregate is determined from data clustering, and subject to some uncertainty depending on the number of clusters chosen.

5.2.1 Week. As the previous sections have explained, the week is central to our analysis and our work is the first to link this to definitions of streaks and breaks. We seek to find common weekly patterns, such as weeks when weekend step count is lower than on weekdays, or when users walked mostly mid-week. For users who tracked for multiple weeks, this allows us to easily model cyclic weekly patterns. By clustering weeks, we desensitize the weekly patterns from occasional or rare exogenous factors. In Fig. 4, the user's 1st, 4th and 5th weeks all are in the same cluster (labelled *d*) – intuitively, d models weeks with many days <10k steps. The 2nd week has no active days (Cluster 0) and the remaining weeks all fall into different clusters (*l*, *f*, *i* and *h*). We explain the cluster types later in Section 7.

5.2.2 Streak. Once we have used clustering to determine similar types of weeks, we can then look for consecutive sequences of similar weeks. We define a streak as a *continuous repetition of similar weeks*. This is intuitive, and may indicate that the user has a **habit** of the same weekly steps activity over multiple weeks. We can then describe user segments based on primary week types and their longevity. Note that our definition is different from Meyer et al. [40,41] who defined a streak as an "uninterrupted series of use days". Users may change streaks for many reasons, but of particular interest for our work is the changes associated with the incentives, as reflected in changes after the prize period. Instead of just considering the frequency of week types

for a user, modeling streaks allows us to see if the user was able to maintain a behavior for consecutive periods of time. In Fig. 4, we see that the 1st week is a 1-week streak (d) and the same streak type covers the 4th and 5th weeks. As each other week is in a different cluster, these map to 1-week steaks.

5.2.3 Phase. We define a phase as a continuous stretch of streaks ended by a long break. Unlike Meyer [41], who defined Phases to end with long breaks, we define a long break as its own Phase, i.e., one with only non-active days (e.g., Week 2 in Fig. 4). In Fig. 4, there are just 3 phases. The first is what we describe as a *very short try* – this covered 12.2% of all phases. The next is what we describe as a *long break* and these covered 8.0% of all phases. (Other phase cluster emerged for longer breaks.) The third phase in the figure, the *long 3 prizes*, is where the user had a long sequence of phases that eventually resulted in that person winning all three prizes. In the figure, the bottom Prizes row shows that user won her first prize in the 6th week, and was still headed towards her next prize.

5.2.4 User. Segmenting users is a key objective of our analysis. By knowing common types of segments, public health officials can construct more informed interventions for different user segments. We model similar users by whether they have similar attributes (e.g., demographics), streaks and phases.

Table 3. Summary of which data features engineered for each time aggregation level. Due to the large numbers of data features, we only described their dimensions and themes here. See Appendix A for details of data features. Not all themes were relevant or necessary to include for each time aggregation level (see table footnote for reasons).

		Data Features @ Time Aggregation			
Dimension	Theme	Weeks	Streaks	Phases	Users
Engagement	Steps Level	✓	✓	✓	✓
	Adherence	✓	✓	*Not included**	*Not included**
	Tracking Duration	*Always 7 days*	✓	✓	✓
Breaks	Break Frequency	*As non-adherence¶*	*As non-adherence¶*	Short (<7d) only	Short & Long
	Break Duration			Short (<7d) only	Short & Long
Habit & Routine	Pattern	By Weekday	Of Weeks	Of Streaks	Of Phases
	Frequency	*Always 1x/week*	✓	✓	✓
	Diversity	*All Weekdays*	✓	✓	✓
Incentives	Sufficiency	✓	✓		
	Proximity	✓	✓	✓	✓
	Sustainability	✓	✓	✓	
	Achievement	*Too short§*	✓	✓	✓
Behavior Change	Steps Level Difference	*Too short§*	*Expect consistent‡*	✓	✓
	Adherence Difference			✓	✓
Demographics†	Gender				✓
	Age				✓
	Body Mass Index				✓

* Modeling with weeks already account for adherence, so not recalculated.

¶ We treat non-active days in short time aggregates as just attributes representing non-adherence, instead of an interruptive break.

§ Weeks too short to see changes, since we measure metrics per week.

‡ Since streaks are consistent, we do not expect to see behavior change within a streak.

† We only included demographic variables for user clustering to independently model behavior at lower levels.

5.3 Applying Dimensions and Themes for Each Time Aggregation Level

We apply the dimensions and themes with the time aggregation levels and combine them in a matrix representation. Table 3 summarizes the dimensions we included for characterizing steps behavior at each time aggregation level. We can see that not all dimensions or themes are applicable at every level. The focus of the Week time unit is to capture basic cycles of repetition or differences in behavior; so it is too short to measure changes in behavior. Instead, some patterns may be repeated over multiple weeks as Streaks. e.g. regular steps

behavior or sustained high-effort behavior to accomplish incentive goals. Note that since we define streaks as consisting of similar weeks, we do not expect to see behavior change in terms of step count. Instead, we seek to detect behavior change *within* each Phase and for the User overall. Finally, to account for the influence of demographics on step count, we include demographics in data features for the clustering. We exclude demographics from lower time levels, since we first want to learn about types of steps behaviors and only afterwards, analyze which user types had different behaviors.

6 BEHAVIOR-BASED DATA ANALYSIS PROCEDURE

We describe our analysis procedure to retrospectively analyze out step count data. Our dataset consists of two tables: one with user *participant* details such as their demographic details and physical attributes (heights and weight), and user *daily step count*, which indicates their step count for each day with the calendar date of the day. We converted calendar dates to Cohort Day, starting from 1 till the last day of each user's participation. This time alignment allows all users to be compared, regardless of the date on which they started. Next, we aggregate the steps time series data iteratively from days to weeks, streaks, phases, and users to aid interpretation. Fig. 5 summarizes the four-stage procedure that is iteratively applied to higher temporal aggregate levels.

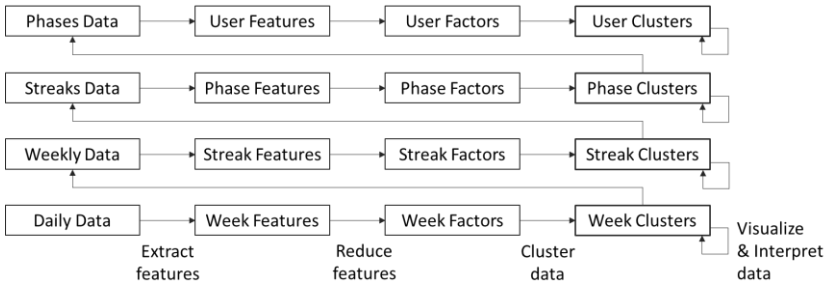


Fig. 5. Data processing flow to extract features, reduce features and cluster data at four time aggregation levels.

6.1 Data Procesing and Analysis for Each Time Aggregation Level

For each time aggregation level, we go through the same multi-step processing and analysis procedure to (i) extract data features, (ii) reduce the features into a set of key factors, (iii) cluster the data into meaningful types, (iv) visualize the clusters in terms of features and factors for interpretation, and (v) verification and iterative refinement of clusters. Although we apply standard data mining techniques at each step, careful curation was needed to identify key factors and tuning with some domain knowledge was needed to generate insightful clusters based on the aforementioned dimensions and themes. We describe the detailed procedure next.

6.1.1 Extract Data Features. We extract data features from each time aggregation level for each defined theme, which are themselves derived from dimensions. Examples include the median active step count level, step count for Mondays, and number of points till a prize is won. See Appendix A for more details. We normalize features to the range 0-1 using min-max normalization; some features are comparable, so they are normalized together. We extracted many features (23-53) to capture properties of the distribution for many variables (e.g., histogram of step count vs. median or mean). Nevertheless, these will be reduced in later PCA preprocessing.

6.1.2 Reduce Features to Principal Factors. To find a reasonable number of meaningful key factors, we next preprocessed the data for dimensionality reduction. Many methods can be chosen for dimensionality reduction, such as principal component analysis (PCA), kernel PCA, non-negative matrix factorization, linear discriminant analysis, and autoencoders. We chose PCA since it is a popular and readily available method that can be used with basic statistical analysis tools, making our method more usable by domain experts in physical activity behavior tracking and interventions, even with limited expertise in machine learning. We chose PCA instead of

exploratory factor analysis (EFA) since our primary goal was dimensionality reduction, instead of identifying a parsimonious representation (with EFA) [18], and to retain more variability [65] in the data for subsequent cluster modeling. We did not employ confirmatory factor analysis (CFA) since we did not want to enforce a fixed set of factors pre-hoc. PCA has a few assumptions, such as linearity, importance of large variances, orthogonality of principal components [53]. Note that there is no requirement for predictor variables (data features, e.g., step count) to be normally distributed, continuous, or even symmetric, since for the purpose of modeling factors, only the outcome variable is considered a random variable [25]. Furthermore, multivariate normal distribution is also not a requirement, but instead features should be linearly related [53]. However, since we use PCA for dimensionality reduction to help with subsequent clustering and we do not require the principal components to be independent, the aforementioned assumptions are not strictly necessary [53]. Nevertheless, we have identified interpretable clusters that replicate earlier work. We later discuss limitations of using PCA which can be mitigated with more sophisticated methods.

We applied PCA on the covariance matrix of data features, instead of the correlation matrix, because some features are on the same scale and we want to keep them comparable. In testing, using PCA on correlation is similar, though with lower explained variance (~75%).. We further processed the principal components with a varimax rotation to simplify for interpretability. We carefully chose the number of factors based on several approaches: (a) the elbow method from a scree plot, (b) ensuring that the total explained variance is >90%, (c) trying to keep the number of factors as small as is reasonable, and (c) ensuring that all dimensions have data features represented by at least one factor. The latter requirement ensures that we can model the steps behavior in terms of themes that we are interested in. Note that some factors may span multiple dimensions; this indicates that the dimensions are correlated. Finally, by observing which data features are positively (or negatively) weighted for each factor, we can interpret their correlation or anti-correlation. This can give some insights into user tendencies and behaviors which we describe in more detail in the next.

6.1.3 Cluster into Higher Level Time Aggregation. We next group the lower-level time data into higher-level time aggregations using hierarchical clustering on the factors with the Ward method. We chose the number of clusters based on several approaches: (a) the elbow method from a scree plot, (b) aiming to retain a small number of clusters to ensure interpretability, and (c) interpreting the clusters to be distinguishable from one another. We performed the latter task with the help of visualization methods described next.

6.1.4 Visualize Features and Factors per Cluster to help to interpret each cluster, using multiple visualization methods: (a) large spreadsheet tables with conditional coloring to highlight differences between clusters, (b) histograms to go beyond summary statistics (e.g., mean, median, standard deviation) to identify distributions and ranges of variables, (c) sample example sequences (of days, weeks, streaks, phases) to speculate on specific instances, and (d) tokenizing sequences using text analytic methods (e.g., word frequency counts) to identify most common chunks (words) and co-occurrences (n-grams).

6.1.5 Verify, Describe and Summarize Clusters. We verified our interpretations of behavior based on prior literature of steps activity or relevant behavioral theories. Finally, we describe each cluster type in Section 7.

Note: Careful attention is needed for feature engineering and factor selection to ensure that the clusters learned are framed in terms of **meaningful** dimensions. We sought to identify groups based on key dimensions and themes. Without specifying these themes in the data features, they may not have driven the clustering. Furthermore, we ensure that the number of clusters is chosen to reflect groups of interest. Hence, we explicitly defined and chose data features, principal factors, and clusters to help us to interpret the data more meaningfully.

7 PATTERNS OF STEPS BEHAVIOR

Using our rigorous, quantitative method, we identify clusters of weekly, streak, and phase steps behavior and user characteristics. As expected, some of these clusters match behaviors reported in previous work [14,40,41] and for consistency, reuse names for relevant patterns. The clusters also provide a descriptive framework to interpret potential reasons for user behavior in an incentive-driven physical activity intervention, as we discuss in Section 8. For brevity, we discuss just key findings here (with more details in Appendix A).

7.1 Week Patterns: Based on Overall Steps Intensity, Daily Consistency and Weekday "Shape" of Activity

Analyzing 1,396,476 weeks, we identified 10 week factors based on the dimensions and themes. The factors selected emphasized the different pre-defined steps levels, revealed weekly patterns by weekday or weekend, and modelled the influence of incentives. The key takeaway is that week types are distinguished by their **overall steps intensity**, the daily **consistency** over the week, and the "**shape**" of activity over the week in terms of which weekday was the user most active. We clustered weeks into 13 types, interpreting the week clusters using a combination of large tables and graph figures. We show some examples in Table 4, Fig. 6 and Fig. 7 and describe two week types in detail.

Week *a* cluster. Reading Table 4 from left to right, (i) *a* Weeks are labeled with a white-red gradient rectangle suggesting that step count is typically low or none in the beginning of the week and increases to a low step count late in the week; (ii) this is summarized that the user typically walks on Sunday and generally has low/no steps for most days, thus taking a long time to reach any prize; (iii) 2.2% (accounting for 32,646) of all weeks are type *a*; (iv) the median daily active step count is moderate 7,752 steps, though (v) there is typically only one active day for such weeks as indicated by the % steps orange heatmap, i.e., mostly (6/7) non-active days and 1/7 days have ≥500 steps; (vi) the green heatmap shows the average adherence for each day of the week, indicating that 99.8% of active days (≥500 steps) fall on Sundays for *a* weeks. Fig. 6 shows average and spread of steps level for different weekdays and (vii) illustrates that, for *a* Weeks, most days have median 0 steps (IQR=0) and Sundays with median 6,888 steps (IQR=3,519-10,579).

Week *f* cluster. Reading Table 4 from left to right, (i) *f* Weeks, labeled with a blue rectangle, have above-average number of days with incentive-level step counts (10-10.5k steps/day); (ii) this is summarized as the week is particularly incentivized to just satisfy the 10k-steps goal; (iii) 11.4% (158,580) of all weeks are type *f*; (iv) the median daily active step count is 14,188 steps and (v) the % steps orange heatmap indicates that the steps levels are typically in the range of 10-10.5k steps (45%), 10.5-15k (29%), and 5-10k (13%); the % day of week green heatmap shows that users are mostly active every weekday. Fig. 6 shows that the step count is typically a little higher on Fridays and Saturdays (median=10,299 and 10,263 steps, respectively). Note that this does not mean that a typical *f* Week has most days with incentive-level step counts, i.e., there may be some arbitrary days with more, fewer, or no steps. Fig. 7 shows the relationship between steps level and consistency and illustrates relatively moderate inconsistency (CoV=0.22). Fig. 8 also shows a specific example of an *f* Week with only two days with 10-10.5k steps and four days with >10.5k steps.

In total, we identify four key patterns for different types of weeks:

- **Consistent weeks** of high steps activity (*g*, *h*) or moderate steps activity (*e*). These intuitive clusters are prevalent - making up 33% of the weeks in the dataset (13.5%, 9.2%, 10.2), as expected More surprising is the 11.4% of *f* weeks where users almost always walked just enough to satisfy the 10k steps/day goal. In Fig. 6 we see that the f cluster is consistently, tightly at the 10K step boundary,
- **Inconsistent weeks** were the 6.1% of very high steps activity (*i*) and 11.0% of low steps activity (*d*). In these weeks users were occasionally very active, or struggle to be active, respectively. In Fig. 6 these cluster have large whiskers, indicating very inconsistent step counts and Fig. 7 shows *d* as the right-most, indicating the most inconsistent and irregular counts.
- **Partially active weeks** where users were more active on weekdays instead of weekends (*k*), or only begin to be more active later in the week (*c*). *k* is not intuitive, but matches the public health literature reporting people as less or non-active on weekends [40,42,61]. Week *c* is not commonly described, but shows up significantly in our dataset (7.9%).
- **Very non-active weeks** with some low step counts later in the week (a, b, *l*) or on Monday (*j*). These could indicate very inactive regular patterns, but also 1-day participation where users tried tracking for only one or a few days before quitting.

Table 4. Summary describing 13 week types of user step activity (N=1.4M weeks). Matrices on the right are heatmaps with more intense colors for higher numeric value. The % Steps heatmap shows the distribution of daily step count levels, and the % Day of Week heatmap shows proportions of active days (steps > 500) for each weekday.

	Week Cluster	% Weeks	Median Active Steps	% Steps: 0	1-500	500-5k	5k-10k	10k-10.5	10.5-15k	15k-20k	20k+	% Day of Week: M	T	W	R	F	S	X
0	Break Week (all non-active steps)	10.3%																
a	Sunday walker, low/no steps, long time to prize	2.2%	7,752															
b	Late Week + Sat Low/Moderate Steps	1.6%	9,555															
c	Mid-week to Weekend Low/Moderate Steps	7.9%	12,764															
d	Tail + Weekend/Mon walker, low steps	11.0%	13,667															
e	Moderate Steps	10.2%	12,204															
f	Incentivized 10k Steps	11.4%	14,188															
g	High Steps	13.5%	15,335															
h	Very High Steps	9.2%	19,861															
i	Extremely High Steps	6.1%	28,640															
j	Monday Walker with Some Very High Steps, then break	2.2%	8,201															
k	Weekday walker, no weekend, inconsistent effort	13.0%	12,262															
l	Mid-Late-week weak walker, no weekend, low steps	1.3%	6,922															

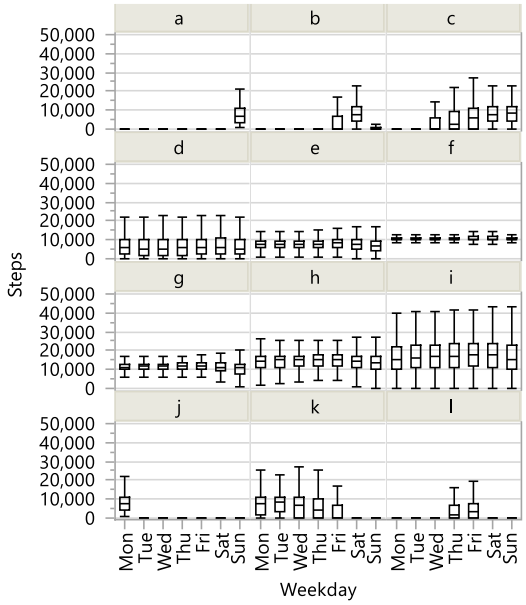


Fig. 6. Box and whisker plots showing different average weekday steps patterns for user weeks in each cluster. End of whiskers indicate 1.5 interquartile range (IQR). Most clusters have somewhat uniform steps levels for all weekdays (e.g., *e*, *f*, *g*, *h*), but higher IQR indicates more inconsistency (*d*, *i*); some weeks are more active on certain weekdays or weekends (*b*, *c*, *k*, l).

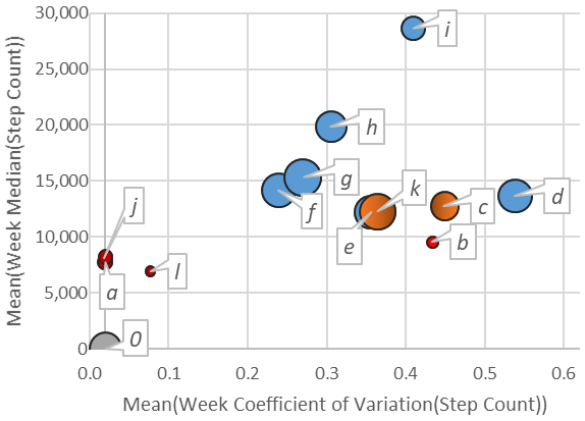


Fig. 7. Bubble chart showing relationship between different week clusters arranged by the mean of Median Step Count and mean CoV of their weeks. Interpretation: larger bubbles indicate larger clusters; higher placed clusters have higher steps level, and clusters towards the right have more inconsistent or irregular steps every day of the week. e.g., week type *i* has the highest steps level, but the steps are not necessarily high every day of the week; *c* and *k* are shaded with orange gradient to highlight that they represent partial active weeks; *a*, *b*, *j*, *l* are shaded red to indicate very short or 1-day active weeks.

We have described some intuitive, expected, and interesting week types based on three key aspects of overall steps intensity, daily consistency, and "shape". To understand whether these are repeatable patterns for each user, we next analyze the sequences of these weeks as streaks in the next section.

7.2 Streak Patterns: Ranging from Very Short and Disengaged to Consistent Series of Weekly High Steps

By identifying consecutive weeks of the same type, we found 962,166 streaks, each spanning one or more week cycles. Most streaks are 1-week long (78.8%), fewer are 2-week (12.4%) or 3-week (4.2%). The longest streak is 42 weeks long. Streaks are mostly reflective of their underlying week type and are labeled similarly (e.g., streaks of *f* weeks are labeled as *f* streaks). Due to limiting the number of clusters, some streaks were merged week types ([*ab*], [*jl*])[2]. We also aimed to identify streaks influenced by incentives and found three streak types when users completed one of the three incentive prize goals ([*fgh*]{2}, [*gfh*]{2}, [*gfh*]{3})[3]. Refer to the Appendix for more details of our identified factors, data features, and clusters. While we can detect healthy or consistent habits at the Streak level, we will have to look at the higher Phase levels to understand how behaviors change and whether the user takes a break or abandons tracking after a streak.

7.3 Phase Patterns: with Varying Duration, Diversity, Difference, and Drive of Steps Activity

By joining consecutive streaks and separating those with long breaks (≥7 days long), we found 254,200 phases, each consisting of one or more types of streaks terminated by a long break. Phase length follows a long-tail distribution. Most phases are short: many are 1-2 weeks long (26.7% and 21.5%, respectively), followed by 3 weeks (10.0%) and 4 weeks (6.8%). The longest phase is 44 weeks long. Similarly, short phases dominate: 1-streak (38.8%), 2-streak (19.7%) and 3-streak (9.5%). Some phases are diverse with many streaks; the most inconsistent phase has 37 streaks. We clustered phases into 15 types based on Phase **duration** in number of days, **diversity** of streaks, whether there was a **difference** in steps level from beginning to end, and whether the high steps level appeared to be **driven** by incentives or not. See Table 11 in Appendix for more details. Key patterns are:

- **Very short trying phases** with low steps level, interrupted with a long break or abandonment (*very short try*, *short try*, *short try high*). The phase name includes "high" to indicate that the average step count was higher than the phase type without "high". These last for only a few days or about 2 weeks, on weekdays, and may contain 1-2 small breaks. We consider that this models users who tried tracking for a while, but then abandoned it after a weak attempt.
- **Very long driven phases**, where users sustained a high steps level for about three months without a long break to win all three prizes in a stretch (*long 3 prizes high*, *long 3 prizes*). However, after winning the final prize, these users take multi-week breaks or abandon tracking; this indicates that these users were incentive-driven. On the other hand, the *self-driven* phase indicates user steps behavior with sustained high step count through winning all prizes and continuing beyond the last prize; this indicates that these users either habitualized and internalized the motivation to be physically active, or did not need the incentives.
- **Incentive driven phases**, where users sustained high steps level for each intermediate prize, and took a long break immediately after winning each prize. This suggests that, in these phases, users were less motivated than the *long 3 prizes* phases and took more time to return to active tracking.
- **Improved phases**, where users increased steps level from the first to last week by having more active days per week and/or increased their weekly step count (*try improved*, *try hard improved*, *p3 won improved*).

The phases we identify have some similarities and differences with the work of Meyer and colleagues [41]; thus, we briefly compare them: our short try phases are similar to their "minor use phase", our long 3 prizes phases are similar to their "very long phase", we do not consider density for phases because we lacked data about intra-day density, we distinguished phases based on whether the steps levels changed over time and whether users achieved goal targets, and we do not have a "restarting phase" which have a break before them. Instead of clustering streaks based on what came before, we can analyze the transition sequences between different types of phases. Interestingly, we found that the *try hard improved* phase is frequently preceded by

[2] We use a regex-like notation for naming streak clusters with multiple week types: [] indicates OR, i.e., the streak consists of only one of the listed week types.
[3] {} indicates the number of occurrences, e.g., {2} indicates that the streak is typically about 2 weeks long. The ordering of the week type in the streak name also indicates the relative dominance of the week type for such streaks, e.g., [*fgh*]{2} streaks are mostly streaks of *f* weeks and [*gfh*]{2} streaks are mostly of *g* weeks.

long breaks (33.0%), *very long breaks* (19.1%), or *extra long breaks* (11.9%); 64.0% for all breaks. Similarly, *p3won improved* and *try improved* phases are normally preceded by various long breaks (58.1% and 57.5%, respectively). This suggests that the "restarting phase" in [41] is equivalent to these three phases.

With the Phase and Streak clusters, we curated common patterns to interpret and summarize user steps activity. This can help us understand *how* users appear to have applied certain behaviors or changed behaviors to accomplish steps goals and outcomes. We defer deeper discussion of how to use these clusters to Section 8.

7.4 User Types: Differentiated by Duration of Participation, (In)Consistency in Steps, and Behavior Change

For the lower level time aggregate levels, we used steps data from all 139,885 users. The user-level analysis we now report was restricted to the 64,329 users (46% of all users) who provided their demographic details when signing up for the NSC. We clustered users into 16 types (see Table 5) based on the user's **duration** of participation, **(in)consistency** of steps level, and **change** in steps activity level.

Key results are summarized in Table 5 with horizontal lines separating five important groups of clusters. Clusters ranged in size from 2.3% (Cluster label, *CD*) to 11.5% (*C2*) of all users (covering 1,500 - 74,000 people). Their median steps on active days ranged from 5,668 (*T1*) to 13,222 (*C5*) and average tracking duration from just 1 day (*T1*) to 196 days (P). To provide a more intuitive view of user behavior, Fig. 8 shows representative users from 12 of the user clusters. We now describe key clusters and their most important properties. Where possible, we use the terms from Meyer et al. [41], highlighting aspects which replicate that work. Key patterns are:

- **Try & Drop users**, the first group in Table 5 and the users at the top of Fig. 8, who tracked only 1 day to ~1 week before abandoning (*T1*, *T2*, *T3*). Their low daily step count and short participation suggests that they did not seriously engage in the program; perhaps they were in early stages of the transtheoretical model (e.g., pre-contemplative and contemplative) [45].
- **Consistent users,** (*C1*-5 in the table, ordered by #Days they participated) who regularly had similar high steps (users who walked more days also tended to have higher median active step counts;). The table also shows a general trend that users who participated longer (#Days) also tended to have higher average step count, indicating an association between these. These clusters represent important user segments.
  - *Overweight*[4] *users* (*CO*) with mean BMI=30.5 - a distinctive feature of this segment (see last column of the table). They were heterogeneous in tracking duration, with 11% at >5 months and 21% at <1 week. They generally had a lower adherence than average (81% for >500 steps, 54% for >5k steps, 30% for >10k steps) and non-notable change in steps level. This group of users are clinically meaningful to target for behavior intervention and also make up a significant 11.5% of the participant pool.
  - *Improved Consistent users* (*CI*) who increased their step count. Interestingly, this cluster had the more women (70% vs. baseline of 60%), a distinctive feature of this cluster suggesting that they were slightly more receptive to this intervention.
  - *Non-sustained Originally Consistent users* (*CD*) who were originally consistent, but their step count eventually deteriorated. On average, these users started out strong for months to win prizes with continuous phases, some winning 3 prizes in 1 stretch (573 users), but they took a break soon after winning all prizes and returned for short periods (<1 week) of moderately high and incentivized step count. They then drop out. Their adherence decreased by 24.6% for >5k steps, 54% for >10k, 25.6% for 15k, and 10.6% for >20k (red in the heatmap in the table).
- **Inconsistent users** who switched to many different types of streaks and had many short and long breaks. One cluster consists of Slow Starters who had months-long breaks after initially tracking (*SS*). Hop-On Hop-Off users alternated between long non-active breaks and active days with moderate to high steps and last very long (several months) in the program. There were also users who improved or deteriorated in their step counts (*II*, *ID*), but unlike consistent users, the changes tend to occur at lower steps levels.

[4] Body mass index (BMI) > 24

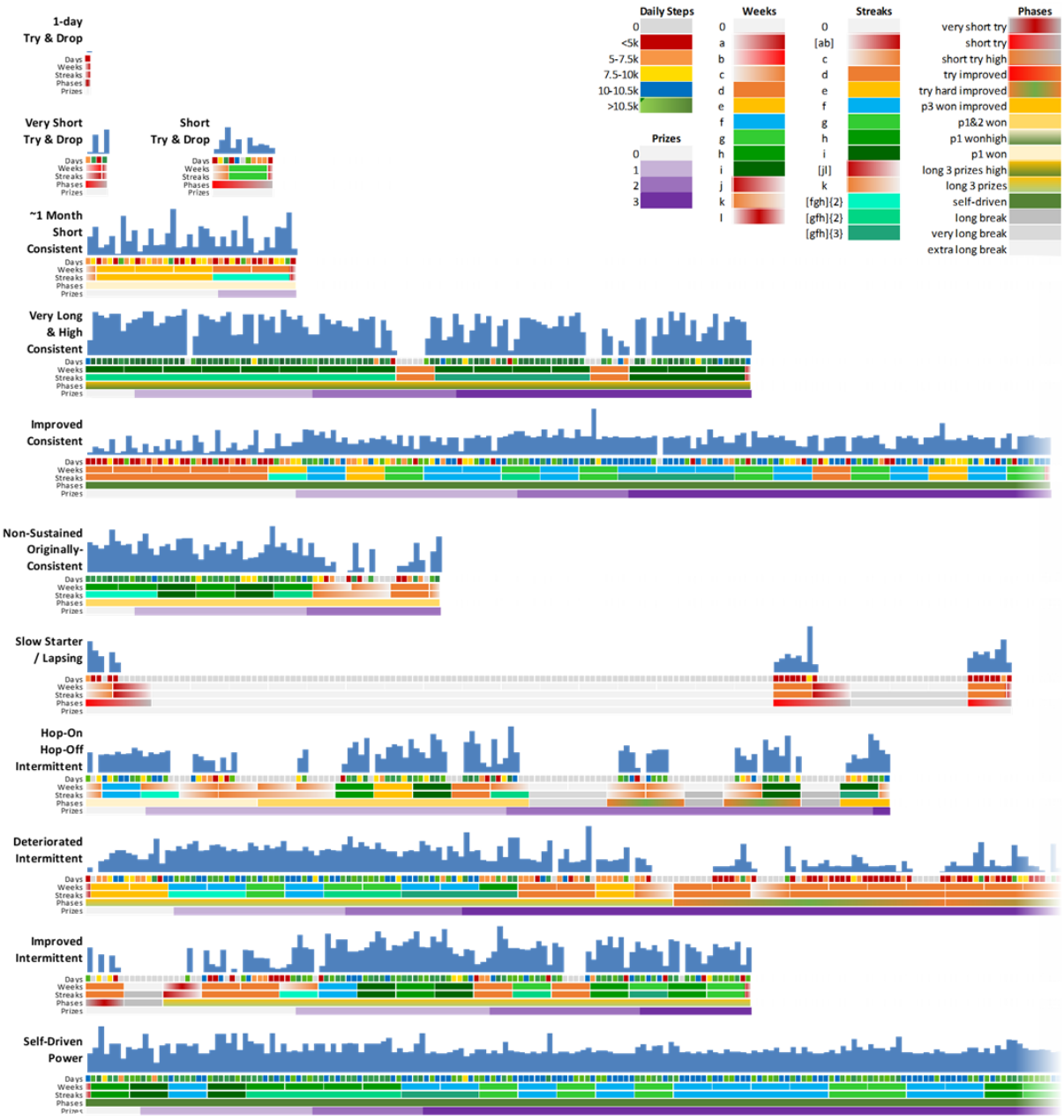


Fig. 8. Time series of users from 12 selected user clusters (4 omitted for brevity). This shows the diversity of step activity. For each user, the bar chart indicates the daily step count (0-50k) from the user's first to last Cohort Day; the multi-color rows below indicate from top to bottom: the daily step count, weeks, streaks, phases, and prizes won. The width is cropped at ~75 days of user participation.

Table 5. Summary of 16 user clusters (segments), with size of each segment (% Users), the median step count, duration of participation (#Days), breaks, a heatmap of the frequency of the phase descriptors (as in the last section), heatmap of change in step count from their 1st to last week, gender expressed as the %age of female users (overall=60%), age and BMI.

						# Phase Type												Diff %Steps Last-First Week							
	User Cluster	% Users	Median Active Steps	# Days	# Breaks	very short try	short try	short try high	try improved	try hard improved	p3 won improved	p1&2 won	p1 won high	p1 won	long 3 prizes high	long 3 prizes	self-driven	long break	very long break	extra long break	≥500	≥5k	≥10k	≥15k	≥20k	G=F	Age	BMI
0	Not clustered	54%	9,829	50	2.8																-2%	-1%	-1%	0%	0%	60%	40.3	20.8
T1	1-day Try & Drop	4.8%	5,668	1	0.0																0%	0%	0%	0%	0%	59%	31.7	22.9
T2	Very Short Try & Drop	7.3%	7,647	4	0.2																0%	0%	0%	0%	0%	54%	31.8	22.8
T3	Short Try & Drop	4.4%	8,689	9	0.7																0%	0%	0%	0%	0%	59%	31.2	21.9
C1	~1-Month Short Consistent	10.3%	9,059	38	2.8																-3%	-3%	-3%	-1%	0%	54%	32.4	22.8
C2	~2-Month Medium Consistent	11.5%	8,997	50	2.6																-3%	-3%	-3%	-1%	0%	60%	54.2	23.2
C3	~3 Month Long Consistent	3.5%	10,043	93	6.7																-10%	-10%	-10%	-3%	-1%	63%	31.8	22.2
C4	~4 Month Very Long Consistent	3.7%	10,892	124	6.5																-7%	-7%	-8%	-1%	0%	56%	31.7	22.5
C5	Very Long & High Steps Consistent	4.3%	13,222	127	5.2																-13%	-14%	-16%	-7%	-5%	61%	42.1	23.3
CO	Overweight Consistent	11.5%	8,479	61	3.7																-5%	-5%	-4%	-1%	0%	62%	37.0	30.5
CI	Improved Consistent	4.6%	11,844	108	4.8																0%	6%	15%	9%	5%	70%	41.4	22.4
CD	Non-Sustained Originally-Consistent	2.3%	10,577	140	6.9																-16%	-24%	-38%	-15%	-6%	59%	37.2	22.8
SS	Slow Starter / Lapsing	5.6%	6,960	130	9.2																-9%	-8%	-4%	-1%	0%	65%	37.2	22.6
HH	Hop-On Hop-Off	3.1%	10,810	156	10.6																-26%	-25%	-23%	-5%	-1%	61%	38.5	23.6
ID	Deteriorated Intermittent	8.9%	9,205	99	7																-32%	-30%	-22%	-5%	-1%	60%	37.2	23.0
II	Improved Intermittent	6.9%	9,022	111	6																27%	26%	17%	4%	1%	63%	39.4	23.8
P	Self-Driven Power	7.3%	11,640	196	6																-5%	-6%	-9%	-4%	-2%	61%	45.4	22.8

- **Self-Driven Power users** who tracked well beyond winning the 3rd prize, tended to have 1 self-driven Phase with many Streaks, have very high adherence (>500 steps), are occasionally somewhat incentivized (14% of days with 10-10.5k steps). There are multiple interpretations of this segment. They may be people who embraced the program from the start and continued to find use tracker beyond the prizes period and demonstrate a sustained high level of activity. They may also be people who were already active before the program but made use of the tracker.

Our clustering analysis has revealed specific segments of activity tracker users, describing their various high activity, lack of activity, improvements or deterioration in steps activity. Opportunities for intervention could target short-term Consistent users to consider how to engage them for longer, and Overweight and Inconsistent users to understand when they have multi-week breaks and determine how to increase their consistency and habituate them towards regular daily high step count, e.g., by taking fewer short break days or increasing the steps of their lowest days weekly. While our analysis stopped at describing 16 user segments, in future work, we could split each cluster into finer groups to identify more specific sub-group behaviors.

8 APPLYING TIME AGGREGATE CLUSTERS TO EXPLAIN STEPS BEHAVIOR

Having developed our framework and derived several time behavioral and user clusters, we now demonstrate their meaningfulness in explaining user steps activity and changes in behavior. Since we did not directly ask users to self-report the reasons for their behavior, our analysis is only retrospective observational and can explain *how* an outcome or activity is achieved instead, not *why*. We repeat the application questions and add clarifying questions for harnessing the clustering results to provide deeper insights:

1. Did the incentives influence and drive steps activity? **How?**
2. Was the steps behavior sustained beyond the incentive period? **How was it (not) sustained?**
3. Did users increase their steps activity? **Can we predict based on the first few days,** which users will improve and sustain their steps activity beyond the incentive period?

We answer these questions in the following subsections. Using our time clusters, we trained decision trees with meaningful descriptions to explain how some users could increase their steps level. Without our behavioral and user clusters, one would be limited to data features from daily step counts and lose the possibility for higher-level understanding. In this section, we evaluate the explanatory and predictive benefits of using the temporal patterns and time-aggregate clusters to analyze user steps behavior. Table 6 summarizes the combinations of data features we investigated. We conducted three evaluations to demonstrate that time clusters can provide new insights explanations by: (1) measuring the performance of classifiers trained with or without time clusters, (2) generating rules and interpreting them, and (3) predicting user behavior from the first days (four weeks) of tracking data. The data features are slightly different for the explanation and prediction modeling. For explanation, we extracted features and counted clusters from the full time period of user participation; but for prediction, we only used data from the first four weeks of the users' data, so this also excludes users who stop tracking within four weeks.

Table 6. Different sets of data features evaluated to show the explanatory and predictive benefits of using temporal patterns and time-aggregate clusters to analyze user steps behavior.

Steps Pattern Features	Description
Daily median steps	Basic data: median daily step count, user age, gender, BMI. These metrics are typically used in public health literature and some HCI/Ubicomp research with quantitative evaluations of interventions (e.g. [10,33]).
+ % Non-active days	Add adherence level of active/non-active days (≥500 steps). This is not commonly considered and was recently emphasized by Tang et al. [57].
+ Steps histogram	Add finer adherence levels as histogram (≥5k, 10k, 15k, 20k). considering more adherence levels from [57], this is also uncommon, since most analysis assumes a Gaussian distribution in step count and typically reports the mean, which we showed in Fig. 1 is insufficient to describe the steps patterns.
+ Week	Add week clusters. This adds features describing the occurrences of each week type, e.g., %(*a*), %(*f*), %(*i*).
+ Week – Steps histogram	Add week clusters, but without daily steps histogram. There is some redundancy in the week clusters containing information about the histogram of step count for each weekday. To investigate the influence of the steps histogram, this feature set that excludes the steps histogram features compared to the Week pattern features (previous row).
+ Streak	Add streak cluster % occurrences, e.g., %(*gfh2*), %(*gfh3*).
+ Phase	Add phase clusters % occurrences, e.g., %(*p1wonhigh*), %(*p1&2won*)
+ User	Add the user cluster, which indicates if the user is in the *Consistent*, *Intermittent*, etc., segment.

We split our user dataset into training (75%) and test (25%) subsets. Since our dataset is imbalanced with respect to the number of users who increased their step counts, we use the area under the precision-recall curve (PR AUC) as our model performance measures instead of accuracy or F1 score [51]. Moreover, F1 scores are less informative, since they are calculated at as specific probability threshold, while PR AUC considers all possible thresholds. We include some F1 score calculations in the Appendix. Furthermore, because of our key goal to understand users' behavior, we chose to keep the dataset representative and did not balance the training dataset. Regarding the prediction task (Question 3), we can balance the classifier decision by applying a cost matrix that can specify a higher prior probability to the majority class or specify which class is costlier to misclassify; this changes the decision probability threshold. We focus on the decision tree explainer because its rules are interpretable and can help derive insights for designing new interventions. We kept 25% of the training dataset as the validation dataset, i.e., 25% of 75% (18.75%) of the total dataset. We built and pruned the decision tree until

the validation accuracy no longer improved. For generalization, we also trained random forest models and found that the relative performance for different data features was similar to that of decision trees. For brevity, we report key results in the rest of this section (with further results in the Appendix). The key takeaway, as shown in Table 7 is that adding our time-aggregate clusters improves the performance of various explainer models. Also notable is that just adding daily steps histogram (fine-grained adherence levels) increased model performance. This may be higher than models trained with behavioral clusters, perhaps, due to the variability of each cluster. However, this comes at a cost of interpretability, based on many splits due to steps levels rather than interpretable time-aggregate clusters.

Table 7. Performance (area under precision-recall curve) of different models trained on different data features to describe user behavior on winning incentive prizes, sustaining activity beyond incentives (Post-P3 2w Median Steps>10k), and increasing steps from the beginning (Diff 2w Median Steps>2k). This shows that temporal features improve performance.

PR AUC	No Prize		Prize 1 Won		Prize 2 Won		Prize 3 Won		Post-P3 2w Median(steps)>10		Diff 2w Median(steps)>2k	
Model with features	DT	RF	DT	RF	DT	RF	DT	RF	DT	RF	DT	RF
Daily median steps	0.987	0.986	0.844	0.826	0.504	0.510	0.906	0.898	0.789	0.753	0.624	0.623
+ % non-active days	0.998	0.997	0.970	0.945	0.745	0.640	0.979	0.958	0.873	0.862	0.693	0.716
+ steps histogram	0.998	0.999	0.976	0.981	0.775	0.791	0.979	0.981	0.885	0.898	0.671	0.723
+ Week w/o histogram	0.998	0.998	0.970	0.958	0.745	0.678	0.979	0.972	0.912	0.914	0.686	0.724
+ Week	0.998	0.999	0.976	0.982	0.774	0.786	0.979	0.981	0.888	0.890	0.669	0.711
+ Streak	1.000	1.000	0.994	0.996	0.955	0.960	0.996	0.998	0.904	0.911	0.760	0.775
+ Phase	1.000	1.000	1.000	0.999	0.999	0.994	1.000	0.999	0.910	0.918	0.779	0.805
+ User	1.000	1.000	1.000	1.000	0.999	0.997	1.000	1.000	0.910	0.921	0.779	0.778

Post-P3 2w Median(steps)>10k: whether the user won prize 3 and had median active step count >10k steps/day during the two weeks after winning that last prize.

Diff 2w Median(steps)>2k: whether the user won prize 3 and had median active step count during the next two weeks with difference of at least 2k steps/day *higher* than during her first two weeks.

DT: Decision Tree, RF: Random Forest

8.1 Did the Incentives Influence and Drive Steps Activity? How Were Their Steps Activity Driven?

Yes, we found that incentives influenced steps behavior such that some users walked to a specific steps level per day and/or took breaks or abandoned tracking around when they won incentive rewards (prizes). We now discuss how analysis at each time aggregation level can reveal behaviors driven by incentives. We observed a spike in **daily** step counts (7% of days) just above 10k steps and modeled this as incentive sufficiency, steps in the range of 10-10.5k. We call these "incentive 10k" days. Subsequently, we found that 11.4% of **weeks** had about 45% of their days as incentive 10k (week type *f*). At the **streak** level, we found prize-winning streaks ([*fgh*]{2}, [*gfh*]{2}, [*gfh*]{3}; see footnote 3) that had more incentive 10k days. Furthermore, these streaks ended soon after the prize was won. This indicates a negative effect that once incentives end, their positive effects are typically not sustained. We also observed this temporal influence of incentives at the **phase** level, where some users delayed taking a long break until they won all three prizes (e.g., *long 3 prizes high* phase), and users did not sustain their high steps beyond the winning prize of 1 onto later prizes. Finally, at the **user** level, we observed that users who won all three prizes tended, obviously, to be consistent in high steps and tended to track for longer periods. On the other hand, the Hop-On Hop-Off users did win prizes, but in a very interrupted manner with short and long breaks.

8.2 Was the Steps Behavior Sustained After Incentives Are Withdrawn? How Was It (not) Sustained?

Partially yes, for a few users who ended with the *self-driven* phase type, which agrees with prior studies showing that increases in steps level has limited sustainability for most people [19]. For details, we focus on the two previously identified improved user types, *Improved Consistent* & *Improved Inconsistent*. *Improved Consistent* users ended their tracking with Phases *self-driven* (19%), *p1 won* (18%), *p1&2 won* (17%), *long 3 prizes* (15%), *p1 won high* (14%) and *try motivated* (12%). Surprisingly, this cluster had the highest proportion of women (70% vs. baseline of 60%), suggesting that women are slightly more receptive to interventions. *Improved Inconsistent* users ended their tracking with Phases *try improved* (21%), *p1 won* (13%), *long 3 prizes* (15%), *try motivated* (11%), *self-driven* (10%), *p1&2 won* (10%). It is encouraging to see some users ending with the *self-driven* Phase, but most of these phases are prize-driven and end soon after the incentive is achieved (non-sustained). Finally, users ending with *try motivated* ended on a short burst of high steps activity, which is even less sustained.

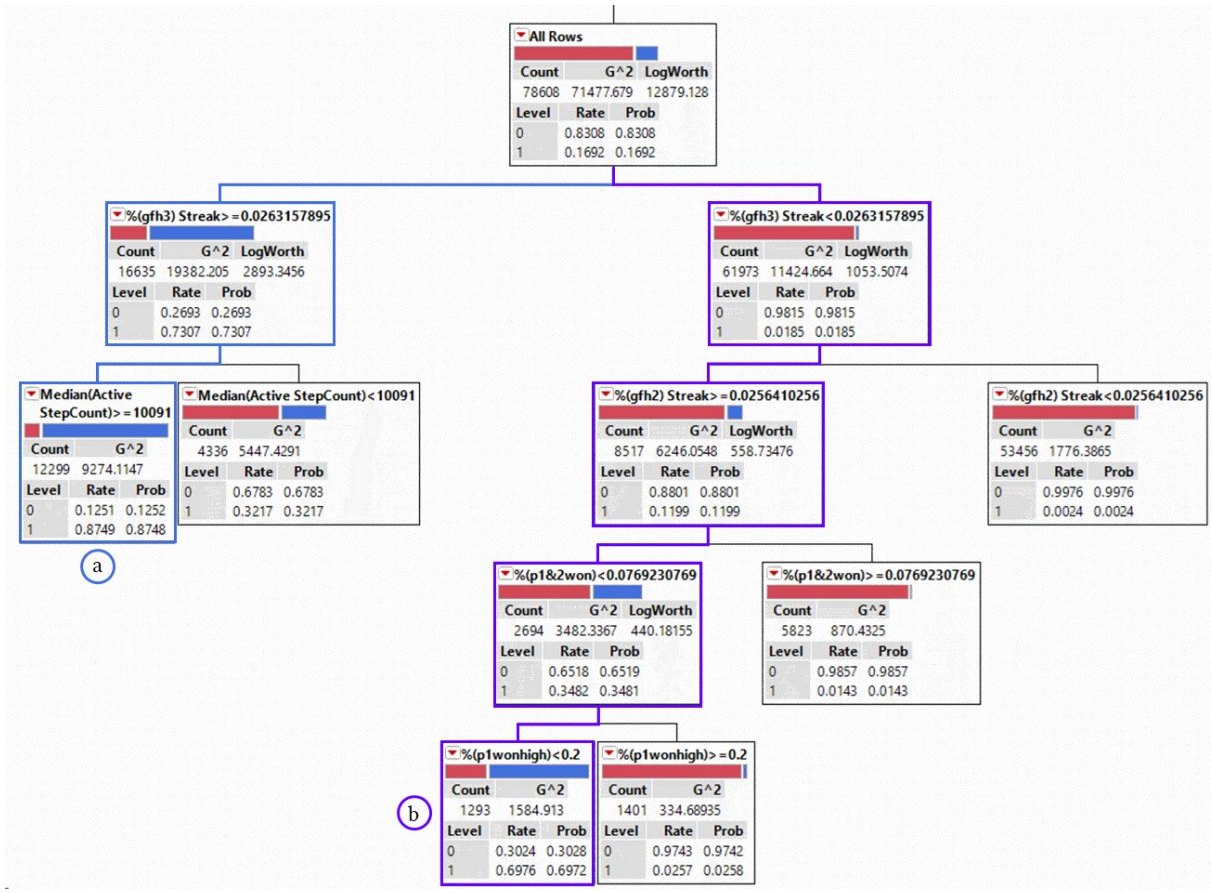


Fig. 9. Partial (pruned) view of trained decision tree with all steps and cluster features for describing which users participated until they won Prize 3 and sustained a median daily step count of at least 10k steps/day in the following two weeks, i.e., Post-P3 2w Median(steps)>10k. Counts show the number of users at each node. Paths to leaves ⓐ and ⓑ are highlighted to illustrate reasons for users to sustain their step count.

Fig. 9 shows a partial decision tree to extract some explanation rules with Weeks, Streaks, Phase and User clusters. Highlighting some key leaves, we see that users are more likely to sustain 10k steps/day if they:

- Match the rules of Leaf ⓐ in the figure, having >2.6% of their streaks as the long and high effort *gfh3* type and high median active step count (at least 10,091 steps/day). This suggests that they may have maintained high step count throughout the intervention too.
- Match Leaf ⓑ, having <20% of their phases as *p1wonhigh*, <7.7% phases as *p1&2won*, >2.6% of their streaks as the moderately long and high effort *gfh2* type. This suggests that these users did not burn out at the start by being too intensive in time or steps level, and sustained to win prize 3 and beyond.

8.3 Did Users Increase Their Steps Activity? Can We Predict Which User Will Improve and Sustain Their Steps Activity Beyond the Incentive Period?

Yes, we had identified increases in steps activity in some phase types (*try improved*, *try hard improved*, *p3won improved*) and for two user types (*Improved Consistent* and *Improved Inconsistent*). In *try improved* phases, users could more easily increase their steps (from Median active steps of 4.9k to 6.9k) because they started with low initial steps. Unfortunately, the other clusters of phases represented temporary increase in steps, namely, for *try hard improved* and *p3won improved*, users had increased their steps just to reach prize 1 and prize 3, respectively, and had multi-week breaks or dropped out after the phase, respectively.

Table 8. Performance of different models to predict user behavior from the first four weeks (4w) of steps data.

PR AUC	No Prize		Prize 1 Won		Prize 2 Won		Prize 3 Won		Post-P3 2w Median(steps)>10		Diff 2w Median(steps)>2k	
Model with features	DT	RF	DT	RF	DT	RF	DT	RF	DT	RF	DT	RF
Daily median steps	0.987	0.986	0.844	0.826	0.504	0.510	0.906	0.898	0.789	0.753	0.624	0.623
+ % non-active days	0.998	0.997	0.970	0.945	0.745	0.640	0.979	0.958	0.873	0.862	0.693	0.716
+ steps histogram	0.998	0.999	0.976	0.981	0.775	0.791	0.979	0.981	0.885	0.898	0.671	0.723
+ Week w/o histogram	0.998	0.998	0.970	0.958	0.745	0.678	0.979	0.972	0.912	0.914	0.686	0.724
+ Week	0.998	0.999	0.976	0.982	0.774	0.786	0.979	0.981	0.888	0.890	0.669	0.711
+ Streak	1.000	1.000	0.994	0.996	0.955	0.960	0.996	0.998	0.904	0.911	0.760	0.775
+ Phase	1.000	1.000	1.000	0.999	0.999	0.994	1.000	0.999	0.910	0.918	0.779	0.805
+ User	1.000	1.000	1.000	1.000	0.999	0.997	1.000	1.000	0.910	0.921	0.779	0.778

Post-P3 2w Median(steps)>10k: whether the user won prize 3 and had median active step count >10k steps/day during the two weeks after winning that last prize.

Diff 2w Median(steps)>2k: whether the user won prize 3 and had median active step count during the next two weeks with difference of at least 2k steps/day *higher* than during her first two weeks.

DT: Decision Tree, RF: Random Forest

We define increased steps activity as an increase of the daily average step count by 2k steps/day [60], and limit our analysis to the difference between the first two weeks of tracking and two weeks after all incentives are won (i.e., prize 3 won, so it is beyond the incentive period). We found that only 5.9% of all users reached the prize 3 milestone *and* increased their steps level by >2k steps; only 23.1% of all users won prize 3. We found that from models trained on the full user data identified longer streaks as influential to modeling sustained and improved steps activity (see Fig. 9). However, this requires several days or weeks of tracking the user to observe. Therefore, we limit the dataset to just the first four weeks of steps activity and included week and streak types. We applied an expanding time window (i.e., 1w, 1-2w, 1-3w, 1-4w) to determine the streak types for each subsequent week. Note that at runtime, we will not know if a user has stopped tracking or is just taking a long break, hence we do not omit users from the dataset who stopped tracking within 4 weeks. Our results (see Table 8) indicate that with histogram and clusters, our model is able to predict user steps maintenance and improvement better than with just basic data features.

9 DISCUSSIONS, LIMITATIONS AND FUTURE WORK

Our work is the first to analyze a population-scale activity tracking dataset using unsupervised and supervised machine learning techniques to quantitatively characterize steps activity in terms of common patterns. By analyzing time series step count data, we have characterized users not only by their profile information (e.g., age, gender, total steps), but also by their daily behaviors. Therefore, our clustering analyses discovered archetypes of steps and wearing behavior for different user segments. For example, highly active Power Users may occasionally have 1-day streaks of low step count. We discuss about users and their behaviors regarding opportunities for health behavior change intervention. We identify limitations, which influence the scope of interpreting our results, and discuss potential for future work.

9.1 Generalizing the Scope of the Framework for Analysis of Large-Scale Tracking Data

We have proposed a framework based on defining dimensions and time aggregation levels, and provided a multi-tier, multi-step processing and analysis method to model steps behavior and different granularity. Our choice of data features and factors can influence the outcome of our clustering results, but as an exploratory analysis method, our method's contribution is the ability to find interesting behaviors for further study. Yet, many of our clusters align with patterns observed in prior work [14,40,56]. In addition, our formal procedure is more transparent than purely qualitative coding techniques that are more subjected to interpretation bias from researchers.

In particular, the foundation of our work is the recognition of the importance of the *week as a fundamental unit of analysis*. We identified distinctive weekly patterns that can be easily applied (e.g., with kNN classification) to other datasets to identify the occurrences of specific week types. The week clusters are also applicable to small datasets that span fewer months (e.g., 12 weeks), since they would still have enough weeks to be classified or clustered. Similarly, our method of determining streaks would apply as long as weeks are repeated. However, our framework may not help with deriving longer phase types specific to a new dataset. Nevertheless, being able to cluster on weeks may provide higher level insights about the dataset.

Furthermore, we intend for analysts to apply the analysis method and/or extend the framework for their own large steps datasets. Many steps tracker datasets may not have incentives or the same types of incentives. Analysts may delete the incentive dimension or retain that theme and modify its themes. Other incentive programs may have different daily or cumulative goals, and these will likely produce different Phase clusters. Datasets also can have different time scales. For example, analyzing intra-day or hourly data [29,41,57], can give further insights into segments of users in terms of when they are physically active, such as only when commuting to/from work, or exercising in the evening. Intra-day data is also useful to further segment users by when they use the wearable during the day (e.g., [28]). On the end of the time scale, one can analyze even longer-term monthly and yearly time cycles [14,56]. Adding these additional time scales may require more aggregation steps, but may be able to detect long-term changes in lifestyle or health, such as developing a chronic disease or physical decline. These time scales may suggest new behavioral and user clusters, and lead to more clusters. However, they may also not be dominant ones when applying hierarchical clustering.

9.2 Insights from Large-Scale Multi-Factorial Data Analysis

Previous work has primarily looked only at a handful of variables, such as usage and engagement (e.g., [29,41]). We have analyzed user steps behavior by five dimensions covering 17 total themes, 55 factors, and 150 data features. The multitude of features, factors, themes and dimensions paints a very rich picture of steps behavior, but also makes it cumbersome to interpret the patterns. Hence, our clustering allows us to identify key prominent and interesting behavior patterns that span arbitrary time lengths. This provides a *more manageable* richer picture and added *flexibility* to understand steps behavior

Compared to their smaller-scale qualitative analysis, our large-scale quantitative analysis can reveal more nuanced clusters, special cases due to analysis on more variables, and quantify the prevalence of different user

types. For example, while Meyers and colleagues found one group of generally consistent users [41], we identified 8 subtypes of consistent users who varied by tracking duration (*C1* to *C5*), as overweight (*CO*) and who improved (*CI*) or deteriorated (*CD*). With our large dataset, we can also identify patterns of users, even if they cover a small portion of the sample. For example, we segmented 5.6% (6,960) of users as Slow Starters (*SS*). This replicates some findings by Meyers and colleagues [41], who only found 5 out of 104 users (5%).

9.3 Generalizations of Steps Behaviors to Other Populations

As behaviors and user segments may vary in different settings, our results may differ if applied to populations, due to the various factors: weather and climate, walkability in the built environment, and car ownership. *Weather and Climate:* The high temperature and humidity year-round in Singapore (Mean=28.2℃/82.8°F [38]) limits outdoor physical activity [3], and the lack of seasons, excludes seasonal effects on physical activity levels. However, in countries with seasons, we expect higher physical activity in the summer than winter [37]. *Walkability:* The built environment can greatly affect the physical activity levels of a population. Compared to rural or suburban locations, cities are typically much more densely packed with amenities (e.g., shops, schools, and apartment buildings) and well-connected by public transportation. Even among cities, there are varying degrees of walkability [31,62] due to factors such as sidewalk access, street connectivity, and land-use density and diversity [35], pedestrian safety and security, convenience and attractiveness, and public policy [31]. Indeed, cities with high walkability support a higher level of physical activity [1]. Finally, the cost of *car ownership* and road tolls are very high in Singapore [5,21], and this strongly limits the amount of driving in the city. These factors contribute to the walking culture in Singapore; the walkability in Singapore is relatively high (~70-80[5]) and is comparable to cities like Boston and Vancouver [63].

9.4 Incentized Steps Behavior

The incentives provided in the NSC encouraged targeted levels of step counts (median steps spikes at 5k, 7.5k, 10k); this is not the case when consumers purchase fitness trackers for personal use [14], or without such incentives [40]. Therefore, we expected participants in our study to have a higher than natural step counts, and with more users just achieving the minimum for each incentive target level. Indeed, users not explicitly given a goal incentive had lower step counts; Althoff et al. reported an average of 4,961 steps for 717,527 users of a freely downloadable commercial mobile app for steps tracking [1]. Furthermore, we observed users *time shifting* their activity with as they were close to winning the next prize, by temporarily increasing steps level, then taking a break soon after winning a prize. To reuse our framework for non-incentivized tracker data, data features and factors related to incentives should be omitted.

There is a selection bias in the recruitment, since participants were recruited specifically for health promotion. Similarly, selection bias is also present for studies of people who purchase fitness trackers to monitor and improve their health [15,17,57]. In contrast, there may be a selection bias against health motivation, due to the much larger recruitment scope, and since the participants were given free activity trackers to encourage participation. Nevertheless, in spite of the difference in behaviors from incentivization, such interventions may play an important role to habitualize people to developing healthy physical activity routines [20].

9.5 Variability in Data-Driven Principle Component Analysis and Clustering

The data mining methods employed allowed us to aggregate large volumes of data into fewer factors (PCA) and groups (clustering); this helps with interpreting general characteristics of behavior or user archetypes. However, there remains variability where some data points (users, phases, streaks) may appear to belong to one or more clusters. Given the relatively high dimensionality of the data features, the reason for assigning a data point to a cluster may not be obvious either. To ensure sufficiently meaningful clustering, we verified the cluster tendency

[5] https://www.walkscore.com/score/the-addison-at-south-tryon-charlotte-8

of our data (all Hopkins Statistics < 0.5) [2] and determined the number of clusters using objective metrics (elbow method of scree plots). Furthermore, since we use PCA primarily for dimensionality reduction, the assumptions can be somewhat relaxed, such as linearity (all pairwise features are linearly related) or that the principal components are orthogonal [53]. While this does not invalidate our analysis, it leads to a sub-optimal modeling of our steps data. Instead, more sophisticated methods could be used for dimensionality reduction, such as Independent Component Analysis (ICA), kernel PCA, and non-metric multidimensional scaling (NMDS), and non-linear data transformation methods that consider skewed, count and zero-inflated variables.

This data-driven clustering approach differs from rule-based segmentation, which may provide more interpretable partitioning of users and their behaviors. For example, to explicitly identify users who were driven by incentives, one could select participants who ended a Streak within 1-5 Active Days after winning a prize. Our clustering results did not specifically isolate this group of Streaks, because there were other more influential factors for grouping Streaks. One could also identify incentive-driven users by selecting users with median step counts within 10-10.5k steps. However, rule-based partitioning is less robust than machine learning methods when handling many variables or data features.

10 CONCLUSION

Our analysis is the first to analyze a massive dataset of 140,000 users of fitness trackers in a 10-month incentives-driven physical activity intervention program. Given the complexity of the large dataset, we developed and applied a framework of dimensions and themes, and multi-tiered time aggregation analysis method to interpret and cluster steps behavior. We have segmented users into 16 user types with 15 phase types, 14 streak types and 13 week types of step behaviors. We have demonstrated how these clusters can be used to provide behavioral insights into different steps outcomes for an incentive-driven activity tracker application. As more tracker-backed physical activity interventions are deployed and big data becomes commonplace, this approach can help researchers to model user steps activity in a scalable way and gain qualitative insights from semantically meaningful clusters of behavioral and user segments.

ACKNOWLEDGMENTS

We thank Dr Mathia Lee for help in discussions about the National Steps Challenge campaign and dataset. This work was carried out at the NUS Institute for Health Innovation and Technology (iHealthtech).

REFERENCES

[1] Tim Althoff, Jennifer L Hicks, Abby C King, Scott L Delp, Zuckerberg Biohub, and San Francisco. 2017. Large-scale physical activity data reveal worldwide activity inequality. *Nature* (2017). DOI:https://doi.org/10.1038/nature23018.Large-scale

[2] Amit Banerjee and Rajesh N. Davé. 2004. Validating clusters using the Hopkins statistic. In *IEEE International Conference on Fuzzy Systems*. DOI:https://doi.org/10.1109/FUZZY.2004.1375706

[3] Catherine B. Chan and Daniel A. Ryan. 2009. Assessing the effects of weather conditions on physical activity participation using objective measures. *Int. J. Environ. Res. Public Health* 6, 10 (2009), 2639–2654. DOI:https://doi.org/10.3390/ijerph6102639

[4] Yu Chen and Pearl Pu. 2014. HealthyTogether: Exploring Social Incentives for Mobile Fitness Applications. In *In Proceedings of Chinese CHI '14*, 25–34. DOI:https://doi.org/10.1145/2592235.2592240

[5] Anthony Chin and Peter Smith. 1997. Automobile ownership and government policy: The economics of Singapore's vehicle quota scheme. *Transp. Res. Part A Policy Pract.* 31, 2 (1997), 129–140. DOI:https://doi.org/10.1016/S0965-8564(96)00012-2

[6] James Clawson, Jessica A Pater, Andrew D Miller, Elizabeth D Mynatt, and Lena Mamykina. 2015. No longer wearing: investigating the abandonment of personal health-tracking technologies on craigslist. In *Proceedings of the 2015 International Conference on Ubiquitous Computing*, 647–658. Retrieved April 15, 2019 from https://dl.acm.org/citation.cfm?id=2807554

[7] Sofie Compernolle, Corneel Vandelanotte, Greet Cardon, Ilse De Bourdeaudhuij, and Katrien De Cocker. 2015. Effectiveness of a Web-Based, Computer-Tailored, Pedometer-Based Physical Activity Intervention for Adults: A Cluster Randomized Controlled Trial. *J. Med. Internet Res.* 17, 2 (February 2015), e38. DOI:https://doi.org/10.2196/jmir.3402

[8] S Consolvo, DW McDonald, … T Toscos - Proceedings of the, and undefined 2008. Activity sensing in the wild: a field trial of ubifit garden. *dl.acm.org*. Retrieved April 15, 2019 from https://dl.acm.org/citation.cfm?id=1357335

[9] S Consolvo, DW McDonald, JA Landay - Proceedings of the SIGCHI, and undefined 2009. Theory-driven design strategies for technologies that support behavior change in everyday life. *dl.acm.org*. Retrieved April 15, 2019 from

https://dl.acm.org/citation.cfm?id=1518766
[10] Sunny Consolvo, Katherine Everitt, Ian Smith, and James A. Landay. 2006. Design requirements for technologies that encourage physical activity. In *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*, 457. DOI:https://doi.org/10.1145/1124772.1124840
[11] Sunny Consolvo, Predrag Klasnja, David W. McDonald, Daniel Avrahami, Jon Froehlich, Louis LeGrand, Ryan Libby, Keith Mosher, and James A. Landay. 2008. Flowers or a robot army?: encouraging awareness & activity with personal, mobile displays. In *Proceedings of the 10th international conference on Ubiquitous computing - UbiComp '08*, 54. DOI:https://doi.org/10.1145/1409635.1409644
[12] Felicia Cordeiro, Daniel A. Epstein, Edison Thomaz, Elizabeth Bales, Arvind K. Jagannathan, Gregory D. Abowd, and James Fogarty. 2015. Barriers and Negative Nudges: Exploring Challenges in Food Journaling. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, 1159–1162. DOI:https://doi.org/10.1145/2702123.2702155
[13] Department of Statistics, Ministry of Trade & Industry, Republic of Singapore. 2017. *Population Trends 2016*. Retrieved from https://www.singstat.gov.sg/-/media/files/publications/population/population2016.pdf
[14] Daniel A. Epstein, Jennifer H. Kang, Laura R. Pina, James Fogarty, and Sean A. Munson. 2016. Reconsidering the device in the drawer. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '16*, 829–840. DOI:https://doi.org/10.1145/2971648.2971656
[15] Daniel A. Epstein, Nicole B. Lee, Elizabeth Bales, James Fogarty, and Sean A. Munson. 2015. Wearables of 2025: Designing Personal Informatics for a Broader Audience. In *CHI 2015 Workshop on "Beyond Personal Informatics: Designing for Experiences of Data."* Retrieved April 15, 2019 from https://homes.cs.washington.edu/~jfogarty/publications/workshop-chi2015.pdf
[16] Daniel A. Epstein, An Ping, James Fogarty, and Sean A. Munson. 2015. A lived informatics model of personal informatics. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15*, 731–742. DOI:https://doi.org/10.1145/2750858.2804250
[17] Daniel Epstein, Felicia Cordeiro, Elizabeth Bales, James Fogarty, and Sean Munson. 2014. Taming data complexity in lifelogs: exploring visual cuts of personal informatics data. In *Proceedings of the 2014 conference on Designing interactive systems - ACM*, 667–676. DOI:https://doi.org/10.1145/2598510.2598558
[18] Leandre R. Fabrigar, Duane T. Wegener, Robert C. MacCallum, and Erin J. Strahan. 1999. Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Methods* 4, 3 (September 1999), 272–299. DOI:https://doi.org/10.1037/1082-989X.4.3.272
[19] Eric A Finkelstein, Benjamin A Haaland, Marcel Bilger, Aarti Sahasranaman, Robert A Sloan, Ei Ei Khaing Nang, and Kelly R Evenson. 2016. Effectiveness of activity trackers with and without incentives to increase physical activity (TRIPPA): a randomised controlled trial. *Lancet Diabetes Endocrinol.* 4, 12 (December 2016), 983–995. DOI:https://doi.org/10.1016/S2213-8587(16)30284-4
[20] Uri Gneezy, Stephan Meier, and Pedro Rey-Biel. 2011. When and Why Incentives (Don't) Work to Modify Behavior. *J. Econ. Perspect.* 25, 4 (November 2011), 191–210. DOI:https://doi.org/10.1257/jep.25.4.191
[21] Mark Goh. 2002. Congestion management and electronic road pricing in Singapore. *J. Transp. Geogr.* 10, 1 (2002), 29–38. DOI:https://doi.org/10.1016/S0966-6923(01)00036-9
[22] Nanna Gorm and Irina Shklovski. 2016. Steps, Choices and Moral Accounting: Observations from a Step-Counting Campaign in the Workplace. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16*, 148–159. DOI:https://doi.org/10.1145/2818048.2819944
[23] Nanna Gorm and Irina Shklovski. 2017. Participant Driven Photo Elicitation for Understanding Activity Tracking. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, 1350–1361. DOI:https://doi.org/10.1145/2998181.2998214
[24] Rúben Gouveia, Evangelos Karapanos, and Marc Hassenzahl. 2018. Activity Tracking in vivo. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–13. DOI:https://doi.org/10.1145/3173574.3173936
[25] Karen Grace-Martin. The Distribution of Independent Variables in Regression Models - The Analysis Factor. *The Analysis Factor*. Retrieved April 15, 2019 from https://www.theanalysisfactor.com/the-distribution-of-independent-variables-in-regression-models-2/
[26] Rebecca Gulotta, Jodi Forlizzi, Rayoung Yang, and Mark Wah Newman. 2016. Fostering Engagement with Personal Informatics Systems. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems - DIS '16*, 286–300. DOI:https://doi.org/10.1145/2901790.2901803
[27] Pedro C Hallal, Lars Bo Andersen, Fiona C Bull, Regina Guthold, William Haskell, Ulf Ekelund, and Lancet Physical Activity Series Working Group. 2012. Global physical activity levels: surveillance progress, pitfalls, and prospects. *Lancet (London, England)* 380, 9838 (2012), 247–57. DOI:https://doi.org/10.1016/S0140-6736(12)60646-1
[28] Erin K. Howie, Anne L. Smith, Joanne A. McVeigh, and Leon M. Straker. 2018. Accelerometer-Derived Activity Phenotypes in Young Adults: a Latent Class Analysis. *Int. J. Behav. Med.* 25, 5 (2018), 558–568. DOI:https://doi.org/10.1007/s12529-018-9721-4
[29] Hayeon Jeong, Heepyung Kim, Rihun Kim, Uichin Lee, and Yong Jeong. 2017. Smartwatch Wearing Behavior Analysis. *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.* 1, 3 (September 2017), 1–31. DOI:https://doi.org/10.1145/3131892
[30] Predrag Klasnja, Sunny Consolvo, and Wanda Pratt. 2011. How to evaluate technologies for health behavior change in HCI research. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, 3063. DOI:https://doi.org/10.1145/1978942.1979396
[31] Holly Virginia Krambeck. 2006. The global walkability index. Massachusetts Institute of Technology. Retrieved April 15, 2019 from http://hdl.handle.net/1721.1/34409
[32] Ian Li, Anind K. Dey, and Jodi Forlizzi. 2011. Understanding my data, myself: supporting self-reflection with ubicomp technologies. In *Proceedings of the 13th international conference on Ubiquitous computing - UbiComp '11*, 405. DOI:https://doi.org/10.1145/2030112.2030166
[33] James J. Lin, Lena Mamykina, Silvia Lindtner, Gregory Delajoux, and Henry B. Strub. 2006. Fish'n'Steps: Encouraging Physical Activity with an Interactive Computer Game. In *Proceedings of the 8th international conference on Ubiquitous Computing*, 261–278.

DOI:https://doi.org/10.1007/11853565_16
[34] Zhicheng Liu, Yang Wang, Mira Dontcheva, Matthew Hoffman, Seth Walker, and Alan Wilson. 2017. Patterns and Sequences: Interactive Exploration of Clickstreams to Understand Common Visitor Paths. *IEEE Trans. Vis. Comput. Graph.* 23, 1 (January 2017), 321–330. DOI:https://doi.org/10.1109/TVCG.2016.2598797
[35] Ria Hutabarat Lo. 2009. Walkability: what is it? *J. Urban. Int. Res. Placemaking Urban Sustain.* 2, 2 (July 2009), 145–166. DOI:https://doi.org/10.1080/17549170903092867
[36] Jill Luoto and Katherine Grace Carman. 2014. *Behavioral economics guidelines with applications for health interventions*. Retrieved April 15, 2019 from https://publications.iadb.org/en/handle/11319/6503
[37] G. R. McCormack, C. Friedenreich, A. Shiell, B. Giles-Corti, and P. K. Doyle-Baker. 2010. Sex- and age-specific seasonal variations in physical activity among adults. *J. Epidemiol. Community Heal.* 64, 11 (November 2010), 1010–1016. DOI:https://doi.org/10.1136/jech.2009.092841
[38] Meteorological Service Singapore. Climate of Singapore. Retrieved from http://www.weather.gov.sg/climate-climate-of-singapore/
[39] Jochen Meyer, Anastasia Kazakova, Merlin Büsing, and Susanne Boll. 2016. Visualization of Complex Health Data on Mobile Devices. In *Proceedings of the 2016 ACM Workshop on Multimedia for Personal Health and Health Care - MMHealth '16*, 31–34. DOI:https://doi.org/10.1145/2985766.2985774
[40] Jochen Meyer, Jochen Schnauber, Wilko Heuten, Harm Wienbergen, Rainer Hambrecht, Hans Jurgen Appelrath, and Susanne Boll. 2016. Exploring Longitudinal Use of Activity Trackers. In *Proceedings - 2016 IEEE International Conference on Healthcare Informatics, ICHI 2016*. DOI:https://doi.org/10.1109/ICHI.2016.29
[41] Jochen Meyer, Merlin Wasmann, Wilko Heuten, Abdallah El Ali, and Susanne C.J. Boll. 2017. Identification and Classification of Usage Patterns in Long-Term Activity Tracking. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, 667–678. DOI:https://doi.org/10.1145/3025453.3025690
[42] Ruth Miller and Wendy Brown. 2004. Steps and sitting in a working population. *Int. J. Behav. Med.* 11, 4 (December 2004), 219–224. DOI:https://doi.org/10.1207/s15327558ijbm1104_5
[43] Cecily Morrison and Gavin Doherty. 2014. Analyzing Engagement in a Web-Based Intervention Platform Through Visualizing Log-Data. *J. Med. Internet Res.* 16, 11 (November 2014), e252. DOI:https://doi.org/10.2196/jmir.3575
[44] Josée Poirier, Wendy L. Bennett, Gerald J. Jerome, Nina G. Shah, Mariana Lazo, Hsin-Chieh Yeh, Jeanne M. Clark, and Nathan K. Cobb. 2016. Effectiveness of an Activity Tracker- and Internet-Based Adaptive Walking Program for Adults: A Randomized Controlled Trial. *J. Med. Internet Res.* 18, 2 (February 2016), e34. DOI:https://doi.org/10.2196/jmir.5295
[45] James O. Prochaska and Wayne F. Velicer. 1997. The transtheoretical model of health behavior change. *Am. J. Heal. Promot.* (1997). DOI:https://doi.org/10.4278/0890-1171-12.1.38
[46] Rachael Purta, Stephen Mattingly, Lixing Song, Omar Lizardo, David Hachen, Christian Poellabauer, and Aaron Striegel. 2016. Experiences measuring sleep and physical activity patterns across a large college cohort with fitbits. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers - ISWC '16*, 28–35. DOI:https://doi.org/10.1145/2971763.2971767
[47] Tom Quisel, Luca Foschini, Alessio Signorini, and David C. Kale. 2017. Collecting and Analyzing Millions of mHealth Data Streams. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*, 1971–1980. DOI:https://doi.org/10.1145/3097983.3098201
[48] Annette Rauner, Filip Mess, and Alexander Woll. 2013. The relationship between physical activity, physical fitness and overweight in adolescents: A systematic review of studies published in or after 2000. *BMC Pediatr.* (2013). DOI:https://doi.org/10.1186/1471-2431-13-19
[49] Cercos Robert and Mueller Florian 'Floyd. 2013. Watch your steps: designing a semi-public display to promote physical activity. In *Proceedings of The 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death*. DOI:https://doi.org/10.1145/2513002.2513016
[50] Geoffrey Rose, Kay-Tee Khaw, and Michael Marmot. 2008. *Rose's Strategy of Preventive Medicine*. Oxford University Press. DOI:https://doi.org/10.1093/acprof:oso/9780192630971.001.0001
[51] Takaya Saito and Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* (2015). DOI:https://doi.org/10.1371/journal.pone.0118432
[52] Patrick C. Shih, Kyungsik Han, Erika Shehan Poole, Mary Beth Rosson, and John M. Carroll. 2015. Use and Adoption Challenges of Wearable Activity Trackers. In *iConference 2015 Proceedings*. Retrieved April 15, 2019 from https://www.ideals.illinois.edu/handle/2142/73649
[53] Jonathon Shlens. 2014. A Tutorial on Principal Component Analysis. *arXiv* (April 2014). DOI:https://doi.org/10.1.1.115.3503
[54] Wally Smith, Bernd Ploderer, Greg Wadley, Sarah Webber, and Ron Borland. 2017. Trajectories of Engagement and Disengagement with a Story-Based Smoking Cessation App. DOI:https://doi.org/10.1145/3025453.3026054
[55] StepItUpUSA. StepItUpUSA. Retrieved September 14, 2017 from https://www.stepitupusa.org/
[56] Lie Ming Tang and Judy Kay. 2017. Harnessing Long Term Physical Activity Data—How Long-term Trackers Use Data and How an Adherence-based Interface Supports New Insights. *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.* 1, 2 (June 2017), 1–28. DOI:https://doi.org/10.1145/3090091
[57] Lie Ming Tang, Jochen Meyer, Daniel A Epstein, Kevin Bragg, Lina Engelen, Adrian Bauman, and Judy Kay. 2018. Defining Adherence: Making Sense of Physical Activity Tracker Data. *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.* 2, 1 (March 2018), 1–22. DOI:https://doi.org/10.1145/3191769
[58] Richard P. Troiano, David Berrigan, Kevin W. Dodd, Louise C. Mâsse, Timothy Tilert, and Margaret Mcdowell. 2008. Physical Activity in the United States Measured by Accelerometer. *Med. Sci. Sport. Exerc.* 40, 1 (January 2008), 181–188. DOI:https://doi.org/10.1249/mss.0b013e31815a51b3

[59] Stewart G. Trost, Russell R. Pate, James F. Sallis, Patty S. Freedson, Wendell C. Taylor, Marsha Dowda, and John Sirard. 2002. Age and gender differences in objectively measured physical activity in youth. *Med. Sci. Sports Exerc.* 34, 2 (February 2002), 350–355. DOI:https://doi.org/10.1097/00005768-200202000-00025
[60] Catrine Tudor-Locke and David R. Bassett. 2004. How Many Steps/Day Are Enough? *Sport. Med.* 34, 1 (2004), 1–8. DOI:https://doi.org/10.2165/00007256-200434010-00001
[61] Catrine Tudor-Locke, David R. Bassett, Ann M. Swartz, Scott J. Strath, Brian B. Parr, Jared P. Reis, Katrina D. DuBose, and Barbara E. Ainsworth. 2004. A preliminary study of one year of pedometer self-monitoring. *Ann. Behav. Med.* (2004). DOI:https://doi.org/10.1207/s15324796abm2803_3
[62] Walk Score. Walk Score Methodology. Retrieved from https://www.walkscore.com/methodology.shtml
[63] Walk Score. 2017 City & Neighborhood Ranking. Retrieved from https://www.walkscore.com/cities-and-neighborhoods/
[64] Gang Wang, Xinyi Zhang, Shiliang Tang, Haitao Zheng, and Ben Y. Zhao. 2016. Unsupervised Clickstream Clustering for User Behavior Analysis. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, 225–236. DOI:https://doi.org/10.1145/2858036.2858107
[65] Keith F. Widaman. 2007. Common Factors Versus Components: Principals and Principles, Errors and Misconceptions. In *Factor analysis at 100: Historical developments and future directions*. Routledge, 177–203. DOI:https://doi.org/10.4324/9780203936764-14
[66] Rayoung Yang, Eunice Shin, Mark W Newman, and Mark S Ackerman. 2015. When fitness trackers don't "fit." In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15*, 623–634. DOI:https://doi.org/10.1145/2750858.2804269
[67] Sha Zhao, Julian Ramos, Jianrong Tao, Ziwen Jiang, Shijian Li, Zhaohui Wu, Gang Pan, and Anind K. Dey. 2016. Discovering different kinds of smartphone users through their application usage behaviors. DOI:https://doi.org/10.1145/2971648.2971696

Received November 2018; revised February 2019; accepted April 2019

A DATA FEATURES AND ROTATED PRINCIPAL COMPONENT FACTORS

These tables describe various data features extracted from each time aggregation layer. We grouped the features by themes which are specific to dimensions of interest. We applied principal components analysis and post-processed with varimax rotation to derive grouped factors that describe key factors to describe user behaviors. We subsequently used these factors for clustering at each time aggregation level.

A.1 Week Data Features and Factors

			Rotated Principal Component Factors									
			Engagement					Habit			Incentives	
Dimension	Theme	Data Feature	Inconsistent & Very Low Step Count	Moderate Step Count	High Step Count	Very High Step Count	Extremely High Step Count	High Steps Mon & Wkend / Low W-F	Weekend Ratio (Sat)	Weekend Ratio (Sun)	Incentivized 10k Step Count	Time Till Prize
			1	2	3	4	5	6	7	8	9	10
Engagement	Steps Level	Median(Active Steps)	0.01	-0.07	0.21	0.18	0.89	-0.03	-0.01	0.00	-0.02	-0.12
		CoV(Active Steps)	0.93	0.08	0.00	-0.13	0.22	-0.07	0.02	-0.01	-0.13	0.00
	Adherence	%Steps(0)	-0.15	-0.45	-0.56	-0.32	-0.25	0.25	0.02	0.00	-0.39	0.27
		%Steps(1-500)	0.11	0.00	-0.10	0.01	-0.06	-0.08	-0.03	-0.06	-0.02	0.07
		%Steps(500-5k)	0.77	0.03	-0.16	0.09	-0.31	-0.05	0.01	0.01	-0.01	0.05
		%Steps(5k-10k)	0.08	0.95	-0.13	-0.15	-0.15	-0.08	-0.03	-0.02	-0.12	-0.05
		%Steps(10.5k-15k)	-0.20	-0.18	0.94	0.00	-0.02	-0.08	-0.01	0.00	-0.01	-0.16
		%Steps(15k-20k)	-0.05	-0.14	0.04	0.81	0.30	-0.13	0.05	0.00	-0.12	-0.13
		%Steps(20k-)	-0.09	-0.05	-0.12	0.08	0.88	0.03	-0.01	0.01	-0.02	-0.04
Habit & Routine	Pattern	Mon Steps	-0.05	-0.03	-0.06	0.15	-0.12	0.76	-0.16	-0.56	-0.03	-0.09
		Tue Steps	0.01	0.04	0.06	-0.16	0.11	0.08	-0.52	-0.18	0.04	-0.08
		Wed Steps	0.03	0.07	0.07	-0.03	0.06	-0.31	-0.46	-0.09	0.06	-0.05
		Thu Steps	0.04	0.06	0.05	0.09	-0.01	-0.53	-0.22	-0.11	0.03	-0.03
		Fri Steps	0.06	0.01	0.00	0.09	-0.05	-0.60	0.18	-0.16	-0.01	0.01
		Sat Steps	0.03	0.02	0.04	-0.17	0.12	-0.12	0.89	-0.03	0.03	0.05
		Sun Steps	-0.06	-0.09	-0.09	0.00	-0.06	0.25	0.16	0.90	-0.06	0.14
		LogRatio Weekend/Weekday Mean Steps	0.01	0.11	0.16	0.06	0.08	0.02	0.64	0.69	0.11	0.05
		LogRatio Weekend/Weekday %Steps(>500)	0.02	0.05	0.10	0.03	0.05	0.02	0.71	0.65	0.07	0.07
Incentives	Sufficiency	%Steps(10k-10.5k)	-0.14	-0.10	0.03	-0.10	-0.06	-0.03	0.00	0.01	0.95	-0.08
	Recency	Days Since Last Prize @ Start	0.03	-0.01	-0.04	0.02	-0.05	-0.04	0.00	-0.01	-0.06	-0.45
	Proximity	Points till Prize 1	0.02	-0.15	-0.22	-0.10	-0.14	-0.08	0.16	0.17	-0.16	0.71
		Points till Prize 2	0.07	-0.02	-0.12	-0.02	-0.10	-0.02	0.07	0.05	-0.08	0.96
		Points till Prize 3	0.08	0.01	-0.10	-0.01	-0.09	-0.01	0.05	0.03	-0.05	0.98

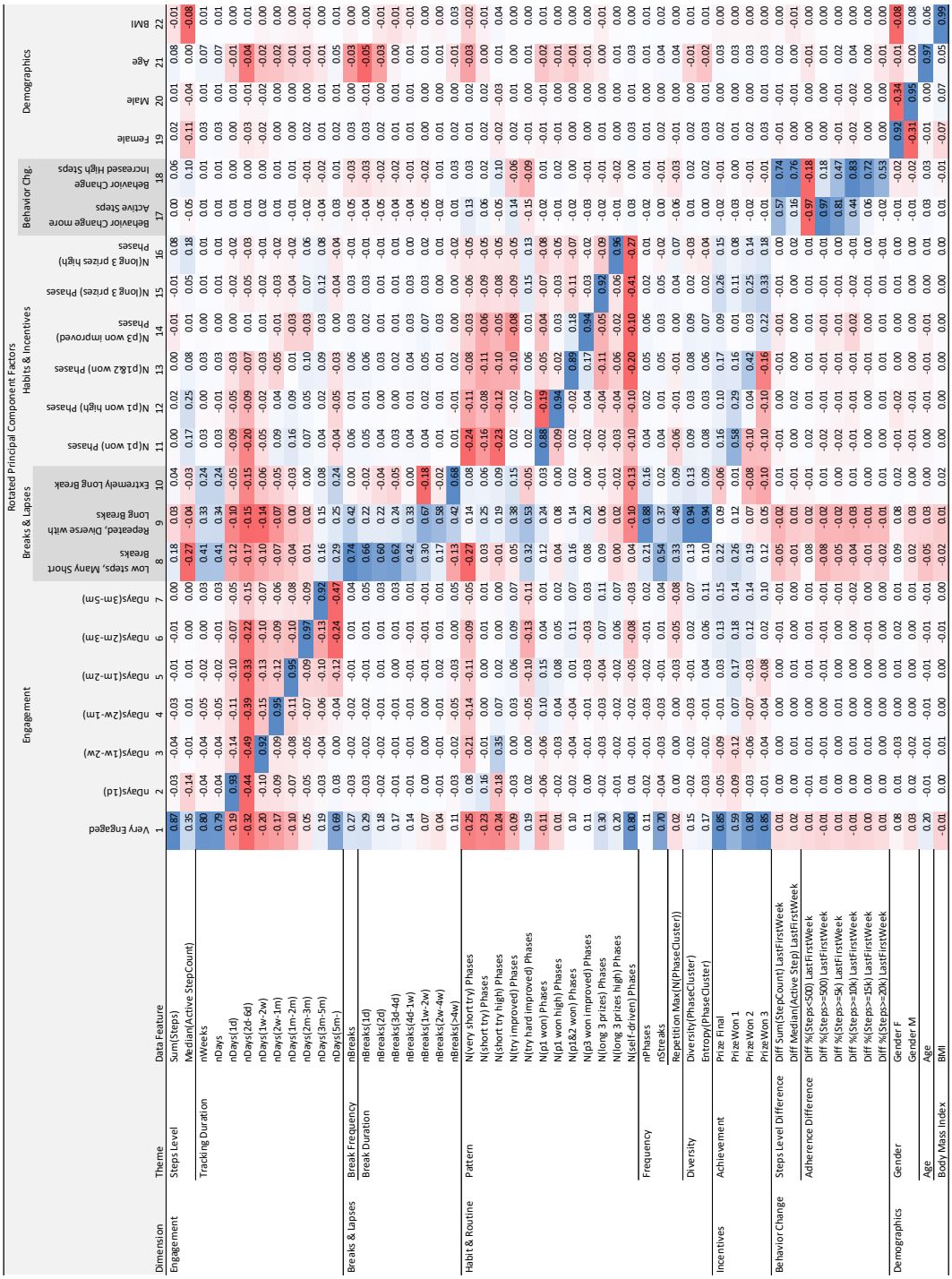
A.2 Streak Data Features and Factors

			Rotated Principal Component Factors												
			Engagement					Habit			Incentives				
			Low Step Count	Moderate Step Count	High Step Count	Very High Step Count	Extremely High Step Count	Mostly Non-Active, Some Weekend Active	Mostly Non-Active, Some Weekday Active	Weekend Ratio	Incentivized 10k Step Count	Time Till Prize	Prize 1 Won	Prize 2 Won	Prize 3 Won
Dimension	Theme	Data Feature	2	3	4	5	6	1	7	8	9	10	11	12	13
Engagement	Steps Level	Median(ActiveSteps)	-0.23	-0.18	0.16	0.33	0.41	-0.08	-0.07	0.01	0.02	-0.15	0.08	0.05	0.08
		Sum(Steps)	-0.06	0.02	0.15	0.18	0.27	-0.07	-0.07	0.05	0.12	-0.05	0.16	0.11	0.21
	Adherence	%Steps(0)	-0.18	-0.39	-0.47	-0.32	-0.25	0.36	0.32	-0.12	-0.34	0.18	-0.08	-0.04	-0.06
		%Steps(1-500)	0.10	0.00	-0.07	-0.04	-0.02	-0.04	-0.03	-0.08	-0.05	0.05	-0.04	-0.03	-0.03
		%Steps(500-5k)	0.55	0.06	-0.12	-0.11	-0.10	-0.10	-0.06	0.01	-0.10	0.12	-0.07	-0.05	-0.07
		%Steps(5k-10k)	0.14	0.79	-0.05	-0.14	-0.11	-0.17	-0.15	0.02	-0.05	-0.04	0.01	-0.01	-0.03
		%Steps(10.5k-15k)	-0.13	-0.12	0.73	0.15	0.09	-0.17	-0.15	0.08	0.13	-0.16	0.07	0.05	0.08
		%Steps(15k-20k)	-0.09	-0.09	0.03	0.79	0.18	-0.06	-0.07	0.05	-0.03	-0.10	0.04	0.03	0.04
		%Steps(20k-)	-0.03	-0.09	-0.05	0.11	0.73	-0.02	-0.03	0.03	-0.06	-0.06	0.04	0.02	0.04
	Tracking Duration	nWeeks	-0.01	0.05	0.08	0.06	0.08	-0.03	-0.04	0.02	0.12	-0.02	0.19	0.14	0.23
Habit & Routine	Pattern	N(a) Weeks	-0.11	-0.12	-0.12	-0.09	-0.07	0.74	-0.11	0.26	-0.09	0.10	-0.01	0.01	0.01
		N(b) Weeks	-0.04	-0.05	-0.06	-0.04	-0.02	0.32	0.04	0.13	-0.04	0.08	-0.05	-0.04	-0.05
		N(c) Weeks	-0.21	-0.27	-0.33	-0.22	-0.17	-0.56	-0.16	0.47	-0.24	0.25	-0.09	-0.05	-0.06
		N(d) Weeks	0.96	-0.06	-0.05	-0.06	-0.04	0.00	-0.02	0.11	-0.06	-0.16	0.05	0.03	0.04
		N(e) Weeks	-0.13	0.94	-0.11	-0.07	-0.08	0.01	-0.01	0.08	-0.08	-0.05	0.04	0.02	0.04
		N(f) Weeks	-0.10	-0.07	-0.07	-0.05	-0.05	-0.03	-0.03	0.07	0.97	-0.06	0.04	0.02	0.03
		N(g) Weeks	-0.15	-0.08	0.94	-0.15	-0.12	-0.03	-0.03	0.08	-0.12	-0.08	0.02	0.03	0.02
		N(h) Weeks	-0.09	-0.07	-0.03	0.98	-0.04	-0.02	-0.02	0.06	-0.09	-0.04	0.03	0.01	0.01
		N(i) Weeks	-0.06	-0.01	-0.02	-0.06	0.98	0.00	-0.01	0.04	-0.05	0.02	0.01	-0.01	-0.02
		N(j) Weeks	-0.07	-0.07	-0.08	-0.06	-0.05	-0.02	0.85	-0.06	-0.06	-0.04	0.03	0.00	-0.03
		N(k) Weeks	-0.12	-0.17	-0.21	-0.14	-0.11	0.04	-0.27	-0.88	-0.15	0.02	-0.04	-0.02	-0.02
		N(l) Weeks	-0.01	-0.03	-0.03	-0.02	-0.02	0.01	0.38	-0.07	-0.02	0.08	-0.05	-0.02	-0.01
		LogRatio Weekend/Weekday Mean Steps	0.03	0.01	-0.01	0.00	-0.01	-0.01	-0.02	0.02	-0.01	-0.50	-0.10	-0.22	-0.26
		LogRatio Weekend/Weekday %Steps(>500)	0.02	0.02	0.04	0.03	0.03	0.26	-0.31	0.89	0.03	0.08	0.00	0.01	0.01
Incentives	Sufficiency	%Steps(10k-10.5k)	-0.09	-0.05	0.09	-0.06	-0.05	-0.07	-0.07	0.05	0.79	-0.07	0.07	0.05	0.06
	Proximity	Points till Prize 1	-0.02	-0.09	-0.14	-0.09	-0.08	0.10	0.03	0.11	-0.09	0.75	-0.03	-0.09	-0.05
		Points till Prize 2	0.03	-0.02	-0.08	-0.06	-0.04	0.05	0.04	0.03	-0.05	0.96	0.08	-0.11	-0.11
		Points till Prize 3	0.05	0.00	-0.06	-0.05	-0.03	0.04	0.04	0.02	-0.04	0.98	0.07	-0.01	-0.11
	Sustainability	Days SinceLastPrize @ Start	0.01	-0.01	0.01	0.01	0.01	0.24	-0.24	0.92	0.01	0.09	0.00	0.00	0.01
	Achievement	PrizeWon 1	-0.05	0.03	0.06	0.06	0.08	-0.05	-0.05	0.01	0.08	0.22	0.96	-0.03	0.02
		PrizeWon 2	-0.04	0.00	0.06	0.03	0.04	-0.04	-0.03	0.02	0.06	-0.03	0.00	0.99	0.00
		PrizeWon 3	-0.04	-0.02	0.03	0.01	0.02	-0.03	-0.01	0.00	0.02	-0.07	-0.04	-0.06	0.98

A.3 Phase Data Features and Factors

			Rotated Principal Component Factors									
			Engag.	Breaks	Habit	Incentives					Behavior C.	
Dimension	Theme	Data Feature	Self-Motivated Long Phase	Many diverse low-steps streaks	Repeating Short Partial Weeks	Time Till Prize	Prize 1 High Burst Effort	Prize 1 Effort	Prize 2 Won	Prize 3 Won	Behavior Change increased Moderate Steps	Behavior Change increased Very High Steps
			1	2	3	4	5	6	7	8	9	10
Engagement	Steps Level	Median(ActiveSteps)	0.17	-0.03	-0.09	-0.05	0.46	0.07	0.11	0.08	-0.08	0.21
		Sum(Steps)	0.85	0.17	-0.04	0.06	0.23	0.09	0.17	0.13	-0.03	0.07
	Tracking Duration	nWeeks	0.86	0.37	-0.01	0.06	0.12	0.15	0.18	0.12	-0.03	0.00
		nDays	0.87	0.36	-0.02	0.03	0.12	0.15	0.18	0.11	-0.01	0.00
Breaks & Lapses	Break Frequency	nBreaks	0.36	0.52	0.04	-0.03	0.01	0.11	0.08	0.00	-0.07	-0.07
	Break Duration	nBreaks(1d)	0.35	0.47	-0.01	0.00	0.02	0.11	0.09	0.01	-0.02	-0.06
		nBreaks(2d)	0.26	0.35	0.05	-0.02	-0.01	0.08	0.05	-0.01	-0.01	-0.05
		nBreaks(3d-4d)	0.25	0.37	0.07	-0.03	-0.01	0.06	0.03	-0.01	-0.02	-0.04
		nBreaks(4d-1w)	0.16	0.28	0.09	-0.07	0.00	0.03	0.01	-0.01	-0.03	-0.03
Habit & Routine	Pattern	N(ab) Streaks	0.01	0.18	-0.01	0.03	-0.05	-0.02	-0.01	0.01	0.09	-0.03
		N(c) Streaks	0.02	0.48	-0.20	0.06	0.16	0.00	-0.02	0.01	-0.16	0.09
		N(d) Streaks	0.49	0.48	0.00	-0.03	-0.07	0.05	0.05	-0.01	0.04	-0.08
		N(e) Streaks	0.41	0.41	0.00	0.01	-0.08	0.11	0.07	0.00	0.03	-0.09
		N(f) Streaks	0.60	0.19	-0.01	0.02	0.00	0.09	0.11	0.08	0.01	-0.01
		N(g) Streaks	0.70	0.23	-0.01	0.03	0.04	0.09	0.11	0.06	-0.01	0.05
		N(h) Streaks	0.60	0.14	-0.01	0.02	0.17	-0.02	0.06	0.05	-0.02	0.08
		N(i) Streaks	0.42	0.13	0.00	0.07	0.34	-0.04	0.07	0.03	-0.03	0.05
		N(jl) Streaks	-0.03	0.00	0.98	0.00	-0.07	-0.04	-0.02	-0.02	-0.08	-0.01
		N(k) Streaks	0.13	0.42	-0.26	0.00	0.02	-0.04	0.00	-0.02	-0.04	-0.06
		N(fgh2) Streaks	0.21	0.32	-0.04	0.20	0.00	0.88	0.11	0.05	-0.11	0.01
		N(gfh2) Streaks	0.43	0.23	-0.01	0.04	0.12	0.10	0.77	0.12	-0.01	-0.03
		N(gfh3) Streaks	0.61	0.07	-0.02	-0.02	0.11	0.05	0.19	0.70	0.00	-0.03
	Frequency	nStreaks	0.76	0.53	0.01	0.05	0.09	0.15	0.18	0.11	-0.03	-0.01
		Repetition Max(N(StreakCluster))	0.83	0.33	-0.01	0.03	0.00	0.08	0.07	0.00	0.00	-0.01
	Diversity	Diversity(StreakCluster)	0.53	0.67	0.05	0.06	0.20	0.24	0.29	0.22	-0.05	-0.01
		Entropy(StreakCluster)	0.34	0.78	0.05	0.02	0.31	0.24	0.23	0.20	-0.05	0.02
Incentives	Proximity	Points till Prize 1 @ Start	0.05	-0.01	-0.01	0.84	0.02	0.09	-0.01	0.01	-0.24	-0.08
		Points till Prize 2 @ Start	0.05	0.00	0.00	0.97	0.00	0.08	0.01	-0.02	-0.15	-0.10
		Points till Prize 3 @ Start	0.04	0.00	0.00	0.98	-0.01	0.06	0.05	-0.02	-0.13	-0.10
	Sustainability	Days SinceLastPrize @ LastActiveDay	0.66	0.16	-0.02	-0.41	-0.08	-0.11	-0.25	-0.23	0.11	0.02
		Days SincePrize3 @ LastActiveDay	0.92	0.05	-0.03	0.03	-0.02	-0.03	-0.04	0.13	0.00	0.03
	Achievement	PrizeWon 1	0.34	0.28	-0.01	0.27	0.52	0.66	0.13	0.03	-0.07	-0.08
		PrizeWon 2	0.55	0.20	-0.02	0.06	0.17	0.11	0.76	0.19	-0.01	-0.02
		PrizeWon 3	0.64	0.08	-0.02	-0.02	0.11	0.06	0.21	0.72	0.01	-0.02
Behavior Change	Steps Level Difference	Diff Sum(StepCount) LastFirstWeek	0.01	-0.05	-0.02	-0.14	0.01	-0.03	-0.01	0.00	0.62	0.70
		Diff Median(Active Steps) LastFirstWeek	0.02	-0.07	0.02	-0.05	0.07	0.01	0.01	-0.01	0.17	0.65
	Adherence Difference	Diff %(Steps<500) LastFirstWeek	0.01	0.03	0.03	0.23	0.06	0.04	0.01	0.00	-0.93	-0.25
		Diff %(Steps>=500) LastFirstWeek	-0.01	-0.03	-0.03	-0.23	-0.06	-0.04	-0.01	0.00	0.93	0.25
		Diff %(Steps>=5k) LastFirstWeek	0.00	-0.03	-0.02	-0.17	-0.02	-0.04	0.00	0.01	0.81	0.48
		Diff %(Steps>=10k) LastFirstWeek	0.00	-0.07	0.00	-0.10	0.02	-0.04	-0.02	-0.01	0.46	0.82
		Diff %(Steps>=15k) LastFirstWeek	0.02	-0.05	-0.01	-0.04	0.04	-0.01	-0.02	-0.01	0.11	0.70
		Diff %(Steps>=20k) LastFirstWeek	0.02	-0.03	-0.01	-0.01	0.04	0.01	0.00	0.00	0.01	0.50

A.4 User Data Features and Factors



A.5 Summary of Factors by Dimensions and Themes

In our data analysis pre-processing, we reduced features using PCA and varimax rotation to derive a set of key factors. We summarize our interpreted factors based on heatmap tables from the previous sections.

	Dimension					
Week Factor	Engagement	Breaks & Lapses	Habit & Routine	Incentives	Behavior Change	Demographics
1 . Inconsistent & Very Low Step Count	Steps Level					
2 . Moderate Step Count	Adherence					
3 . High Step Count	Adherence					
4 . Very High Step Count	Adherence					
5 . Extremely High Step Count	Steps Level, Adherence					
6 . High Steps Mon & Weekend / Low W-F			Pattern			
7 . Weekend Ratio (Sat)			Pattern			
8 . Weekend Ratio (Sun)			Pattern			
9 . Incentivized 10k Step Count				Sufficiency		
10 . Time Till Prize				Proximity		

Week Factor	Description that week has …
1 . Very Low Step Count & Inconsistent	Many days with very low steps (500-5k) with high variation.
2 . Moderate Step Count	Most days with moderate steps (5k-10k), and few non-wearing (steps=0) days.
3 . High Step Count	Most days with high steps (10.5k-15k), and few non-wearing days.
4 . Very High Step Count	Many days with very high steps (15k-20k), and few non-wearing days.
5 . Extremely High Step Count	Many days with extremely high steps (20k-).
6 . High Steps Mon & Weekend / Low W-F	More steps on Mon and/or fewer steps on Fri or Sat.
7 . Weekend Ratio (Sat)	More steps on Sat and/or fewer steps in mid-week (Tue or Wed).
8 . Weekend Ratio (Sun)	More steps on Sun and/or fewer steps on Mon.
9 . Incentivized 10k Step Count	Most steps just at the incentivized 10k level (10k-10.5k) and few non-wearing days.
10 . Time Till Prize	A long time to go before the user wins the next prize.

Streak Factor	Dimension					
	Engagement	Breaks & Lapses	Habit & Routine	Incentives	Behavior Change	Demographics
1 . Low Step Count	Adherence		Pattern			
2 . Moderate Step Count	Adherence		Pattern			
3 . High Step Count	Adherence		Pattern			
4 . Very High Step Count	Adherence		Pattern			
5 . Extremely High Step Count	Adherence		Pattern			
6 . Mostly Non-Active, Some Weekend Active			Pattern			
7 . Mostly Non-Active, Some Weekday Active			Pattern			
8 . Weekend Ratio			Pattern			
9 . Incentivized 10k Step Count				Sufficiency		
10 . Time Till Prize				Proximity		
11 . Prize 1 Won			Pattern	Achievement		
12 . Prize 2 Won			Pattern	Achievement		
13 . Prize 3 Won			Pattern	Achievement		

Streak Factor	Description that streak has …
1 . Low Step Count	Mostly *d* -type week(s) with low steps (500-5k).
2 . Moderate Step Count	Mostly *e* -type week(s) with moderate steps (5k-10k).
3 . High Step Count	Mostly *g* -type week(s) with high steps (10.5k-15k).
4 . Very High Step Count	Mostly *h* -type week(s) with very high steps (15k-20k).
5 . Extremely High Step Count	Mostly *i* -type week(s) with extremely high steps (20k-).
6 . Mostly Non-Active, Some Weekend Active	Mostly *a* -type week(s), fewer *c* -type weeks(s), active mainly on weekends.
7 . Mostly Non-Active, Some Weekday Active	Mostly *j* -type week(s), active mainly on weekdays.
8 . Weekend Ratio	More active weekend days than weekdays.
9 . Incentivized 10k Step Count	Mostly *f* -type week(s) with incentivized 10k steps (10k-10.5k).
10 . Time Till Prize	A long time to go before the user wins any next prize.
11 . Prize 1 Won	The user winning Prize 1
12 . Prize 2 Won	The user winning Prize 2
13 . Prize 3 Won	The user winning Prize 3

Phase Factor	Dimension					
	Engagement	Breaks & Lapses	Habit & Routine	Incentives	Behavior Change	Demographics
1 . Self-Motivated Long Phase	Duration		Pattern, Frequency	Sustainability		
2 . Many diverse low-steps streaks	Duration	Frequency, Duration	Frequency, Diversity			
3 . Repeating Short Partial Weeks			Pattern			
4 . Time Till Prize				Achievement		
5 . Prize 1 High Burst Effort	Steps Level		Pattern, Diversity	Sustainability, Achievement		
6 . Prize 1 Effort			Pattern, Diversity	Achievement		
7 . Prize 2 Won			Pattern, Diversity	Sustainability, Achievement		
8 . Prize 3 Won			Pattern, Diversity	Sustainability, Achievement		
9 . Behavior Change increased Moderate Steps					Steps Level, Adherence	
10 . Behavior Change increased Very High Steps					Steps Level, Adherence	

Phase Factor	Description that phase has …
1 . Self-Motivated Long Phase	Tracked over a long duration, with more *g*, *f*, and *h* type streaks, more repeated streaks, and many days since winning prize 3 (therefore, self-driven).
2 . Many diverse low-steps streaks	Diverse with many different lower-steps level streaks (mostly *c*, *d*, *e*, *k* types) and short breaks.
3 . Repeating Short Partial Weeks	Mostly short very low steps streaks (*[ji]* type).
4 . Time Till Prize	A long time to go before the user wins any next prize.
5 . Prize 1 High Burst Effort	High median active steps in order to quickly win Prize 1
6 . Prize 1 Effort	Many *[fgh](2)* -type streaks with users trying various high-steps streaks in an effort to win Prize 1.
7 . Prize 2 Won	Many *[gfh](2)* -type streaks with users walking various very high-steps streaks with a higher chance to win Prize 2.
8 . Prize 3 Won	Many [gfh](3)-type streaks with users walking various longer, very high-steps streaks with a higher chance to win Prize 3.
9 . Behavior Change increased Moderate Steps	The user increasing the number of active days or days with moderate steps level.
10 . Behavior Change increased Very High Steps	The user increasing the number of days with high steps level from the first to last week.

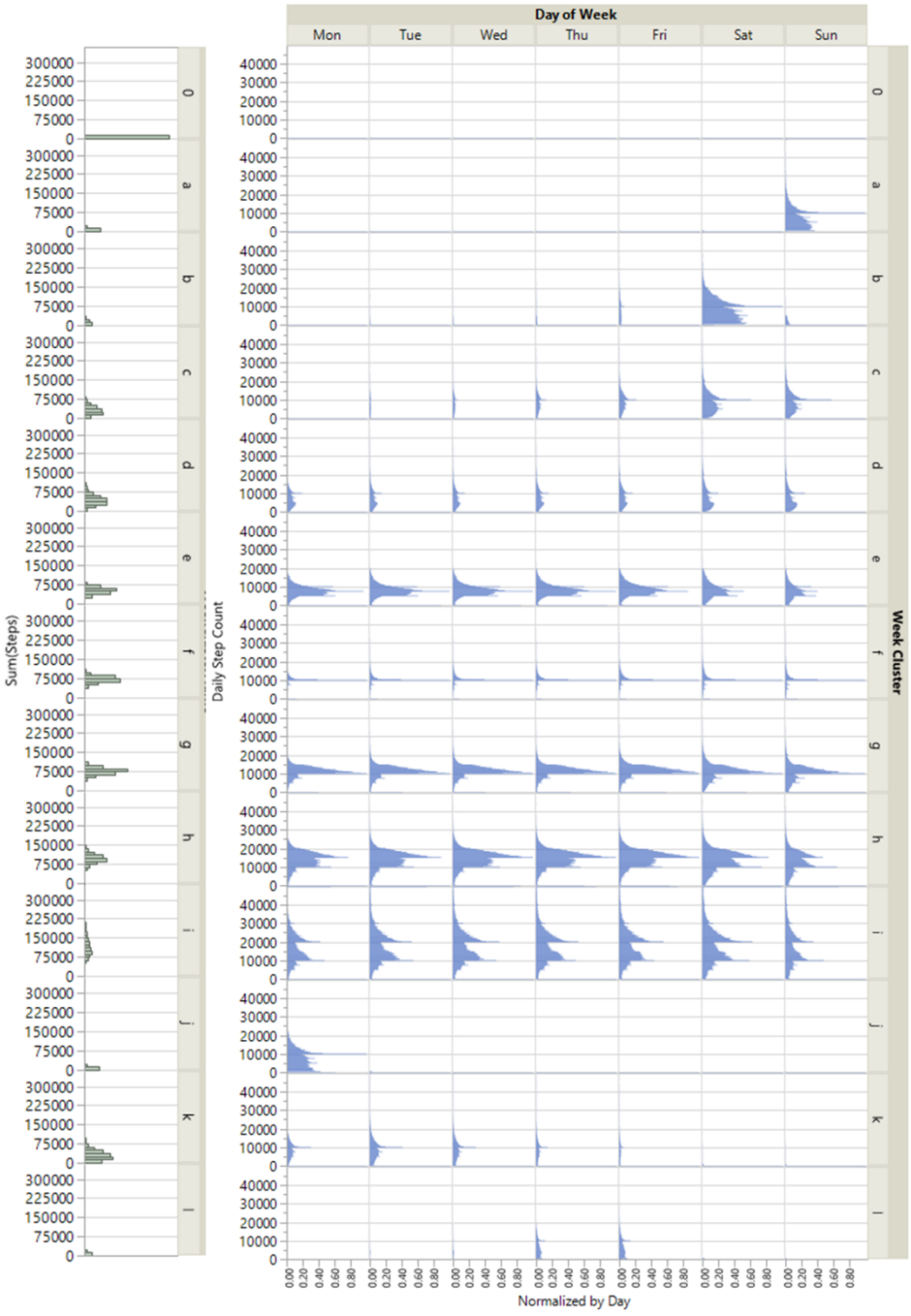
User Factor	Dimension					
	Engagement	Breaks & Lapses	Habit & Routine	Incentives	Behavior Change	Demographics
1 . Very Engaged	Steps Level, Duration		Pattern, Frequency	Achievement		
2 . nDays(1d)			Pattern			
3 . nDays(1w-2w)			Pattern			
4 . nDays(2w-1m)			Pattern			
5 . nDays(1m-2m)			Pattern			
6 . nDays(2m-3m)			Pattern			
7 . nDays(3m-5m)			Pattern			
8 . Low steps, Many Short Breaks	Steps Level, Duration	Frequency, Duration	Pattern, Frequency			
9 . Repeated, Diverse with Long Breaks	Duration	Frequency, Duration	Pattern, Frequency, Diversity			
10 . Extremely Long Break		Duration	Pattern			
11 . N(p1 won) Phases			Pattern	Achievement		
12 . N(p1 won high) Phases			Pattern	Achievement		
13 . N(p1&2 won) Phases			Pattern	Achievement		
14 . N(p3 won improved) Phases			Pattern	Achievement		
15 . N(long 3 prizes) Phases			Pattern	Achievement		
16 . N(long 3 prizes high) Phases			Pattern	Achievement		
17 . Behavior Change more Active Steps					Steps Level, Adherence	
18 . Behavior Change Increased High Steps					Steps Level, Adherence	
19 . Female						Gender
20 . Male						Gender
21 . Age						Age
22 . Body Mass Index (BMI)						BMI

User Factor	Description that user has/is …
1 . Very Engaged	Tracked for very long (>5 months) with the self-driven phase, has high median active steps, and tends to win all prizes.
2 . nDays(1d)	Tracked for only 1-day.
3 . nDays(1w-2w)	Tracked for 1-2 weeks.
4 . nDays(2w-1m)	Tracked for 2 weeks to 1 month.
5 . nDays(1m-2m)	Tracked for 1-2 months.
6 . nDays(2m-3m)	Tracked for 2-3 months.
7 . nDays(3m-5m)	Tracked for 3-5 months.
8 . Low steps, Many Short Breaks	Low median active steps and has several streaks and tends to take short breaks (<1 week).
9 . Repeated, Diverse with Long Breaks	Several repeated and/or diverse phases separated by long breaks (>1, >2 or >4 weeks).
10 . Extremely Long Break	Lapsed with an extremely long break (>4 weeks).
11 . N(p1 won) Phases	*P1 won* Phase instead of short try Phases.
12 . N(p1 won high) Phases	P1 won high Phase instead of P1 won, to differenciate whether high effort is taken.
13 . N(p1&2 won) Phases	*P1&2 won* Phase instead of *self-driven* Phase.
14 . N(p3 won improved) Phases	*P3won improved* Phase instead of *self-driven* Phase.
15 . N(long 3 prizes) Phases	*Long 3 prizes* Phase instead of *self-driven* Phase.
16 . N(long 3 prizes high) Phases	*Long 3 prizes high* Phase instead of self-driven Phase.
17 . Behavior Change more Active Steps	The user increasing the number of active days or days with moderate steps level.
18 . Behavior Change Increased High Steps	The user increasing the number of days with high steps level from the first to last week.
19 . Female	Female
20 . Male	Male
21 . Age	Older
22 . Body Mass Index (BMI)	More overweight

	User Cluster	Notes
0	*Not clustered*	Incomplete demographics, so omitted from clustering
T1	*1-day Try & Drop*	1 -day trial only and also very low step count (Median=5.67k) with 100% for >500 steps, 48% for >5k, 17% for 10k).
T2	*Very Short Try & Drop*	<1 week. More males
T3	*Short Try & Drop*	<2 weeks
C1	*~1-Month Short Consistent*	Contains phases that aimed for P1 prize with high steps effort. Sometimes demonstrating improvement in step count at the end of the Phase (e.g., P.?). Mostly won P1 only and more male users than average (46%
C2	*~2-Month Medium Consistent*	Consistent with very few breaks (Median=2.6, mostly 1-day long) over a moderate period of days (Median=50 days). This is a heterogeneous cluster with users spanning 1 day to several months. Step count
C3	*~3 Month Long Consistent*	Higher step counts than earlier cluster (Median active steps 10.0k), higher adherence (81% for >=500 steps, 71% for >=5k steps, 46% for >= 10k steps. Takes a lot of streaks S.k which means more of them tend to rest
C4	*~4 Month Very Long Consistent*	~3-6-months or longer. Higher step counts than earlier cluster (Median active steps 10.8k) and even more days with steps at the 10-10.5k incentive level (14%), higher adherence (85% for >=500 steps, 78% for >=5k
C5	*Very Long & High Steps Consistent*	Users are "sprinting" towards winning all 3 prizes by applying higher effor with high steps (more >20k steps/day), but drop out with a short streak soon after winning the last prize. Adherence levels are good
CO	*Overweight Consistent*	Heterogeneous cluster of mostly obese (BMI>30 cutoff) or overweight (BMI>24) [WHO] with some users lasting many months (11% at 5m+) or few days (21% at <1 week). Adherence lower than average (81% for >500 steps, 54% for >5k steps, 30% for >10k steps), negligible change in steps level.
CI	*Improved Consistent*	More females than average (69.6% vs. 59.7%). Increased steps at high adherence levels (17.1% increase for >5k steps, 39.5% for 10k, 24.4% for 15k, and 12.6% for 20k). Average of median active steps increased by 5.8k
CD	*Non-Sustained Originally-Consistent*	Started out strong for months to win prizes with continuous phases, some winning 3 prizes in 1 stretch (573 users), but they took a break soon after winning all prizes and returned with short periods (<1 week) of somewhat motivated and high steps. They then drop out. Their adherence decreased by 24.6% for >5k steps, 54% for >10k, 25.6% for 15k, and 10.6% for >20k.
SS	*Slow Starter / Lapsing*	Took 1-2 long breaks (0's) over a long period (many days)
HH	*Hop-On Hop-Off*	Many diverse phases and streaks. They all end up winning all 3 prizes, but with a break in between by "hopping on" to have an effortful phase (+P1) at the beginning to win Prize 1, then "hopping off" for a break
ID	*Deteriorated Intermittent*	Started with moderate steps level (Median Active steps = 9.4k) and decreased in their last week (Median Active steps = 7.4k). Adherence decreased by 56.6% for >500 steps, 52.2% for >5k, 33.7% for >10k).
II	*Improved Intermittent*	Slightly more females than average (62.8%). Increased steps at high adherence levels (58.7% increase for >500 steps, 55.1% for >5k, 32.2% for 10k). Average of median active steps increased by 2.7k steps. They mostly participated for more than 2 weeks, most at least 2 months or longer, some win all 3 prizes, but most only win some and some users do not win any prize. This is a heterogeneous cluster with many users taking short attempt from the beginning and some becoming more motivated (higher steps) in some later streaks, yet some users also take long or extremely long breaks. They achieved their increased steps level from a very low level (Median active steps 6.2k to 9.0k) to a moderate level.
P	*Self-Driven Power*	Tended to have 1 *self-driven* Phase with many Streaks. Somewhat incentivized (14% of days with 10-10.5k steps). Many weeks with high or incentive high steps (W.g mean=6.9 weeks, W.f mean 5.2 weeks). Very high adherence.

All differences described as statistically significant from a means comparison analysis (Tukey HSD) with Bonferonni correction (i.e., α=0.001).

B DISTRIBUTION OF DAILY STEP COUNT IN WEEKS BY WEEK CLUSTER



C DETAILS OF STATISTICAL ANALYSIS ON DEMOGRAPHIC FACTORS

We divided age into discrete bins to allow for ANOVA analysis. There was a significant fixed effect with gender and age for both step count and life span: men walked more per day than women ($F(1,120438)= 205.5$, $p<0.0001$), but women participated in the program for longer than men ($F(1,120438)=156.0$, $p<0.0001$); older users both walked more and for longer than younger users (Daily active steps: $F(4,120438)=361.4$, $p<.0001$; Life span: $F(4,120438)=2397.2$, $p<.0001$). Furthermore, there was an interaction effect where younger women (<50 years old) walked fewer daily steps than younger men, but there was no difference for older women and men; $F(4,120438)=18.5$, $p<.0001$.

D ADDITIONAL FIGURES OF STEP COUNT DISTRIBUTION

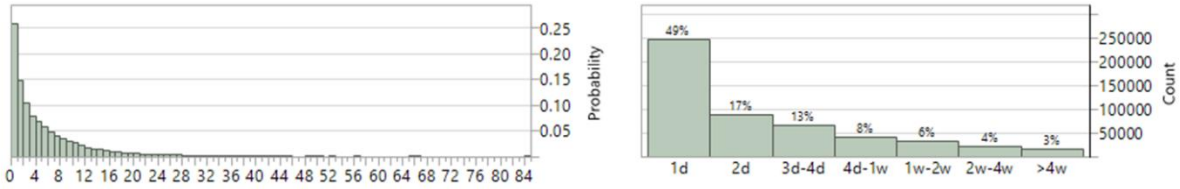


Fig. 10. Distribution of number of breaks (Left) and break lengths (Right). Break length in days (d) and weeks (w): 1d, 2d, 3d-4d (half week), 4d-1w (<1 week), 1w-2w (long), 2w-4w (very long), 4w+ (extra long).

Table 9. (Left) Average of median daily active steps and participation life span of users by their gender and age.

		Median Daily Active Steps	Participation Life Span (Days)
Gender	Female	7720	47.8
	Male	8309	39.5
Age	<30	8015	43.7
	30-40	8255	69.8
	40-50	8393	79.5
	50-60	9277	94.3
	≥60	9597	105.6

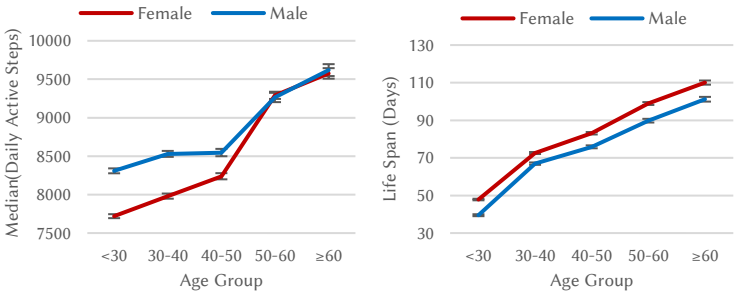


Fig. 11. Mean of differences in median daily active steps (Middle) and participation life span (Right) by age and gender.

E ADDITIONAL DETAILS OF PATTERNS OF STEPS BEHAVIOR

For brevity, we had omitted detailed descriptions of step pattern clusters. We provide the details in this section.

E.1 Streak Patterns: Ranging from Very Short and Disengaged to Consistent Series of Weekly High Steps

We identified 13 streak factors. We chose data features such that the streaks will be closely related to their underlying week type, but also consider streak length (number of weeks), and incentive proximity and achievement. Although we could bin the streak length into 5 evenly sized bins, this would lead to 13×5=65 combinations of streaks, which could make subsequent clustering difficult. Therefore, streak clustering serves the purpose of dimensionality reduction for the later phase clustering.

Streak clustering resulted in 13 clusters (see Table 10). Note that each streak in these clusters actually consists purely of one type of week. As designed, streak types follow a similar pattern to week types and we labeled them with similar names to represent this correspondence.

Key patterns are:

- Most streak types contained streaks of the same week type but varying lengths, indicating that streak length was not an influential feature for grouping clusters (see the diagonal pattern in the # Week Type heatmap in Table 10). The streaks are Moderate Partial-Week Streaks (*c*, *k*), Moderate Full Week-Streak (*d*, *e*), Highly Active Streaks (*f*, *g*, *h*, *i*).
- Very short and low steps level week types are similar and merged into the same streak types (*[ab]*, *[jl]*). These tend to be one-off with no weekly repetition, indicating rapid loss of engagement before a break or abandonment.
- Prize-winning streaks, when users completed one of the three incentive goals at the end of the streak (*[fgh]{2}*, *[gfh]{2}*, *[gfh]{3}*). These streaks tend to consist of weeks with consistent incentivized or high step count (*f*, *g*, *h*), and stretch for multiple weeks. This suggests healthy habitual behavior.

Table 10. Summary describing 13 streak clusters of user step activity. Matrices on the right indicate heatmap of values. Streak clusters represent pure occurrences of week clusters which are either repeated once or over multiple weeks.

	Streak Cluster	% Streaks	Median Active Steps	# Week Type: 0	a	b	c	d	e	f	g	h	i	j	k	l	# Weeks: 1	2	3-4	4-8	9+
0	Break Streak, consisting of 1 or more Break weeks	4.8%	-																		
[ab]	1 week of *a* or *b*types (very low steps, weekend)	5.5%	7,362																		
c	1 week of *c*-type (low steps, late-week)	10.4%	8,890																		
d	1-2 weeks of *d*-type (very low & low steps)	10.5%	7,858																		
e	1-2 weeks of *e*-type (high steps)	8.1%	8,001																		
f	1-2 weeks of *f*-type (incentive 10k steps)	6.4%	10,474																		
g	1-2 weeks of *g*-type (high steps)	9.5%	11,324																		
h	1-2 weeks of *h*-type (very high steps)	6.5%	14,076																		
i	1-2 weeks of *i*-type (extremely high steps)	4.9%	15,134																		
[jl]	1 week of *j* or *l* types (very low steps, early week)	4.9%	7,624																		
k	1 week of *k*-type (moderate steps, early week)	15.0%	9,179																		
[fgh]{2}	1-2 weeks of *f*, *g*, or *h* types; prize 1 won	6.3%	10,703																		
[gfh]{2}	1-2 weeks of *g*, *f*, or *h* types; prize 2 won	4.1%	11,711																		
[gfh]{3}	1-3 weeks of *g*, *f*, or *h* types; prize 3 won	3.1%	11,970																		

Regex-like notation for naming streak clusters: [] indicates OR, i.e., the streak consists of any listed week type; {} indicates number of occurrences, {2} means ~2 weeks.

E.2 Phase Patterns: with Varying Duration, Diversity, Difference, and Drive of Steps Activity

We identified 10 phase factors relating to whether the phase was many days long, had a high diversity of streak types, indicated self-driven behavior of persistent and sustained high steps level, whether a prize was won or will soon be won, and whether the steps level changed from the beginning to the end of the phase.

Table 11. Summary describing 15 phase clusters of user step activity. Matrices on the right indicate heatmap of values.

Phase Cluster	% Phases	Median Active Steps	#	# Breaks	# Streak Type *ab*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*jl*	*k*	*f2*	*g2*	*g3*	Diff %Steps Last-First Week ≥500	≥5k	≥10k	≥15k	≥20k
very short try	12.2%	7,714	4	0.4														0%	0%	0%	0%	0%
short try	10.1%	6,991	10	1.2														-1%	-1%	0%	0%	0%
short try high	11.9%	8,130	13	1.3														-8%	-4%	-1%	0%	0%
try improved	3.2%	6,944	21	1.7														59%	40%	14%	3%	1%
try hard improved	9.1%	9,368	23	1.8														24%	20%	13%	4%	1%
p3 won improved	1.2%	10,486	33	2.7														34%	31%	19%	6%	2%
p1&2 won	5.6%	10,456	66	4.4														4%	4%	2%	1%	1%
p1 wonhigh	4.5%	12,491	44	3.0														-6%	-4%	-3%	0%	0%
p1 won	12.4%	9,914	38	3.1														-6%	-5%	-4%	0%	0%
long 3 prizes high	2.0%	13,453	100	3.7														-10%	-8%	-9%	-3%	-1%
long 3 prizes	5.1%	11,205	96	4.6														-4%	-3%	-4%	-1%	1%
self-driven	4.4%	11,735	184	5.7														4%	5%	2%	1%	0%
long break	8.0%	-	7																			
very long break	5.5%	-	16																			
extra long break	4.7%	-	53																			

Shorthand for phase types: ab=[ab], jl=[jl], f2=[fgh]{2}, g2=[gfh]{2}, g3=[gfh]{3}.

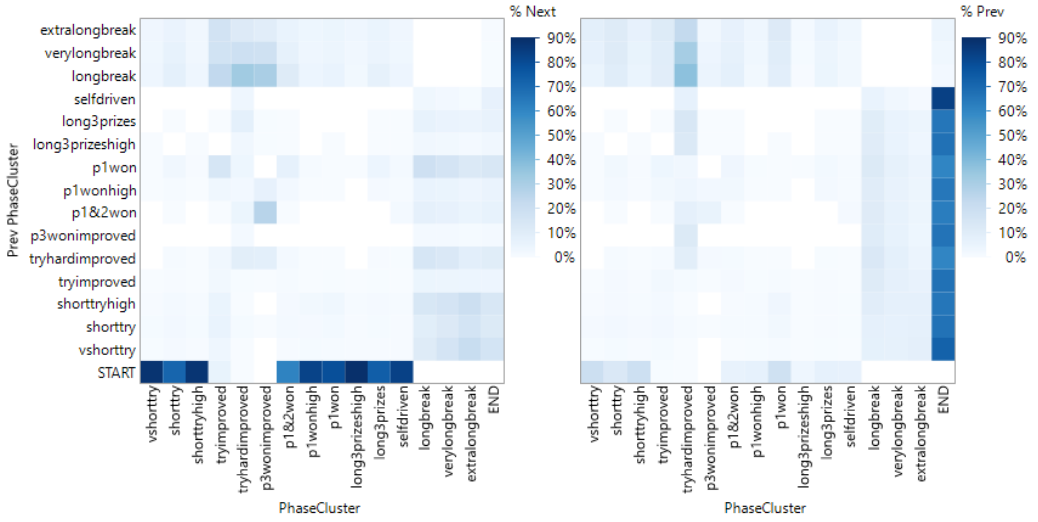


Fig. 12. Transition matrices showing % of previous phases (row) that are followed by current phases (column). The Left figure is normalized by column, and the right figure is normalized by row. E.g., the *p3wonhigh* phase is normally preceded by *p1&2won* or a *longbreak*; longbreaks are normally followed by the *tryhardimproved* phase.

E.3 User Types: Differentiated by Duration of Participation, (In)Consistency in Steps, and Behavior Change

We identified 22 user factors based on a combination of all five dimensions of engagement, breaks, incentives, behavior change and demographics. By considering the histograms of tracking duration and break lengths, we can identify factors indicative of users who tracked for extremely long times (~3-5 months) and users who had extremely long breaks (>4 weeks). Among all the types of phases, mainly phases regarding prize winning were strongly correlated to our user factors. Feature correlations suggest that users who had phases that won just Prize 1 or 2 tend to avoid "short try" or similar phases.

F ADDITIONAL FIGURES OF TRAINED DECISION TREES TO DESCRIBE AND PREDICT PARTICIPANT BEHAVIOR

Table 12. Performance (area under precision-recall curve and F1-score) of different models trained on different data features with the full dataset. This shows that including temporal features improves model accuracy for explaining which prizes users won.

Post-P3 2w Median(steps)>10k

	PR AUC		F1 Score (@p=.5)		
Model with features	Decision Tree	Random Forest	Decision Tree	Random Forest	Features description
Daily median steps	0.789	0.753	0.749	0.707	Basic data: median daily step count, user age, gender, BMI
+ % non-active days	0.873	0.862	0.807	0.787	Add adherence level of active/non-active days (≥500 steps)
+ steps histogram	0.885	0.898	0.810	0.815	Add finer adherence levels as histogram (≥5k, 10k, 15k, 20k)
+ Week	0.888	0.890	0.803	0.804	Add week clusters
+ Streak	0.904	0.911	0.844	0.841	Add streak clusters
+ Phase	0.910	0.918	0.853	0.849	Add phase clusters
+ User	0.910	0.921	0.853	0.851	Add user cluster

Post-P3 2w Median(steps)>10k: whether the user won prize 3 and had median active step count >10k steps/day during the two weeks after winning that last prize.
Diff 2w Median(steps)>2k: whether the user won prize 3 and had median active step count during the next two weeks with difference of at least 2k steps/day *higher* than

Table 13. Performance (area under precision-recall curve) of different models trained on different data features. This shows that including temporal features improves model accuracy for explaining which prizes users won.

PR AUC	No Prize Won		Prize 1 Won		Prize 2 Won		Prize 3 Won	
Model with features	DT	RF	DT	RF	DT	RF	DT	RF
Daily median steps	0.9873	0.9863	0.844	0.826	0.504	0.510	0.906	0.898
+ % non-active days	0.9980	0.9972	0.970	0.945	0.745	0.640	0.979	0.958
+ steps histogram	0.9984	0.9992	0.976	0.981	0.775	0.791	0.979	0.981
+ Week w/o histogram	0.9980	0.9977	0.970	0.958	0.745	0.678	0.979	0.972
+ Week	0.9984	0.9993	0.976	0.982	0.774	0.786	0.979	0.981
+ Streak	0.9996	0.9998	0.994	0.996	0.955	0.960	0.996	0.998
+ Phase	1.0000	1.0000	0.9998	0.9995	0.999	0.994	1.000	0.999
+ User	1.0000	1.0000	0.9998	0.9995	0.999	0.997	1.000	1.000

DT: Decision Tree, RF: Random Forest

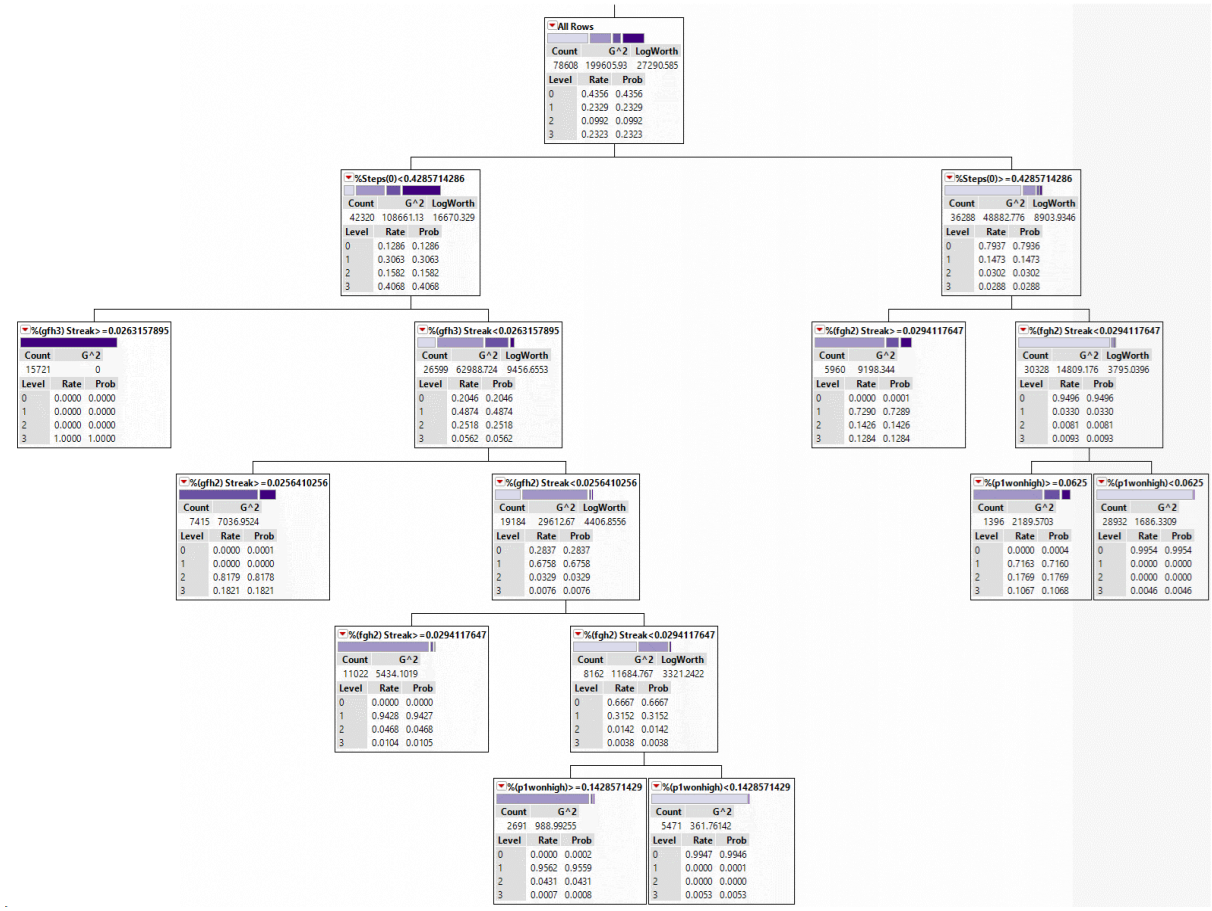


Fig. 13. Partial (pruned) view of trained decision tree with all steps and cluster features for describing which final prize participant users won.

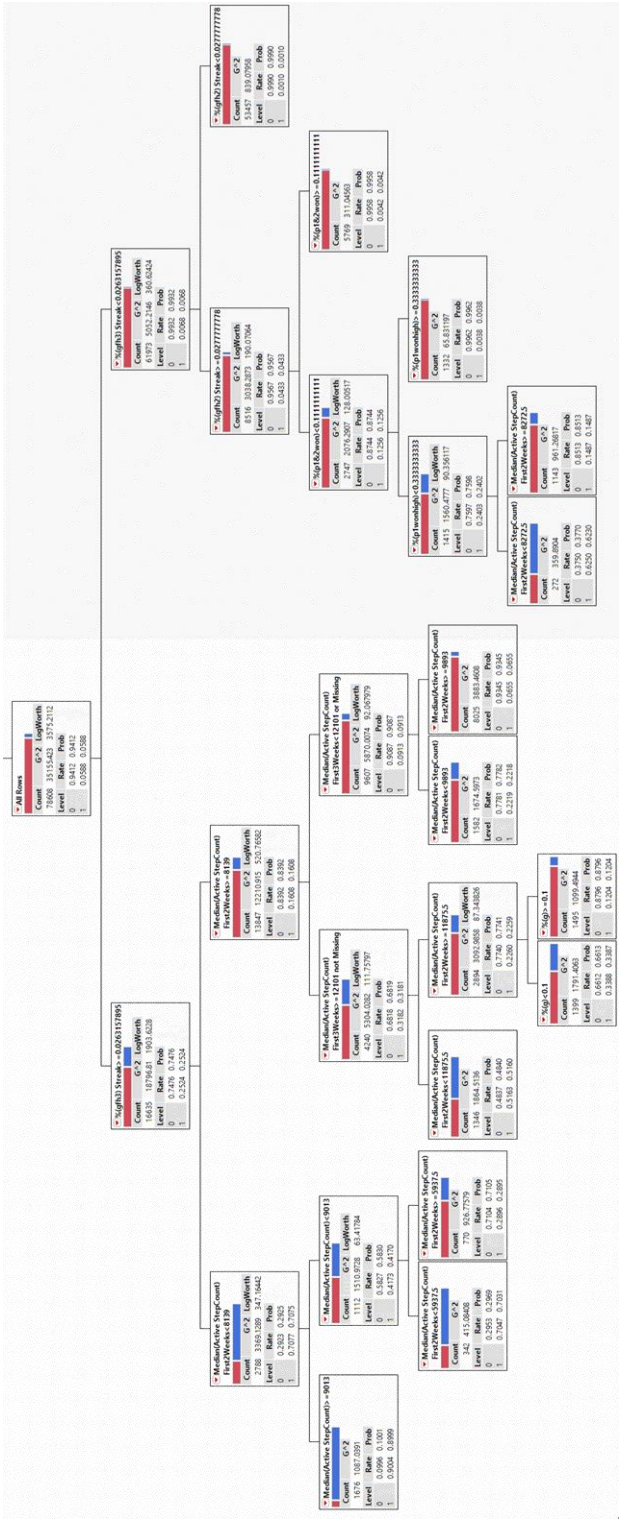


Fig. 14. Partial (pruned) view of trained decision tree with all steps and cluster features for describing which users participated until they won prize 3 and had a higher median daily step count of at least 2k steps/day during the following two weeks compared to their first two weeks of tracking, i.e., Diff 2w Median(steps)>2k.

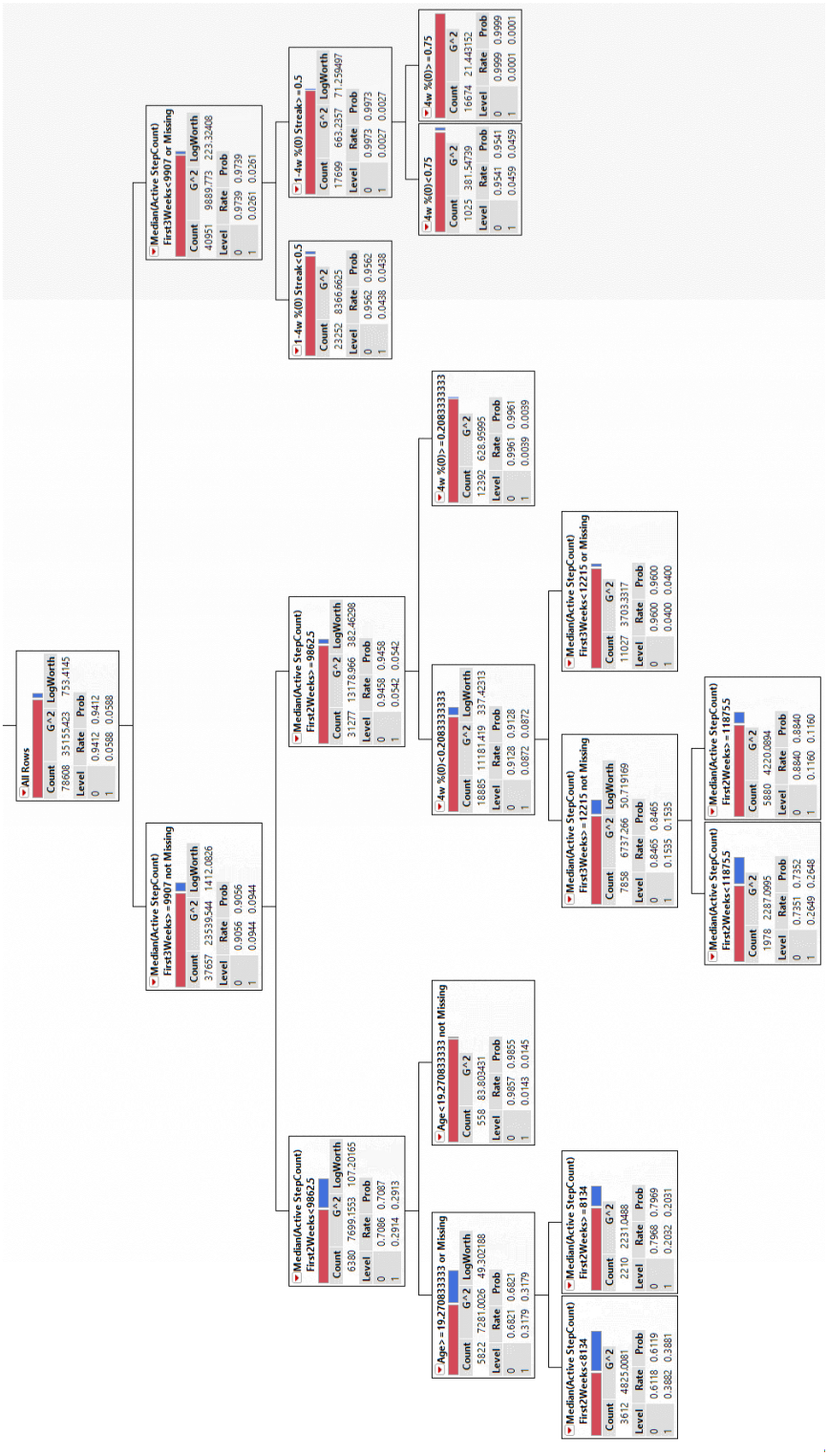


Fig. 15. Partial (pruned) view of trained decision tree with all steps and cluster features from only the first four weeks of data for predicting if a users will participate until she wins prize 3 and has a higher median daily step count of at least 2k steps/day for the following two weeks compared to their first two weeks of tracking, i.e., Diff 2w Median(steps)>2.